A note on the interpretation and simulation of reparameterized intercepts in constrained versions of the nominal response model

Carl F. Falk [a] ✉ ⓘ

[a]McGill University

Abstract ■ This is a brief expository paper on reparameterized intercepts under constrained variants of the nominal response model, including the generalized partial credit and partial credit models. Such parameterizations are commonly found in item response theory software packages such as flexMIRT®, IRTPRO, and OpenMx / *rpf*, and both these models are highly popular in educational and psychological testing. A heuristic graphical interpretation is provided. We give examples of how intercepts may be easily generated for Monte Carlo simulation studies, including a brief study to increase generalizability and explore limitations of a recently developed information matrix test to detect misspecification when collapsing adjacent response categories.

Keywords ■ Generalized partial credit model; Nominal response model; Monte Carlo simulations; Category collapsing..

✉   carl.falk@mcgill.ca

doi   10.20982/tqmp.17.4.p345

Introduction

As the nominal response model (NRM; Bock, 1972; Thissen & Cai, 2016; Thissen, Cai, & Bock, 2010) itself is complicated, tutorials often overlook interpretation of the intercepts in favor of other aspects (Falk & Ju, 2020; Revuelta, Maydeu-Olivares, & Ximenez, 2019). In this paper, we focus on interpretation of reparameterized intercepts under the Thissen et al. (2010) NRM, and in particular for constrained versions of the NRM such as the generalized partial credit model (GPCM; Muraki, 1992) and partial credit model (PCM; Masters, 1982). We first present various parameterizations of the NRM. Next a heuristic graphic interpretation of the intercept parameters under Thissen et al. (2010) in the context of the GPCM along with code for an example of simulating items. In the final section, a small Monte Carlo study is conducted to increase generalizability and explore limitations of an information matrix test developed by Harel and Steele (2018) to detect misspecification when collapsing adjacent response categories under the PCM.

Nominal, generalized partial credit, and partial credit models

A general form for the category response functions (CRFs) of NRM for an item is:

$$T(k|\theta) = \frac{\exp(z_k)}{\sum_{m=1}^{(K-1)} \exp(z_m)} \tag{1}$$

where $k = 0, 1, \ldots, K - 1$ are categories and θ is the latent trait. Conceptually, Equation (1) traces the probability of category k for an item at values of the latent trait, θ.

There are multiple ways to write z_k. In the original NRM, Bock (1972) used $z_k = a_k\theta + c_k$. Thus, each category had its own specific slope (a_k) and intercept (c_k). In recent unidimensional versions of the NRM (e.g., Thissen et al., 2010), the following is used:

$$z_k = a^* s_k \theta + c_k \tag{2}$$

where a^* is a slope parameter that can be interpreted similarly to a factor loading, s_k is a scoring function value for category k, and c_k is an intercept for category k. For identification, c_k may be subject to a constraint (e.g., sum equal

to zero or the first $c_0 = 0$) or parameterized as described shortly. Similarly, constraints are necessary on scoring function values for identification, but are described elsewhere (**tbc10**). When the scoring function values are fixed to ordered integers, $\mathbf{s} = \begin{bmatrix} s_0 & s_1 & \cdots & s_{K-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & \cdots & K-1 \end{bmatrix}$, the items are ordinal with respect to the latent trait and the NRM reduces to the GPCM, which represents a popular item response models for polytomous items in educational and psychological measurement contexts (Penfield, 2014). Further constraining slopes equal across items results in the PCM, and additional constraints on intercepts lead to the rating scale model (Andrich, 1978).

Traditional parameterizations of the GPCM often have a threshold (b) analogous to overall difficulty and step parameters (d_j) that control cross-over among CRFs:

$$z_k = \sum_{j=0}^{k} 1.7a(\theta - b + d_j) \tag{3}$$

where 1.7 is the usual scaling constant (for a review, see Savalei, 2006). For example, by combining b and d_j, the point along θ where adjacent CRFs cross can be determined. Alternatively, these cross-over points can be determined by the b_j parameters in yet another parameterization of the GPCM:

$$z_k = \sum_{j=0}^{k} 1.7a(\theta - b_j) \tag{4}$$

Note that constraints on d_j in (3) or on b_j in (4) are also necessary for identification, usually $d_0 = 0$ and $\sum_{j=1}^{k} d_j = 0$ for (3), and $b_0 = 0$ or $\sum_{j=0}^{k} b_j = 0$ for (4).

There are at least two reasons why enhanced understanding of the intercepts in (2) is desirable. First, variants of Equation (2) are used in recent software packages such as flexMIRT® (Cai, 2017), IRTPRO (Cai, Thissen, & du Toit, 2011), and the *rpf* module that works with OpenMx (Neale et al., 2016; Pritikin, 2020b; Pritikin & Falk, 2020; Pritikin, Hunter, & Boker, 2015) in part because (2) is more easily generalized to measure multiple latent constructs than is (3) and (4). Second, it can be cumbersome for Monte Carlo simulation studies to simulate step parameters or cross-over parameters since a value for one parameter may affect whether a value for another parameter is realistic (e.g., leading to cross-over of CRFs that are atypical). Thus, sometimes a very small variance or limited range (e.g., Zhou & Huggins-Manley, 2020) or even fixed values (e.g., Kim & Paek, 2017) for each parameter is used, which adversely affects generalizability and makes it difficult to study how variability in CRFs may affect the performance of studied approaches.

Understanding reparameterized intercepts

The aforementioned sofware packages implement a version of the NRM that further reparameterizes the intercepts, $\mathbf{c} = \mathbf{T}_c \boldsymbol{\gamma}$, where $\mathbf{c}$ ($K \times 1$) contains all intercepts for an item, $\boldsymbol{\gamma}$ is ($K-1$) $\times 1$ and contains all estimated parameters, and $\mathbf{T}_c$ is a $K \times (K-1)$ matrix. Thus, $\boldsymbol{\gamma}$ contains parameters, $\gamma_1, \gamma_2, \ldots, \gamma_K$, that are all freely estimated. These values are transformed into intercepts for calculation of Equation (2).

In our experience, students often have difficulty unpacking concise descriptions of the "gamma" parameters: "... the linear-Fourier basis separates the parameter space into a (first) component for $b = -\gamma_1/a^*$ and a remainder that parameterizes 'spacing' among the thresholds or crossover points of the curves" (Thissen et al., 2010, p. 61)[1]. Direct interpretation of the intercepts is similarly difficult. The goal of this brief expository paper is to elucidate how such parameters may be interpretable when using a linear-Fourier basis for $\mathbf{T}_c$ as recommended by Thissen et al. (2010) and as the default in these software packages.[2] The matrix is typically constructed as follows (see Thissen et al., 2010) :

$$\mathbf{T} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & t_{22} & \cdots & t_{2(K-1)} \\ 2 & t_{32} & \cdots & t_{3(K-1)} \\ \vdots & \vdots & & \vdots \\ K-1 & 0 & \cdots & 0 \end{bmatrix} \tag{5}$$
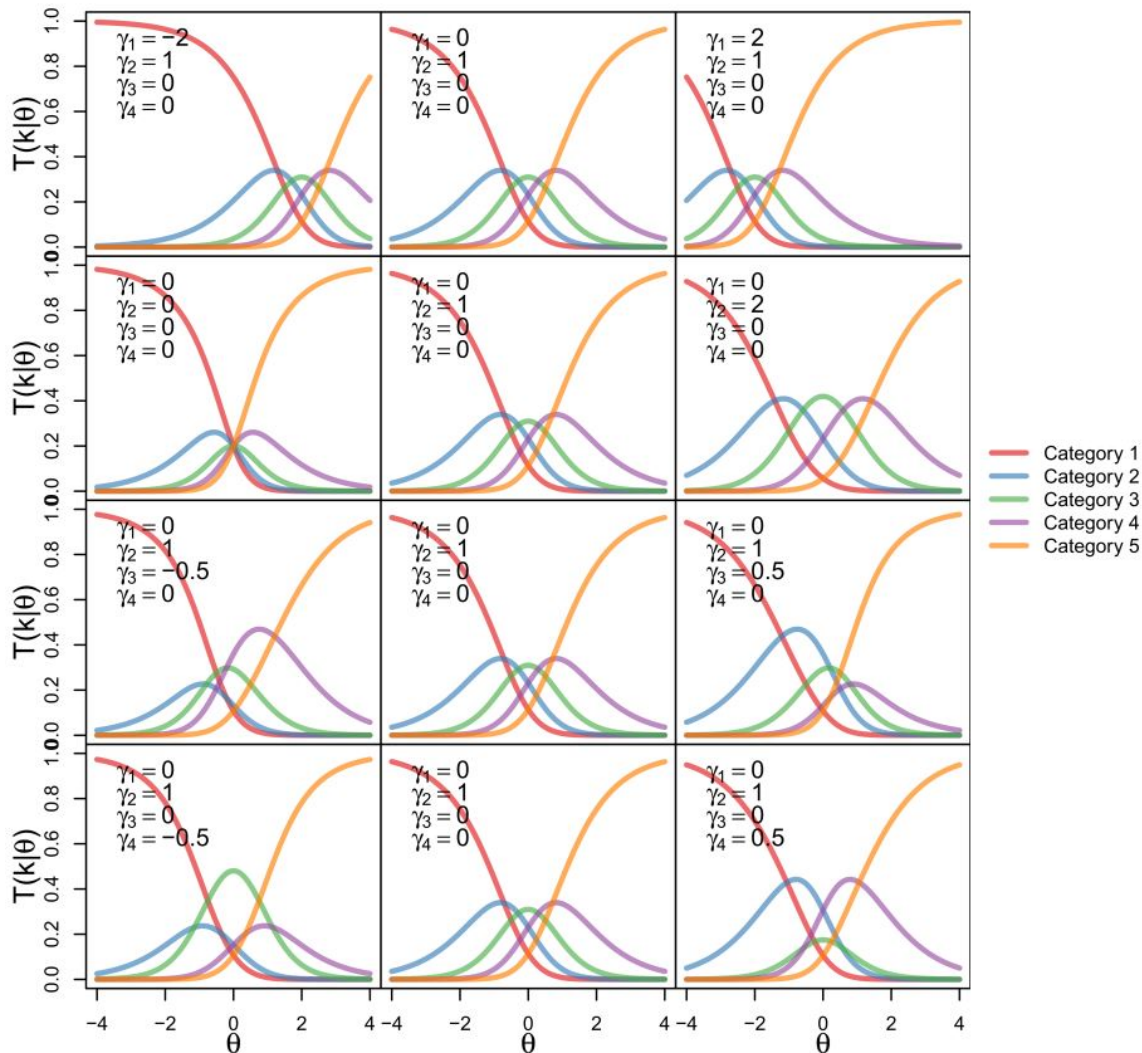
with $t_{kk'} = \sin(\pi(k'-1)(k-1)/(K-1))$

To explain, we rely on a 5-category item and its matrix for the intercepts:

$$\mathbf{T}_c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & .7071 & 1 & .7071 \\ 2 & 1 & 0 & -1 \\ 3 & .7071 & -1 & .7071 \\ 4 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

$\mathbf{T}_c$ has a linear first column, with remaining columns determined by a Fourier series. Conceptually, these additional columns take the places of "quadratic and higher-order polynomial terms" (Thissen et al., 2010, p. 93). The columns from the second onward are mutually orthogonal (their inner product is zero). Likewise, we may think of the γ parameters as providing information about the strength and direction of each trend, and for the most part can be independently interpreted and manipulated in simulations.

A graphical, heuristic interpretation is apparent by varying only one parameter at a time for a 5-category

---

[1]Subscripts dropped from original quote to match notation in the current manuscript.
[2]This is the "Trend" option for $\mathbf{T}_a$ and $\mathbf{T}_c$ matrices in flexMIRT®, IRTPRO, and *rpf*.

Figure 1 ∎ Manipulation of one gamma parameter at a time.

GPCM item in Figure 1. Each row changes a single γ (with $a^* = 1$ for all plots). For instance, the first row varies γ_1, but keeps all other γ's constant. In this way, it is easy to see that γ_1 controls the difficulty of the item, which was noted by Thissen et al. (2010). The CRF shapes remain the same, but shift in response to changes in γ_1 – with negative values resulting in more difficult items. Conceptually, γ_2 controls the overall spread in where the CRFs cross – with $\gamma_2 = 0$ resulting in all CRFs crossing at the same point, and positive values (e.g., $\gamma_2 = 2$) resulting in CRFs that cross in order but are more spread out. In our opinion, these latter CRFs seem more typical of Likert-type or partial credit items on cognitive tests. These first two trends will be typical of all ordered polytomous items. The re-

maining γ's may depend on the total number of categories. For this 5-category item, γ_3 appears to control asymmetry in how close the spacing is between crossover points, with negative γ_3 leading to closer crossover points on the lower end of θ, and positive values leading to this same effect at the lower end of θ. In addition, we can see submersion of the second and fourth categories for negative and positive γ_3, respectively. Sometimes such a pattern results in real data, yet simulation of step parameters from a limited range may avoid such CRFs altogether. Finally, γ_4 appears to vary spacing of cross-over points such that alternating categories are more/less dominant or submerged. For instance, on the left-hand side of the bottom row in Figure 1, the first, third, and last categories appear to be domi-

nant response options (when $\gamma_4 = -.5$), while the opposite is true for the right-hand side of the bottom row (when $\gamma_4 = .5$).

For those who wish to experiment more regarding visualization, we recommend two resources. First, CRFs for the parameterization in (2) can be plotted using either R code in flexMIRT®support page (https://vpgcentral.com/software/irt-software/support/) or using the itemModel-Explorer() function in the ifaTools R package (Pritikin, 2020a; Pritikin & Schmidt, 2016). To use the latter for the GPCM, the "nominal" model should be chosen as the item model, "trend" matrices for both "T.a" and "T.c" options, "alf1" should be set to 1 and the remaining "alf" parameters should be set to 0. To explain briefly, $\mathbf{s} = \mathbf{T}_a\boldsymbol{\alpha}$ is typically used to parameterize scoring functions in (2), with $\mathbf{T}_a$ constructed similarly to $\mathbf{T}_c$. Setting the first α to 1 with the remaining 0 will provide the ordinal scoring function values of the GPCM (the first column of $\mathbf{T}_a$). The user may then manipulate the slope ("a") and γ ("gam") parameters as desired.

Finally, the interpretation for the γ parameters here holds when the linear-Fourier series contrast matrix is used. Thissen et al. (2010) present an identity-based matrix that could be utilized or the user may employ a custom contrast matrix. In these cases, the γ parameters will have a different interpretation. In particular, with the identity-based matrix the γ's would have the same interpretation as the intercepts in (2): the first intercept is fixed to zero and the remainder of the intercepts (c_1 to c_k) translate directly to the γ parameters (γ_1 to γ_k) without modification (e.g., $\gamma_1 = c_1$). Further examples of use of the identity-based matrix, albiet in the context of slope parameters, are given by Preston, Reise, Cai, and Hays (2011). We would presume the user has another purpose if a custom contrast matrix is used.

Simulating GPCM item parameters: Example code

Given that the γ's essentially represent independent trends, this may make simulation of CRFs easier.[3] Figure 2 presents plots of 20 simulated GPCM items. Relevant R code (see Listing 1 at the end) utilizes the *rpf* package (Pritikin, 2020b) and demonstrates generation of both item parameters and item responses for 5 category items, using the same data generating distributions as mentioned here.[4] Slopes were generated from a lognormal distribution, $a^* \sim log\mathcal{N}(0, .15^2)$. We also desired items that were slightly difficult to endorse, $\gamma_1 \sim \text{unif}(-1.25, .25)$, CRFs that were usually spread out instead of crossing in the same place, $\gamma_2 \sim \text{unif}(.5, 2)$, and allowed some mild heterogeneity with respect to other features, $\gamma_3 \sim$

unif$(-.6, .3)$, $\gamma_4 \sim \text{unif}(-.3, .3)$. Such CRFs provide heterogeneity with some CRFs crossing out of order (e.g., item 8), but while still appearing realistic. Examination of real data could also be used to inform a choice of generating distribution for each parameter. As a concise example, we fit a between-item multidimensional GPCM to the same Big Five personality data (Goldberg, 1992) as analyzed by Jeon and De Boeck (2019) from the Open-Source Psychometrics Project ($N = 7,899$; https://openpsychometrics.org/_rawdata/). Across the 50 items (with 5 categories each), the mean values for a^* ($M = 1.03, SD = .39$), γ_2 ($M = 1.64, SD = .59$), γ_3 ($M = -.10, SD = .11$), and γ_4 ($M = .10, SD = .09$) were similar to the center of the generating distributions listed above, and only γ_1 ($M = .56, SD = .67$) indicated that items were slightly more difficult to endorse. In educational testing contexts, it may be desirable to generate a^* and γ_1 with some negative dependency (Lord, 1975). We emphasize that the above generating distributions may not be representative of all applications and testing contexts. In practice, checking for reasonable information functions and marginal probabilities is also recommended. For these 20 items, for example, test information peaks at about 24 near $\theta = .5$, and assuming a standard normal θ, marginal reliability is approximately .93 and the lowest marginal probability is approximately .035 for the highest category on item 9 - meaning that about 3.5% of respondents from a standard normal θ would select this category.

Simulation Studies

To illustrate the utility of being able to simulate a greater variety of CRF shapes, we provide a small set of simulations inspired by Harel and Steele (2018). During test construction, there are several reasons why a psychometrician may desire to collapse and recode adjacent categories (for a review, see Harel & Steele, 2018). For example, one might suspect that category boundaries are not clearly distinguishable by participants (Preston et al., 2011). In such a case, collapsing may be well justified and not result in misspecification. If very few responses are observed in some categories, collapsing may be done to avoid estimation difficulty (Rose et al., 2014), even if it technically consitutes misspecification. Harel and Steele (2018) developed a new information matrix test (IMT) to detect misspecification when adjacent categories for the PCM are collapsed into a single category. Our studies are an exploration of how features of the collapsed item and other items on the test may affect statistical power to detect misspecification.
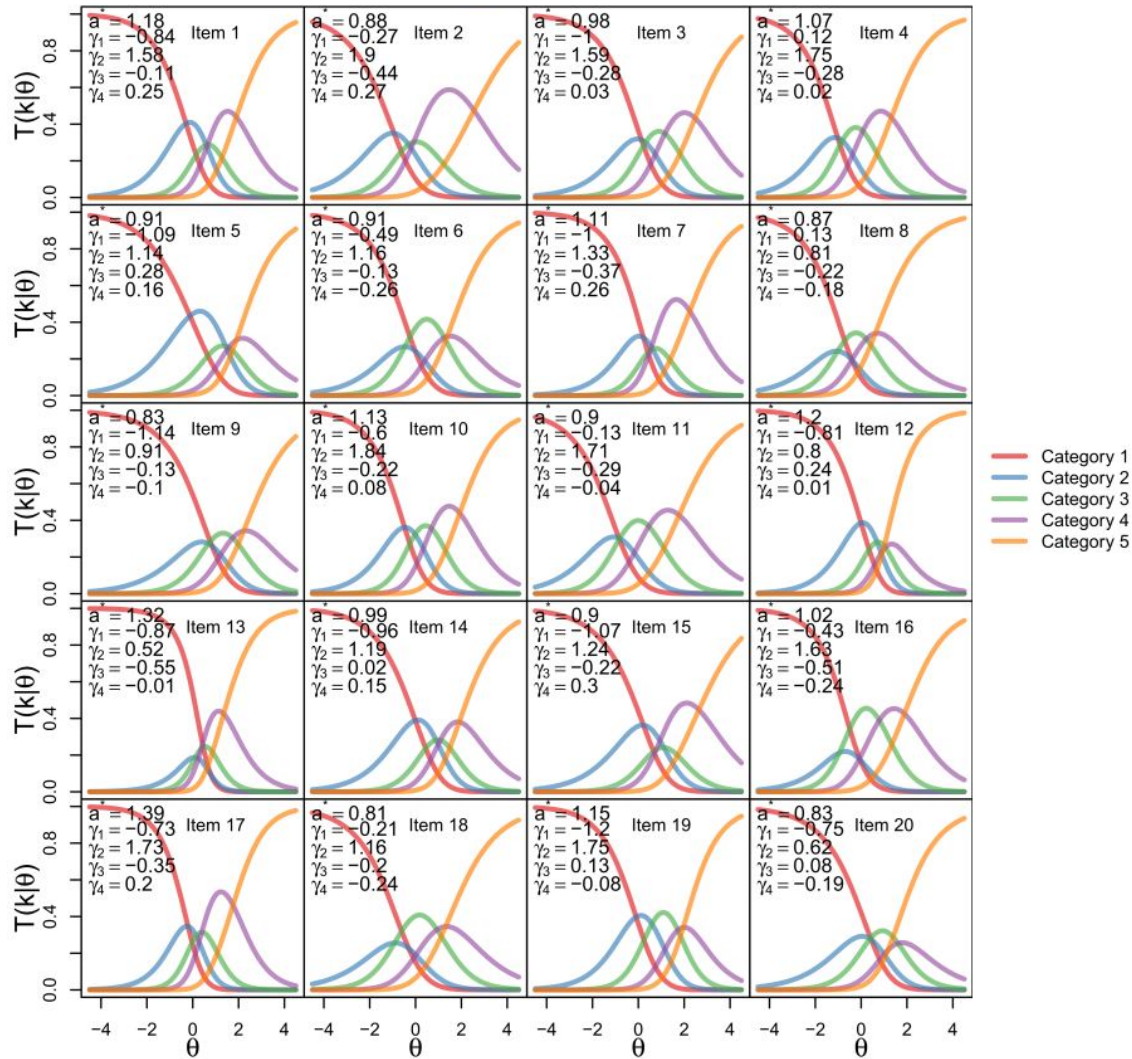
Harel and Steele (2018) investigated Type I error (false positives) and power of the IMT versus several alternatives

---

[3]In our experience, the distribution of slopes may affect choices for the other parameters.
[4]The code is slightly more compact than that used for Figure 2.

Figure 2 ■ Simulated GPCM items.

(Preston et al., 2011; Kang & Chen, 2008, i.e.,) in simulations. In any given replication, a single item was subject to the IMT, with all items following the PCM:

$$z_k = \sum_{j=0}^{k} (\theta - b_j) \qquad (7)$$

which is equivalent to Equation (4) with $1.7a = 1$, and $b_0 = 0$ for identification, or to the NRM in Equation (2) with $a^* = 1$ and ordinal scoring functions. Latent traits were generated from a standard normal distribution.

The tested item followed one of four different shapes

(Figure 3). See Table 1 for both threshold (b's) and gamma (γ's) parameters.[5] We will focus on two (out of four) conditions that were crossed with item type: 1) a condition where no collapsing was done; and 2) a condition where the middle and 4th categories were collapsed. The first condition results in no misspecification and can be used to assess Type I error, whereas the second condition results in misspecification as the recoded item no longer follows the PCM.

All untested items had threshold parameters (b's) "selected to be a series of quantiles of a standard normal distribution, ranging from -1.96 to 1.96, with the items de-

---

[5]To convert from thresholds to gammas, note that $c_k = -\sum_{j=0}^{k} b_j/(1.7a)$ under (4) or $c_k = -\sum_{j=0}^{k} b_j$ under (7). Once intercepts are obtained, from $\mathbf{c} = \mathbf{T}_c \boldsymbol{\gamma}$ we can solve for $\boldsymbol{\gamma}$ via the matrix operations: $\boldsymbol{\gamma} = (\mathbf{T}_c' \mathbf{T}_c)^{-1} \mathbf{T}_c' \mathbf{c}$.
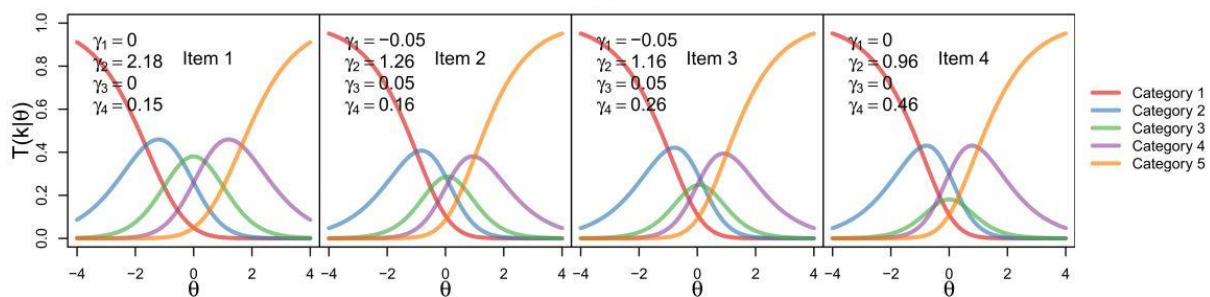
Table 1 ■ Item parameters for the tested item under four item shape conditions used by Harel and Steele ([2018](#)) and our Study 1.

Item Shape	b_1	b_2	b_3	b_4	γ_1	γ_2	γ_3	γ_4
1	-1.645	-.385	.385	1.645	0	2.18	0	.15
2	-1	0	.2	1	-.05	1.26	.05	.16
3	-1	.2	0	1	-.05	1.16	.05	.26
4	-1	.5	-.5	1	0	.96	0	.46

Note. Slope parameter is fixed to 1. Parameters from both the PCM and NRM are provided.

Figure 3 ■ Tested items from Harel and Steele ([2018](#))

signed to interlace, and thus each cover a large portion of the range of expected θ values" (Harel & Steele, [2018](#), p. 218) A similar description is given in describing Item 1 in Figure [3](#). We take this to mean that item parameters were fixed across replications and provide information across the latent trait but primarily towards the middle of the latent trait distribution.

Based on these items, we make two observations. First, the choice to collapse is likely to occur when few respondents endorse a category (i.e., the category has low marginal probability). The tested items involved increasing submersion of the middle category, which leads to a marginal probability as low as .12 under item 4. However, the middle category is still located at the middle of the latent trait distribution, where most respondents were generated. Harel and Steele ([2018](#)) report higher power when this middle category has lower marginal probability (e.g., the power under item 4 was greater than under item 1). We suppose that it is more common that low marginal probabilities that prompt collapsing more often occur towards the endpoint categories – outside of the most concentrated part of the latent trait distribution – and that this pattern regarding marginal probability and power may not hold generally. Second, the remaining items on the test, if actually similar to item 1, may provide information primarily in the middle of the latent trait generating distribution and also near the middle collapsed category. To investigate whether a variety of items adversely affect power, we generated a greater variety of PCM items by manipulating intercepts using the NRM parameterization.

Method

We designed two simulation studies. In all cases, latent traits were generated from a standard normal distribution, the total number of items was fixed at 6 (only 1 item was tested), and sample size ($N = 100, 250, 500$) was manipulated. Although we generated items using a constrained NRM, the PCM was used for model fitting and construction of the IMT (with $\alpha = .05$). Conceptually, the IMT relies on the equivalence of the negative expected Hessian (i.e., the matrix of second-order derivatives of the log-likelihood) and expected outer product of the score vector (i.e., first-order derivatives of the log-likelihood) under a correctly specified model (e.g., Ranger & Kuhn, [2012](#); Yuan, Cheng, & Patton, [2014](#)). A discrepancy therefore may indicate misfit and can be used to form a test statistic. Harel and Steele ([2018](#)) form such a test statistic from these matrices relevant for a subset of item parameters (i.e., b_j's) surrounding or adjacent to the collapsed categories. Under a correctly specified model, the IMT asymptotically follows a chi-square distribution with degrees of freedom equal to the number of unique elements in the relevant subset of the aforementioned matrices. For further technical details, we refer the reader to their original paper, and note that the code used to implement the test is available in the R

package *IMTest* (Harel, 2017).[6]

Study 1

In Study 1, item parameters for the tested items were identical to those used in Harel and Steele (2018) (Table 1). The 5 untested items were PCM items randomly generated across replications and were intended to be slightly difficult, $\gamma_1 \sim \text{unif}(-.75, .25)$, interlacing, $\gamma_2 \sim \text{unif}(.5, 2)$, and varying remaining features, $\gamma_3 \sim \text{unif}(-.5, .3)$ and $\gamma_4 \sim \text{unif}(-.3, .3)$. Type I error (no collapsing) and power based on collapsing the middle and 4th categories as in the original Harel and Steele (2018) study were examined. We conducted 10,000 replications per cell of the design. Aside from untested item shapes, these conditions overlap with those from Harel and Steele (2018). This Study was aimed at extending the generalizability of the IMT under conditions where the other items on the test were varied.

Study 2

In Study 2, the same data generating conditions as in Study 1 were used for non-tested items. We varied the approach for generating the tested items and categories. In Condition 1, the tested item was fairly difficult such that the highest category had lower marginal probability due to being far from the center of the latent trait distribution, $\gamma_1 \sim \text{unif}(-1.25, .25)$, $\gamma_2 \sim \text{unif}(.75, 3)$, $\gamma_3 \sim \text{unif}(-.25, .75)$, and $\gamma_4 \sim \text{unif}(-.75, .75)$. We investigated collapsing of the highest two categories, with the IMT involving the highest threshold. In Condition 2, generated items were roughly in the middle of the latent trait distribution, $\gamma_1 \sim \text{unif}(-.75, .75)$, but other features were varied such that the lowest two categories sometimes had low marginal probability, also in part to being located far from the center of the latent trait distribution. In particular, spacing between crossover points was high, $\gamma_2 \sim \text{unif}(2, 3)$, and some asymmetry may lead to lower marginal probability of the first and second categories, $\gamma_3 \sim \text{unif}(-1, -.25)$, and $\gamma_4 \sim \text{unif}(-.75, .5)$. 10,000 replications per cell were conducted with tested items re-used for each of the three sample size conditions. Thus, Study 2 was designed to test endpoint categories while varying item features to result in somewhat low marginal probability of the involved categories. On average in Condition 1, the 4th and 5th categories had marginal probabilities of .13 and .08, respectively, and under Condition 2, the 1st and 2nd categories had .11 and .12 marginal probabilities.

Results and Discussion

For Study 1, power and Type I error (false positive) rates (Table 2) were similar to those reported by Harel and Steele

(2018). In most cases, Type I error was close to the nominal rate (.05). Comparable rates of power were found – around .1 or below at $N = 100$ and between .48 and .64 at $N = 500$. We also observed higher power for Item 4 than for the other items, in the same pattern as previously observed.

Under Study 2, Type I error rates were very close to the nominal rate (Table 3). However, power did not increase as rapidly with sample size as in Study 1, with power slightly below .1 at $N = 100$ and reaching .26 and .32 for Conditions 1 and 2, respectively, at $N = 500$. We employed an ad-hoc approach to further probe for whether our proposal that varying the marginal probability of the relevant categories resulted in different power. In particular, we fit two generalized linear mixed models (separate models for Condition 1 and 2) with a binomial link function to Study 2 data. The random intercept was based on replication number as the same item parameters for the tested item were repeated at each sample size. Predictors included sample size (dummy coded, with $N = 100$ as the reference group), and a composite that was the sum of the marginal probabilities of the two tested categories. In general, the pattern of results (Table 4) suggested that marginal probability of the relevant categories was positively related to power under Condition 1, but was unrelated under Condition 2. This result is in contrast to the pattern apparent from Study 1 and originally found by Harel and Steele (2018). Thus, there are likely other non-trivial features of tested or untested items that affect power of the IMT. We suspect that a similar pattern may be found if other approaches to detect misspecification were studied.

Conclusion

In conclusion, it is hoped that this paper will aid in interpreting "gamma" parameters that may commonly appear in software output such as that from flexMIRT®, IRTPRO, and the item factor analysis module of *OpenMx*. Even if a traditional parameterization of the GPCM or PCM is provided as output by some of these software programs, it may still be sometimes necessary or useful to simulate data by using the "gamma" parameters. The strategy we have proposed may then be used, and we expect to be useful in both unidimensional and multidimensional GPCM and PCM applications where each items loads on one factor at a time.

Authors' note

---

[6]The IMT also requires Monte Carlo methods to approximate the covariance matrix among the residual terms. We used the same 100,000 simulations per test as used by Harel and Steele (2018).

Table 2 ■ Type I error and power under Study 1.

Item Shape	N	Type I Error	Power
1	100	.051	.077
2	100	.051	.079
3	100	.050	.092
4	100	.051	.103
1	250	.054	.212
2	250	.051	.234
3	250	.052	.251
4	250	.050	.298
1	500	.050	.480
2	500	.051	.526
3	500	.053	.567
4	500	.049	.642

Table 3 ■ Type I error and power under Study 2.

Condition	N	Type I Error	Power
1	100	.051	.083
2	100	.052	.092
1	250	.051	.152
2	250	.050	.190
1	500	.052	.260
2	500	.052	.319

Table 4 ■ Generalized linear mixed effects models for power in Study 2.

Condition 1				
Fixed Effect	Estimate	SE	z	p
Intercept	-3.12	.06		
N250	.73	.05	15.42	<.001
N500	1.47	.05	32.22	<.001
MargProb	2.15	.18	11.92	<.001
Condition 2				
Fixed Effect	Estimate	SE	z	p
Intercept	-2.55	.06		
N250	.90	.05	20.06	<.001
N500	1.68	.04	38.01	<.001
MargProb	-.02	.18	-.13	.894

Note. Sample size is dummy coded, with $N = 100$ as the reference group. MargProb = sum of the marginal probability of the two collapsed categories. Variance of the random intercept was .58 under Condition 1, and .67 under Condition 2.

Address all correspondence to: Carl F. Falk, Department of Psychology, McGill University, 2001 McGill College, 7th Floor, Montreal, QC H3A 1G1, Canada. Email: carl.falk@mcgill.ca.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. doi:10.1007/BF02291411

Cai, L. (2017). *FlexMIRT version 3.51: Flexible multilevel item factor analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *11*, 72–99. doi:10.3389/fpsyg.2020.00072

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42. doi:10.1037/1040-3590.4.1.26

Harel, D. (2017). Information matrix test for generalized partial credit models (Version 1.0.0). IMTest. Retrieved from https://CRAN.R-project.org/package=IMTest

Harel, D., & Steele, R. J. (2018). An information matrix test for the collapsing of categories under the partial credit model. *Journal of Educational and Behavioral Statistics*, *43*(6), 721–750. doi:10.3102/1076998618787478

Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an irtree approach. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 517–537. doi:10.1111/bmsp.12182

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*, 391–406. doi:https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kim, T., & Paek, I. (2017). A comparison of item parameter and standard error recovery across different R packages for popular unidimensional IRT models. In A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *L* (pp. 421–430). Quantitative psychology . Cham: Springer International Publishing.

Lord, F. M. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika*, *40*, 205–217. doi:10.1007/BF02291567

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi:10.1002/j.2333-8504.1992.tb01436.x

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., . . . Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549. doi:10.1007/s11336-014-9435-8

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36–48. doi:10.1111/emip.12023

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, *71*, 523–550. doi:10.1177/0013164410382250

Pritikin, J. N. (2020a). ifaTools: Toolkit for item factor analysis with 'openMx' (Version 1.0). Retrieved from https://CRAN.R-project.org/package=ifaTools

Pritikin, J. N. (2020b). rpf: Response probability functions (Version 1.0.0). Retrieved from https://CRAN.R-project.org/package=rpf

Pritikin, J. N., & Falk, C. F. (2020). OpenMx: A modular research environment for item response theory method development. *Applied Psychological Measurement*, *44*, 561–562. doi:10.1177/0146621620929431

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for item factor analysis. *Educational and Psychological Measurement*, *75*(3), 458–475. doi:10.1177/0013164414554615

Pritikin, J. N., & Schmidt, K. M. (2016). Model builder for item factor analysis with openMx. *R Journal*, *8*, 1–11.

Ranger, J., & Kuhn, J.-T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*(3), 247–268. doi:10.1111/j.1745-3984.2012.00174.x

Revuelta, J., Maydeu-Olivares, A., & Ximenez, C. (2019). *Factor analysis for nominal (first choice) data*. doi:10.1080/10705511.2019.1668276

Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology*, *67*, 516–526. doi:10.1016/j.jclinepi.2013.10.024

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, *71*, 763–767. doi:10.1007/s11336-004-1237-y

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume 1: Models* (pp. 51–73). Boca Raton, FL: Taylor & Francis.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 43–75). New York, NY: Taylor & Francis.

Yuan, K.-H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*(2), 232–254. doi:10.1007/s11336-013-9334-4

Zhou, S., & Huggins-Manley, A. C. (2020). The performance of the semigeneralized partial credit model for handling item-level missingness. *Educational and Psychological Measurement*, *80*, 1196–1215. doi:10.1177/00131644120918392

Citation

Listing 1 follows

Listing 1 ■ Simulating GPCM item parameters

```r
# This example simulates item parameters and responses
# for twenty-five items with 5 categories per item,
# unidimensional GPCM

library(rpf) # load rpf
set.seed(1239) # set random number seed

n<-25 # number of items
ncat<-5 # number of categories

# simulate item parameters
a<-rlnorm(n,0,.15)
alf1<-rep(1,25)
alf2<-alf3<-alf4<-rep(0,25)
gam1<-runif(n,-1.25,.25)
gam2<-runif(n,.5,2)
gam3<-runif(n,-.6,.3)
gam4<-runif(n,-.3,.3)

# combine into matrix, 1 item per row
pars<-cbind(a,alf1,alf2,alf3,alf4,gam1,gam2,gam3,gam4)

# create nrm.item object
nrm.item<-rpf.nrm(ncat, factors=1, T.a="trend",T.c="trend")

# generate data
N<-500 # sample size
thetas<-rnorm(N,0,1) # standard normal latent traits

# simulate responses
dat<-rpf.sample(thetas, replicate(n,nrm.item), t(pars))

# Result is matrix with "ordinal" factors as columns for use with OpenMx
# If merely numeric data is required with other software, simply do:
dat<-apply(dat,2,as.numeric)

# and if it is also desired that categories start from 0, 1, 2, 3, etc.
# instead of 1,2,3,4, etc, do:
dat<-dat-1

# dat is now ready for use
```