A Graded Response Model Framework for Questionnaires With Uniform Response Formats

Applied Psychological Measurement
2019, Vol. 43(4) 290–302
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618789394
journals.sagepub.com/home/apm

Dirk Lubbe¹ and Christof Schuster¹

Abstract

Questionnaires with uniform-ordered categorical response formats are widely applied in psychology. Muraki proposed a modified graded response model accounting for the items' uniform response formats by assuming identical threshold parameters defining the category boundaries for all items. What is not well known is that there is a set of closely related models, which similarly assume identical thresholds. The present article gives a framework illustrating the differences between these models and their utility for understanding questionnaire responses in detail. The models are explained as constrained cases of a one-dimensional factor model for ordered categorical data. Furthermore, the authors show that the models can be written and fitted as structural equation models, which allows for a very flexible and general purpose use. Instructions on implementing the models in Mplus and SAS PROC NL MIXED are given.

Keywords

polytomous item response, graded response model, category thresholds, latent trait, item response theory, slope–scale model

Introduction

Responses to questionnaire items are prevalently scored using ordered categorical response scales. The categories of these scales are organized in terms of agreement or preference, such as “strongly disagree,” “disagree,” “undecided,” “agree,” and “strongly agree.” Usually, a uniform response format is used for all items. This suggests that response categories of all items may be treated similarly—at least to some degree.

Likert (1932) introduced an approach regarding the psychometric properties of all items as identical. In a Likert-type scale, all response categories are assigned consecutive positive integer values. The unweighted sum of these values are used as test score. Proceeding in this way assumes (a) that intervals between response categories are identical for all items and categories and (b) that item discriminations are also identical. However, these assumptions are seldom justified for ordered categorical data. Especially, assuming identical intervals between response categories is critical.

¹Justus-Liebig-Universität Giessen, Germany

Corresponding Author:

Dirk Lubbe, Department of Psychology, Justus-Liebig-Universität Giessen, Otto-Behaghel-Straße 10 F1, 35394 Gießen, Germany.

Email: Dirk.Lubbe@psychol.uni-giessen.de

As an alternative to the Likert-type scale approach, there are various latent trait models available, specifically accounting for an ordered categorical response format. A prominent example is the *graded response model* (GRM, Samejima, 1969). As opposed to the Likert-type scale approach, the GRM does not make any assumptions on the intervals between response categories. It estimates one threshold parameter between consecutive response categories of each item, thus achieving a saturated description of their locations, in the sense that no constraints are imposed on them. Although this may seem preferable in comparison with not accounting for different intervals between response categories, a saturated description might overfit the systematic variation of intervals between thresholds. Especially, if either the number of response categories is large or the sample is small, modeling too many parameters may affect the reliability and reproducibility of the estimates.

A compromise between the Likert-type scale approach and the GRM is the so-called “rating scale parametrization” (e.g., Muraki, 1990). Similar to the GRM, threshold parameters are estimated, allowing for different intervals between response categories. However, the same thresholds are applied to all items. In this way, the uniform response format across items as well as the graded measurement are accounted for. In addition, the number of parameters is considerably reduced in comparison with the saturated description of threshold locations of the GRM.

Clearly, assuming identical thresholds across items is a feasible approach for modeling a uniform category structure. However, depending on the model’s parametrization, different results are obtained. Specifically, the item discrimination parametrization is of central importance in this regard.

There are two ways to parametrize the discriminations. Depending on statistical tradition and software they are modeled either as (a) slope parameters or (unstandardized) loadings, or (b) as scale parameters. As a matter of fact, the authors will explain below that discriminations can also be considered as the ratio of slope and scale (see Equation 5). Although slope parameters are common in factor analysis, scale parameters are generally used in item response theory.

In the present context, the distinction between the two parametrizations has not drawn much attention. One reason is that the choice of either parametrization is arbitrary for most models of ordered categorical responses. Consequently, slopes and scales are usually treated as interchangeable. An early example for this are models by Muraki (1990) and Skrondal (1996). The two proposed similar but distinct modifications to the GRM by specifying identical thresholds across items. Although Muraki used a scale parameter, Skrondal used a slope parameter. The authors refer to Muraki’s modified GRM as MGRM1 and to Skrondal’s model as MGRM2. Although the two proposals are different, Skrondal, referring to Muraki, did not claim having proposed a new model.

In this article, the authors will explain why slopes and scales are distinct, if the same thresholds are assumed for all items. In addition, they will derive a third model that contains both slope and scale parameters, which is denoted as slope–scale model (SSM). Although such a model has never been formally published, Skrondal and Rabe-Hesketh (2004) have pointed out that the original proposal of Muraki (1990) might have been too restrictive. They showed that if slope parameters are estimated freely in connection with constrained thresholds, the model is still identified if the error variance of one item instead of all items is fixed. However, they did not elaborate further on this model and did not provide information on model fitting and parameter interpretation.

On the basis of the three models (MGRM1, MGRM2, and SSM), the authors will outline the implications of differently specified discriminations. They consider the additional choice of models with differing complexity useful for two reasons. First, there are more options for finding an economic, yet, well-fitting description of the data. Especially in cases in which the sample is small or the number of response categories is large, different types of constrained models

should assist that purpose. Second, depending on similarities in item phrasing or content, different models might suggest themselves. The more homogeneous these aspects are, the more likely will a model having a restrictive threshold structure fit the data well. In this regard, comparing different models can be useful to investigate similarities among items more clearly.

This article is organized as follows. First, the functions of slope and scale parameters are illustrated within the well-known underlying variable approach for ordered categorical data (e.g., Bartholomew, Martin, & Irini, 2011). Second, the specific models are explained and their differences are demonstrated. Third, parameter estimation and fit assessment are considered. Fourth, an empirical data set of questionnaire responses is analyzed. Finally, the results are summarized and discussed. In addition, a simulation study investigating the reliability of parameter estimates and inferential conclusions using robust weighted least squares (WLS) estimation as well as program code for Mplus and SAS PROC NL MIXED are provided in the online supplements.

Factor Analytic Model for Ordered Categorical Responses

As basis for illustrating the dependency between the item discrimination parametrization and assuming identical thresholds across items, the authors use the underlying variable approach for ordered categorical responses as described by Muthén and Asparouhov (2002). Its advantage is that all models regarded in this article can be expressed as linear factor models with different constraints on either slope or scale parameters.

For modeling observed categorical responses within the underlying variable approach, a continuous latent response is assumed. This latent response may be regarded as true response, from which only the categorical format is observed. A one-dimensional factor model for latent responses y_j^* of item j , with $j = 1, \dots, p$, is

$$y_j^* = \lambda_j \eta + \sigma_j \varepsilon_j, \quad (1)$$

where λ_j is the *slope* for y_j^* on trait η , and σ_j is the *scale* of residual ε_j . Standard distributional assumptions are (a) $\eta \sim N(0, 1)$, (b) $\varepsilon_j \sim N(0, 1)$, and (c) mutual independence of trait and residuals, for $j = 1, \dots, p$. It follows that $y_j^* \sim N(0, \lambda_j^2 + \sigma_j^2)$.

To relate y_j^* to an observed $(q + 1)$ categorical response y_j , the y_j^* scale is subdivided using q threshold parameters τ_{jk} , with $k = 1, \dots, q$. In addition, it holds that $\tau_{j0} = -\infty$ and $\tau_{j(q+1)} = +\infty$. Assigning the response categories with consecutive integer values from 0 to q , threshold τ_{jk} lies between adjacent categories $k - 1$ and k . The conditional probability of selecting category k or any category above k equals the probability of the latent response exceeding the lower threshold τ_{jk} , that is,

$$P(y_j \geq k | \eta) = P(y_j^* > \tau_{jk} | \eta). \quad (2)$$

For the factor model in Equation 1, these conditional probabilities can be derived by replacing y_j^* with the equation's right side and solving for ε_j :

$$P(y_j \geq k | \eta) = P(\varepsilon_j > \sigma_j^{-1}(-\lambda_j \eta + \tau_{jk}) | \eta).$$

Because ε_j is assumed standard normal, the expression can be simplified further, using $P(\varepsilon > z) = P(\varepsilon < -z)$, which applies to symmetrical distributions around zero. Consequently, the following is obtained:

Figure 1. Linear relation between the latent ability η and the latent response y^* .

Note. Scale σ gives the residual distribution's standard deviation, λ is the regression line's slope, and τ_1 and τ_2 mark the category thresholds on the y^* axis.

$$P(y_j \geq k|\eta) = \Phi\left(\sigma_j^{-1}(\lambda_j\eta - \tau_{jk})\right), \quad (3)$$

where Φ is the cumulative distribution function of the normal distribution.

Finally, the probability of selecting a specific category k equals the difference between the probability of exceeding its lower and its upper threshold, that is,

$$P(y_j = k|\eta) = P(y_j \geq k|\eta) - P(y_j > k|\eta). \quad (4)$$

Because category probabilities are obtained as probability difference, models based on this approach are commonly classified as difference models (e.g., Thissen & Steinberg, 1986).

Figure 1 summarizes the outlined principles for an exemplary three-categorical item. Consequently, there are two τ parameters marking the category thresholds on the y^* axis. The relation between η and y^* is displayed by the regression line with slope λ . The scale σ is displayed as the standard deviation of the residual distribution for $\eta = 0$. The probability of selecting a specific category, given $\eta = 0$, is displayed by differently shaded areas under the residual distribution.

Both, scale σ and slope λ measure the item's discrimination. If σ is large, then the item's discrimination is small, because category probabilities do not change considerably for different levels of η . On the contrary, the smaller the value of σ gets, the more closely a specific category is linked to a specific trait-level. The same relation can be expressed using slope λ . If λ is small, then the linear relation between trait and response is small accordingly. Consequently, similarly to a large-scale parameter, different levels of η do not correspond to an unambiguous

preferences for specific categories. A large λ , on the contrary, denotes a stronger differentiation between different response categories, as does a small-scale parameter. Thus, slope and scale are inversely related. If both parameters are free to vary, then they jointly determine the item discrimination a , as given by the following:

$$a = \frac{\lambda}{\sigma} \quad (5)$$

Slope and Scale Models

Models using either slopes or scales are identical, if the following relations hold true for all items:

$$\lambda_j = \frac{1}{\sigma_j} \text{ and } \tau_{jk} = \frac{\tilde{\tau}_{jk}}{\sigma_j}, \quad (6)$$

where $\tilde{\tau}_{jk}$ is threshold of a model with scales. This follows from Equation 3, if first σ_j is set equal to 1.0 and then λ_j is set equal to 1.0. Equating these results and comparing coefficients yields Equation 6. It is commonly known that the GRM cannot distinguish between slopes and scales.

For specifying identical thresholds across items, Muraki (1990) and Skrandal (1996) used the following constraint:

$$\tau_{jk} = \mu_j + \tau_k,$$

where the $p \times q$ thresholds are expressed by p item-specific location parameters μ_j and q threshold parameters τ_k , which are identical across items. For model identification 1, τ_k needs to be fixed. For instance, one can either fix their sum to 0.0 or fix a single threshold to zero instead. Thus, only $q - 1$ threshold parameters are free to vary. The relations given in Equation 6 now become

$$\lambda_j = \frac{1}{\sigma_j}, \mu_j = \frac{\tilde{\mu}_j}{\sigma_j}, \text{ and } \tau_k = \frac{\tilde{\tau}_k}{\sigma_j}, \quad (7)$$

where again the tilde on top of a parameter indicates that it belongs to a model having a scale rather than a slope parameter. The last equation shows that unless σ_j is constant across items, this equation cannot be true, because the thresholds are required to be uniform across items. Thus, if σ_j (or similarly λ_j) is not identical for all items, models using the two parametrizations yield different model fits to a given data set.

It follows that MGRM1 and MGRM2 are distinct models. Specifically, the two models can be written using Equation 3 by imposing constraints on either slopes or scales.

MGRM1 is obtained by fixing all λ_j to 1.0.

$$\Phi\left(\sigma_j^{-1}\left(\eta - \mu_j - \tau_k\right)\right). \quad (8)$$

MGRM2 is obtained by fixing all σ_j to 1.0 instead.

$$\Phi\left(\lambda_j\eta - \mu_j - \tau_k\right). \quad (9)$$

Figure 2. Category probabilities of two three-categorical items for MGRM1 and MGRM2 for given trait η .
 Note. The solid line denotes the category characteristic of Category 1, the dashed line of Category 2, and the dotted line of Category 3. MGRM = Modified graded response model.

In addition, this implies the possibility to estimate the parameters of a hybrid model, the SSM, jointly. The SSM is

$$\Phi\left(\sigma_j^{-1}\left(\lambda_j\eta - \mu_j - \tau_k\right)\right), \quad (10)$$

where the first item's scale is fixed to 1.0, for model identification.

To illustrate the differences between models in more detail, Figure 2 gives the category characteristic curves (CCC) of MGRM1 and MGRM2 for two three-category items. The authors begin with the parameters pertaining to MGRM1. Its item parameters are $\sigma_1 = 2$ and $\tilde{\mu}_1 = -1/2$ for Item 1 and $\sigma_2 = 4/3$ and $\tilde{\mu}_2 = -1/2$ for Item 2. The threshold parameters for MGRM1 are $\tilde{\tau}_1 = 0$ and $\tilde{\tau}_2 = 2$. For MGRM2, the item parameters were selected to match those of MGRM1. Using the relations given in Equation 7, they are $\lambda_1 = 1/2$ and $\mu_1 = -1/4$ for Item 1 and $\lambda_2 = 3/4$ and $\mu_2 = -3/8$ for Item 2. Because the discriminations of the two items differ, MGRM2 threshold parameters τ_1 and τ_2 cannot be chosen in a way that the relation in the third equation in Equation 7 is satisfied for both items. To illustrate, the authors use the first item's scale $\sigma_1 = 2$ to calculate the threshold parameters pertaining to MGRM2, yielding $\tau_1 = 0/2 = 0$ and $\tau_2 = 2/2 = 1$. Now, all three equations of Equation 7 are true for Item 1. Consequently,

CCCs of MGRM1 and MGRM2 are identical for Item 1. However, for Item 2, the third equation of Equation 7 is not true. Thus, CCCs for Item 2 differ between MGRM1 and MGRM2.

CCCs of the two models differ in two ways. First, consider the dotted vertical lines. They denote the value of η for which the probability of selecting Category 1 versus 2 and 3 (left line), and Categories 1 and 2 versus 3 (right line) is 0.5, respectively. These values are obtained by transforming thresholds from the y^* scale onto the η scale. For MGRM1, in which the slope of the regression line of y_j^* on η is 1.0 for all items, their value is identical to the values of $\tilde{\tau}_k + \tilde{\mu}_j$. For MGRM2, their value changes depending on the item discrimination. Specifically, they are defined as $(\tau_k + \mu_j)/\lambda_j$.

Clearly, the two models differ regarding the positioning of thresholds in relation to trait η . For MGRM1, the difference in η values between two consecutive thresholds τ_k and τ_{k+1} is identical for all items. For the present example, this difference is $\tilde{\tau}_1 - \tilde{\tau}_2 = 2$. For MGRM2, these differences change depending on the item's discrimination. For Item 1 it is 2 and for Item 2 the difference between thresholds on the η scale is 4/3. Consequently, smaller differences in trait are required to exceed consecutive thresholds if the item discrimination is large and vice versa.

The second distinction between CCCs of MGRM1 and MGRM2 is more subtle. It pertains to the shapes of CCCs for different item discriminations. For MGRM2, the shape of all CCCs are expanded or contracted versions of each other, which yield perfectly identical category probabilities for all items at the corresponding thresholds. In the example in Figure 2, CCCs of MGRM2 of Item 1 would perfectly match those of Item 2, if the former were contracted by a factor of 2/3. For MGRM1, this relation does not apply. Clearly, the shapes of the CCCs of both items cannot be matched by contracting or expanding them.

Both MGRM1 and MGRM2 are rather restrictive in different ways. Although thresholds of MGRM1 are identical on the trait scale, MGRM2 assumes identically shaped CCCs for all items, irrespective of their discriminations. Theoretically, it is difficult to determine which assumptions are better suited for item responses in general. On one hand, it might be considered an advantage, if all τ_k have the same value on y^* and $\eta - \mu$, as is assumed by MGRM1. This way, a specific increase in trait level corresponds to an identical change of response categories for all items. Especially for uniformly labeled response scales, this may be regarded as a plausible relation between trait and observed response. On the other hand, the assumptions of MGRM2, in which intervals between thresholds are not assumed identical on the η scale for all items, are viable too. However, if assumptions of both models appear too restrictive, it might be advisable to use a more flexible parametrization, such as that of the SSM.

The SSM requires $p - 1$ additional parameters compared with either MGRM1 or MGRM2. In addition, it constitutes a special case among these models because slopes and scales assume complementary functions. To illustrate, the authors rewrite the SSM using discrimination parameter a_j , that is

$$\Phi\left(a_j\eta - \sigma_j^{-1}\left(\mu_j - \tau_k\right)\right).$$

In this notation, σ_j exclusively scales the thresholds' location for each item. The discrimination parameter a_j , on the contrary, exclusively measures the linear relation between trait and latent response. Consequently, the contribution of σ_j is restricted to modeling the items' threshold structure, whereas a_j models the item's covariance structure. As a consequence, thresholds are not fixed on the trait scale, as is case for MGRM1. Furthermore, the CCCs' shapes are no longer identical between items, because thresholds are expanded/contracted by item-specific scale parameter σ_j . For MGRM1 and MGRM2, on the contrary, slopes and scales contribute to measuring both, threshold *and* covariance structure, simultaneously.

Parameter Estimation and Model Fit Assessment

Parameter estimates for MGRM1, MGRM2, and SSM can be obtained by analyzing the polychoric item correlations and empirical thresholds using weighted or unweighted least square estimators. The models may also be fitted using marginal maximum likelihood (MML) estimation, as for instance implemented in SAS PROC NL MIXED.

Although maximum likelihood estimation is well established in the item response literature using the response function approach (Bartholomew et al., 2011), least square estimation is more prevalent analyzing ordered categorical responses based on the underlying variable approach. A benefit of least square estimation is that it is computationally less demanding. Furthermore, given the currently available software, it can be used more flexibly. Specifically, the presented models can be used in connection with virtually any type of structural equation model. Particularly models that consider mean structure, as for instance structural equation models with grouping factors or latent growth curve models, might benefit from the additional flexibility for modeling threshold structures.

Note that maximum likelihood and least squares do not differ with respect to the scope of models that can be realized with them. However, currently, the present models (with exception of the GRM) can only be fitted using least square estimators in general purpose structural equation modeling software. Consequently, for the purpose of this article, the focus is on least square estimation.

Because the models can be fitted as structural equation models, there are two viable options for assessing and comparing the models' goodness of fit: χ^2 difference testing and comparing goodness-of-fit indices. Note that similar testing options are also available using limited information goodness-of-fit tests as proposed by Maydeu-Olivares and Joe (2006). Because MGRM1, MGRM2, and SSM are nested within the GRM, and MGRM1 and MGRM2 are nested within the SSM, χ^2 difference testing appears to be the natural approach in the present context. However, there are some instances in which comparing fit indices might also be useful.

When using weighted least square estimation, it is generally recommended to use the so-called robust estimation approach that uses a less complex weight matrix, which, compared with a full weight matrix, is reduced to its diagonal elements, and yields robust standard errors and χ^2 tests of model fit (Flora & Curran, 2004). However, using this approach, it is not recommendable to perform χ^2 difference tests on the basis of the simple difference of χ^2 values and corresponding degrees of freedom. Instead an adjusted testing procedure should be applied (Satorra & Bentler, 1999). These adjusted χ^2 difference tests are implemented in common structural equation modeling software. An instructive example using Mplus code is given below in connection with a data example.

When testing the models against each other, it is useful to proceed in the following sequence. First, the SSM should be tested against the GRM. Because both models only differ with respect to their threshold structure, this is a specific test for the assumption of identical thresholds across items. Thus, if the GRM yields a superior fit, the only source of potential misfit is the thresholds. If the data suggest that the SSM is not inferior, the assumption of identical thresholds is supported. Second, MGRM1 and MGRM2 should be tested against the SSM, if the latter was not rejected by the previous test. This way, different parametrizations of the item discriminations can be tested specifically. Although it would be also feasible to test MGRM1 or MGRM2 against the GRM directly, these tests are less informative, because they cannot differentiate whether the source of misfit is the threshold structure or the choice of item discrimination parameter, if the more restrictive models are suggested to be inferior.

Although χ^2 difference testing is a useful method for comparing models, it is widely known that if the sample is large, the power to detect even miniscule differences in model fit will

increase accordingly (e.g., Browne & Cudeck, 1992). Because it is unlikely that the *true* thresholds will perfectly match any of the constrained assumptions, the GRM will most likely yield the best fit for $n \rightarrow \infty$. In such cases, comparing goodness-of-fit indices may also yield useful results when assessing the items threshold structure.

A viable goodness-of-fit statistic, in our opinion, is the root mean square error of approximation (RMSEA, Steiger & Lind, 1980). The RMSEA is a measure of discrepancy due to approximation $\Sigma - \Sigma(\theta)$, where Σ denotes the true covariance- and threshold structure of the items and $\Sigma(\theta)$ is the covariance- and threshold structure implied by the true model parameters, collected in the θ vector, that is $\theta = (\lambda_1 \cdots \lambda_p, \sigma_1 \cdots \sigma_p, \mu_1 \cdots \mu_p, \tau_2 \cdots \tau_k)'$. Thus, it states how well the model would have fitted the population covariance- and threshold structure with optimally chosen model parameters. The (true) discrepancy due to approximation does not depend on the sample size. The RMSEA is calculated using the χ^2 value of model fit, that is,

$$\text{RMSEA} = \sqrt{\max \left[\frac{(\chi^2 - \text{df})}{(\text{df} \cdot n)}, 0 \right]},$$

where df denotes the degrees of freedom of the χ^2 test and n is the sample size. Values below 0.08 indicate an acceptable and values below 0.05 a good model fit (e.g., Hu & Bentler, 1999). Compared to other goodness-of-fit indices, the RMSEA has the advantage that it is usually provided together with a 90% confidence interval. This way, the sampling error can be accounted for when comparing results of different models.

In addition to the RMSEA, the authors will consider the comparative fit index (CFI, Bentler, 1990), that is,

$$\text{CFI} = 1 - \frac{\max[(\chi^2 - \text{df}), 0]}{\max[(\chi_I^2 - \text{df}_I), 0]},$$

and the Tucker–Lewis index (TLI; Tucker & Lewis, 1973), that is,

$$\text{TLI} = 1 - \frac{\left(\frac{\chi^2}{\text{df} - 1} \right)}{\left(\frac{\chi_I^2}{\text{df}_I - 1} \right)},$$

where χ_I^2 and df_I are the χ^2 value and degrees of freedom of the independence model. As may be inferred from their equations, CFI and TLI evaluate the relative achievement of a model compared with the assumption of independent variables. CFI or TLI values larger than 0.95 indicate a good model fit (e.g., Hu & Bentler, 1999).

Of course, comparing goodness-of-fit indices is not a replacement for a goodness-of-fit test. However, in terms of a heuristic model comparison, the use of goodness-of-fit indices is quite common in applied research and should also provide meaningful results with respect to comparative model evaluation. The authors investigated this property (among others) using Monte Carlo simulation studies. The results are provided in the online supplements to this article. Specifically, the simulation studies show that (a) consistent parameter estimates are obtained using robust WLS estimation, (b) that χ^2 difference testing is reliable and adheres to a predefined significance level, and (c) that goodness-of-fit indices are a viable option for comparative model assessment.

Empirical Example

As a concise instruction on model fitting and parameter interpretation, the authors now present results of an empirical example. The authors analyzed a publicly available data sample of Reif and Melich (2015). The data can be accessed via the *ltm*-package in R (R Core Team, 2017), where it is referred to as “SCIENCE” data. The sample comprises responses of 392 subjects to seven items measuring the attitude toward science and technology. Items were scored on a four-category scale with responses “strongly disagree,” “disagree to some extent,” “agree to some extent,” and “strongly agree.”

The authors selected items 3, 4, 6, and 7 for further analysis, because for them the assumption of a unidimensional polychoric correlation structure holds. By eliminating items that lead to a misfit based on inadequately fitted correlations, any remaining misfit may be attributed more clearly to the models’ different threshold assumptions. GRM, SSM, MGRM1, and MGRM2 were fitted with Mplus 8 using robust weighted least square (WLSMV) estimation. The corresponding program code is provided in the supplementary materials.

Overall, 18 sample moments were fitted: $(p \times (p - 1))/2 = 6$ polychoric correlations and $q \times p = 12$ empirical thresholds. The GRM required 16 parameters: four σ_j parameters and three τ_{jk} parameters for each of the four items. The SSM required 13 parameters: four λ_j parameters, three σ_j parameters, four μ_j parameters, and two τ_k parameters. MGRM1 and MGRM2 required three parameters less than the SSM. As identification constraint for SSM, MGRM1, and MGRM2, the first threshold, τ_1 , was fixed to 0.0.

Table 1 gives the parameter estimates of the four models. For better comparability, slopes and scales were transformed into discrimination parameters using Equation 5. Clearly, there is a strong dissimilarity between discrimination estimates of GRM and SSM, on one hand, and MGRM1 and MGRM2, on the other hand. This already indicates that the latter two models might not fit the data well. Specifically, discriminations of Items 2 and 3 were very heterogeneous between models. Furthermore, estimates of μ_j and τ_k were very similar between SSM and MGRM2, whereas estimates pertaining to MGRM1 were somewhat more extreme.

Next, the authors consider the model fit. Table 2 gives the χ^2 test of model fit as well as RMSEA, CFI, and TLI of the four models. Considering the χ^2 statistic, GRM and SSM yielded a good fit, whereas MGRM1 and MGRM2 did not. The same conclusion can be drawn from the other goodness-of-fit indices. Cutoff criteria for these indices were used according to Hu and Bentler (1999). Specifically, RMSEA values lower than 0.08 were considered as acceptable, and values lower than 0.05 as good, and CFI or TLI values larger than 0.95 were considered as good also. Accordingly, the fit of the SSM was considered as good and the fit of the GRM ranged between good and acceptable. For MGRM1 and MGRM2, the fit clearly was poor.

Also, the authors considered the results of the χ^2 difference tests as proposed by Satorra and Bentler (1999). Testing SSM against GRM yielded $\chi^2 = 3.348$, $df = 3$, and $p = .341$. Thus, the SSM was not inferior to the GRM. Comparing MGRM1 and MGRM2 to the SSM yielded $\chi^2 = 48.234$, $df = 3$, $p < .001$ and $\chi^2 = 66.545$, $df = 3$, $p < .001$, respectively. Thus, the assumptions of MGRM1 and MGRM2 are too restrictive for the present sample.

Finally, the authors compare the above results with results that were obtained using MML estimation. The models were fitted using SAS PROC NLMIXED with adaptive Gauss–Hermite quadrature and quasi-Newton optimization. The number of quadrature points was fixed to seven for all analyses. The model code is provided in the supplementary materials. Note that MML estimation requires considerably more computing time than robust least square estimation. In case of many items, maximum likelihood may run into numerical difficulties.

Table 3 gives the parameter estimates of the four models analogously to Table 1. Clearly, the estimates obtained from robust weighted least square estimation and MML estimation are very

Table 1. Parameter Estimates With Standard Errors for the Science Data Using WLSMV Estimation.

	GRM	SSM	MGRM1	MGRM2
a_1	0.589 (0.094)	0.589 (0.094)	0.631 (0.053)	0.661 (0.079)
a_2	2.605 (1.940)	2.605 (1.940)	0.691 (0.055)	0.809 (0.087)
a_3	0.110 (0.061)	0.110 (0.061)	0.441 (0.040)	0.178 (0.073)
a_4	0.450 (0.079)	0.451 (0.079)	0.596 (0.049)	0.581 (0.080)
μ_1	—	-1.648 (0.109)	-2.682 (0.213)	-1.606 (0.085)
μ_2	—	-2.081 (0.130)	-3.358 (0.249)	-2.058 (0.094)
μ_3	—	-2.585 (0.151)	-4.375 (0.309)	-2.364 (0.085)
μ_4	—	-1.841 (0.116)	-3.004 (0.232)	-1.779 (0.084)
τ_2	—	1.194 (0.082)	1.939 (0.165)	1.169 (0.066)
τ_3	—	2.874 (0.155)	4.691 (0.318)	2.756 (0.086)

Note. WLSMV = weighted least square means and variance adjusted; GRM = graded response model; SSM = slope-scale model; MGRM = Modified graded response model.

Table 2. Goodness-of-Fit Statistics for the Science Data.

	GRM	SSM	MGRM1	MGRM2
χ^2	4.705	7.916	67.560	75.263
df	2	5	8	8
p	.095	.161	<.001	<.001
RMSEA	0.059	0.039	0.138	0.146
90% CI	[0.000, 0.130]	[0.000, 0.087]	[0.109, 0.169]	[0.117, 0.178]
CFI	0.986	0.984	0.683	0.642
TLI	0.957	0.981	0.762	0.731

Note. GRM = graded response model; SSM = slope-scale model; MGRM = Modified graded response model; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index.

Table 3. Parameter Estimates With Standard Errors for the Science Data Using MML Estimation.

	GRM	SSM	MGRM1	MGRM2
a_1	0.610 (0.095)	0.605 (0.092)	0.629 (0.064)	0.601 (0.087)
a_2	2.394 (1.126)	2.417 (1.106)	0.686 (0.077)	0.723 (0.089)
a_3	0.111 (0.068)	0.110 (0.067)	0.459 (0.041)	0.197 (0.101)
a_4	0.455 (0.080)	0.456 (0.079)	0.597 (0.058)	0.586 (0.088)
μ_1	—	-1.651 (0.111)	-2.658 (0.234)	-1.594 (0.084)
μ_2	—	-2.086 (0.134)	-3.345 (0.279)	-2.026 (0.093)
μ_3	—	-2.593 (0.163)	-4.251 (0.345)	-2.412 (0.089)
μ_4	—	-1.841 (0.123)	-2.967 (0.255)	-1.772 (0.086)
τ_2	—	1.186 (0.081)	1.902 (0.172)	1.155 (0.062)
τ_3	—	2.883 (0.158)	4.651 (0.357)	2.759 (0.081)

Note. MML = marginal maximum likelihood; GRM = graded response model; SSM = slope-scale model; MGRM = Modified graded response model.

similar, considering the sample size of the present data. This is not surprising, because both estimation approaches yield consistent parameter estimated if their respective assumptions are satisfied.

Comparing the fit of the four models using Akaike's information criterion (AIC) also yields the same conclusions as the χ^2 difference tests based on robust weighted least square

estimation. The SSM fits the data best, having an AIC value of 3,461.1. The second best fit is yielded by the GRM, having an AIC value of 3,464.2. MGRM1 and MGRM2 fit distinctly worse. Their AICs are 3,493.3 and 3,501.1, respectively.

Of course, these results should not be overgeneralized. In some cases, a more restrictive model might prove more useful, whereas there might be other data sets for which even the SSM might be too restrictive. However, the present example similarly shows that there are instances in which the slightly more flexible structure of the SSM may prove useful.

Discussion

The present article demonstrates that the uniform response format of questionnaire items can be accounted for and tested by a systematic hierarchy of models. These models have been shown to be conveniently specified as constrained cases of a one-dimensional factor model for ordered categorical data. A benefit of this conceptualization is that they can be fitted using structural equation modeling software.

The authors have shown that testing the GRM against the SSM is a viable test to determine whether the assumption of identical thresholds across items is valid. If this assumption holds, then more restricted assumptions can be investigated by testing the SSM against MGRM1 and MGRM2. Note that this general procedure is not restricted to testing the assumption of identical thresholds. For instance, one might also test for symmetrical thresholds around μ_j or equidistant thresholds.

Aside from investigating the threshold structure, the models may also be used for item selection. To facilitate later item analyses and ability scoring, one might prefer selecting items that fit a certain model structure. Furthermore, the presented models do not have to be restricted to the one-dimensional case. One advantage of being able to fit them in a structural equation environment is that they may be conveniently generalized to multidimensional cases or included in more complex structural equation models.

As the empirical example has shown, the complexity of the SSM is well suited for analyzing questionnaire responses, whereas the assumptions of MGRM1 and MGRM2 may be too restrictive. However, in general, it is difficult to know in advance which model is suited best for analyzing questionnaire responses. Aside from the obvious dependency of sample size and model complexity on the fit, it has to be decided for each data set individually which model should be preferred. For this purpose the proposed testing procedure has proven useful.

Although, in the present article, the authors have focussed on robust weighted least square estimation, this should not reflect a general preference over maximum likelihood estimation. However, given the currently available software, least square estimation allows for a more flexible and convenient use of the models in our opinion. Furthermore, as has been shown, robust WLS yields consistent parameter estimates and reliable inference, which should satisfy the requirements of most practical applications. Specific differences between ML and WLS are often more of a technical nature. Although the numerical integration makes ML estimation computationally more demanding, researchers may nevertheless prefer this approach. ML estimation also offers possibilities to relax some distributional assumptions, as for instance the Bock and Aickin (1981) proposal to using the nonparametric maximum likelihood estimator (NPMLE) approach to estimating the latent trait distribution.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

- Bartholomew, D. J., Martin, K., & Irini, M. (2011). *Latent variable models and factor analysis: A unified approach* (2nd ed.). Hoboken, NJ: John Wiley. doi:10.1002/9781119970583
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246. doi:10.1037/0033-2909.107.2.238
- Bock, D. R., & Aickin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459. doi:10.1007/BF02293801
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230-258. doi:10.1177/0049124192021002005
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55. doi:10.1080/10705519909540118
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713-732. doi:10.1007/s11336-005-1295-9
- Muraki, E. J. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59-71. doi:10.1177/014662169001400106
- Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Retrieved from <https://www.statmodel.com/download/webnotes/CatMGLong.pdf>
- R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from <https://www.R-project.org/>
- Reif, K., & Melich, A. (2015). *Euro-barometer 38.1: Consumer protection and perceptions of science and technology, November 1992*. Inter-University Consortium for Political and Social Research. doi:10.3886/ICPSR06045.v2
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplements No. 17.
- Satorra, A., & Bentler, P. (1999). *A scaled difference chi-square test statistic for moment structure analysis* (Tech. Rep.). University of California, Los Angeles. Retrieved from <http://preprints.stat.ucla.edu/260/chisquare.pdf>
- Skrondal, A. (1996). *Latent trait, multilevel and repeated measurement modelling with incomplete data of mixed measurement levels*. Oslo, Norway: Section of Medical Statistics, University of Oslo.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation modeling* (1st ed.). Boca Raton, FL: Chapman & Hall.
- Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. In *Paper presented at the annual meeting of the Psychometric Society*. Iowa City, IA.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577. doi:10.1007/BF02295596
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10. doi:10.1007/BF02291170