PSYCHOMETRIKA—VOL. 87, NO. 4, 1238–1269
DECEMBER 2022
https://doi.org/10.1007/s11336-022-09865-7

ITEM RESPONSE THRESHOLDS MODELS: A GENERAL CLASS OF MODELS FOR VARYING TYPES OF ITEMS

GERHARD TUTZ

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

A comprehensive class of models is proposed that can be used for continuous, binary, ordered categorical and count type responses. The difficulty of items is described by difficulty functions, which replace the item difficulty parameters that are typically used in item response models. They crucially determine the response distribution and make the models very flexible with regard to the range of distributions that are covered. The model class contains several widely used models as the binary Rasch model and the graded response model as special cases, allows for simplifications, and offers a distribution free alternative to count type items. A major strength of the models is that they can be used for mixed item formats, when different types of items are combined to measure abilities or attitudes. It is an immediate consequence of the comprehensive modeling approach that allows that difficulty functions automatically adapt to the response distribution. Basic properties of the model class are shown. Several real data sets are used to illustrate the flexibility of the models

Key words: thresholds model, latent trait models, item response theory, graded response model, Rasch model.

Modern item response theory provides a variety of models for the measurement of abilities, skills or attitudes, see, for example, Lord and Novick (1968), Van der Linden (2016a), Mair (2018). The history of its evolution has been traced back carefully by Van der Linden (2016b) and Thissen and Steinberg (2020).

Essential components of item response theory are that items can be located on the same scale as the latent trait and that the latent trait accounts for observed interrelationships among the item responses (Thissen and Steinberg, 2020). In addition, it is essential that the responses are random and have to be described by a probabilistic model to explain their distributions (Van der Linden, 2016b). These features distinguish item response theory from classical test theory (Lord and Novick, 1968), which uses an a priori score on the entire test by assuming an additive decomposition of an observed test score into a true score and a random error.

Item response models are typically tailored to the type of item. For binary items Rasch models and normal-ogive models are in common use (Rasch, 1961; Birnbaum, 1986), for ordered models the graded response model (Samejima, 1995, 2016), the partial credit model (Masters, 1982, Glas & Verhelst, 1989) and the sequential model (Tutz, 1989) have been used. For count data items, among others, Rasch's Poisson count model and extensions as the Conway–Maxwell–Poisson model (Rasch, 1960; Forthmann et al., 2020) have been proposed. Continuous response models have been considered by Samejima (1973), Müller (1987), and Mellenbergh (2016). For taxonomies of item response models see Thissen and Steinberg (1986) and Tutz (2020).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-022-09865-7.

Correspondence should be made to Gerhard Tutz, Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799Munich, Germany. Email: tutz@stat.uni-muenchen.de

The threshold model proposed here advances a unifying approach. Rather than developing different models for different types of responses a common response model for all sorts of responses is considered. In the model, each item has its own item difficulty function that determines the distribution of the response. Since item difficulty functions are item-specific, the form of the distribution can vary across items. The model class is rather general, it comprises various commonly used models as the binary Rasch model, the normal-ogive model and the graded response model, for the latter it offers a sparser parameterization. It also provides a genuine latent trait model for continuous responses, which can be seen as a latent trait version of classical test theory. In addition to providing a common framework for existing and novel models, it offers a way to combine different types of items in one test, what has been described as mixed item-formats. Instead of using linkage methods (Kim and Lee, 2006; Kolen and Brennan, 2014) to combine different items, the model itself accounts for the different sorts of items.

Major advantages of the approach are:

- The model provides a common framework for several models in common use.
- A genuine latent trait model for continuous responses as an alternative to classical test theory is contained as a special case.
- The model is very flexible and allows for quite different response distributions.
- Items can have different formats, they can be continuous, binary or polytomous, and the common model automatically accounts for the distributional differences. The model links performance on items that can differ in distributional form to person abilities.

The threshold model and basic concepts are introduced in Sect. 1. It is in particular demonstrated how difficulty functions can be used to model the distribution of responses. In Sect. 2, the case of discrete responses is considered and it is shown that common binary models and the graded response model are special cases of threshold models. In Sect. 3, further properties and alternative modeling approaches are considered. Section 4 is devoted to mixed item formats. In Section 5, a more flexible way of specifying difficulty functions is given, which allows to let the data determine which function fits best. The computation of estimates is considered in Sect. 6, although illustrative applications are given already in the previous sections. In the "Appendix", results that are mentioned in the text are given in a more formal way together with proofs.

1. Thresholds Models: Basic Concepts

Let Y_{pi} denote the response of person p on item i ($p \in \{1, \ldots, P\}, i \in \{1, \ldots, I\}$) having support S. The general thresholds model we propose is given by

$$P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(y))), \tag{1}$$

where $F(.)$ is a strictly monotonically increasing distribution function, θ_p is a person parameter, α_i a discrimination parameter, and $\delta_i(.)$ is a non-decreasing item-specific function, called *item difficulty* function, which is defined on the support S. The function $F(.)$ is a *response function*, which to a degree determines the distribution of the response. Since $F(.)$ is increasing for fixed threshold y, the probability of a response larger than y increases with increasing person parameter θ_p. Thus, θ_p can be seen as an ability or attitude parameter, which indicates the tendency of a person to obtain a high score. Higher values of θ_p are associated with a greater chance of a correct or affirmative response to each item. The name of the model refers to the modeling of the threshold y. It is not a threshold on the latent scale, which are the thresholds that are usually considered in latent trait modeling but on the *observable* scale.

In addition to the link between the latent variables and the observable response specified in equation (1), conditional independence of observable variables given the latent variables is assumed, which is a typical assumption in item response theory often referred to as local independence, e.g., Lord (1980). Conditional independence together with the latent monotonicity makes the model a monotone latent variable model in the sense of Holland and Rosenbaum (1986). Latent monotonicity as defined by Holland and Rosenbaum (1986) means that the probability $P(Y_{pi} > y|\theta_p, \delta_i(.))$ is a nondecreasing function of the person parameter for all items. Since the response function $F(.)$ and the item difficulty functions are monotone, latent monotonicity holds for the thresholds model.

The specifics of the thresholds model follow from the functions that are chosen. The item difficulty function $\delta_i(.)$ contains the properties of the item, in particular if it is easy or hard to score high. It also determines the concrete form of the response distribution, which is only partially determined by $F(.)$. In the following, it is shown that the model allows for quite different distributions of responses although the function $F(.)$ is chosen fixed.

Latent trait models have been extensively discussed for binary or other categorical responses. Nevertheless, we start with the less familiar case of continuous responses and first investigate the potential of the thresholds model as a latent trait model for continuous responses.

1.1. Linear Item Difficulty Functions

A particular interesting item difficulty function is the linear one, which allows for some simplifications. Let Y_{pi} be a continuous response variable and the item difficulty be linear, $\delta_i(y) = \delta_{0i} + \delta_i y$, $\delta_i > 0$. Then, one has a parametric model with item parameters δ_{0i}, δ_i. One obtains that the expectation and variance of Y_{pi} are given by

$$\mathrm{E}(Y_{pi}) = \gamma_i\theta_p - \gamma_{0i}, \quad \mathrm{var}(Y_{pi}) = c\gamma_i^2/\alpha_i^2, \tag{2}$$

where $\gamma_i = 1/\delta_i$, $\gamma_{0i} = (\delta_{0i} + d/\alpha_i)/\delta_i$, with constants d, c that are determined by the distribution function $F(.)$, for the concise form of constants and a proof, see "Appendix". In addition, for symmetric response function $F(.)$ the distribution function of Y_{pi} is a shifted and scaled version of $F(.)$.

It is immediately seen that high ability θ_p indicates a tendency to high responses. The item parameter γ_i is a scaling parameter, and γ_{0i} is the location on the latent scale. It represents the 'basic' difficulty of the item; if γ_{0i} is large, the expected response is small, and vice versa. The specific choice $\gamma_i = 1$ (equivalent to $\delta_i = 1$) and $\alpha_i = 1$ yields the simpler forms $\mathrm{E}(Y_{pi}) = \theta_p - \gamma_{0i}$, $\mathrm{var}(Y_{pi}) = c$, which means that the response is simply determined by the difference between ability θ_p and item difficulty γ_{0i}, a property that is familiar from the binary Rasch model or the normal-ogive model without a slope parameter. If $F(.)$ is symmetric, $d = 0$, which means the expectation does not depend on the discrimination parameter. It, however, determines the variance such that large values of the discrimination parameter are associated with small variances of the response.

1.2. The Person Threshold and the Item Characteristic Function

The link between the person and the difficulty functions can be described and visualized in several ways. An important function is the *person threshold function* (PT function), which for fixed θ_p is defined by

$$g_{i,\theta_p}(y) = P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(y))),$$

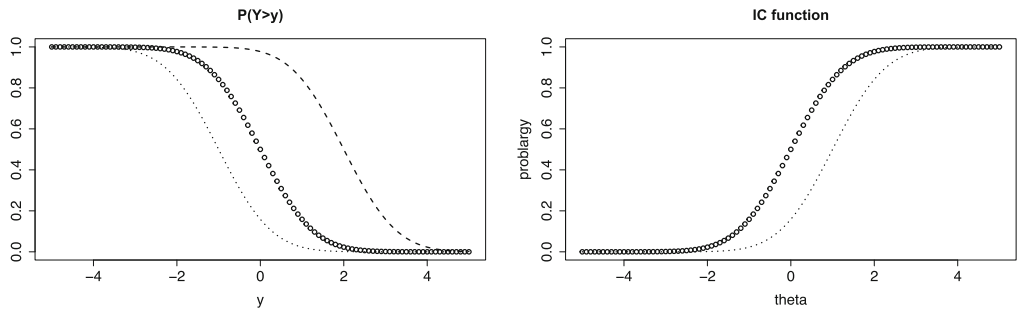

FIGURE 1.
Left: Person threshold functions, $P(Y > y)$, for values $\theta = 0$ (circles), $\theta = 2$ (dashed), $\theta = -1$ (dotted); right: item characteristic functions for $y = 0$ (circles) and $y = 1$ (dotted)

It shows the probability of an response above y for a specific person with ability θ_p. It is strongly related to the distribution function of Y_{pi}, which is simply given by $F_{pi}(y) = 1 - F(\alpha_i(\theta_p - \delta_i(y)))$. The distribution function is denoted with the subscript pi to distinguish it from the response function $F(.)$ (which itself is a distribution function).

The second function is the general *item characteristic function* (IC function). It is an extended form of the item characteristic function commonly used in binary item response theory, and is defined by

$$IC_{i,y}(\theta_p) = P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(y))).$$

It shows the probability for an response above a fixed value y for varying abilities. In contrast to binary models where only the value $y = 0$ is interesting, for responses with more than two possible values one has more than one function. If the response is continuous any value y can occur. Thus, the functions depend on the item i *and* y.

For illustration we first consider the simple case of linear difficulty functions ($\alpha_i = 1$). The left picture in Fig. 1 shows the person threshold function for three values of θ if $F(.)$ is the normal distribution function and the threshold function is linear, $\delta(y) = y$. It is seen that a person with $\theta = 2$ (dashed lines) has higher probability of a response above y than a person with $\theta_p = 0$ (circles) for all values y. The right picture shows the IC function for two values of y, $y = 0$ (circles) and $y = 1$ (dotted). It is seen that the probability of a response above y is strictly increasing with ability θ. In this simple case, the IC functions for different y are just shifted versions of the same basic normal distribution function. This changes with the parameters of the item difficulty function.

Therefore, let us consider the parameters of the difficulty function in more detail. The first parameter δ_{0i} in the difficulty function $\delta_i(y) = \delta_{0i} + \delta_i y$ determines the location of the item. The corresponding mean of Y_{pi} is $-\delta_{0i}/\delta_i$ (for $\theta_p = 0$ and symmetric function $F(.)$). Thus, the PT function is shifted to the left for large location parameter $\delta_{0i} > 0$, which represents the basic difficulty. The second parameter determines the variance of Y_{pi}, large δ_i means that the variance is small. The left picture in Fig. 2 shows the PT functions for the simple function $\delta_i(y) = y$ (circles) and $\delta_i(y) = 2 + 3y$ (dashed). It is seen that for the latter difficulty function the PT function is shifted to the left and the variance is much smaller, which is seen from the steep decrease of the dashed function. The right picture shows the corresponding item characteristic functions.

If the item discrimination parameter is the same for all items, the IC functions have the same form for all items, namely that of the distribution function $F(.)$. This is immediately seen from the definition of the function $IC_{i,y}(\theta_p) = F(\alpha_i(\theta_p - \delta_i(y)))$ since for fixed value y the value of

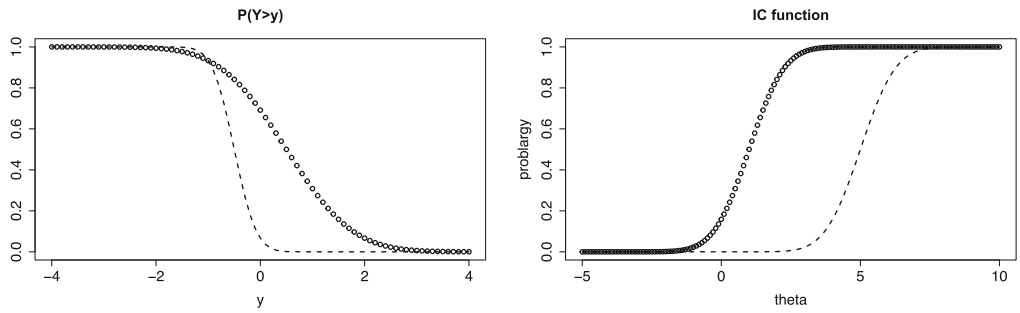

FIGURE 2.
Left: Person threshold functions, $P(Y > y)$, for value $\theta = 0.5$ and $\delta_i(y) = y$ (circles), $\delta_i(y) = 2 + 3y$ (dashed); right: item characteristic functions for the two items for $y = 1$

$\delta_i(y)$ is fixed. It is an important aspect regarding interpretation. For any y the item characteristic functions are increasing, and never do cross. That means a person with larger θ_p than another person has always a larger probability of a response above threshold y. This property, which is well known from binary Rasch models, also holds for the continuous thresholds model with fixed α_i. It holds in spite of the scaling of the person parameter in the term $\gamma_i\theta_p$ of Eq. (2). If the binary Rasch model is extended to the 2PL model, often referred to as Birnbaum or 2PL model, item characteristic functions typically cross, as they do in the general thresholds model with a discrimination parameter.

The simplicity of the IC functions for fixed discrimination parameters has an additional advantage. Since all IC functions are just shifted (and possibly scaled) versions of the same function, they show which items are harder, and which are easier to solve. One obtains an ordering without having to investigate the item parameters.

The essential properties of the model, which hold for all sorts of responses to be considered later, can be described by the following functions, which refer to different aspects of the model.

- The item difficulty function, which characterizes the difficulty of the item over the whole range of possible outcomes.
- The person threshold function, which represents the distribution of the responses. The concrete form of the distribution as well as the support (see below) depend on the difficulty functions. The distributions may take quite different forms for different difficulty functions.
- The form of the item response function is kept fixed, for all items the probability of scoring above the threshold y increases in the same way with the ability. However, it depends on the threshold y, respectively, the corresponding $\delta_i(y)$, how large the probability of an response above y is.

It is essential to distinguish between two specifications regarding the complexity of the difficulty function. Let, more general, the difficulty functions be given by $\delta_i(y) = \delta_{0i} + \delta_i g(y)$, where $g(.)$ is a monotonically increasing function. Then, a simplifying assumption is that the difficulty functions have common slopes, that is, $\delta_1 = \cdots = \delta_I = \delta$. Without this restriction slopes may vary across items. For linear difficulty functions and $\alpha_i = 1$ the assumption of common slopes simply means that for all responses one assumes the same variance.

If difficulty functions have the form $\delta_i(y) = \delta_{0i} + \delta_i g(y)$ with fixed function $g(.)$, the model is parametric with $\alpha_i, \delta_{0i}, \delta_i, i = 1, \ldots, I$, representing the item parameters, and θ_p, $p = 1, \ldots, P$, representing the person parameters. To obtain identifiability, some restrictions are needed. One can, for example, choose fixed values for one discrimination parameter and one person parameter

(e.g. $\alpha_1 = 1, \theta_1 = 0$). In general, restrictions can depend on the support of the response, see Proposition 8.2.

Difficulty functions of the form $\delta_i(y) = \delta_{0i} + \delta_i g(y)$ contain the slopes δ_i. The concept of slopes should be distinguished from the concept that is sometimes used in binary and polytomous models. In binary models as the 2PL model, $P(Y_{pi} = 1) = F(\alpha_i(\theta_p - \delta_i))$, the parameter α_i is a discrimination parameter, but is also often referred to as slope parameter. It has a quite different meaning than the slope in thresholds models. Large discrimination parameters have the effect that the increase in probability is stronger when θ_p increases than for smaller discrimination parameters. For fixed θ_p, δ_i larger values of α_i mean that response probabilities are more extreme than for smaller values (closer to 1 if $\theta_p - \delta_i > 0$, closer to 0 if $\theta_p - \delta_i < 0$). The slopes in the difficulty functions of thresholds models have a different effect, they refer to the difficulty of items. As seen from equation (2) in the case of linear difficulty functions if the slope δ_i increases the expectation of responses decreases. Larger slopes means smaller expected responses, although also the variance changes. To keep the two concepts apart, we always refer to α_i as a discrimination parameter, and the notion of "slope" refers to the slope of the difficulty function.

As in other IRT models, alternative parameterizations can be used. The predictor $\eta_{pi} = \alpha_i(\theta_p - \delta_{0i} - \delta_i g(y))$ can also be given in the form $\eta_{pi} = \alpha_i\theta_p - \tilde{\delta}_{0i} - \tilde{\delta}_i g(y)$, where $\tilde{\delta}_{0i} = \alpha_i\delta_{0i}$, $\tilde{\delta}_i = \alpha_i\delta_i$. In the alternative parameterization, only the first term contains the person parameter, the intercept and slope are built from the item discrimination parameter and the original intercept and slope. The parameterization is helpful to clarify the role of slopes and discrimination parameters. The expectation of the response given in equation (2) suggests that the item slope (of the original parameterization) is the essential scaling parameter that acts as a discrimination parameter or factor loading, and α_i seems superfluous. But for linear difficulty functions (and symmetric response function) one obtains in the alternative parameterization $\mathrm{E}(Y_{pi}) = (\alpha_i\theta_p - \tilde{\delta}_{0i})/\tilde{\delta}_i$, which shows that θ_p is weighted by $\alpha_i/\tilde{\delta}_i$. Thus, the expectation is determined by α_i and $\tilde{\delta}_i$ if one uses the alternative parameterization. The original parameterization $\alpha_i, \delta_{0i}, \delta_i$ has the advantage that it is closer to parameterizations that are typically used in traditional binary and multi-categorical models. The alternative parameterization is useful in extensions to multi-dimensional structures to be considered later.

1.3. Links to Classical Test Theory

The thresholds model for continuous responses is a genuine latent trait model. It has all the attributes of a latent trait model, items are located on the same scale as the latent ability, the latent variable accounts for observed interrelationship among the item responses and responses are described by a probabilistic model to explain their distribution. In contrast, classical test theory, which is often used for continuous data, is not a latent trait model in this sense. It is a regression type model, in which an a priori score on the entire test is chosen by assuming an additive decomposition of an observed test score into a true score and a random error; it can be traced back to Spearman (1904), an extensive presentation is found in Lord and Novick (1968).

A similar decomposition is obtained for the thresholds model with linear difficulty functions. If it holds one has

$$Y_{pi} = E(Y_{pi}) + E_{pi},$$

where $\mathrm{E}(Y_{pi}) = \gamma_i\theta_p - \gamma_{0i}$, $\mathrm{var}(E_{pi}) = c\gamma_i^2/\alpha_i^2$ and Y_{pi} has distribution function $F(.)$. Random sampling of individuals yields

$$Y_{*i} = E(Y_{*i}) + E_{*i},$$

which corresponds to the decomposition into a true-score and an error-score random variable (see Section 2.6 Lord and Novick, 1968) with the true score depending only on the measurement instrument. The error-score E_{pi} follows distribution function $F(.)$, and has expectation 0 and variance $c\gamma_i^2/\alpha_i^2$. In classical test theory, the variance of the error-score is often assumed to be the same for all responses, which means that in addition $\gamma_i = \gamma, \alpha_i = \alpha$ holds for all i.

Models for continuous responses have also been considered by Mellenbergh (2016) including Spearman's one factor model and the model for congeneric measurements and Noel and Dauvier (2007), who proposed a beta item response model. The models considered there are, as the classical test theory, rather restrictive since the response is assumed to be a linear function of the latent traits.

1.4. Alternative Item Difficulty Functions

If difficulty functions are linear, the responses follow the distribution function $F(.)$. However, responses come with quite different distributions. They can be strictly positive, for example if the response time is an indicator of the ability of a person, or they are restricted to specific intervals, for example if a person scores in a given interval continuously or approximately continuously by using numbers, say $1, 2, \ldots, 100$. In both cases a normal distribution is inadequate, although in the latter case with numbers $1, 2, \ldots, 100$ investigators typically use a normal distribution in spite of the problems that occur at the boundaries of the interval.

A strength of the thresholds model is that it allows to account for the support of the response by using specific difficulty functions. Let the response function $F(.)$ again be the standard normal distribution and the item difficulty be given by $\delta(y) = \log(y)$. Figure 3 (left, first row) shows the person threshold functions ($\alpha = 1$) for persons with parameters $\theta = 0$ (circles), $\theta = 1$ (dashed) and $\theta = -1$ (dotted). Although a normal distribution is assumed for the response function, the response is strictly positive, and definitely not normally distributed, as is seen from the corresponding densities (right, first row).

In practice, test data are always restricted to specific finite values. A prominent case is Likert-type responses on 5 or 7-point scales. Although values are definitely discrete, often they are considered as continuous and common distributions as the normal distribution are assumed. The problem that responses at the boundary cannot follow a normal distribution is typically ignored. For truly continuous response, scales represented by continuous line segments Samejima (1973) extended the graded response model to responses to an open line segment, and Müller (1987) extended the rating scale model to responses to a closed line segment. Both extensions are derived as limiting cases of discrete response models.

The thresholds model offers an alternative way to account for the fact that data are restricted to a fixed interval, and specify a proper distribution for which the support is the interval in which responses are observed. Without loss of generality, one can choose the interval $[0, 1]$ because data can always be transformed into that interval. Then, an attractive difficulty function that can be used is the inverse function $\delta(y) = aF^{-1}(y)$ with some constant a. The second row of Fig. 3 shows the person threshold functions and the corresponding densities if the response is restricted to the interval $[0, 1]$, and $a = 1$. It is seen that densities have support $[0, 1]$ and are not normally distributed although the normal response function $F(.)$ generates the distribution. For large θ_p the distribution is shifted to the right, but still within support $[0, 1]$. There is no mis-specification of the distribution for very large or small values of θ_p, as occurs if one assumes a normal distribution for the response itself (instead of using a normal response function and appropriate difficulty functions in the thresholds model).

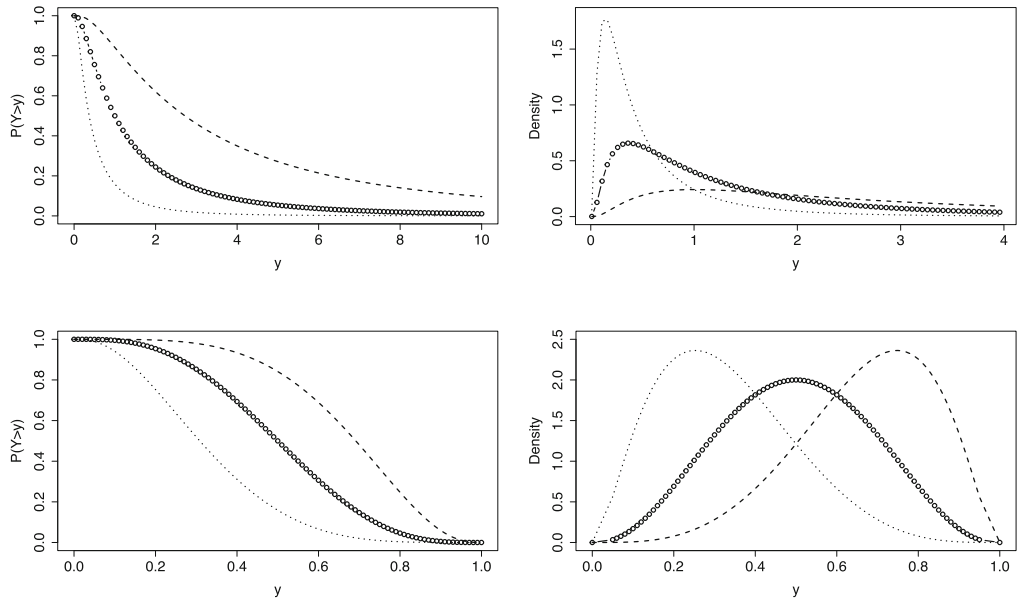

FIGURE 3.
In the first row the item difficulty is $\delta(y) = \log(y)$ (for non-negative responses), in the second row the item difficulty is $\delta(y) = F^{-1}(y)$ (responses in [0, 1]). Left column shows $P(Y > y)$ for values $\theta = 0$ (circles), $\theta = 1$ (dashed), $\theta = -1$ (dotted); right column shows the corresponding densities

1.5. Illustrative Application: Cognition Data

For illustration, we use the data set *Lakes* from the R package *MPsychoR* (Mair, 2018). It is a multi-facet G-theory application taken from Lakes and Hoyt (2009). The authors used the response to assess children's self-regulation in response to a physically challenging situation. The scale consists of three domains, cognitive, affective/motivational, and physical. We use the cognitive domain only. Each of the 194 children was rated on six items on his/her self-regulatory ability with ratings being on a scale from 1 to 7. Mair (2018) used the data to illustrate concepts of classical test theory implicitly assuming a metric scale level.

We fit a thresholds model with linear difficulty functions, normal response function and fixed discrimination parameters ($\alpha_i = 1$). The first row of Fig. 4 shows the person threshold functions for $\theta_p = 0$, under the assumption of common slopes in the difficulty functions (left) and with possibly varying slopes (right). The numbers in the curves denote the items. It is seen that items 3 and 4 are hardly distinguishable, items 2 and 6 are harder and items 1 and 5 easier. It is seen from the right picture (varying slopes) that the variance of responses is smaller for items 2 and 6 when compared to the other items, which corresponds to the large estimated slopes of items 2 and 6 in Table 1. The second row of Fig. 4 shows the corresponding IC functions. It is seen that the distance between the pairs of items $\{3, 4\}$ and items $\{2, 6\}$ is larger if the model allows for varying slopes. The last row shows the difficulty functions. They are strictly parallel in the case of a common slope. For varying slopes the pairs of items are still close to each other but the items 2 and 6 have larger slopes. Table 1 shows the estimated parameters, standard errors for the intercept were between 0.152 (item 3) and 0.329 (item 1), for the slope between 0.043 (item 3) and 0.052 (item 5) in the model with varying slopes.

Since one has nested models, it is of interest if the model with varying slopes can be simplified to the model with common slopes in the difficulty functions. The corresponding log-likelihood test is 108.34 on 5 df, which clearly indicates that the simplified model is not adequate. As mentioned

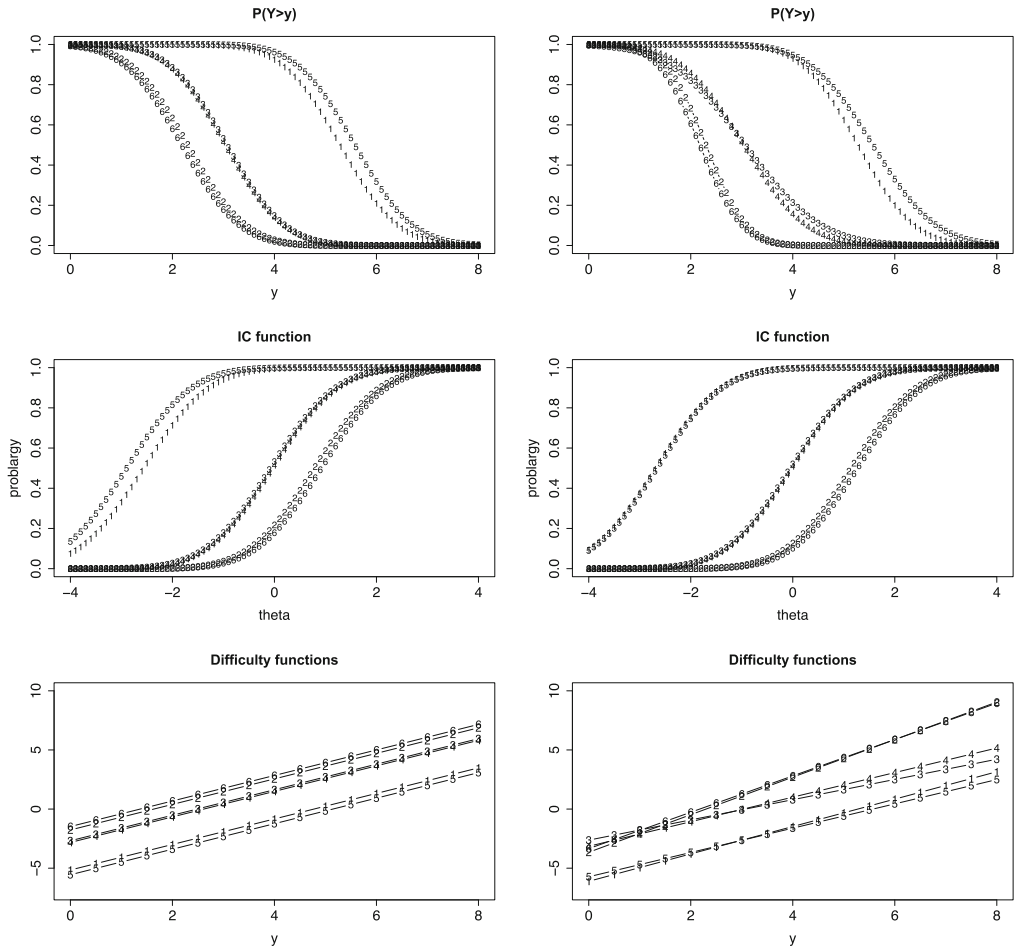

FIGURE 4.
First row: person threshold functions, $P(Y > 1)$, for cognition data ($\theta_p = 0$) and linear difficulty functions; second row: IC functions for $y = 3$; left: common slopes are assumed, right: varying slopes; third row: difficulty functions

before, testing that slopes are constant means testing if the variances of the error score are constant, which seems not to be the case. We focused on the model with fixed discrimination parameters since varying discrimination parameters did not improve the fit. More concise, the log-likelihood for the model with varying discrimination was -1445.832, which is the same value as for the model with fixed discrimination parameter. The AIC for the latter is 2917.664, for the model with varying discrimination parameter one obtains 2927.664, which clearly favors the model with fixed discrimination parameters. With the exception of item 2, for all items the parameter estimates for the more general estimates were the same as for the model with fixed discrimination parameter. For item 2 the estimates of the intercept and slope were -2.236756 and 0.9694296, and the estimate of the discrimination parameter was 1.630, for all other items the estimate was 1.

2. Discrete Responses

In the following, first it is shown that classical models for binary and ordered responses can be represented as thresholds models. Then, models with infinite support are considered.

TABLE 1.
Estimated parameters for cognition data

Item	Common slope		Varying slopes	
	Intercept	Slope	Intercept	Slope
1	−5.935405	1.123857	−6.302662	1.1934046
2	−2.592028	1.123857	−3.646508	1.5804284
3	−3.452920	1.123857	−2.635715	0.8577245
4	−3.407838	1.123857	−3.149895	1.0391432
5	−6.269576	1.123857	−5.747873	1.0304311
6	−2.451351	1.123857	−3.367810	1.5432459
Log-lik	−1500.006		−1445.832	

2.1. Binary and Ordered Categorical Responses

Let us start with the simplest case of a binary response variable $Y_{pi} \in \{0, 1\}$. Then, the only relevant value of the function $\delta_i(y)$ is $\delta_i(0) = \delta_{0i}$ because $P(Y_{pi} > 0) = P(Y_{pi} = 1)$, and if $P(Y_{pi} = 1)$ is known, all response probabilities are known. The thresholds model yields immediately the binary response model

$$P(Y_{pi} = 1|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_{0i})).$$

Thus, if $F(.)$ is chosen as normal distribution one obtains the normal-ogive model, if $F(.)$ is the logistic distribution function one obtains the 2PL model, which simplifies to the binary Rasch model if $\alpha_i = 1$ (Rasch, 1961).

The binary case makes it clear why the difficulty function is defined on the support S of Y_{pi} rather than on the whole field of real numbers. For $Y_{pi} \in \{0, 1\}$ one has to consider only $\delta_i(0)$ and $\delta_i(1)$. For the latter one has $\delta_i(1) = \infty$ since $P(Y_{pi} > 1) = 0$. For the general case see Proposition 8.2 in the "Appendix".

Let now $Y_{pi} \in \{0, \ldots, k\}$ be a response variable with ordered categories, and let the difficulty functions $\delta_i(y)$ be restricted only by the assumption that it is a strictly monotonically increasing function. Let parameters be defined by $\delta_{ir} = \delta_i(r - 1)$. Then, one obtains the thresholds model

$$P(Y_{pi} \geq r|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_{ir})), \quad r = 1, \ldots, k,$$

which is a well-known model, namely Samejima's graded response model (Samejima, 1995, 2016).

To obtain the graded response model without further constraints, it is essential that the form of the difficulty functions is restricted by the monotonicity assumption only. The monotonicity assumption itself is indispensable because otherwise the thresholds model would not be defined. Nevertheless, it is again interesting to consider the model with a pre-specified threshold function. If $\delta_i(y) = \delta_{0i} + \delta_i y$ holds, one obtains that differences between adjacent item parameters are constant, $\delta_{ir} - \delta_{i,r-1} = \bar{\delta}_i$. In this simplified version of the graded response model, each item is characterized by just three parameters, the item discrimination, the location δ_{0i} and the slope δ_i. It reduces the number of parameters in a similar way as the Rasch rating scale model (Andrich, 1978, 2016) reduces the number of parameters in the partial credit model. Simplified versions also result from using alternative *fixed* difficulty functions, for example, the log function $\delta_i(y) =$

$\delta_{0i} + \delta_i \log(y)$, which has been used above to obtain $Y_{pi} \geq 0$, or the inverse function, which can been used to restrict responses to fixed intervals.

In particular, if the number of categories is large or medium sized, as for example in a 9-point rating scale, it is tempting to assume that responses are (approximately) continuous and use corresponding modeling approaches, a strategy that is often found in applied research, see also Robitzsch (2020). The graded response model takes the support seriously, it is a model that explicitly assumes that the response is discrete and therefore follows a multinomial distribution. The thresholds model, which contains the graded response model as a special case, is quite flexible concerning the assumption of the support. In the general model formulation, $P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(y)))$, only the effect of the ability and the item difficulty function on the probability of a response above threshold y is fixed. It applies to continuous as well as discrete data. Of course, when estimating by maximum likelihood methods, one has to distinguish between the discrete and the continuous case since the densities have to be specified. However, in practice the estimated difficulty functions are very similar (see next section), the crucial part is indeed the specification of the response function $F(.)$ and the difficulty function.

The thresholds model can be seen as bridging the gap between continuous and discrete responses. The bridging can be made more explicit in the case of the graded response function. As shown in the "Appendix", there is a strong link between the continuous thresholds model and the graded response model since the thresholds model $P(Y_{pi} > y) = F(\alpha_i(\theta_p - \delta_i(y)))$ holds for continuous response Y_{pi} if and only if the graded response model holds for all categorizations

$$Y_{pi}^{(c)} = r \quad \Longleftrightarrow \quad Y_{pi} \in (\tau_r, \tau_{r+1}],$$

where $\tau_1 < \cdots < \tau_k$ are any ordered thresholds. Since the graded response model itself is a thresholds model, this means that thresholds models are stable under categorization, that is, they also hold if one considers categorized versions of the response. It should be noted that observable responses are considered, the result differs from the usual result that the graded response is a categorized version of a *latent* variable.

2.2. *Political Fears*

As an illustrating example, we consider data from the German Longitudinal Election Study (GLES), which is a long-term study of the German electoral process (Rattinger et al., 2014). The data we are using originate from the pre-election survey for the German federal election in 2017 and are with political fears. The participants were asked: "How afraid are you due to the ..."—(1) refugee crisis?—(2) global climate change?—(3) international terrorism?—(4) globalization?—(5) use of nuclear energy? The answers were measured on Likert scales from 1 (not afraid at all) to 7 (very afraid). The model is fitted under the assumption that fear is the dominating latent trait, which is considered as unidimensional. We use 200 persons sampled randomly from the available set of observations.

Figure 5 shows the person threshold functions obtained when using logarithmic difficulty functions, varying slopes and discrimination parameters. The left picture shows the fitted functions when assuming a discrete, multinomial distribution, the right picture when assuming a continuous distribution. It is seen that the fitted person threshold functions are rather similar. In both cases, varying slopes are needed (likelihood ratio test yields 39.66 for discrete distribution, 32.14 for continuous distribution on 4 df). Also varying discrimination parameters seem more appropriate (likelihood ratio test 8.844 for continuous distribution, 9.536 for discrete distribution on 4 df). Table 2 shows the estimates for the model with varying slopes and fixed discrimination parameters and the model with varying discrimination parameters.

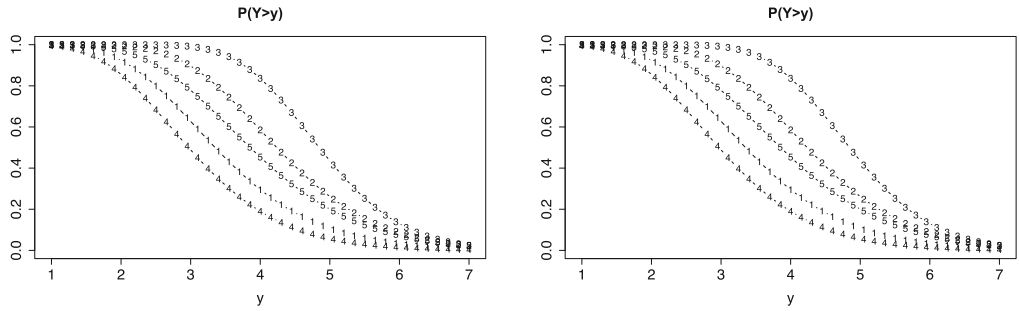

FIGURE 5.
PT functions for political fear data with logarithmic difficulty functions, varying slopes and discrimination parameters, left: discrete distribution, right: continuous distribution

TABLE 2.
Estimated parameters for fears data

Item	Fixed discrimination			Varying discrimination		
	Intercept	Slope	α_i	Intercept	Slope	α_i
1	−4.840	3.056	1	−4.353	2.749	1.086
2	−6.661	3.807	1	−5.399	3.086	1.225
3	−7.569	4.093	1	−4.730	2.557	1.868
4	−4.571	3.040	1	−3.661	2.434	1.241
5	−5.716	3.400	1	−5.424	3.226	1.000
Log-lik	−1979.395			−1974.973		

2.3. *Discrete with Infinite Support: Count Data*

Measurement of cognitive abilities often uses count data, for example, the number of remembered stimuli (Süß et al., 2002), or the number of generated ideas in a fixed time interval (Forthmann et al., 2017), for an overview see also Forthmann et al. (2020). In all these cases, the responses are counts with $Y_{pi} \in \{0, 1, 2, \dots\}$. A classical model that has been used for this kind of data is Rasch's Poisson count model Rasch (1960), which has been extended to the Conway–Maxwell–Poisson model by Forthmann et al. (2020).

The thresholds model is a flexible alternative to these models. An attractive choice of a fixed difficulty function is the log-function in the form $\delta_i(y) = \log(y + 1)$. Figure 6 shows the person threshold functions and the densities for two values of person parameters, $\theta = 1$ (bold) and $\theta = 0$ (gray), where $F(.)$ is the standard normal distribution function. It is seen that the PT function for $\theta = 1$ is always larger than the PT function for $\theta = 0$. The densities show that the persons with $\theta = 1$ tend to score higher than persons with $\theta = 0$. The IC functions are not shown since by construction they have the form of a normal distribution.

The flexibility of the count thresholds model is comparable to the Conway–Maxwell–Poisson model if the difficulty functions are specified by $\delta_i(y) = \delta_{0i} + \delta_i \log(y + 1)$ since the slope δ_i allows for additional variability of the response across items.

2.4. *Verbal Fluency Data*

Forthmann et al. (2020) used a data set with four commonly used verbal fluency tasks, which they were so kind to let us use for illustration. The data set includes two semantic fluency tasks,

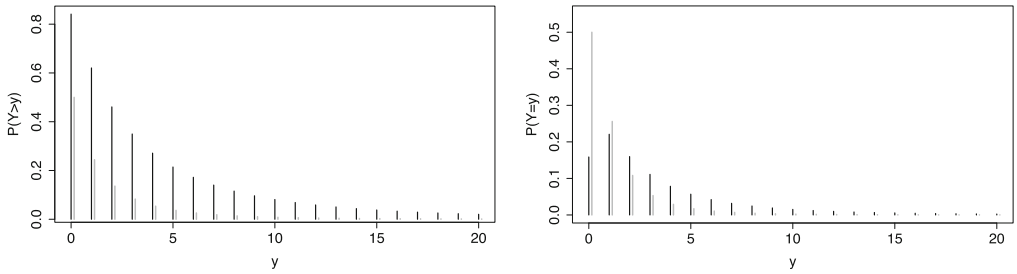

FIGURE 6.
Left: $P(Y > y)$ for values $\theta = 1$ (bold), $\theta = 0$ (gray) for count data with item difficulty function $\delta_i(y) = \log(y + 1)$; on the right-hand side the corresponding probability mass functions are shown

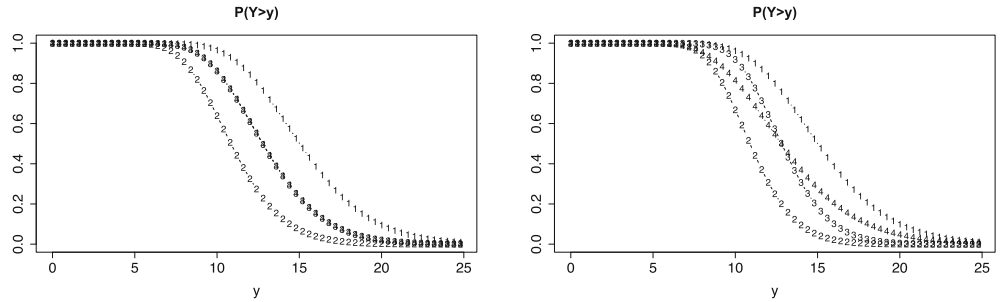

FIGURE 7.
Person threshold functions, $P(Y > y)$, for value $\theta = 0$ and $\delta_i(y) = \delta_{0i} + \delta_i \log(1 + y)$ for verbal fluency data assuming a discrete distribution; left: common slope, right: varying slopes

namely animal naming (item 1) and naming things that can be found in a supermarket (item 4) and two letter fluency tasks, words beginning with letter f (item 2) or letter s (item 3). The 202 participants had one minute to complete each of the verbal fluency tasks.

Figure 7 shows the person threshold functions for common slope (left) and for varying slope (right) if count data are considered as discrete (loglik = -2065.42, $\sigma_\theta = 1.04$ for common slope, loglik = -2038.58, $\sigma_\theta = 1.09$ for varying slopes). It is seen that under the assumption of a common slope items 3 and 4 have virtually the same threshold function, item 1 allows for higher responses, item 2 is harder, and counts tend to be lower. If slopes are allowed to vary over items, the order of items remains the same, but items 3 and 4 have slightly different functions. Item 3 shows a more distinct decrease indicating smaller dispersion than item 4. We also fitted the model with varying discrimination parameters, but there is no indication that they are needed (loglik = -2036.004, $\sigma_\theta = 0.92$).

The functions given in Fig. 7 are obtained by explicitly using the support $\{0, 1, 2, \dots\}$, and therefore assuming a discrete distribution. Thus, the curves should be interpreted only at values $\{0, 1, 2, \dots\}$, only for simplicity of presentation they were shown as continuous functions.

Since counts are on a metrical scale one could also think of fitting a model that assumes a continuous response, and consider it as an approximation. We fitted the corresponding thresholds model and obtained virtually the same functions as given in Fig. 7, which are therefore not shown. Of course the likelihood values differ from the values obtained by using a discrete model. However, inference yields similar results. The likelihood ratio test that compares the model with common slopes in the difficulty functions to the model with varying slopes is 53.68 for the discrete model and 54.08 for the continuous model (on 3 df). Thus, in both cases the model with varying slopes turns out to be more appropriate.

The thresholds model is an alternative to more classical approaches to model count data as Rasch's Poisson count model (Rasch, 1960). In additive parameterization, the model specifies the expected response μ_{pi} for person p on item i by $\mu_{pi} = \exp(\theta_p - \delta_i)$, where θ_p is the person ability and δ_i the item difficulty. For the distribution, a Poisson distribution is assumed. Fitting of the model yields the log-likelihood -2072.40 and for AIC 4154.58. Comparison with the thresholds model with varying slopes (log-likelihood -2038.06 and AIC 4094.11) shows that the thresholds model shows superior fit.

3. Further Properties and Alternative Modeling Approaches

3.1. Choice of Difficulty Functions

Item responses can be continuous or discrete, in the latter case the number of categories can be finite or infinite. As has been illustrated in the previous sections, in particular the difficulty function determines the range of the response. It therefore has to be adapted to the item type. Table 3 gives an overview of possible combinations of item types and difficulty functions, where it is assumed that the response function $F(.)$ is symmetric and continuous with support $\mathbb{R}$.

If there is reason to assume that responses follow a specific continuous distribution function $F_{\text{resp}}(.)$, for example the normal distribution, the thresholds model with response function $F(y) = F_{\text{resp}}(y)$ and linear difficulty functions is a natural choice. Then, the distribution of Y_{pi} has distribution function $F_{\text{resp}}(.)$ with the mean and the variance determined in a simple way (see Sect. 1.1). For the density, one obtains the simple form $f((y-\mu_{pi})/\sigma_{pi})/\sigma_{pi}$, where $f(.) = F'(.)$ is the density of the response, and μ_{pi}, σ^2_{pi} are the mean and variance of Y_{pi} (Proposition 8.1 in "Appendix"). Although the form of the density reminds of the normal distribution, it holds for any symmetric distribution.

For items with responses $Y_{pi} > 0$, difficulty functions should be chosen such that $\lim_{y\to 0} \delta_i(y) = -\infty$ to ensure that $Y_{pi} > 0$ holds. As an example, the logarithmic function is given in Table 3. If responses are from a known interval (a, b), difficulty functions should fulfill $\lim_{y\to a} \delta_i(y) = -\infty$, $\lim_{y\to b} \delta_i(y) = \infty$. After a transformation of the responses into the interval $(0, 1)$, natural choices are inverse distribution functions, for example, $g(y) = aF^{-1}(y)$ or the logit transformation $g(y) = \log(y/(1-y))$.

In general, for continuous responses the difficulties determine the distributions of the response. They do not necessarily follow classical response distributions. However, some classical distributions can be obtained by choosing specific difficulty functions. If one chooses the linear difficulty function, one obtains the distribution that is assumed for the response function. In particular, if the response function is the normal distribution, responses are normally distributed. A combination that also yields a classical distribution is the normal response function together with the logarithmic difficulty function. Then, one obtains for the responses the log-normal distribution with density $1/(y\sqrt{2\pi}\bar{\sigma}_i)\exp(-(\log(y) - \bar{\mu}_{pi})^2/(2\bar{\sigma}_i^2))$, where $\bar{\mu}_{pi} = (\theta_p - \delta_{0i})/\delta_i$, $\bar{\sigma}_i = 1/(\alpha_i\delta_i)$. In other combinations of response and difficulty functions, one specifies the difficulty function instead of choosing a response distribution as in more traditional modeling approaches.

If responses are discrete and have finite support, the response distribution is always the multinomial distribution. The only values of the difficulty function that enter the model are $g(0), \ldots, g(k-1)$ (for $Y_{pi} \in \{0, 1, \ldots, k\}$). In the binary case, the choice of the difficulty function is irrelevant since $\delta_{0i} + \delta_i g(0)$ can always be condensed into one intercept parameter $\tilde{\delta}_{0i} = \delta_{0i} + \delta_i g(0)$. A simple function that has been used in applications is the logarithmic function $g(y) = \log(1+y)$, which is evaluated at $0, 1, \ldots, k-1$. However, alternative functions could be constructed. One alternative is the adapted logit function $g(y) = \log((1+y)/(k-y))$.

TABLE 3.
Item types and difficulty functions

Item type	Support	Difficulty function, $\delta_i(y) = \delta_{0i} + \delta_i g(y)$	
Continuous	$Y_{pi} \in \mathbb{R}$	Linear	$g(y) = y$
	$Y_{pi} \geq 0$	Logarithmic	$g(y) = \log(y)$
	$Y_{pi} \in (0, 1)$	Inverse	$g(y) = aF^{-1}(y)$
Discrete	$Y_{pi} \in \{0, 1, \dots\}$	Logarithmic	$g(y) = \log(1 + y)$
	$Y_{pi} \in \{0, 1, \dots, k\}, k > 1$	Logarithmic	$g(y) = \log(1 + y)$
		Logit	$g(y) = \log((1 + y)/(k - y))$
	$Y_{pi} \in \{0, 1\}$		$g(y)$ not used

It is symmetric around $m = (k - 1)/2$, such that $g(m + a) = -g(m - a)$ and is steeper than $\log(1+k)$ for large values of y. The symmetry makes it attractive since it entails more symmetric distributions, for example, one obtains $P(Y_{pi} = 0) = P(Y_{pi} = k)$ for $\theta_p = \delta_{0i}$. It extends more naturally to values beyond the support. While for both functions $\lim_{y\to-1} g(y) = -\infty$ holds, at the right boundary one has $\lim_{y\to k} g(y) = \infty$ for the transformed logit function while $\log(1+k)$ is a finite value. For count data there is no right boundary and there is no need for symmetry, which means that symmetric functions have no advantage over the logarithmic function.

3.2. *Item Information*

In traditional IRT models, item information is considered a useful concept. In general item information (or Fisher information) for item i can be defined as the expectation

$$I(\theta_p) = \mathrm{E}\left(-\frac{\partial^2 l_i(Y;\theta_p)}{\partial\theta_p\partial\theta_p}\right),$$

where $l_i(Y;\theta_p)$ is the log-likelihood for item i and observation Y. For discrete $Y \in \{0, 1, \dots\}$, the log-likelihood is given by $l_i(Y;\theta_p) = \sum_r Y_r \log(\pi_{ir}(\theta_p))$, where $Y_r = 1$ if $Y = r$, $Y_r = 0$ otherwise, and $\pi_{ir}(\theta_p) = P(Y = r|\theta_p, \alpha_i, \delta_i(.))$ is the probability of a response in category r on item i. One obtains

$$I(\theta_p) = \sum_r \frac{\pi'_{ir}(\theta_p)^2 - \pi_{ir}(\theta_p)\pi''_{ir}(\theta_p)}{\pi_{ir}(\theta_p)},$$

where $\pi'_{ir}(\theta_p)$ and $\pi''_{ir}(\theta_p)$ are the first and second derivatives of $\pi_{ir}(\theta_p)$ with respect to θ_p. For example, for the Rasch model one obtains the simple form $I(\theta_p) = \pi_{i0}(\theta_p)\pi_{i1}(\theta_p) = (1 - \pi_{i1}(\theta_p))\pi_{i1}(\theta_p)$. The information in more general binary models has been considered, for example, by Lord (1980), Magis (2013).

If the number of categories is finite, $Y \in \{0, \dots, k\}$, a simpler form can be derived, in which second derivatives are not needed. By using $\pi_{i0}(\theta_p) = 1 - \pi_{i1}(\theta_p) - \cdots - \pi_{ik}(\theta_p)$ one obtains

$$I(\theta_p) = \sum_{r=1}^{k} \frac{\pi'_{ir}(\theta_p)^2}{\pi_{ir}(\theta_p)} + \frac{(\pi'_{i1}(\theta_p) + \cdots + \pi'_{ik}(\theta_p))^2}{1 - \pi_{i1}(\theta_p) - \cdots - \pi_{ik}(\theta_p)}.$$

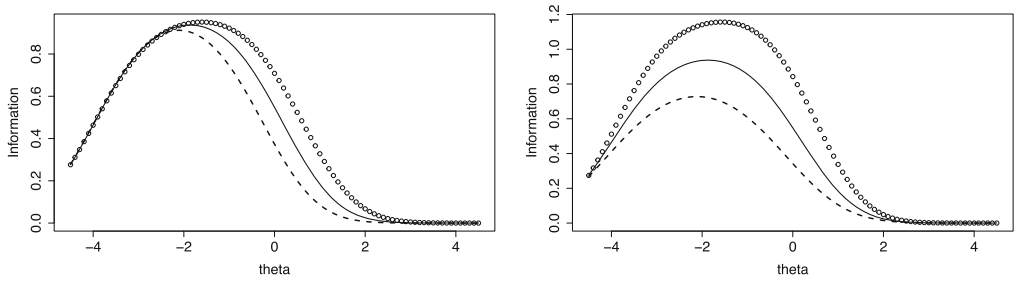

FIGURE 8.
Information functions for three items with intercepts $\delta_{0i} = -3$, item 1: dotted line, item 2: drawn line, item 3: dashed line; left: varying item slopes, $(\delta_1, \delta_2, \delta_3) = (1.2, 1.0, 0.8)$, item discrimination fixed, $\alpha_i = 1$; right: varying item discrimination, $(\alpha_1, \alpha_2, \alpha_3) = (1.2, 1.0, 0.8)$

For illustration, Fig. 8 shows the information functions for three items with 10 response categories and logarithmic difficulty functions. Since varying intercepts yield just shifted versions of the information function we let all the intercepts be the same ($\delta_{0i} = -3$). The left picture shows the information functions for item slopes $(\delta_1, \delta_2, \delta_3) = (1.2, 1.0, 0.8)$ and $\alpha_i = 1$ for all items (item 1: dotted line, item 2: drawn line, item 3: dashed line). Since item one is the hardest its peak is at larger θ-values than for the other two items. The right side shows the information function if in addition the α_is vary, $(\alpha_1, \alpha_2, \alpha_3) = (1.2, 1.0, 0.8)$. The discrimination parameter changes the range of the information function. Item 1, which has the largest discrimination parameter, yields the largest values. As has been shown for linear difficulty functions large values of α_i are linked to small variances of the response, an effect which is also present for nonlinear functions and explains why information is larger for large discrimination parameters.

For continuous responses, one obtains a closed form for the observed information only. It is given by $I_{\text{obs}}(\theta_p, Y) = -\partial^2 l_i(Y; \theta_p)/\partial\theta_p\partial\theta_p = \alpha_i^2((f'(\eta_{ipY})/f(\eta_{ipY}))^2 - f''(\eta_{ipY})/f(\eta_{ipY}))$, where $\eta_{ipY} = \alpha_i(\theta_p - \delta_i(Y))$, $f(.)$ is the derivative of $F(.)$, and $f'(.)$, $f''(.)$ the first and second derivatives. The expected observation can be obtained by numerical integration. For normal distribution function $F(.)$, it can be computed explicitly, yielding $I(\theta_p) = I_{\text{obs}}(\theta_p, Y) = \alpha_i^2$. Then, the information depends on the discrimination parameter only.

3.3. Differential Item Functioning

Differential item functioning (DIF) is the well-known phenomenon that the probability of a correct response among equally able persons differs in subgroups. For example, the difficulty of an item may depend on the membership to a racial, ethnic or gender subgroup. Then, the performance of a group can be lower because these items are related to specific knowledge that is less present in this group. Various methods have been developed to avoid the potential measurement bias and discrimination, see, for example, Millsap and Everson (1993), Zumbo (1999), Rogers (2005), Osterlind and Everson (2009) and Magis et al. (2010).

For thresholds models, DIF can be investigated in a similar way as in approaches that have been used for the Rasch model, namely by including covariates in the model. Let $\boldsymbol{x}_p$ a person-specific vector of covariates that contains, for example, gender, race, but also metric covariates like age. In a generalized thresholds model, the person parameter θ_p is replaced by $\theta_p + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i$ yielding the differential item functioning thresholds model

$$P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p + \boldsymbol{x}_p^T\boldsymbol{\gamma}_i - \delta_i(y))).$$

The parameter $\boldsymbol{\gamma}_i$ is item-specific and indicates the presence of DIF if it is unequal zero. The hypothesis $H_0 : \boldsymbol{\gamma}_i = \mathbf{0}$ can be tested by using likelihood ratio test, which is easy to do in particular if one considers DIF in items one at a time. For Rasch models, which are special cases of the thresholds model, DIF detection of this type has been considered, for example, by Paek and Wilson (2011), however restricted to binary covariates, which distinguish between a focal and a reference group. If $\boldsymbol{x}_p$ is vector-valued, and one wants to model DIF in all items simultaneously, simple marginal estimates will be hard to obtain for a larger number of items. More recently, penalty approaches have been proposed that also work for a larger number of items in Rasch models (Tutz and Schauberger, 2015). In penalty approaches, the likelihood is replaced by penalized likelihood with penalty terms that enforce selection of covariates. They work in a similar way as the penalty methods considered in Sect. 5 but with different penalty terms. The penalty terms used in these approaches can also be used in the estimation of thresholds models yielding a general concept of DIF modeling in thresholds models.

The differential item functioning thresholds model models DIF within a specific item response model as does the Rasch differential item functioning model considered by Paek and Wilson (2011). The approach has advantages over more traditional DIF detection approaches as Lord's χ^2 test (Lord, 1980) and the logistic regression method (Swaminathan and Rogers, 1990; Magis et al., 2015). The latter approach uses the test score of respondents to act as matching variable and as a proxy for respondent's ability. It assumes that test scores do represent the respondent's ability, which might hold in special models but certainly not in general.

3.4. Alternative Modeling Strategies

3.4.1. Nonparametric Item Response Models A flexible class of models are nonparametric item response models (Mokken, 1971; Junker and Sijtsma, 2001; Sijtsma and Molenaar, 2016). The binary nonparametric homogeneity model assumes only local independence, unidimensionality, and monotonicity and therefore encompasses binary thresholds models. Thresholds models are more restrictive since they assume a fixed response function $F(.)$, while in the homogeneity model the response functions can have any form provided they do not decrease. The flexibility of thresholds models refers to the distribution of the responses. In particular in models with more general difficulty functions to be considered later, the form of the distributions is hardly restricted. For binary responses, this flexibility is not exploited since the form of the response distribution is fixed to be a Bernoulli distribution.

One strength of the thresholds model is that it preserves the essential components of IRT models, but let distributions take various forms. It is able to not only fit binary responses but also count data and continuous responses, and, crucially, link it to the same latent construct. While nonparametric models provide maximal flexibility in characterizing the relationship between latent construct and item score, thresholds models do not aim at the item score, they aim at linking various possible distributions in a flexible way to person abilities.

3.4.2. Multidimensional IRT Models Multidimensional IRT models provide an alternative extension of more classical response models, see, for example, Swaminathan and Rogers (2018); Chalmers (2012). The R package *mirt* (Chalmers, 2012) allows to fit binary and multi-categorical models as the graded response model with a multi-dimensional structure. The basic concept is to replace the unidimensional trait θ_p by an m-dimensional trait (latent factors) $\boldsymbol{\theta}_p^T = (\theta_{p1}, \ldots, \theta_{pm})$. Then, in the simple binary model one uses $P(Y_{pi} = 1) = F(\alpha_{i0} + \boldsymbol{\alpha}_i^T \boldsymbol{\theta}_p)$, where α_{i0} is an item-specific intercept and $\boldsymbol{\alpha}_i$ an item-specific parameter vector. In multi-categorical models, the intercepts are category-specific, and the model contains more than one threshold.

It is straightforward to use multi-dimensional structures in thresholds models. Instead of the predictor $\eta_{pi} = \alpha_i(\theta_p - \delta_{i0} - \delta_i g(y)) = \alpha_i\theta_p - \tilde{\delta}_{i0} - \tilde{\delta}_i g(y)$, where $\tilde{\delta}_{i0} = \alpha_i\delta_{i0}$, $\tilde{\delta}_i = \alpha_i\delta_i$,

one uses $\eta_{pi} = \boldsymbol{\alpha}_i^T \boldsymbol{\theta}_p - \tilde{\delta}_{i0} - \tilde{\delta}_i g(y)$. Then, if $g(.)$ is kept flexible, for example by using basis functions, one obtains for categorical responses the multi-dimensional models used by Chalmers (2012). An advantage is that within this framework one also obtains multi-dimensional models for count data and continuous responses, and mixed item formats can be used within a test.

A caveat is that multi-dimensional models can yield paradoxical results (Jordan & Spiess, 2012), which might also be the case in multi-dimensional thresholds models. Nevertheless, specific multi-dimensional models turned out to be very useful, for example to model response styles, see Johnson and Bolt (2010), Wetzel and Carstensen (2017), Plieninger (2016), Henninger and Meiser (2020).

4. Mixed Item Formats

Tests often contain a mixture of different item formats. When measuring proficiency, the efficiency of tests can be increased by including binary items, polytomous items, continuous ones as well as count data items. The formats of items in a mixed-format test are often categorized into two classes: multiple choice (MC) and constructed response (CR). As Kim and Lee (2006) noted, typically, MC items are dichotomously scored (DS) and CR items are polytomously scored (PS).

There is a considerable body of methods of scale linking for mixed-format tests. The methods are inspired by the linkage methods for data obtained from two groups of examinees through common items (Kim & Lee, 2006). Common methods are mean/mean, mean/sigma, and Stocking–Lord linkage, see, for example, (Hanson & Béguin, 2002; Ogasawara, 2001; Kim & Hanson, 2002; Kolen & Brennan, 2014).

The thresholds model addresses the problem of different item formats in a quite different way. By construction, it assumes that there is a common latent trait that determines the outcome for all items. The model itself does not distinguish between continuous, polytomous or binary items. The only implicit assumption is that it contains order information.

The difference in item formats is captured in the difficulty functions. They determine which responses can be expected given a fixed ability parameter, and what distributional form the responses have. In the mixed formats case, it is not sensible to assume a common slope in the difficulty functions, instead slopes should vary freely, then item difficulty functions automatically adapt to the item. Resulting item functions can be quite different for, say, a dichotomous item and an item with five categories. The interpretation has to account for the type of item. For the dichotomous item, only the value $\delta(0)$ is relevant, while for a five-categories item the set $\delta(j)$, $j = 0, \ldots, 4$ determines the response. For continuous functions, the whole difficulty function is interpretable. In contrast to the case of homogeneous item-types, it is less instructive to look at the corresponding item characteristic functions since when considering $P(Y_{pi} > y)$ the value y has quite different meaning for different items.

For illustration of mixed item-formats, we consider the cognition data, in which responses range between 1 and 7. We changed the formats of two items, item 1 and item 5, to make them three-categories items by using the thresholds 4 and 6. More precisely, for the items the response is 0 if $Y_{pi} \leq 4$, 1 if $4 < Y_{pi} \leq 6$, and 2 if $Y_{pi} > 6$. Figure 9 shows the estimated difficulty and PT functions. It is seen that the difficulty functions of the other items (left picture) remain virtually the same as for the original items shown in Fig. 4 (lower right picture). As expected the difficulty functions for items 1 and 5 have changed since now different values of y are relevant. Therefore, they are given separately (right picture). Figure 9 also shows the corresponding person thresholds functions. Again the curves for item 1 and item 5 are quite different from the curves in Fig. 4 because for these items the support is different, namely 0,1,2 (corresponding to 0,4,6 on the original y-scale), but for the other items the curves are almost the same.

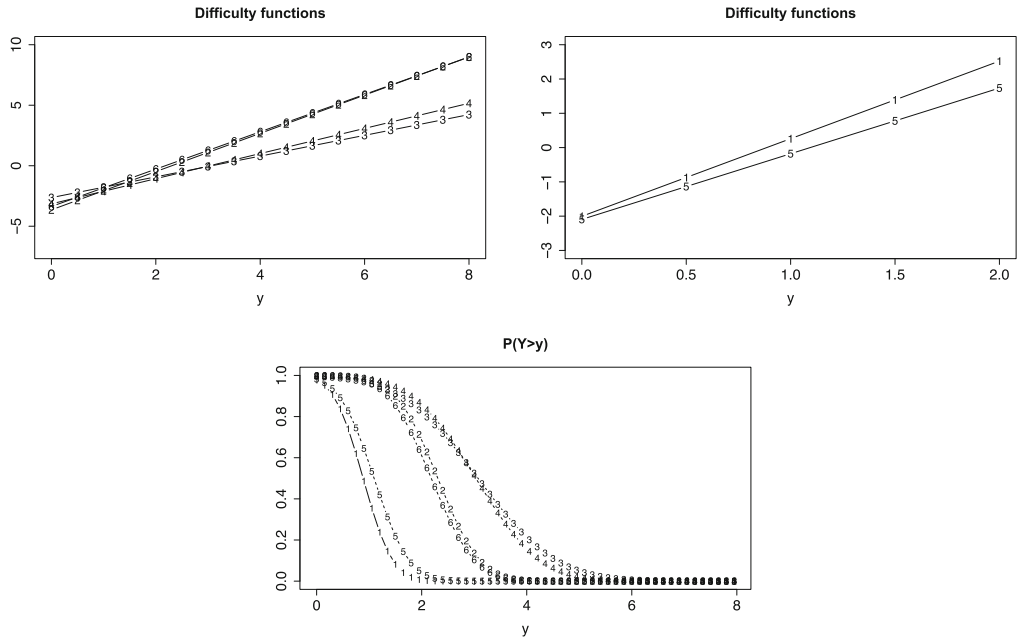

FIGURE 9.
Cognition data with items 1 and 5 as three-categories items. First row: difficulty functions, below: person threshold functions for $\theta = 0$

A similar experiment was made for the verbal fluency data which have support $0, 1, \ldots$. Items 2 and 3 have been changed to three-categories items with support 0,1,2 by using thresholds 9 and 14. Figure 10 shows the difficulty and PT functions for the mixed-formats case. As for the cognition data the PT functions for the unchanged items are very similar to the fits for the original items (see Fig. 7, right picture), but the PT functions for the items 2 and 3 have distinctly changed since the new y-values are 0,1,2, which correspond to 0,9,14 on the original response scale.

5. More General Models: Flexible Difficulty Functions

The choice of the difficulty function determines the response distribution beyond the choice of the response function. As shown before, it can in particular be used to restrict the support of the response. A fixed choice, for example by using linear difficulty functions, assumes that items differ only by intercepts and slopes (of the difficulty function). A fixed choice has the advantage that each item is determined by just two parameters δ_{0i}, δ_i, a disadvantage is that the true difficulty function and the distribution of the responses, which depends on the difficulty function, are typically unknown.

A more flexible approach that avoids that one has to choose a specific type of function, and lets the data themselves decide is obtained by letting difficulty functions be determined by basis functions, an approach that has been extensively used in statistics and machine learning (Vidakovic, 1999; Wood, 2006a; 2006b; Ruppert et al. 2009; Wand, 2000). Let us assume that the difficulty functions are given by

$$\delta_i(y) = \sum_{l=0}^{M} \delta_{il} \Phi_{il}(y), \tag{3}$$

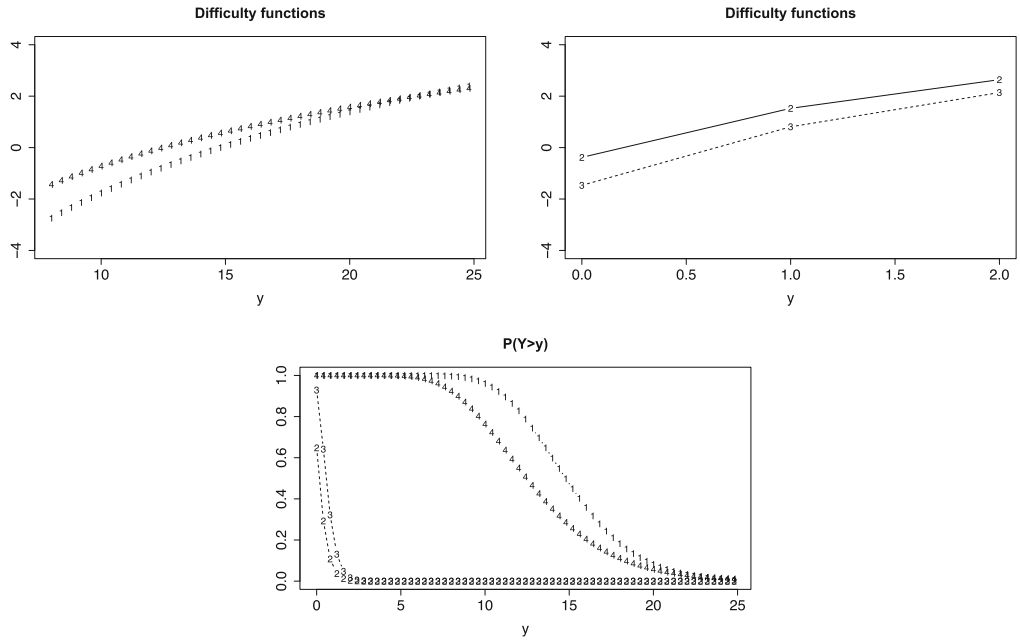

FIGURE 10.
First row: difficulty functions for fluency data with items 2 and 3 changed to three categories items, below: person threshold functions for $\theta = 0$

where $\Phi_{il}(.), l = 0, \ldots, M$ are chosen basis functions. The simple choice $\Phi_{i0}(y) = 1$, $\Phi_{i1}(y) = y$, $M = 1$ means that the item difficulty functions are linear. Much more flexible models are obtained by alternative functions as radial basis functions or spline functions. A particular attractive choice is B-splines as propagated and motivated extensively by Eilers and Marx (1996, 2021). They are very flexible and can closely approximate a variety of functions. In the literature, they were typically used to approximate functions of observable variables; here they are used to specify the unobservable difficulty functions. If difficulty functions are expanded in basis functions, they have to fulfill that they are non-decreasing, which typically calls for some restrictions. In the case of B-splines, a restriction that ensures that functions are non-decreasing is that $\delta_{i0} \leq \cdots \leq \delta_{iM}$.

If a basis, for example, B-splines have been chosen, there are two basic strategies to select the number of basis functions. One is to choose a relatively small number of basis functions, say 6 or 8, which often is enough to provide the needed flexibility. An alternative strategy is to choose a relatively large number, say 30 to 40 basis functions. Then, the number of coefficients to be estimated increases strongly, and simple log-likelihood fitting is no longer appropriate since it typically yields overfitting. Instead of fitting by maximizing the log-likelihood, one has to use penalization methods, that is, one maximizes a penalized log-likelihood, in which the differences between coefficients of adjacent basis functions are restricted to not vary too strongly, see Eilers and Marx (1996, 2021). Disadvantages are that one has to choose a tuning parameter, for example, by cross-validation, and that one has to deal with a penalized log-likelihood instead of the usual likelihood. Therefore, we use the former strategy in the examples. There is one case where penalization can not be avoided, namely when fitting flexible functions that are supposed to be the same for all items (see Sect. 6). For more general regularization methods that could be adapted to the smoothing of difficulty functions, see also Eilers and Marx (1996), Hastie and Tibshirani 1986) and Bühlmann and Van De Geer (2011).

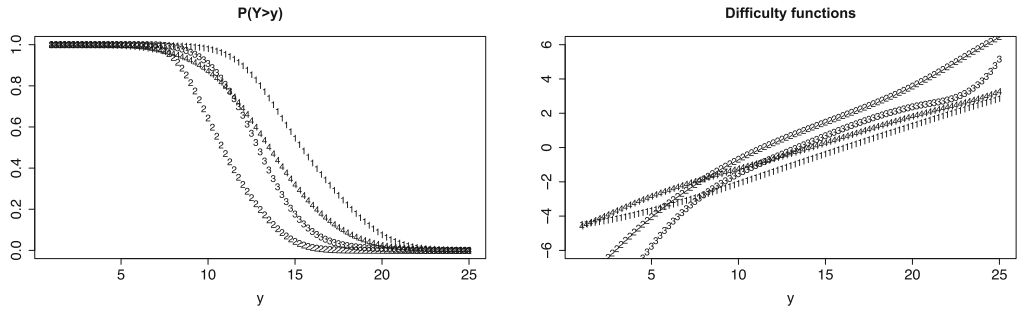

FIGURE 11.
Fitted PT (left) and difficulty (right) functions with B-spline-based difficulty functions for fluency data

If difficulty functions have the form (3), the model is parametrized by $\alpha_i, \delta_{i1}, \ldots, \delta_{iM}$, $i = 1, \ldots, I$, as item parameters, and θ_p, $p = 1, \ldots, P$, as person parameters.

Although computation is more demanding, flexible difficulties can be used as a diagnostic tool to investigate if a fixed difficulty function is appropriate. It can also be used to investigate if single items have quite different distributions. One should distinguish between two cases, difficulty function as specified in Eq. (3), which vary freely across items, and a slightly more restrictive approach, which assumes that only the location varies across items. The latter uses the simpler expansion

$$\delta_i(y) = \delta_{0i} + \sum_{l=0}^{M} \delta_l \Phi_l(y). \tag{4}$$

It assumes that the location δ_{0i} is item-specific, but the form of the function is the same for all items.

Figure 11 shows the PT functions and the fitted difficulty functions for the verbal fluency data if difficulties are not restricted (6 cubic spline functions). It is seen that the obtained PT functions are very similar to the functions obtained for fixed logarithmic difficulty functions (Fig. 7, right picture). The difficulty functions deviate somewhat from logarithmic functions but are not far away in the middle range where observations are located (not shown). Nevertheless, the AIC criterion suggests that the more flexible model should be preferred (4039.59 for the splines fit, 4094.11 for the model with logarithmic difficulty functions). However, the correlation between posterior estimates of person parameters obtained for the splines and the logarithmic model was 0.991, and estimates were very similar. Thus, one might conclude that there is no substantial improvement over the model with logarithmic difficulty functions.

Figure 12 shows the PT functions and the fitted difficulty functions for the fear data if B-splines (6 cubic spline functions) generate the difficulty functions, and a discrete distribution is used. For comparison, the second row shows the PT and difficulty functions for logarithmic difficulty functions. Though the order of the items remains the same, the form of the difficulty functions changes if splines are used instead of the logarithmic function. Also the AIC (3484.07) is distinctly smaller than the value obtained for logarithmic difficulty functions (3980.79). The correlation between posterior estimates of person parameters obtained for the splines and the logarithmic model was 0.959, and therefore smaller than for the fluency data.

Figure 13 shows the PT functions and the fitted difficulty functions for the cognition data if B-splines (6 cubic spline functions) generate the difficulty functions. It is seen that difficulty functions differ from functions obtained for linear functions (see Fig. 4) though the grouping in pairs of items is quite similar. It suggests that the response distributions deviate from the normal distribution, which is implicitly assumed by using linear difficulty functions. AIC for splines

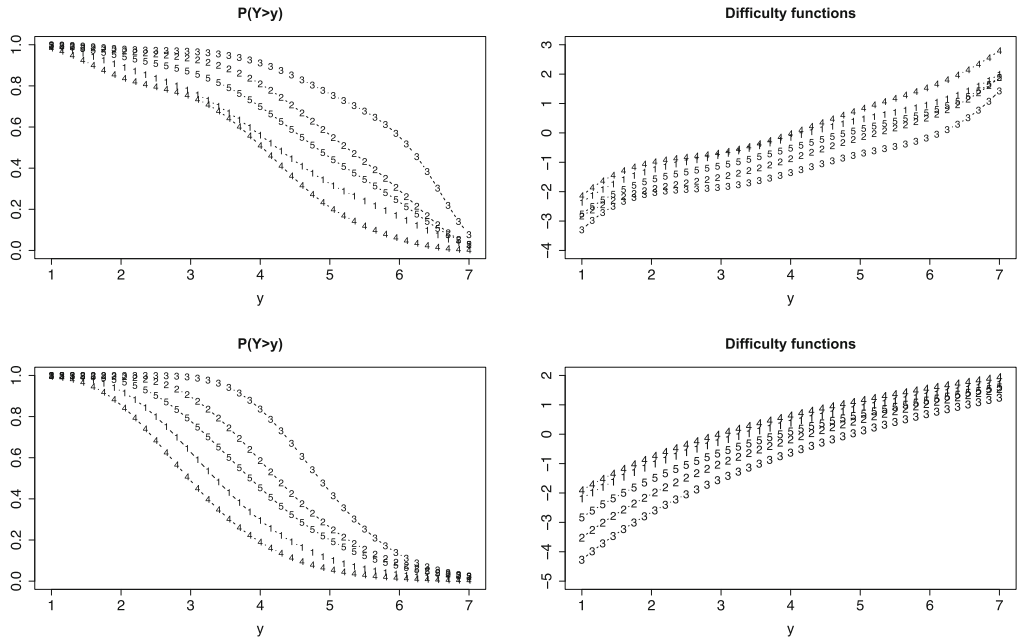

FIGURE 12.
Fitted PT (left) and difficulty (right) functions with B-spline-based difficulty functions for fear data; for comparison second row shows the difficulty functions for logarithmic difficulty functions

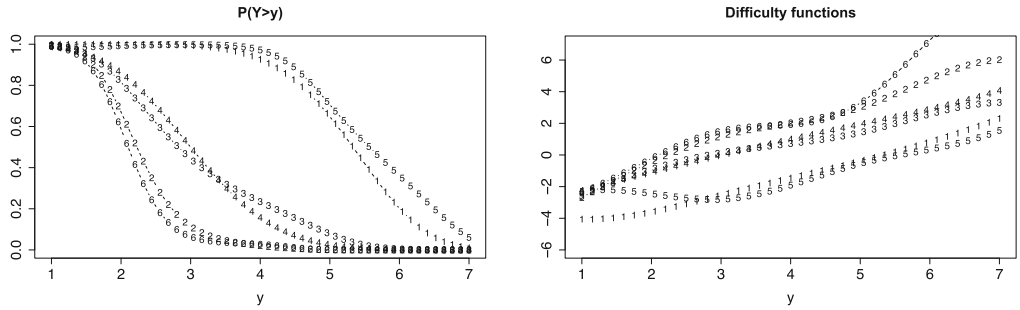

FIGURE 13.
Fitted PT (left) and difficulty (right) functions with B-spline-based difficulty functions for cognition data

was 2803.62, for varying coefficients with fixed difficulty function was 2917.66, the correlation between posterior estimates of person parameters obtained for the splines and the fixed model was 0.956.

6. Obtaining Estimates and Inference

In the following, marginal maximum likelihood methods for the estimation of item parameters and posterior estimation of person parameters are considered under the usual assumption of conditional independence of observable variables given the latent variables.

6.1. Marginal Maximum Likelihood Estimation

Let the general thresholds model hold. Then, the distribution function for observation Y_{pi} has the form:

$$F_{pi}(y) = P(Y_{pi} \leq y) = 1 - F(\alpha_i(\theta_p - \delta_i(y))).$$

For *continuous* responses, one obtains the density by building derivatives yielding

$$f_{pi}(y) = \frac{\partial F_{pi}(y)}{\partial y} = f(\alpha_i(\theta_p - \delta_i(y)))\alpha_i\delta_i'(y),$$

where $f(.)$ is the density corresponding to $F(.)$, and $\delta_i'(y) = \partial\delta_i(y)/\partial y$ is the derivative of the threshold function.

For *discrete* responses $Y_{pi} \in \{0, 1, \dots\}$, the probability mass function is obtained by building differences. Then, one has the discrete density function

$$\begin{aligned} f_{pi}(0) &= 1 - P(Y_{pi} > 0) = 1 - F(\alpha_i(\theta_p - \delta_i(0))), \\ f_{pi}(r) &= P(Y_{pi} > r - 1) - P(Y_{pi} > r) \\ &= F(\alpha_i(\theta_p - \delta_i(r-1))) - F(\alpha_i(\theta_p - \delta_i(r))), r = 1, 2, \dots \end{aligned}$$

where $\sum_r f_{pi}(r) = 1$. For simple binary responses, one obtains

$$f_{pi}(0) = 1 - F(\alpha_i(\theta_p - \delta_i(0))) \quad f_{pi}(1) = F(\alpha_i(\theta_p - \delta_i(0))),$$

where $\delta_{0i} = \delta_i(0)$ is the familiar difficulty parameter.

If difficulties are expanded in basis functions, they have the form:

$$\delta_i(y) = \sum_{l=0}^{M} \delta_{il}\Phi_{il}(y) = \boldsymbol{\Phi}_i(y)^T\boldsymbol{\delta}_i,$$

where $\boldsymbol{\Phi}_i(y)^T = (\Phi_{i0}(y), \dots, \Phi_{iM}(y))$, $\boldsymbol{\delta}_i^T = (\delta_{i0}, \dots, \delta_{iM})$. The corresponding derivative is given by

$$\delta_i'(y) = \sum_{l=0}^{M} \delta_{il}\Phi_{il}'(y) = \boldsymbol{\Phi}_i'(y)^T\boldsymbol{\delta}_i,$$

where $\boldsymbol{\Phi}_i'(y)^T = (\Phi_{i0}'(y), \dots, \Phi_{iM}'(y))$ is the vector of derivatives of basis functions.

Let now observations be given by $y_{pi}, i = 1, \dots, I, p = 1, \dots, P$. The estimation method that is used is marginal likelihood by assuming that person parameters are normally distributed, $\theta_p \sim N(0, \sigma_\theta^2)$. Maximization of the marginal log-likelihood can be obtained by integration techniques. We use numerical integration by Gauss–Hermite integration methods. Early versions for univariate random effects date back to Hinde (1982) and Anderson and Aitkin (1985).

Let $\boldsymbol{\delta}_i$ denote the vector of all parameters linked to the difficulty function of item i. For fixed difficulty functions, it has length two, for expansions in basis functions it is, more generally,

$M+1$. The vector $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_I)$ collects the item slopes. With $\boldsymbol{\delta}^T = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_I^T, \boldsymbol{\alpha}^T, \sigma_\theta)$ denoting the set of all item parameters and $f_{0,\sigma_\theta}(.)$ denoting the density of the normal distribution $N(0, \sigma_\theta^2)$, the marginal likelihood has the form:

$$L(\boldsymbol{\delta}) = \prod_{p=1}^{P} \int \prod_{i=1}^{I} f_{pi}(y_{pi}) f_{0,\sigma_\theta}(\theta_p) \mathrm{d}\theta_p,$$

yielding the log-likelihood

$$l(\boldsymbol{\delta}) = \log(L(\boldsymbol{\delta})) = \sum_{p=1}^{P} \log \left(\int \prod_{i=1}^{I} f_{pi}(y_{pi}) f_{0,\sigma_\theta}(\theta_p) \mathrm{d}\theta_p \right).$$

The score function $s(\boldsymbol{\delta}) = \partial l / \partial \boldsymbol{\delta}$, which takes on the value of 0 for maximum likelihood estimates, has components

$$\begin{aligned}
\frac{\partial l}{\partial \delta_{ij}} &= \sum_{p=1}^{P} \int \frac{\partial f_{pi}(y_{pi})}{\partial \delta_{ij}} \prod_{l \neq i} f_{pi}(y_{pl}) f_{0,\sigma_\theta}(\theta_p) \mathrm{d}\theta_p / c_p, \\
\frac{\partial l}{\partial \alpha_i} &= \sum_{p=1}^{P} \int \frac{\partial f_{pi}(y_{pi})}{\partial \alpha_i} \prod_{l \neq i} f_{pi}(y_{pl}) f_{0,\sigma_\theta}(\theta_p) \mathrm{d}\theta_p / c_p, \\
\frac{\partial l}{\partial \sigma_\theta} &= \sum_{p=1}^{P} \int \prod_{i=1}^{I} f_{pi}(y_{pi}) \frac{\partial f_{0,\sigma_\theta}(\theta_p)}{\partial \sigma_\theta} \mathrm{d}\theta_p / c_p,
\end{aligned}$$

where $c_p = \int \prod_{i=1}^{I} f_{pi}(y_{pi}) f_{0,\sigma_\theta}(\theta_p) \mathrm{d}\theta_p$. The derivation uses that the order of integration and differentiation can be interchanged, which holds if densities are continuous and continuously differentiable, in particular it holds if $F(.)$ is the normal distribution function.

The form of the derivatives depends on the distribution of the responses. For continuous responses, one obtains

$$\begin{aligned}
\frac{\partial f_{pi}(y_{pi})}{\partial \delta_{ij}} &= \alpha_i \Phi'_{ij}(y_{pi}) f(\alpha_i(\theta_p - \delta_i(y_{pi}))) - \alpha_i^2 f'(\alpha_i(\theta_p - \delta_i(y_{pi}))) \Phi_{ij}(y_{pi}) \boldsymbol{\Phi}'_i(y_{pi})^T \boldsymbol{\delta}_i, \\
\frac{\partial f_{pi}(y_{pi})}{\partial \alpha_i} &= \boldsymbol{\Phi}'_i(y_{pi})^T \boldsymbol{\delta}_i \{ f(\alpha_i(\theta_p - \delta_i(y_{pi})) + \alpha_i(\theta_p - \boldsymbol{\Phi}_i(y_{pi})^T \boldsymbol{\delta}_i) f'(\alpha_i(\theta_p - \delta_i(y_{pi}))) \},
\end{aligned}$$

with $f'(.)$ denoting the derivative of $f(.)$. For discrete responses, one has

$$\begin{aligned}
\frac{\partial f_{pi}(y_{pi})}{\partial \delta_{ij}} &= -\alpha_i f(\alpha_i(\theta_p - \delta_i(y_{pi} - 1))) \Phi_{ij}(y_{pi} - 1) + \alpha_i f(\alpha_i(\theta_p - \delta_i(y_{pi}))) \Phi_{ij}(y_{pi}), \\
\frac{\partial f_{pi}(y_{pi})}{\partial \alpha_i} &= (\theta_p - \delta_i(y_{pi} - 1)) f(\alpha_i(\theta_p - \delta_i(y_{pi} - 1))) - (\theta_p - \delta_i(y_{pi})) f(\alpha_i(\theta_p - \delta_i(y_{pi}))),
\end{aligned}$$

where $\Phi_{ij}(-1)$ is defined by $\Phi_{ij}(-1) = 0$ and $\delta_i(-1) = -\infty$.

For simple difficulty functions, the score functions simplify accordingly. For example, when the difficulty functions are linear, one has $\mathbf{\Phi}_i(y)^T = (1, y)$, and $\mathbf{\Phi}_i'(y)^T = (0, 1)$. An approximation of the covariance of the estimate, $\text{cov}(\hat{\boldsymbol{\delta}})$, is obtained by the observed information $-\partial^2 l/\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T$.

Some caution is needed when fitting the model (4) with a common difficulty function expanded in B-splines. Since B-splines sum up to 1 at any given value, the parameters in model (4) are not identified. This can be fixed by choosing a fixed value for one of the parameters δ_{0i}, for example, $\delta_{01} = 0$. One can also use the general form (3) and use a tailored penalty. Instead of maximizing the log-likelihood, one maximizes the penalized log-likelihood $l(\{\delta_i\}) = l(\{\delta_i\}) - P_\lambda(\{\delta_i\})$ with penalty term

$$P_\lambda(\{\delta_i\}) = \lambda \sum_{i=2}^{I} \sum_{l=2}^{M} [(\delta_{il} - \delta_{i,l-1}) - (\delta_{i-1,l} - \delta_{i-1,l-1})]^2.$$

The choice of λ determines the estimates. For $\lambda = 0$, one maximizes the usual log-likelihood. For $\lambda \longrightarrow \infty$, the differences of adjacent parameters become the same for all items, that is, $\delta_{il} - \delta_{i,l-1} = \delta_{i-1,l} - \delta_{i-1,l-1}$, otherwise the penalty term would also go to infinity. Then, the levels of the functions can differ but not the form of the function. Therefore, using a large value of λ automatically yields shifted difficulty functions.

For the computation of estimates, we used Gauss–Hermite integration, which works rather well since all integrals are unidimensional. The written R program uses the computation of derivatives in combination with the R function *optim*.

6.2. *Illustrative Simulation*

For illustration, we show the results of a small simulation study. Figure 14 shows the estimates for count data with varying slopes for $I = 5$, $P = 200$ and $\sigma_\theta = 1$. The dots show the true values of the parameters. The first row shows estimates if there is no variation in discrimination parameters, $\alpha_i = 1$, which is also assumed when estimating parameters. The second and third row shows estimates for varying discrimination parameters. It is seen that the parameters are estimated rather well if discrimination parameters are fixed. If they are varying, estimation becomes less accurate but is able to separate discrimination parameters from intercepts and slopes. There is no variation in the last discrimination parameter since it was fixed ($\alpha_5 = 1$).

6.3. *Estimating Person Parameters*

If estimates of item parameter are found, posterior mode or mean estimation yields estimates of person parameters. For given item responses $\boldsymbol{y}^T = (y_1, \ldots, y_I)$, the posterior is given by

$$f(\theta|\boldsymbol{y}, \boldsymbol{\delta}) = \frac{\prod_{i=1}^{I} f(\alpha_i(\theta_p - \delta_i(y_i)))\alpha_i\delta_i'(y_i) f_{0,\sigma_\theta}(\theta_p)}{\int \prod_{i=1}^{I} f(\alpha_i(\theta_p - \delta_i(y_i)))\alpha_i\delta_i'(y_i) f_{0,\sigma_\theta}(\theta_p)\mathrm{d}\theta_p}.$$

Replacing the parameter $\boldsymbol{\delta}$ by its estimate $\hat{\boldsymbol{\delta}}$ allows to compute the mode of the posterior $\hat{\theta}_m$ or the posterior mean

$$\hat{\theta}_m = \mathrm{E}(\theta|\boldsymbol{y}, \boldsymbol{\delta}) = \int \theta f(\theta|\boldsymbol{y}, \boldsymbol{\delta})\mathrm{d}\theta.$$

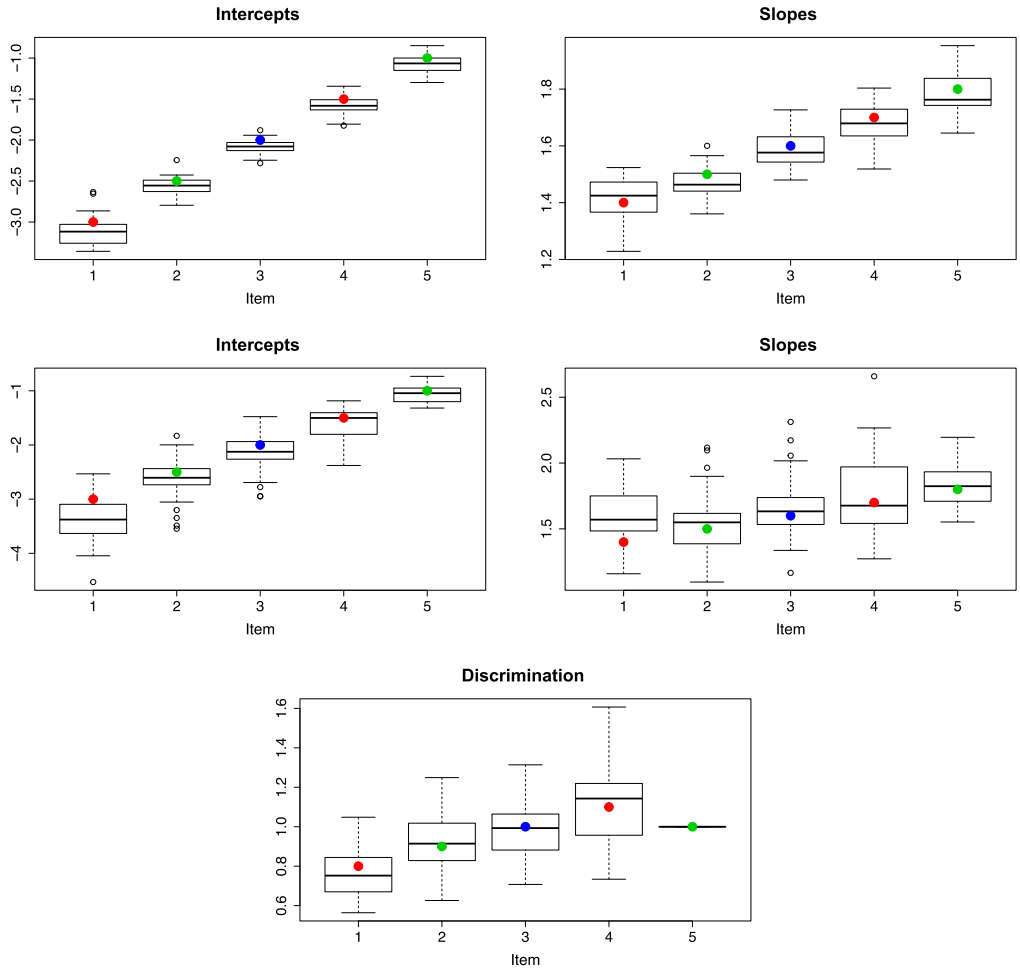

FIGURE 14.
Estimates for simulated count data; first row: fixed discrimination parameters ($\alpha_i = 1$); second and third row: varying discrimination parameters, which are also estimated

Figure 15 illustrates that posterior estimates are close to true values. Data were generated for 10 items with linear difficulty functions. The intercepts were chosen as $\delta_{i0} = -2.25 + (i-1)0.5$, $i = 1, \ldots, 10$, yielding $-2.25, -1.75, \ldots, 2.25$. The slopes of the first four items were $\delta_i = 1$, for the next four items $\delta_i = 2$, and for the remaining two items $\delta_i = 3$. For $P = 50$ and $P = 100$ with θ_p drawn randomly from $N(0, 1)$, one obtains the plots of true person parameters against estimated parameters shown in Fig. 15.

7. Concluding Remarks

The comprehensive class of thresholds models has been introduced and illustrated in examples. Also basic properties of the model class have been shown. Future research might be devoted to extensions of the model class and further investigations of its properties. As already mentioned, it is straightforward to include explanatory variables by using the additive term $\theta_p + \boldsymbol{x}_p^T \boldsymbol{\beta} - \delta_i(y)$ instead of the simple term $\theta_p - \delta_i(y)$, where $\boldsymbol{x}_p$ is a person-specific explanatory variable and $\boldsymbol{\beta}$

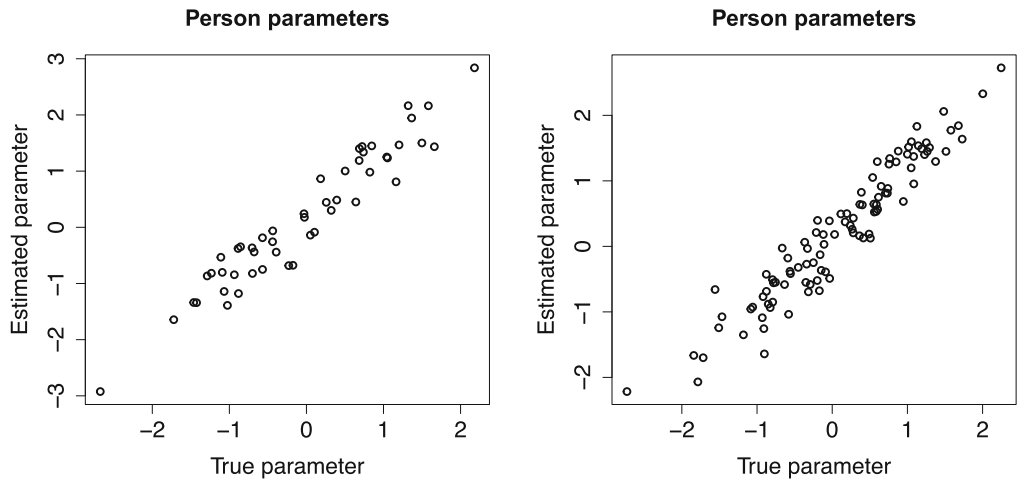

FIGURE 15.
True person parameters plotted against fitted values for simulation data for P=50 (left) and P=100 (right)

the corresponding weight. The latter can also be item-specific. The incorporation of explanatory variables can be useful to investigate sources of heterogeneity in a response scale, and has been propagated, for example, by Jeon and De Boeck (2016). Also the extension to multidimensional models is a possible topic of further research. Various general methods of model checking for categorical responses have been proposed (Swaminathan et al., 2006; Maydeu-Olivares, 2013; Haberman et al., 2013). They can also be applied to thresholds models, which, in the case of categorical responses are specific graded response models. Similar approaches might be developed for continuous responses and count data taking the specific distributions into account. It might also be useful to exploit that all thresholds models become simply structured, familiar binary models if responses are dichotomized. Then, model checking for binary models can be used, but one has to find ways how to combine the results obtained for the dichotomizations.

We restricted consideration to symmetric response functions $F(.)$. The use of a normal or a logistic response function yields very similar results, although the scaling is different. However, the use of non-symmetric distributions as, for example, the extreme value distribution might make a difference. In principle also discrete response functions could be used, the extreme case being a zero-one function as in the Guttman model; however, they include jumps that might be less realistic when assuming a continuous latent trait.

Software for the computation of marginal maximum likelihood estimates will be made available on Github.

Acknowledgments

I want to thank Pascal Jordan for all his helpful comments on a very early version of the paper. I am also grateful to the associate editor and four anonymous reviewers for their various suggestions and corrections.

Funding Information Open Access funding enabled and organized by Projekt DEAL.

Appendix

Proposition 8.1. *If the item difficulty function in the thresholds model with continuous distribution function $F(.)$ and corresponding density $f(y) = \partial F(y)/\partial y$ is linear, $\delta_i(y) = \delta_{0i} + \delta_i y$, $\delta_i \geq 0$ one obtains for the expectation and the variance*

$$\mu_{pi} = \mathrm{E}(Y_{pi}) = (\theta_p - \delta_{0i} - E_F/\alpha_i)/\delta_i, \tag{5}$$

$$\sigma^2_{pi} = \mathrm{var}(Y_{pi}) = var_F/(\alpha_i\delta_i)^2, \tag{6}$$

where $E_F = \int yf(y)\mathrm{d}y$ is the expectation corresponding to distribution function $F(.)$, and $var_F = \sigma^2_F = \int (y - E_F)^2 f(y)dy$ is the variance linked to $F(.)$
If, in addition, $F(.)$ is symmetric the distribution function of Y_{pi} is a shifted and scaled version of $F(.)$ and the density is given by $\sigma_F f(\sigma_F(y - \mu_{pi})/\sigma_{pi})/\sigma_{pi}$, which simplifies to $f((y - \mu_{pi})/\sigma_{pi})/\sigma_{pi}$ if $\sigma_F = 1$.

Proof. For linear item function, the thresholds model has the form $P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_{0i} - \delta_i y))$. The corresponding distribution function is

$$F_{Y_{pi}}(y) = P(Y_{pi} \leq y) = 1 - F(\alpha_i(\theta_p - \delta_{0i} - \delta_i y)).$$

The density is given by

$$f_{Y_{pi}}(y) = \frac{\partial F_{Y_{pi}}(y)}{\partial y} = f(\alpha_i(\theta_p - \delta_{0i} - \delta_i y))\alpha_i\delta_i,$$

yielding the expectation

$$\mathrm{E}(Y_{pi}) = \alpha_i\delta_i \int yf(\alpha_i(\theta_p - \delta_{0i} - \delta_i y))\mathrm{d}y.$$

With $\eta = \alpha_i(\theta_p - \delta_{0i} - \delta_i y)$ and $\mathrm{d}\eta/\mathrm{d}y = -\alpha_i\delta_i$ one obtains

$$\mathrm{E}(Y_{pi}) = -\int_{\infty}^{-\infty} \frac{\theta_p - \delta_{0i} - \eta/\alpha_i}{\delta_i} f(\eta)\mathrm{d}\eta = \frac{\theta_p - \delta_{0i} - E_F/\alpha_i}{\delta_i}$$

where $E_F = \int yf(y)\mathrm{d}y$ is a parameter that depends on F but not on i.

The variance is given by

$$\mathrm{var}(Y_{pi}) = \int \left(y - \frac{\theta_p - \delta_{0i} - E_F/\alpha_i}{\delta_i} \right)^2 f(\alpha_i(\theta_p - \delta_{0i} - \delta_i y))\alpha_i \delta_i \mathrm{d}y = \int (\frac{\eta - E_F}{\alpha_i \delta_i})^2 f(\eta)\mathrm{d}\eta$$
$$= \mathrm{var}_F / (\alpha_i^2 {\delta_i}^2),$$

where $\mathrm{var}_F = \int (\eta - E_F)^2 f(\eta)\mathrm{d}\eta$.
If $F(.)$ is symmetric, $E_F = 0$ and therefore $\mu_{pi} = (\theta_p - \delta_{0i})/\delta_i$. With $\sigma_F = \sqrt{\mathrm{var}_F}$, one obtains

$$f_{Y_{pi}}(y) = f(\alpha_i(\theta_p - \delta_{0i} - \delta_i y))\alpha_i \delta_i = f(\alpha_i \delta_i(\mu_{pi} - y))\alpha_i \delta_i = \sigma_F f(y - \mu_{pi})/\sigma_{pi})/\sigma_{pi}.$$

For standardized response function ($\sigma_F = 1$), one obtains the simple form $f_{Y_{pi}}(y) = f((y - \mu_{pi})/\sigma_{pi})/\sigma_{pi}$. □

Proposition 8.2. *For the thresholds model with continuous response function $F(.)$, let the difficulty functions be defined on the support S taking values from $\mathbb{R} \cup \{\infty\}$.*

(1) Person parameters and item functions are identifiable if one discrimination parameter and one person parameter are fixed, for example, $\alpha_1 = 1, \theta_1 = 0$.
(2) Person parameters and item functions are identifiable if one discrimination parameter is fixed, for example $\alpha_1 = 1$, and $\delta_i(y_0)$ is fixed for one value $y_0 \in \tilde{S}$, where $\tilde{S} = S$ if S is infinite and $\tilde{S} = \{m_1, \ldots, m_{k-1}\}$ for finite support $S = \{m_1, \ldots, m_k\}$. In the latter case, $\delta_i(m_k) = \infty$ holds for all items.

Proof. Since the response function is strictly increasing, one has

$$\alpha_i(\theta_p - \delta_i(y)) = \tilde{\alpha}_i(\tilde{\theta}_p - \tilde{\delta}_i(y)) \tag{7}$$

for all items and values $y \in S$.
Let now for two parameterizations $\theta_p, \delta_i(.)$ and $\tilde{\theta}_p, \tilde{\delta}_i(.)$ one discrimination parameter and one person parameter be fixed by $\alpha_1 = \tilde{\alpha}_1 = 1, \theta_1 = \tilde{\theta}_1 = 0$. If one chooses $\theta_1 = 0$, and accordingly $\tilde{\theta}_1 = 0$ one obtains $\alpha_i \delta_i(y) = \tilde{\alpha}_i \tilde{\delta}_i(y)$ for all items and values $y \in S$. For item 1 one obtains $\delta_1(y) = \tilde{\delta}_1(y)$, and therefore $\theta_p = \tilde{\theta}_p$. Using $\tilde{\delta}_i(y) = (\alpha_i/\tilde{\alpha}_i)\delta_i(y)$ in (7) yields $\alpha_i \theta_p = \tilde{\alpha}_i \tilde{\theta}_p$ and therefore $\alpha_i = \tilde{\alpha}_i$, from which $\delta_i(y) = \tilde{\delta}_i(y)$ follows.
Let now $\alpha_1 = \tilde{\alpha}_1 = 1$ and $\delta_i(y_0) = \tilde{\delta}_i(y_0) = 0$ for one value $y_0 \in S$ for infinite support S. Then, one obtains from (7) for item 1 and $y = y_0$ that $\theta_p = \tilde{\theta}_p$ holds for all persons. Equation (7) yields $(\alpha_i - \tilde{\alpha}_i)\theta_p = \alpha_i \delta_i(y) - \tilde{\alpha}_i \tilde{\delta}_i(y)$, which for $y = y_0$ yields $(\alpha_i - \tilde{\alpha}_i)\theta_p = 0$ and therefore $\alpha_i = \tilde{\alpha}_i$. Then, one also has $\delta_i(y) = \tilde{\delta}_i(y)$.
If the support is finite, one has to choose $\delta_i(y_0) = 0$ for $y_0 \in \{m_1, \ldots, m_{k-1}\}$ since for $\delta_i(m_k)$ one always has $\delta_i(m_k) = \infty$ because $0 = P(Y_{pi} > m_k) = F(\theta_p - \delta_i(m_k))$ has to hold for any θ_p. If $\delta_i(y_0) = 0$ is chosen accordingly (and $\alpha_1 = \tilde{\alpha}_1 = 1$), the derivation is the same as in the case of infinite support. □

Proposition 8.3. *The thresholds model $P(Y_{pi} > y) = F(\alpha_i(\theta_p - \delta_i(y)))$ holds for continuous response Y_{pi} iff the graded response model holds for all categorizations*

$$Y_{pi}^{(c)} = r \iff Y_{pi} \in (\tau_r, \tau_{r+1}],$$

where $\tau_1 < \cdots < \tau_k$ are any ordered thresholds.

Proof. Let the thresholds model $P(Y_{pi} > y|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(y)))$ hold for continuous response Y_{pi} for some increasing function $\delta_i(y)$. Let a categorized version be defined by

$$Y_{pi}^{(c)} = r \text{ if } Y_{pi} \in (\tau_r, \tau_{r+1}].$$

for any partition $\tau_0 = -\infty < \tau_1 < \cdots < \tau_k$.
One obtains $P(Y_{pi} > \tau_r|\theta_p, \alpha_i, \delta_i(.)) = F(\alpha_i(\theta_p - \delta_i(\tau_r)))$ and therefore with $\delta_{ir} = \delta_i(\tau_r)$

$$P(Y_{pi}^{(c)} \geq r) = F(\alpha_i(\theta_p - \delta_{ir})), \tag{8}$$

which is the graded response model for discrete response $Y_{pi}^{(c)}$.
Let now the discretized version (8) hold for all discretizations. Let us consider the discretization $\tau_1 < \cdots < \tau_k$ with response $Y_{pi}^{(c)}$ and parameters δ_{ir}, and the discretization $\tau_1 + \Delta < \cdots < \tau_k$ with response $Y_{pi}^{(c_\Delta)}$ and parameters $\delta_{ir}^{(c_\Delta)}$, where $\Delta < \tau_2 - \tau_1$. Let the difficulty function be defined by $\delta_i(\tau_1) = \delta_{i1}$, $\delta_i(\tau_1 + \Delta) = \delta_{i1}^{(c_\Delta)}$ to obtain

$$P(Y_{pi} > \tau_1) = F(\alpha_i(\theta_p - \delta_i(\tau_1))), \quad P(Y_{pi} > \tau_1 + \Delta) = F(\alpha_i(\theta_p - \delta_i(\tau_1 + \Delta))),$$

Since this holds for any values τ_1, Δ one obtains the thresholds model for continuous response Y_{pi}. □

References

Anderson, D. A., & Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society Series B, 47*, 203–210.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Andrich, D. (2016). Rasch rating-scale model. In W. Van der Linden (Ed.), *Handbook of modern item response theory* (pp. 75–94). Springer.

Birnbaum, A. (1986). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Addison-Wesley.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer-Verlag.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(1), 1–29.

Eilers, P. H., & Marx, B. D. (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–121.

Forthmann, B., Gühne, D., & Doebler, P. (2020). Revisiting dispersion in count data item response theory models: The Conway–Maxwell–Poisson counts model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 32–50.

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal, 29*(3), 257–269.

Glas, C., & Verhelst, N. (1989). Extensions of the partial credit model. *Psychometrika, 54*(4), 635–659.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika, 78*(3), 417–440.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models (c/r: p. 310–318). *Statistical Science, 1*, 297–310.

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods, 25*(5), 560.

Hinde, J. (1982). Compound Poisson regression models. In R. Gilchrist (Ed.), *GLIM 1982 international conference on generalized linear models* (pp. 109–121). Springer-Verlag.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085.

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics, 35*(1), 92–114.
Jordan, P., & Spiess, M. (2012). Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika, 77*(1), 127–152.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.
Kim, J.-S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*(3), 255–270.
Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53–76.
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer.
Lakes, K. D., & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research? *Journal of Clinical Child & Adolescent Psychology, 38*(1), 144–165.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.
Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304–315.
Magis, D., Bèland, S., Tuerlinckx, F., & Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862.
Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics, 40*, 111–135.
Mair, P. (2018). *Modern psychometrics with R*. Springer.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71–101.
Mellenbergh, G. J. (2016). Models for continuous responses. In W. Van der Linden (Ed.), *Handbook of item response theory* (Vol. One, pp. 181–192). Chapman and Hall/CRC.
Millsap, R., & Everson, H. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297–334.
Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Walter de Gruyter.
Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52*(2), 165–181.
Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement, 31*(1), 47–73.
Ogasawara, H. (2001). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25*(4), 373–383.
Osterlind, S., & Everson, H. (2009). *Differential item functioning* (Vol. 161). Sage Publications, Inc.
Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023–1046.
Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement, 77*, 32–53.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333).
Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., & Wolf, C. (2014). Pre-election cross section (GLES 2013). *GESIS Data Archive, Cologne ZA5700 Data file Version 2.0.0*.
Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In *Frontiers in education* (Vol. 5, pp. 177). Frontiers.
Rogers, H. (2005). Differential item functioning. In *Encyclopedia of Statistics in Behavioral Science*. Wiley Online Library.
Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics, 3*, 1193–1256.
Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203–219.
Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*(4), 549–572.
Samejima, F. (2016). Graded response model. In W. Van der Linden (Ed.), *Handbook of item response theory* (pp. 95–108). Chapman and Hall/CRS.
Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. Van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 331–350). Chapman and Hall/CRC.
Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.
Süß, H.-M., Oberauer, K., Wittmann, W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence, 30*(3), 261–288.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. *Handbook of Statistics, 26*, 683–718.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361–370.
Swaminathan, H., & Rogers, H. J. (2018). Normal-ogive multidimensional models. In W. Van der Linden (Ed.), *Handbook of item response theory*, (pp. 167–187). Chapman and Hall/CRC.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577.
Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation, 1*(1), 5.
Tutz, G. (1989). Sequential item response models with an ordered response. *British Journal of Statistical and Mathematical Psychology, 43*, 39–55.
Tutz, G. (2020). A taxonomy of ordinal item response models. Technical report, https://arxiv.org/abs/2010.01382
Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*, 21–43.
Van der Linden, W. (2016). *Handbook of item response theory*. Springer.
Van der Linden, W. (2016). Introduction. In W. Van der Linden (Ed.), *Handbook of modern item response theory* (pp. 1–9). Chapman and Hall/CRC.
Vidakovic, B. (1999). *Statistical modelling by wavelets. Wiley series in probability and statistics*. Wiley.
Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics, 15*, 443–462.
Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*, 352–364.
Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics, 48*, 445–464.
Wood, S. N. (2006). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B, 65*, 95–114.
Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.

Manuscript Received: 24 MAY 2021
Final Version Received: 24 MAR 2022
Published Online Date: 27 APR 2022