LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz & Gunther Schauberger

Uncertainty as Response Style in Latent Trait Models

Technical Report Number 217, 2018
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Uncertainty as Response Style in Latent Trait Models

Gerhard Tutz & Gunther Schauberger

Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

September 24, 2018

Abstract

It is well known that the presence of response styles can affect estimates in item response models. Various approaches to account for response styles have been suggested, in particular the tendency to extreme or middle categories has been included in the modelling of item responses. A response style that has been rarely considered is the noncontingent response style, which occurs if persons have a tendency to respond randomly and nonpurposefully, which might also be a consequence of indecision. A model is proposed that extends the Rasch model and the Partial Credit Model to account for a response style that accounts for subject-specific uncertainty when responding to items. It is demonstrated that ignoring the subject-specific uncertainty may yield biased estimates of model parameters. Uncertainty as well as the underlying trait are linked to explanatory variables. The parameterization allows to identify subgroups that differ in response style and underlying trait. The modeling approach is illustrated by using data on the confidence of citizens in public institutions.

Keywords: Rasch model; Partial credit model; Rating scales; Response styles; Ordinal data; Heterogeneity, Dispersion; Differential Item Functioning.

1 Introduction

Response styles are a problem in psychological measurement since ignoring their presence typically yields biased estimates and can affect the validity of scale scores. Models that explicitly account for response styles and model the heterogeneity in the population are able to reduce the bias and yield better inference, in particular if response styles are linked to explanatory variables.

Various methods for investigating response styles in latent trait theory have been proposed. Bolt and Johnson (2009), Johnson and Bolt (2010), Bolt and Newton (2011), and Falk and Cai (2016) use the multi-trait model to investigate the presence of a response style dimension. Johnson (2003) considered a cumulative type model for extreme response styles, Wetzel and Carstensen (2017), Plieninger (2016) and Tutz et al. (2018) proposed partial credit models that account for specific response styles. An alternative strategy for measuring response style is the use of finite mixtures. Eid and Rauber (2000) considered a mixture of partial credit models. It is assumed that the whole population can be divided into disjunctive latent classes. After classes have been identified it is investigated if item characteristics differ between classes, potentially revealing differing response styles. Finite mixture models for item response data were also considered by Gollwitzer et al. (2005) and Maij-de Meij et al. (2008). Related latent class approaches were used by Moors (2004), Kankaraš and Moors (2009), Moors (2010) and Van Rosmalen et al. (2010).

More recently, tree-based methods to investigate response styles have been proposed. In tree-based methods one assumes a nested structure, first a decision about the direction of the response is modelled and then the strength. Models of this type have been considered by Suh and Bolt (2010), De Boeck and Partchev (2012), Thissen-Roe and Thissen (2013), Jeon and De Boeck (2016), Böckenholt (2012), Khorramdel and von Davier (2014), Plieninger and Meiser (2014), Böckenholt (2017) and Böckenholt and Meiser (2017).

The focus of the present paper is on the noncontingent response style (NCR), which seems to have been neglected in the literature. The noncontingent response style is found if persons have a tendency to respond to items carelessly, randomly, or nonpurposefully (Van Vaerenbergh and Thomas, 2013; Baumgartner and Steenkamp, 2001). Although tree-based methods are strong tools it seems hard to capture this response style by tree-based methods. Given their hierarchical nature they are more appropriate to model extreme response styles or the preference for specific categories. Finite mixture model that are in common use typically fit different item response models in the components without specifying a specific structure. However, if one does not specify a response style structure in the components it is hard to identify specific response styles after fitting. Finite mixture models that should be mentioned because they do assume a specific structure in one of the components to account for uncertainty are the so-called CUB models, which have been propagated in a series of papers by Piccolo (2003), D'Elia and Piccolo (2005), Iannario and Piccolo (2016), Gottard et al. (2016), Tutz et al. (2017), Simone and Tutz (2018). The basic assumption is that the choice of a response category is determined by a mixture of a distinct preference and uncertainty. The latter is represented by a uniform distribution over the response categories. But CUB models are designed as regression models without assuming repeated measurements, they are not latent trait models, uncertainty is linked to explanatory variables, and they do not account for subject-specific

response styles.

The modelling strategy proposed in the following is the explicit modelling of the noncontingent response style by introducing subject-specific parameters that are consistent throughout items and might be determined by external explanatory variables. The proposed model explicitly aims at modelling the heterogeneity in the population. We consider in detail extensions of the partial credit model, which contains the binary Rasch model as the most important member.

2 Unobserved Heterogeneity and the Occurrence of Invalid Parameters

In the following we consider a specific form of unobserved heterogeneity that can cause severe problems in latent trait models. It is of interest because it can be seen as one of the sources of a noncontingent response style and a motivation for the model that is proposed. For simplicity we consider the binary Rasch model although the same problems are found in latent trait models with more than two response categories. The binary Rasch model assumes that the response $Y_{pi} \in \{0, 1\}$ of person p when meeting item i is determined by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}, \quad p = 1, \ldots, P, \quad i = 1, \ldots, I.$$

In achievement tests, θ_p typically represents the ability of the person and δ_i the difficulty of the item. In questionnaires θ_p may represent the attitude and δ_i an item-specific threshold on the latent scale. In both cases it is assumed that the parameters are on the same latent scale. For the identification of problems that may arise when using the Rasch model it is instructive to consider the derivation of the model from the assumption of latent random variables. When person p meets item i one assumes:

- The ability or attitude is determined by the continuous random variable $Y^*_{pi} = \theta_p + \sigma\varepsilon_{pi}$, where θ_p is a fixed parameter linked to the person, ε_{pi} is a random variable that represents the variability of the response and σ is a dispersion parameter.

- The link between the unobserved variable Y^*_{pi} and the observed response is given by

  $$Y_{pi} = 1 \quad \text{if} \quad Y^*_{pi} \geq \delta_i, \tag{1}$$

  which means that one observes $Y_{pi} = 1$ if the latent variable is larger than the item-specific threshold δ_i.

If one assumes that the noise variable ε_{pi} has the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$, it is straightforward to derive the model

$$P(Y_{pi} = 1) = \frac{\exp((\theta_p - \delta_i)/\sigma)}{1 + \exp((\theta_p - \delta_i)/\sigma))}, \quad p = 1, \ldots, P, \quad i = 1, \ldots, I.$$

Since the parameters in this representation are not identifiable, constraints on the parameters are needed. Typically one uses the *scale constraint* $\sigma = 1$ and a *location constraint* by choosing a fixed value for one of the parameters, for example, $\theta_1 = 0$ or $\delta_1 = 0$. Then the model is equivalent to the Rasch model with a location constraint, which is always needed and is assumed to be fixed in the following.

The derivation uses implicitly that the dispersion parameter σ is the same for all persons. However, this is a strong assumption that does not have to hold. Let us assume more generally that the latent variable is given by $Y^*_{pi} = \theta_p + \sigma_p \varepsilon_{pi}$ with person-specific dispersion σ_p. To keep things simple let us first consider the case where the dispersion takes only two values, depending on a binary trait like gender or age group (young/old). This can be represented by $\sigma_p = \exp(x_p\gamma)$, where x_p is a group indicator with values $x_p \in \{0, 1\}$. Then one obtains

$$\sigma_p = \begin{cases} \exp(\gamma) & \text{if } x_p = 1 \\ 1 & \text{if } x_p = 0. \end{cases}$$

If one derives the observed response in the same way as previously as a dichotomized version of latent variables one obtains different parameters for the two groups. More concrete, one obtains

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \theta_p - \delta_i, \quad \text{in the group} \quad x_p = 0$$

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \frac{\theta_p}{e^\gamma} - \frac{\delta_i}{e^\gamma}, \quad \text{in the group} \quad x_p = 1.$$

This entails peculiar effects if one wants to compare parameters. Actually one has two Rasch models, one that holds in the subpopulation $x_p = 0$ and one in the subpopulation $x_p = 1$. Formally these can be given by

$$\log\left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)}\right) = \theta_p^{(s)} - \delta_i^{(s)}, \quad s = 0, 1, \tag{2}$$

with $s = 0$ representing $x_p = 0$ and $s = 1$ representing $x_p = 1$. In the group $x_p = 0$ one has the original parameters

$$\theta_p^{(0)} = \theta_p, p = 1, \ldots, P, \quad \delta_i^{(0)} = \delta_i, i = 1, \ldots, I,$$

whereas in the group $x_p = 1$ one has the parameters

$$\theta_p^{(1)} = \frac{\theta_p}{e^\gamma}, p = 1, \ldots, P, \quad \delta_i^{(1)} = \frac{\delta_i}{e^\gamma}, i = 1, \ldots, I.$$

It is essential that in both subpopulations simple Rasch models hold. However, comparison of parameters between groups may be strongly misleading. For illustration, let x_p refer to gender with $x_p = 1$ coding females and $x_p = 0$ males. Let

us consider two persons, one female with parameter θ_f, one male with parameter θ_m, which have the same strength parameter, that is, $\theta_f = \theta_m$. If one compares the Rasch model parameters of the two persons one obtains

$$\frac{\theta_m^{(0)}}{\theta_f^{(1)}} = \frac{\theta_m}{\theta_f} e^{\gamma} = e^{\gamma}.$$

That means, if $\gamma > 0$, although the underlying abilities are the same ($\theta_f = \theta_m$), the comparison of the Rasch model parameters measured by the Rasch model parameters $\theta_p^{(s)}$ indicates that the ability of the male person is larger than the ability of the female person. The reason is that the female person is confronted with "simpler" items δ_i/e^{γ} than the male persons. Consequently, the ability of females measured in terms of the Rasch model parameters is considered to be lower for females.

It should be noted that the Rasch model does not hold in the total population. However, it holds in each subpopulation and can be legitimately fitted within subpopulations. But parameters (and parameter estimates) can not be compared since parameters in each subpopulation are scaled by using the scale constraint $\sigma = 1$ in each subpopulation.

Even if one does not want to compare parameter estimates it is obvious that one runs into problems if one ignores heterogeneity and fits a simple Rasch model to the total population. The heterogeneity of the person parameters is less severe because although the persons come from different subpopulations each person has his/her own parameter. However, estimates of item parameters tend to be biased because persons from different subpopulations meet items with different difficulty parameters. For males the difficulties are δ_i and for females δ_i/e^{γ}, which may be seen as a specific form of differential item functioning, which will be discussed later.

Similar problems with unobserved heterogeneity have been found for binary and ordinal regression models, Allison (1999) showed that misleading parameter estimates can occur if one fits a binary logit model in separate groups. Some methods to correct parameter estimates in regression were considered by Williams (2009), Mood (2010), Karlson et al. (2012), Breen et al. (2014), and Tutz (2018).

3 Heterogeneity and Response Styles

In the following we consider models that are able to avoid the occurrence of biased estimates caused by unobserved heterogeneity. We will consider the family of Rasch models represented by the partial credit model. First we briefly consider the partial credit model.

3.1 The Partial Credit Model

Let $Y_{pi} \in \{0, 1, \ldots, k\}$, $p = 1, \ldots, P$, $i = 1, \ldots, I$ denote the ordinal response of person p on item i. The *partial credit model* (PCM) assumes for the probabilities

$$P(Y_{pi} = r) = \frac{\exp(\sum_{l=1}^{r} \theta_p - \delta_{il})}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} \theta_p - \delta_{il})}, \quad r = 1, \ldots, k,$$

where θ_p is the person parameter and $(\delta_{i1}, \ldots, \delta_{ik})$ are the item parameters of item i. For notational convenience the definition of the model implicitly uses $\sum_{k=1}^{0} \theta_p - \delta_{ik} = 0$. With this convention an alternative form is given by

$$P(Y_{pi} = r) = \frac{\exp(r\theta_p - \sum_{k=1}^{r} \delta_{ik})}{\sum_{s=0}^{k} \exp(\sum_{k=1}^{s} \theta_p - \delta_{ik})}.$$

The PCM was proposed by Masters (1982), see also Masters and Wright (1984).

The defining property of the partial credit model is seen if one considers adjacent categories. The resulting presentation

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)}\right) = \theta_p - \delta_{ir}, \quad r = 1, \ldots, k$$

shows that the model is locally (given response categories $r - 1$, r) a binary Rasch model with person parameter θ_p and item difficulty δ_{ir}. It is immediately seen that for $\theta_p = \delta_{ir}$ the probabilities of adjacent categories are equal, that is, $P(Y_{pi} = r) = P(Y_{pi} = r - 1)$.

3.2 An Extended Partial Credit Model

The extended version of the partial credit model that is proposed has the form

$$P(Y_{pi} = r) = \frac{\exp(\sum_{l=1}^{r} e^{\alpha_p}(\theta_p - \delta_{il}))}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} e^{\alpha_p}(\theta_p - \delta_{il}))}, \quad r = 1, \ldots, k. \tag{3}$$

Thus, the usual predictor in the PCM, $\eta_{pir} = \theta_p - \delta_{ir}$, which distinguishes between category $r - 1$ and r, is replaced by the more general predictor

$$\eta_{pir} = e^{\alpha_p}(\theta_p - \delta_{ir}), \quad r = 1, \ldots, k,$$

which contains the additional subject-specific parameter α_p. As is discussed in the following, the parameter α_p can be seen as a subject-specific response style parameter, which describes a tendency to a specific response pattern.

Interpretation of Subject-Specific Parameters

Let us start with the simplest case of a binary response ($k = 1$). Then it is easily seen that the following holds.

> If $\alpha_p = 0$ for all p one obtains the binary Rasch model.
>
> If $\alpha_p > 0$ the person p is a strong discriminator, he/she has a distinct preference for specific categories. For $\alpha_p \to \infty$ one obtains $P(Y_{pi} = 1) = 1$ if $\theta_p > \delta_{i1}$, and $P(Y_{pi} = 0) = 1$ if $\theta_p < \delta_{i1}$.
>
> If $\alpha_p < 0$ the person p is a weak discriminator, For $\alpha_p \to -\infty$ one obtains $P(Y_{pi} = 1) = 0.5$ for all abilities/attitudes θ_p. The person shows a non-contingent response style (NCR), which means he/she has a tendency to respond to items randomly, or nonpurposefully.

In the general PCM one has to distinguish between two cases, ordered thresholds and un-ordered thresholds. In the case of *ordered thresholds* ($\delta_{ir} \leq \delta_{i,r+1}$) one obtains the following:

> If $\alpha_p = 0$ for all p one obtains the traditional PCM.
>
> For $\alpha_p \to \infty$ one obtains for a person with $\theta_p \in (\delta_{ir}, \delta_{i,r+1})$ the probability $P(Y_{pi} = r) = 1$, one observes a distinct response, the person knows exactly which category he/she prefers. The property holds for all k if one defines in addition $\delta_{i0} = -\infty$, $\delta_{i,k+1} = \infty$.
>
> For $\alpha_p \to -\infty$ one obtains $P(Y_{pi} = r) = 1/(k+1)$ for all abilities/attitudes θ_p. The person's response has a discrete uniform distribution over the response categories, which means simple guessing.

For illustration, the impact of the parameter α_p is visualized in Figure 1. It shows the response probabilities for a PCM with four categories for five different values of α_p. For $\alpha_p = 0$ one obtains the response probabilities given in the middle, which represent the response probabilities for the traditional PCM without subject-specific heterogeneity. It is seen that for decreasing α_p one comes closer to a uniform distribution across categories, whatever the parameter θ_p is, for increasing α_p the preference for categories becomes very distinct depending on the value of θ_p. The chosen parameters are rather large/small so that the impact becomes obvious.

In the case of three response categories ($k = 2$), which are considered for simplicity, and *reverse thresholds* $\delta_{i2} < \delta_{i1}$ one obtains the following behaviour.

For $\alpha_p \to \infty$ the probabilities are given by:

> For all persons one obtains $P(Y_{pi} = 1) = 0$, that is, the middle category is never chosen.

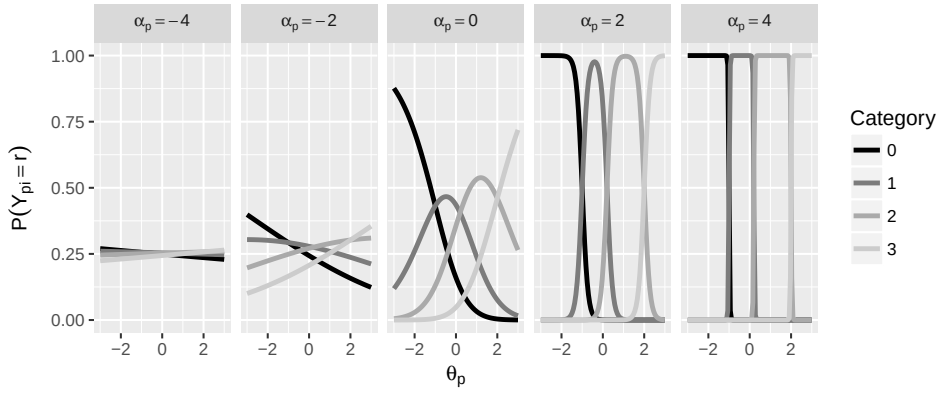

FIGURE 1: *Response probabilities in an extended PCM for four values of α_p (ordered thresholds).*

For person $\theta_p < \delta_{i2}$ one obtains $P(Y_{pi} = 0) = 1$.

For person $\theta_p > \delta_{i1}$ one obtains $P(Y_{pi} = 2) = 1$.

Thus the inverse structure of thresholds yields a more distinct avoidance of the middle category than the traditional PCM.

For $\alpha_p \to -\infty$ one obtains again $P(Y_{pi} = r) = 1/(k+1))$ for all abilities/attitudes θ_p. The person has a discrete uniform distribution over the response category, which means simple guessing.

As has been demonstrated the parameter α_p can be seen as modelling the subject-specific decisiveness or discriminatory power. For large α_p the person has distinct preferences, for small α_p the person tends to a choose one of the response categories at random which can be seen as noncontingent response style or indecision. Since it is not possible to determine if indecisiveness or carelessness is the reason we will, more generally, refer to the subject-specific effect e^{α_p} as *uncertainty effect*. Although the used terminology primarily refers to attitude measurement or personality questionnaires uncertainty may also come into play in achievement tests. The uncertainty may refer to a nonpurposeful response representing a person's ability to work concentrated or distractedly. Without specifying the specific source we consider the term e^{α_p} as representing uncertainty and call the extended model (3) the *uncertainty partial credit model* (UPCM).

The uncertainty can also explain the occurrence of response patterns that are unlikely in a unidimensional model in which uncertainty is ignored. The responses of a person with high uncertainty is hardly predictable, since he/she shows random behaviour. Therefore response patterns might occur that appear

strange in a unidimensional model that does not account for heterogeneity of uncertainty.

It should be noted that the response style parameter α_p is strongly linked to the unobvserved heterogeneity considered in Section 2. In the special case of binary responses ($k = 1$) the parameter $e^{-\alpha_p}$ in the extended model represents the unobserved dispersion σ_p in the latent variable $Y^*_{pi} = \theta_p + \sigma_p \varepsilon_{pi}$ With $\sigma_p = e^{\gamma_p}$ model, one has $\gamma_p = -\alpha_p$. This interpretation is also possible in the general PCM. If one derives the PCM from latent variables locally (given categories $r-1$, r) the same reasoning applies as for the binary Rasch model. While e^{γ_p} represents the distinctiveness of the response $e^{-\gamma_p}$ represents the uncertainty of person p.

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. In particular, 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. The effect is measurement bias and possibly discrimination. More generally, including ordinal and nominal responses, DIF is present if the response probabilities among persons with equal trait differ in subgroups. Various forms of differential item functioning have been considered in the literature, see Magis et al. (2010) for an instructive overview of DIF detection methods.

Differential item functioning usually aims at identifying those items that have different difficulties in differing subgroups. It is typically assumed that just *some* of the available items show this property. This is different in the case considered here when persons have varying uncertainty represented in the factor e^{α_p}. If the parameter α_p is linked to a binary variable like gender one obtains that the effective parameters of all *all* items are modified in one subgroup. Similar as in Section 2 let us consider the binary model and let $\alpha_p = \alpha$ if $x_p = 1$ and $\alpha_p = 0$ if $x_p = 0$. Then the 'effective' person and item parameters ($e^{\alpha_p}\theta_p$ and $e^{\alpha_p}\delta_i$) in group $x_p = 0$ are

$$\theta_1, \ldots, \theta_P \quad \delta_1, \ldots, \delta_I,$$

and in group $x_p = 1$

$$e^{\alpha}\theta_1, \ldots, e^{\alpha}\theta_P \quad e^{\alpha}\delta_1, \ldots, e^{\alpha}\delta_I.$$

Thus, even when the underlying abilities θ_p are the same in both groups the probability of a correct response differs in the groups, which corresponds to the general definition of DIF that the probability of a correct response among equally able persons differs in subgroups. Nevertheless, it should be seen as a specific form of DIF, which could be called *uniform DIF across items*.

One further consequence of the modification of person and item parameters is the reduced possibility of comparing item difficulties. If persons have different

response styles, that is, different parameters α_p, the person parameters θ_p cannot be compared directly. Only for persons p and $\tilde{p}$, which have the same slope parameter ($\alpha_p = \alpha_{\tilde{p}}$), the log-odds can be compared directly by

$$\frac{P(Y_{pi} = r)/P(Y_{pi} = r-1)}{P(Y_{\tilde{p}i} = r)/P(Y_{\tilde{p}i} = r-1)} = e^{\theta_p - \theta_{\tilde{p}}}.$$

The Generalized Partial Credit Model

It is noteworthy that the extended partial credit model considered here differs from the *generalized partial credit model* as considered by Muraki (1992), Muraki (1997). It assumes

$$\log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r-1)}\right) = a_i(\theta_p - \delta_{ir}), \quad r = 1, \ldots, k,$$

which includes an item-specific slope parameter a_i, not a subject-specific parameter. In the generalized partial credit model the items have differing discriminatory power. In contrast the uncertainty partial credit model considered here allows a subject-specific uncertainty parameter, which means that discriminatory power varies across persons.

3.3 Including Subject-Specific Characteristics

In the extended PCM each person has its own response style parameter α_p, which yields a large number of parameters. Thus, for estimation it is useful to assume that they are random effects. In the light of differential item functioning it is of special interest to investigate if response styles (or equivalently dispersion heterogeneity) is determined by subject-specific covariates. To this end we let the response style parameter depend on a vector of subject-specific covariates $\boldsymbol{x}_p$ in the form

$$\alpha_p = \alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha},$$

and assume that α_{p0} follows the normal distribution $\mathbb{N}(0, \sigma^2)$. In the same way one can include explanatory variables for the trait parameter by using

$$\theta_p = \theta_{p0} + \boldsymbol{x}_p^T \boldsymbol{\xi}.$$

Thus the general uncertainty partial credit model (UPCM) we consider is

$$P(Y_{pi} = r) = \frac{\exp(\sum_{l=1}^{r} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}}(\theta_{p0} + \boldsymbol{x}_p^T \boldsymbol{\xi} - \delta_{il}))}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}}(\theta_{p0} + \boldsymbol{x}_p^T \boldsymbol{\xi} - \delta_{il}))}, \quad r = 1, \ldots, k.$$

Figure 2 shows the resulting response probabilities if a binary predictor (male: $x_p = 1$, female: $x_p = 0$) is included in the location part and the dispersion part

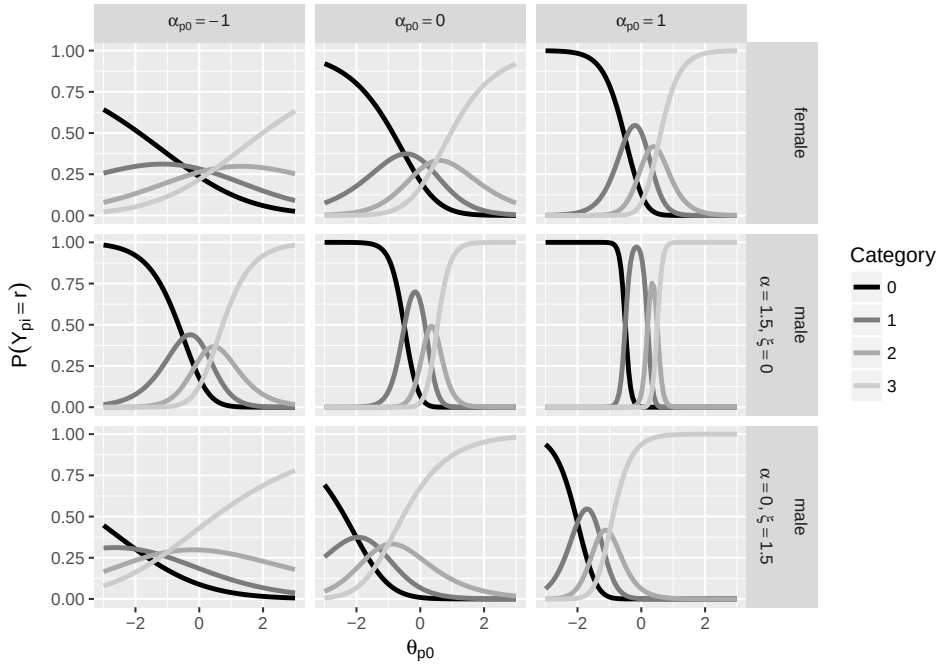

FIGURE 2: *Response probabilities in an extended PCM with a binary predictor.*

of an extended PCM with parameters $\alpha = 1.5$ and $\xi = 0$ for the middle row of the plot and parameters $\alpha = 0$ and $\xi = 1.5$ for the bottom plot. The first row shows the effect of α_{p0}, larger values increase the distinctness, smaller values decrease distinctness. In the middle row one sees the probabilities resulting from an additional dispersion effect $\alpha = 1.5$, which makes all responses more distinct in the male population. In the third row the location/trait effect $\xi = 1.5$ is visualized. It increases the probabilities for higher categories since the trait is stronger in the male population.

4 Estimation

To reduce the number of parameters one assumes that the uncertainty parameters are drawn from a normal distribution $N(0, \sigma_\alpha^2)$. The corresponding marginal likelihood with $\boldsymbol{\delta}^T = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_I^T)$ is

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma^2) = \prod_{p=1}^{P} \int P(\{Y_{p1}, \ldots, Y_{pI}\}) f(\alpha_{p0}) d\alpha_{p0},$$

where $f(\alpha_{p0})$ is the density $N(0, \sigma_\alpha^2)$ of the random effects, $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_{P-1})$, $\boldsymbol{\delta}_i^T = (\delta_{i1}, \ldots, \delta_{ik})$. The corresponding log-likelihood simplifies to

$$l(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\delta}, \sigma^2) =$$
$$\sum_{p=1}^{P} \log \left(\int \prod_{i=1}^{I} \prod_{r=1}^{k} \{ \frac{\exp(\sum_{l=1}^{r} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}} (\theta_{p0} + \boldsymbol{x}^T \boldsymbol{\xi} - \delta_{il}))}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}} (\theta_{p0} + \boldsymbol{x}^T \boldsymbol{\xi} - \delta_{il}))} \}^{y_{pir}} f(\alpha_{p0}) d\alpha_{p0} \right),$$

where $y_{pir} = 1$ if $Y_{pi} = r$ and $y_{pir} = 0$ otherwise.

Maximization of the marginal log-likelihood can be obtained by integration techniques.

Typically one first wants to obtain good estimates of the item parameters and estimate person parameters later for the validated test tool. Therefore, one also assumes a distribution for the person effects, which yields the marginal likelihood

$$L(\boldsymbol{\delta}, \boldsymbol{\Sigma}) = \prod_{p=1}^{P} \int P(\{Y_{p1}, \ldots, Y_{pI}\}) f(\alpha_{p0}, \theta_{p0}) d\alpha_{p0} \; d\theta_{p0},$$

where $f(\alpha_{p0}, \theta_{p0})$ now denotes the two-dimensional density of the person parameters, $N(\mathbf{0}, \boldsymbol{\Sigma})$. The diagonals of the matrix $\boldsymbol{\Sigma}$ contain the variance of the response style parameters σ_γ^2 and the variance of the person effects, σ_α^2, the off diagonals are the covariances between response style and location effects, $\text{cov}_{\alpha\theta}$. The corresponding log-likelihood is

$$l(\boldsymbol{\delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\Sigma}) =$$
$$\sum_{p=1}^{P} \log \left(\int \prod_{i=1}^{I} \prod_{r=1}^{k} \{ \frac{\exp(\sum_{l=1}^{r} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}} (\theta_{p0} + \boldsymbol{x}^T \boldsymbol{\xi}_p - \delta_{il}))}{\sum_{s=0}^{k} \exp(\sum_{l=1}^{s} e^{\alpha_{p0} + \boldsymbol{x}_p^T \boldsymbol{\alpha}} (\theta_{p0} + \boldsymbol{x}^T \boldsymbol{\xi}_p - \delta_{il}))} \}^{y_{pir}} f(\alpha_{p0}, \theta_{p0}) d\alpha_{p0} \; d\theta_{p0} \right).$$

The embedding into the framework of generalized mixed models allows to use methods that have been developed for this class of models. One strategy is to use joint maximization of a penalized log-likelihood with respect to parameters and random effects appended by estimation of the variance of random effects, see Breslow and Clayton (1993) and McCulloch and Searle (2001). However, joint maximization algorithms tend to underestimate the variances and, therefore, the true values of the random effects. An alternative strategy, which is used here, is numerical integration by Gauss-Hermite integration methods. For an overview on estimation methods for generalized mixed model see McCulloch and Searle (2001) and Tutz (2012). The likelihood can be maximized numerically, and also the corresponding Hessian can be approximated numerically for the final parameter estimates. This allows for the calculation of (numerically approximated) standard errors.

5 Simulations

We conducted a small simulation study to evaluate the performance of the method and the possible consequences of ignoring the response style. We used $n = 300$ observations on $I = 10$ items with each item having $k = 5$ categories. The data were simulated under the assumption that the uncertainty partial credit model (UPCM) holds. As explanatory variables we used one binary variable and one continuous variables drawn from a standard normal distribution. We fix the respective effects of the explanatory variables to $\boldsymbol{\xi}^T = (0.2, -0.1)$ for the trait effects and $\boldsymbol{\alpha}^T = (-0.2, 0)$ for the response style effects. Furthermore, the covariance matrix of the random effects was fixed to

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}.$$

The simulation was conducted with 100 replications. Figure 3 compares the estimates of the item parameters of a regular PCM to the item parameter estimates obtained for the UPCM. The boxplots show the respective estimates together with the true values, separately for each item and separately for PCM and UPCM. True values are highlighted by (red) crosses. It can be seen, that in contrast to the UPCM, the regular PCM estimates are biased.

Figure 4 displays the estimates of the random effects covariance matrix $\boldsymbol{\Sigma}$. Again the estimates can be compared to estimates from the regular PCM, however obviously the PCM only provides estimates for the random effect of the trait. While the PCM clearly underestimates the variance of the trait effects, the UPCM estimates all parameters reasonably well. Figure 5 displays the estimates of all covariate effects, both for trait and response style effects. All effects are estimated rather well by the UPCM model.

6 An Application

For illustration, we consider data from the ALLBUS, the general survey of social science carried out by the German institute GESIS (`http://www.gesis.org/allbus`). The data contain the answers of 2535 respondents from the questionnaire in 2012. In particular, we consider 8 items that refer to the degree of confidence the participants have in public institutions and organizations. These institutions are the federal court, the Bundestag (parliament), the justice system, TV, press, government, police and political parties. The items are measured on a scale from 1 (no confidence at all) to 7 (excessive confidence). As explanatory variables for the trait and for the response style effects we used the following person characteristics:

Age: Age of participant in years

Gender: 0: male; 1: female

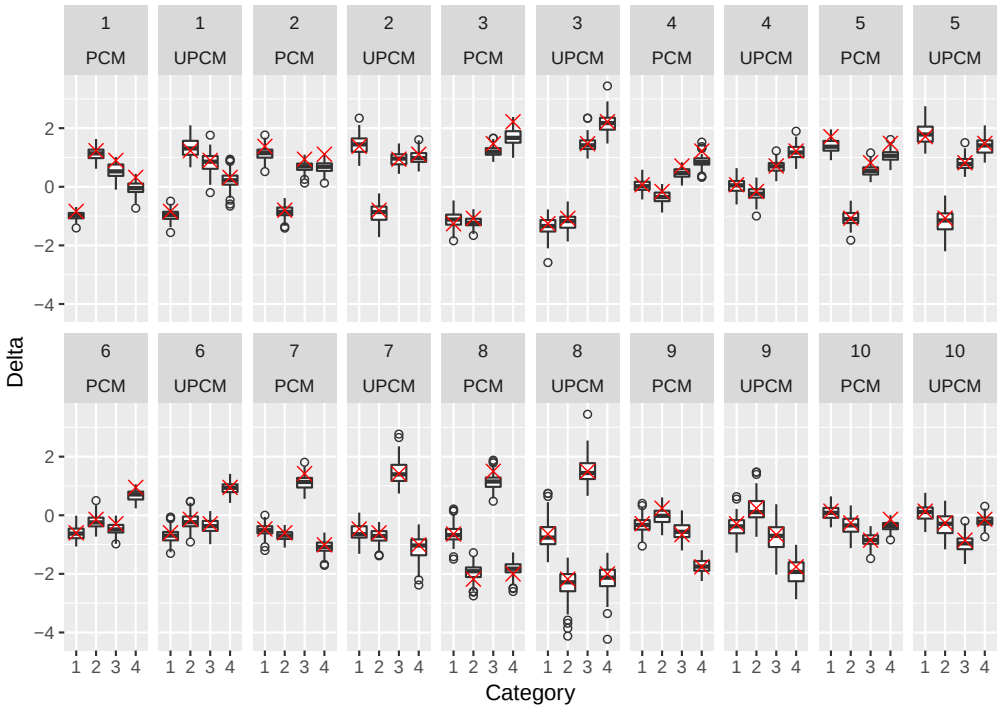

FIGURE 3: *Boxplots for estimates of the four threshold parameters δ_{ir} together with true values (red crosses). Estimates are displayed separately for all 10 items and both for the regular PCM model and the UPCM model.*

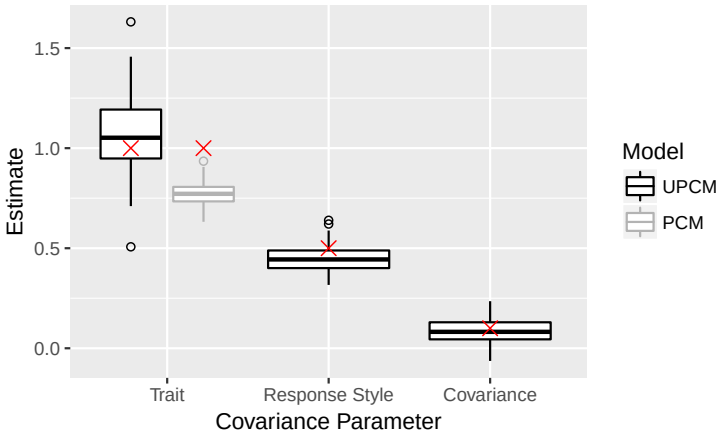

FIGURE 4: *Boxplots for estimates of random effects covariance parameters from covariance $\boldsymbol{\Sigma}$ together with true values (red crosses). PCM only entails a random effect for trait effects, the respective estimates are shown for comparison.*

Income: Income of participant in Euros

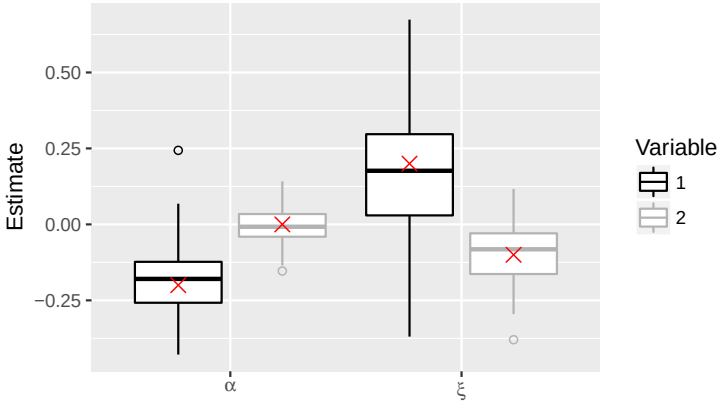

FIGURE 5: *Boxplots for estimates of covariate effects ξ for trait effects and α for response style effects together with true values (red crosses) and separately for both explanatory variables.*

WestEast: 1: East Germany/former GDR; 0: West Germany/former FRG

To ensure that all covariate effects are comparable in their size all variables were standardized. We applied both a simple Partial Credit Model (PCM) and the UPCM to the data. While in the PCM the variance of the random effect for the trait parameters was estimated to be $\hat{\sigma}^2 = 0.736$, when fitting the UPCM the covariance matrix was estimated as

$$\hat{\Sigma} = \begin{pmatrix} 0.917 & 0.039 \\ 0.039 & 0.423 \end{pmatrix}.$$

While there seems to be no correlation between both random effects it seems that the random response style effect can not be neglected. It should be noted that this refers only to the random effect response style component. This effect is modelled in addition to the effects of the covariates on the response style.

Figure 6 displays the estimates of the item parameters of both the simple PCM and the proposed UPCM. It can be seen that in particular the estimates for the exterior thresholds differ between both models while the estimates for the inner thresholds are rather similar.

Table 1 collects the parameters estimates of both the trait effects and the response style effects of the explanatory variables together with the corresponding p-values. It is seen that with the exception of the gender and age effects on confidence all effects have to be considered as relevant.

For the interpretation of the effects we propose a visualization tool, which is in particular helpful, when many explanatory variables are available. For the motivation let us consider again the uncertainty partial credit model, which can

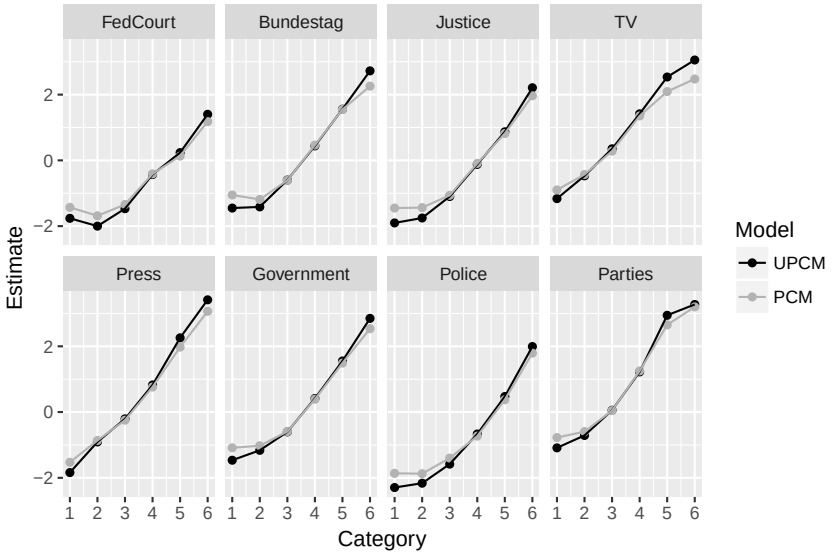

FIGURE 6: *Item parameter estimates for confidence data, separately for simple PCM and the proposed UPCM.*

	ξ (Response style)	α (Trait)
Income	0.051 (0.000)	0.056 (0.004)
Gender	-0.004 (0.847)	0.044 (0.020)
Age	-0.034 (0.092)	-0.056 (0.002)
WestEast	-0.156 (0.000)	-0.039 (0.040)

TABLE 1: *Parameter estimates for effects of explanatory variables (together with p-values), both for trait effects ξ and for response style effects α.*

be given by

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right) = e^{\alpha_{p0}+\boldsymbol{x}_p^T\boldsymbol{\alpha}}(\theta_{p0}+\boldsymbol{x}_p^T\boldsymbol{\xi}-\delta_{ir}).$$

From this representation it is seen that the person and item parameters determine the log-odds of observing category r rather than category $r-1$. One obtains

a multiplicative effect e^{α_j} if the j-th variable increases by one unit, and

a location effect that shifts the second part of the predictor by ξ_j if the j-th variable increases by one unit

We plot for each variable the effect point (ξ_j, e^{α_j}) together with 0.95 confidence intervals in both direction, which yields stars (Figure 7). The no-effects

reference point is $(0, e^0) = (0, 1)$. The abscissa represents the effect on traits, values on the right (larger than zero) indicate that the trait increases with increasing variable values, values on the left (below zero) indicate that the trait decreases when the variable increases. It is seen that higher income increases confidence in public institutions and people living in the former east (WestEast=1) tend to have reduced confidence. It is also seen that age tends to reduce the confidence, although the effect is not significant at the 0.05 level, since the star crosses the zero line $\xi_j = 0$. The ordinate represents the uncertainty or random behaviour. Large values (above 1) indicate distinctness of the response, small values (below 1) indicate indecision. Income increases distinctness, also females have a tendency to a more distinct response. Increasing age reduces the distinctness of the response, also people from the former west show higher uncertainty. In summary, only income and WestEast appear to have distinct effects on the general trait level while all variables show significant effects with respect to the response style.

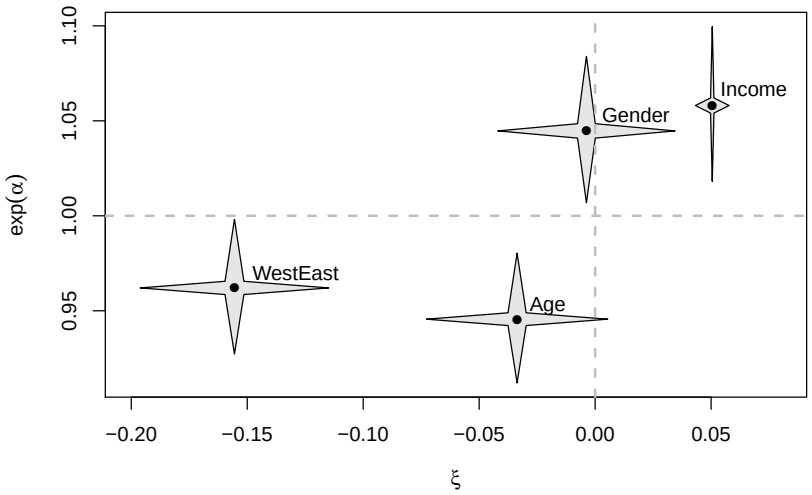

FIGURE 7: *(Exponential) effects of explanatory variables in ALLBUS data together with confidence intervals both for trait effects ξ and response style effects α*

.

7 Alternative Item Response Models

The partial credit model is an extension of the binary Rasch model, but not the only one. Also Samejima's graded response model (Samejima, 2016) and the sequential model (Tutz, 1989; Verhelst et al., 1997) are extensions of the binary model, which contain the Rasch model as special cases. In the same way as

the partial credit model these models can be extended to contain an additional subject-specific uncertainty component. It is less interesting in the sequential model, which assumes a step wise solving of items, but is sensible in the case of the graded response model, which can be derived from an underlying latent trait and works well in personality questionnaires and attitude scales. The graded response model has the form

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

where $F(.)$ again is a cumulative distribution function, typically chosen as the logistic function. The extended version assumes for the probabilities

$$P(Y_{pi} \geq r) = F(e^{\alpha_p}(\theta_p - \delta_{ir})), \quad r = 1, \ldots, k, \tag{4}$$

with e^{α_p} representing the subject-specific factor. However, some caution is warranted when interpreting the subject-specific term. It differs from the corresponding term in the partial credit model. The way the subject-specific term modifies the response probabilities is seen best when looking at the extreme cases. One obtains the following properties.

For $\alpha_p = 0$ one obtains the traditional graded response model.

For $\alpha_p \to \infty$ one obtains for a person with $\theta_p \in (\delta_{ir}, \delta_{i,r+1})$ the probability $P(Y_{pi} = r) = 1$, that means a person knows exactly what he/she wants.

For $\alpha_p \to -\infty$ one obtains $P(Y_{pi} = 0) = P(Y_{pi} = k) = 0.5$.

In particular the last case ($\alpha_p \to -\infty$) shows that the subject-specific term has a different meaning in the graded response model. Persons with $\alpha_p \to -\infty$ choose one of the extreme categories, which means they show what is called an extreme response style (ERS). Thus when going through the continuum between $\alpha_p = -\infty$ to $\alpha_p = \infty$

one covers the continuum between an extreme response style and a distinct response.

For the partial credit model with a subject-specific term

one covers the continuum between a uniform distribution, which means uncertainty, and a distinct response.

The difference in interpretation is caused by the specific property of the partial credit model that modification of the local responses (given $Y \in \{r-1, r\}$) modifies automatically all the other response probabilities. The extended graded response model is in itself of interest but refers to a different response style and

is not further investigated here. The graded response model with a subject-specific factor as given in (4) was considered previously by Ferrando (2009), for alternative models see also Ferrando (2014).

An interesting model is the binary Rasch model with a subject-specific term, which is a special case of both extensions. For the binary Rasch model $\alpha_p \to -\infty$ means that $P(Y_{pi} = 0) = P(Y_{pi} = 1) = 0.5$. This is hardly an extreme response style, it means a simple random choice from the alternatives $Y \in \{0, 1\}$. Therefore, the interpretation is in line with the interpretation of the extended partial credit model, not the extended graded response model. The underlying reason is that for binary responses the notion of an extreme response style is not sensible.

It should be noted that subject-specific factors for binary models were considered before. The model proposed by Reise (2000) has been critically discussed by Conijn et al. (2011). The latter investigated in particular problems with the representation as a multilevel logistic regression model. More recently, Ferrando (2016) proposed a normal-ogive model that contains item and person discrimination parameters. The presence of two factors makes more difficult estimation procedures necessary. Ferrando (2016) proposes a two-step approach, which works only under rather specific assumptions.

The model proposed here differs from the models proposed by Ferrando and others in several respect. We consider extensions of the partial credit model, not the graded response model. Moreover, we include explanatory variable and use marginal estimation methods that allow that the slope parameters can be correlated with content related parameters. The additional parameters are considered as response style parameters, the model is embedded into the framework of continuous response style modeling.

8 Concluding Remarks

The extended uncertainty partial credit model that has been proposed adds a subject-specific uncertainty component to the traditional PCM. It can be used to investigate if response styles are determined by person characteristics. Ignoring the uncertainty component can yield biased estimates.

The model differs from multi-trait models that account for response styles by using a linear predictor with some of the components describing response styles, models that have been proposed, among others, by Plieninger (2016) and Wetzel and Carstensen (2017). In contrast to these model a multiplicative predictor is specified with one of the factors representing the response style, the other the difference between trait and item parameter. The multiplicative structure is also found in Muraki's extended partial credit model, but in a quite different way. The UPCM specifies a subject-specific uncertainty whereas in Muraki's model the slope parameters may vary across items. Consequently, estimation methods

for the two models are quite different.

References

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological methods & research 28*(2), 186–208.

Baumgartner, H. and J.-B. E. Steenkamp (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research 38*(2), 143–156.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods 17*(4), 665–678.

Böckenholt, U. (2017). Measuring response styles in likert items. *Psychological Methods* (22), 69–83.

Böckenholt, U. and T. Meiser (2017). Response style analysis with threshold and multi-process irt models: A review and tutorial. *British journal of mathematical and statistical psychology 70*(1), 159–181.

Bolt, D. M. and T. R. Johnson (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement 33*(5), 335–352.

Bolt, D. M. and J. R. Newton (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement 71*(5), 814–833.

Breen, R., A. Holm, and K. B. Karlson (2014). Correlations and nonlinear probability models. *Sociological Methods & Research 43*(4), 571–605.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Conijn, J. M., W. H. Emons, M. A. van Assen, and K. Sijtsma (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research 46*(2), 365–388.

De Boeck, P. and I. Partchev (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software 48*(1), 1–28.

D'Elia, A. and D. Piccolo (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis 49*, 917–934.

Eid, M. and M. Rauber (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment 16*(1), 20.

Falk, C. F. and L. Cai (2016). A flexible full-information approach to the modeling of response styles. *Psychological methods 21*(3), 328.

Ferrando, P. J. (2009). A graded response model for measuring person reliability. *British Journal of Mathematical and Statistical Psychology 62*(3), 641–662.

Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate behavioral research 49*(4), 390–405.

Ferrando, P. J. (2016). An IRT modeling approach for assessing item and person discrimination in binary personality responses. *Applied psychological measurement 40*(3), 218–232.

Gollwitzer, M., M. Eid, and R. Jürgensen (2005). Response styles in the assessment of anger expression. *Psychological assessment 17*(1), 56.

Gottard, A., M. Iannario, and D. Piccolo (2016). Varying uncertainty in CUB. *Advances in Data Analysis and Classification 10*(2), 225–244.

Iannario, M. and D. Piccolo (2016). A comprehensive framework of regression models for ordinal data. *Metron 74*(2), 233–252.

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

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika 68*(4), 563–583.

Johnson, T. R. and D. M. Bolt (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.

Kankaraš, M. and G. Moors (2009). Measurement equivalence in solidarity attitudes in europe insights from a multiple-group latent-class factor approach. *International Sociology 24*(4), 557–579.

Karlson, K. B., A. Holm, and R. Breen (2012). Comparing regression coefficients between same-sample nested models using logit and probit: a new method. *Sociological Methodology 42*(1), 286–313.

Khorramdel, L. and M. von Davier (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research 49*(2), 161–177.

Magis, D., S. Bèland, F. Tuerlinckx, and P. Boeck (2010). A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior Research Methods 42*(3), 847–862.

Maij-de Meij, A. M., H. Kelderman, and H. van der Flier (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*.

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

Masters, G. N. and B. Wright (1984). The essential process in a family of measurement models. *Psychometrika 49*, 529–544.

McCulloch, C. and S. Searle (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European sociological review 26*(1), 67–82.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. a multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review 20*(4), 303–320.

Moors, G. (2010). Ranking the ratings: A latent-class regression model to control for overall agreement in opinion research. *International Journal of Public Opinion Research 22*(1), 93–119.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series 1992*(1).

Muraki, E. (1997). A generalized partial credit model. *Handbook of modern item response theory*, 153–164.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica 5*, 85–104.

Plieninger, H. (2016). Mountain or molehill? a simulation study on the impact of response styles. *Educational and Psychological Measurement 77*, 32–53.

Plieninger, H. and T. Meiser (2014). Validity of multiprocess irt models for separating content and response styles. *Educational and Psychological Measurement 74*(5), 875–899.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research 35*(4), 543–568.

Samejima, F. (2016). Graded response model. In W. Van der Linden (Ed.), *Handbook of item response theory*, pp. 95–108.

Simone, R. and G. Tutz (2018). Modelling uncertainty and response styles in ordinal data. *Statistica Neerlandica*.

Suh, Y. and D. M. Bolt (2010). Nested logit models for multiple-choice item response data. *Psychometrika 75*(3), 454–473.

Thissen-Roe, A. and D. Thissen (2013). A two-decision model for responses to likert-type items. *Journal of Educational and Behavioral Statistics 38*(5), 522–547.

Tutz, G. (1989). Sequential item response models with an ordered response. *British Journal of Statistical and Mathematical Psychology 43*, 39–55.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.

Tutz, G. (2018). Binary response models with underlying heterogeneity: Identification and interpretation of effects. *European Sociological Review, published online (https://doi.org/10.1093/esr/jcy001)*.

Tutz, G., G. Schauberger, and M. Berger (2018). Response styles in the partial credit model. *Applied Psychological Measurement, published online*.

Tutz, G., M. Schneider, M. Iannario, and D. Piccolo (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification 11*(2), 281–305.

Van Rosmalen, J., H. Van Herk, and P. Groenen (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research 47*(1), 157–172.

Van Vaerenbergh, Y. and T. D. Thomas (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research 25*(2), 195–217.

Verhelst, N. D., C. Glas, and H. De Vries (1997). A steps model to analyze partial credit. In *Handbook of modern item response theory*, pp. 123–138. Springer.

Wetzel, E. and C. H. Carstensen (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment* (33), 352–364.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research 37*(4), 531–559.