Order-Constrained Estimation of Nominal Response Model Parameters to Assess the Empirical Order of Categories

Educational and Psychological
Measurement

2018, Vol. 78(5) 826–856

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417714296

journals.sagepub.com/home/epm

Miguel A. García-Pérez¹

Abstract

Bock's nominal response model (NRM) is sometimes used to identify the empirical order of response categories in polytomous items but its application tags many items as having disordered categories. Disorderly estimated categories may not reflect a true characteristic of the items but, rather, a numerically best-fitting solution possibly equivalent to other solutions with orderly estimated categories. To investigate this possibility, an order-constrained variant of the NRM was developed that enforces the preassumed order of categories on parameter estimates, for a comparison of its outcomes with those of the original unconstrained NRM. For items with ordered categories, order-constrained and unconstrained solutions should account for the data equally well even if the latter solution estimated disordered categories for some items; for items with truly disordered categories, the unconstrained solution should outperform the order-constrained solution. Criteria for this comparative analysis are defined and their utility is tested in several simulation studies with items of diverse characteristics, including ordered and disordered categories. The results demonstrate that a comparison of order-constrained and unconstrained calibrations on such criteria provides the evidence needed to determine whether category disorder estimated on some items by the original unconstrained form of the NRM is authentic or spurious. Applications of this method to assess category order in existing data sets are presented and practical implications are discussed.

¹Universidad Complutense, Madrid, Spain

Corresponding Author:

Miguel A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, Pozuelo de Alarcón, 28223 Madrid, Spain.

Email: miguel@psi.ucm.es

Keywords

item response theory, polytomous models, ordered responses, category order

Introduction

Psychological tests and patient-reported outcome measures often use items with response categories in a preassumed order. Rasch-type item response theory models for ordered responses (i.e., the type referred to as *polytomous Rasch models* by Andrich, 2013a) are often fitted to these data. These models share the characteristics that (1) they assume a specific order for the response categories, (2) the conditional probability of a response in category k given that the response is in categories $k - 1$ or k is a two-parameter logistic function of trait level, (3) the slope parameter of such logistic function has the same value for all pairs of consecutive categories on all items, and (4) only category location parameters (referred to as *thresholds*) may vary across items. When these models are fitted to data, within-item thresholds often turn up disordered (see, e.g., Ashley, Smith, Keding, Jones, Velikova, & Wright, 2013; Bee, Gibbons, Callaghan, Fraser, & Lovell, 2016; Bokhary, Suttle, Alotaibi, Stapleton, & Boon, 2013; Bourke, Wallace, Greskamp, & Tormoehlen, 2015; Brogårdh, Lexell, & Lundgren-Nilsson, 2013; das Nair, Moreton, & Lincoln, 2011; Dougherty, Nichols, & Nichols, 2011; Oluboyede & Smith, 2013; Wetzel, Hell, & Pässler, 2012). Disordered thresholds seem to question the empirical validity of the preassumed order of categories for the incumbent items and, hence, the adequacy of a scoring system that assigns increasingly higher scores to higher categories in their preassumed order.

Some action would be needed if the order of estimated thresholds indicated the empirical order of categories and overrode their preassumed order (a position advocated by some; e.g., Andrich, 2004, 2013a, 2013b; Andrich, de Jong, & Sheridan, 1997). Because parameter estimates are affected by sampling error, an immediately obvious approach is to use statistical methods to test the hypothesis that true parameters are indeed ordered (Salzberger, 2015). Yet the most common action is to remove the incumbent items or to collapse response categories until ordered threshold estimates are obtained (see the references cited above). Guidance for collapsing categories may be sought via analysis of the extent to which adjacent categories discriminate trait levels (e.g., Preston et al., 2015; Preston & Reise, 2014; Preston, Reise, Cai, & Hays, 2011), a strategy that is applicable even when thresholds are ordered.

An alternative position justified and advocated by others (e.g., Adams, Wu, & Wilson, 2012; Annoni, Weziak-Bialowolska, & Farhan, 2013; García-Pérez, 2017; Linacre, 1999, 2004; Wang, Deutscher, Yen, Werneke, & Mioduski, 2014; Wetzel & Carstensen, 2014) is that category parameters in Rasch-type models for ordered responses are unrelated to category order. Under this stance, some method is still needed to assess a potential mismatch between empirical and preassumed orders of categories. The nominal response model (NRM; Bock, 1972, 1997) is a suitable candidate for such query, given that it does not preassume any order and its parameters can yet be interpreted in the context of ordered categories. In fact, the NRM differs

from Rasch-type models for ordered responses in that (1) it does not preassume any order for the categories although (2) the conditional probability of a response in some arbitrary category given that the response is in it or in another arbitrary category is still a two-parameter logistic function of trait level but (3) the slope and location of each of those functions may vary across pairs of categories within and across items. As discussed later, an order for the categories can nevertheless be estimated post hoc, which places the NRM in the general context of models for ordered categories although not quite of the Rasch type: The slopes of the conditional probability functions just mentioned are not constant.

In the NRM for K -category items, the option response function (ORF) describing the probability p_k of a response in category k (for $1 \leq k \leq K$) as a function of trait level θ is

$$p_k(\theta) = \frac{\exp[a_k\theta + c_k]}{\sum_{m=1}^K \exp[a_m\theta + c_m]}, \quad (1)$$

subject to the identifiability constraint $\sum_{k=1}^K a_k = \sum_{k=1}^K c_k = 0$. (A subscript denoting the item is omitted on the understanding that parameters vary across items.) In the NRM, the order of categories in an item is given by the order of the a parameters (Bock, 1972, 1997; Muraki, 1992; Samejima, 1996; Thissen, Steinberg, & Fitzpatrick, 1989), as shown next. For any arbitrary pair of categories (say, k and l) the conditional probability of a response in category l given that the response is in categories k or l is

$$\frac{p_l(\theta)}{p_k(\theta) + p_l(\theta)} = \frac{1}{1 + \exp[(a_k - a_l)\theta - (c_l - c_k)]}.$$

This function is monotone increasing if $a_k < a_l$ and, in the context of ordered categories, this implies that category k precedes category l . This property allows post hoc estimation of the order of categories via pairwise analyses, which invariably leads to the result that the order of categories is exclusively given by the order of the a s. Thus, if the response categories of actual items are empirically ordered as preassumed, estimated NRM parameters should satisfy $a_k < a_{k+1}$ for $1 \leq k < K$.

This NRM feature has been used to assess the order of categories in many studies (e.g., González-Romá & Espejo, 2003; Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2013; Hernández, Espejo, & González-Romá, 2006; Maydeu-Olivares, 2005; Murray, Booth, & Molenaar, 2016; Preston et al., 2011, 2015; see also Bock, 1997; Samejima, 1972, 1996; Thissen & Cai, 2016). These studies show that the NRM tags a relatively large number of items as having disordered categories, something that seems suspicious. Extensive simulations later confirmed that NRM calibration massively returns purportedly disordered categories for items whose categories are indeed ordered (García-Pérez, 2017). This outcome questions the utility of NRM calibration for assessing the empirical order of categories, but it deserves further scrutiny.

Figure 1. Top row: Option response functions (ORFs) under the nominal response model for five-category items with ordered (left panel) or disordered (right panel) category parameters (a_s) and with disordered crossings (τ_s) in both cases (see the tables on the right of each panel; crossings are computed from a_s and c_s via Equation 7). Locations of the crossings of consecutive categories in their preassumed order are indicated by arrows in the panels. Bottom row: Product of the ORFs in the top panels and a unit-normal density function. The area under the product functions gives the expected marginal probability distribution of responses, indicated in the tables on the right of each panel.

Although part of these results may be due to sampling error, the most common cause of disorderly estimated categories was found to be excess flexibility and insufficient restraint in parameter estimation. Because the NRM is unconstrained as to order, parameter estimates contradicting the preassumed order of categories can be obtained when an alternative solution with category parameters in the pre-assumed order is equally tenable. Figure 1 illustrates this point in a simplified situation that does not misrepresent the actual problem in any respect. The top row shows item ORFs for two alternative parameter sets, which differ as to the purported order of its five categories. On the left, categories are ordered as preassumed (i.e., $a_k < a_{k+1}$ for $1 \leq k \leq 4$); on the right, categories are portrayed as disordered (i.e., $a_3 < a_2$ and $a_3 < a_1$) and the preassumed middle category seems to be first. The two sets of ORFs differ almost only at the lower range of θ (say, for $\theta < -1$), where data that could resolve which parameters best suit the item are scarce under a unit-normal distribution of θ : Only about 16% of the respondents will have θ s in that range and their responses will be distributed across several categories, hampering the estimation of the shape of the ORFs in that range.

To make the problem more conspicuous, the bottom row of Figure 1 shows the product of the ORFs with a unit-normal probability density function, a crucial step in marginal maximum likelihood estimation of item parameters. The product functions are virtually identical in both cases and, naturally, integration renders also a nearly identical expected distribution of responses across categories. Clearly, in both cases the expected marginal distribution will fit nearly identically the empirical marginal

distribution that might have given rise to both sets of parameter estimates. In other words, orderly or disorderly estimated categories on NRM calibration may only be associated with inconsequential aspects of the ORFs at extreme ranges of θ , when nothing else in the data actually indicates disordered categories. Yet NRM calibration seems too prone to producing item parameters purporting disordered categories when an alternative account with ordered categories seems equally tenable.

When disordered estimates arise for Likert-type items whose categories express levels of agreement with a statement or frequency of occurrence of a symptom, the extraordinary claim of disordered categories requires strong support. Empirical evidence to back up such claim is hard to obtain, but one can alternatively ask whether item parameters exist that account for the data equally well without contradicting the preassumed order of categories. If the answer is positive, there is certainly no reason to claim disordered categories. To answer this question, an amended calibration procedure is needed that enforces the preassumed order of categories by estimating NRM parameters under the order constraint $a_k \leq a_{k+1}$. Such an option is unavailable in existing software but it can be easily implemented as described in the next section. Existing item response theory models for ordered responses enforce additional assumptions beyond the order constraint and they cannot be used for this purpose (see the “Discussion” section).

The main goal of this research was to investigate the outcomes of conventional NRM calibration (i.e., unconstrained as to order) and compare them with those of order-constrained NRM calibration of the same data. Items whose categories are ordered as preassumed will ideally not be purported to have disordered categories in an unconstrained calibration but, if the latter occurs, constrained calibration should offer an analogous account of the data. At the same time, items whose categories are not ordered as preassumed should be poorly fitted in a constrained calibration. These issues are addressed in a number of simulation studies that cover large and small samples of respondents as well as sets of items with different types of true characteristics, including ordered or disordered categories. Empirical illustrations are also provided using existing data sets.

In a nutshell, the results show that items with ordered categories are often purported to have disordered categories in unconstrained calibration (even with large samples), but order-constrained calibration provides an identical account of the data. In contrast, items whose categories are truly disordered cannot be portrayed in an order-constrained calibration. These results document the validity of a method that assesses the empirical order of categories by comparing the outcomes of order-constrained and unconstrained NRM calibrations.

Constraining the NRM to Enforce the Preassumed Order of Categories

Incorporating the constraint $a_k \leq a_{k+1}$ implies estimating parameters subject to the conditions

$$\begin{cases} a_1 < 0 \\ a_k \in \left[a_{k-1}, \frac{-\sum_{m=1}^{k-1} a_m}{K+1-k} \right], & 2 \leq k < K. \\ a_K = -\sum_{k=1}^{K-1} a_k \end{cases} \quad (2)$$

The first condition arises because the zero-sum constraint $\sum_{k=1}^K a_k = 0$ requires a_1 to be negative-valued; the second condition incorporates the order constraint by bounding each intermediate a_k (for $2 \leq k < K$) between a minimum at a_{k-1} and a maximum at the largest value that it might have if all subsequent a_m (for $k < m \leq K$) were to be equal to a_k ; the third condition simply states that a_K is not a free parameter under the zero-sum constraint. On the other hand, c parameters are still estimated subject only to their zero-sum constraint.

Estimates of a parameters must thus satisfy the linear inequalities

$$\begin{cases} a_k \leq a_{k+1}, & 1 \leq k < K-1 \\ a_{K-1} \leq \frac{-\sum_{k=1}^{K-2} a_k}{2} \end{cases}, \quad (3)$$

with the nonfree parameter a_K computed as $a_K = -\sum_{k=1}^{K-1} a_k$.

A custom program was written in MATLAB for marginal maximum likelihood estimation of NRM item parameters with a choice to enforce or not to enforce the order constraint. This program was used in all the analyses presented next but MULTILOG (du Toit, 2003) was also used to confirm that our program does not produce spurious outcomes not observed when unconstrained calibration is obtained with a conventional software package. Note that MULTILOG was only used for the stated purpose, as it cannot estimate NRM parameters under an order constraint on the a s. Many other programs are available for unconstrained NRM estimation but comparative analyses have shown that they all yield analogous results (see, e.g., Kim, Bao, Horan, Kim, & Cohen, 2015; Peterson et al., 2014; Preston & Reise, 2014).

Assessing Category Order by a Comparative Analysis of Calibrations

Unconstrained calibration may return disorderly estimated categories for items with truly ordered categories, but a constrained calibration should give an analogous account of the data with ordered categories. On the other hand, constrained and unconstrained calibrations are expected to differ meaningfully for items with truly disordered categories, and in these cases the unconstrained calibration should account for the data better. Only some items (the *suspect* items) may turn up with disorderly estimated categories in unconstrained calibration and a procedure is needed to determine whether such disorder is spurious or authentic. Judging whether or not estimated ORFs only differ across calibrations in inconsequential ranges of θ (as in the illustration of Figure 1) seems appropriate, but such judgments are not dependable.

Criteria are discussed next for assessing whether or not both calibrations account for the data equally well.

Item parameters estimated in either calibration can be used to obtain expected a posteriori estimates $\hat{\theta}$ of the respondents' trait levels as

$$\hat{\theta}_i = \frac{\int_{-\infty}^{\infty} \theta \varphi(\theta) \prod_{j=1}^J P_{R_{ij}}(\theta) d\theta}{\int_{-\infty}^{\infty} \varphi(\theta) \prod_{j=1}^J P_{R_{ij}}(\theta) d\theta}, \quad (4)$$

where φ is the unit-normal density function and R_{ij} is the i th respondents' response to the j th item (with $1 \leq R_{ij} \leq K$ and $1 \leq j \leq J$). If both calibrations render virtually identical trait estimates despite the presence of suspect items, differences between ORFs (which are involved in trait estimation; see Equation 4) must be understood as inconsequential and, hence, the disorderly estimated categories of suspect items must be regarded as spurious. On the other hand, trait estimates from either calibration will not be tightly related if the disorderly estimated categories of suspect items are authentic. The utility of this criterion will be demonstrated below, even though it obviously mixes up suspect and nonsuspect items.

Goodness-of-fit indices can also be obtained for each individual item in each calibration by splitting the sample of N respondents into G groups according to their estimated trait levels, computing the observed count O_{jkg} of responses in each category in each group, computing the expected count E_{jkg} of responses in each category in each group, and finally computing the likelihood-ratio statistic

$$G_j^2 = 2 \sum_{g=1}^G \sum_{k=1}^K O_{jkg} \log \left(\frac{O_{jkg}}{E_{jkg}} \right), \quad (5)$$

which asymptotically has a χ^2 distribution with $(G - 2)(K - 1)$ degrees of freedom. In our present use of this criterion, the number G of groups was a minimum of 5 and a maximum of 15 and it varied with N and K so as to minimize the chances of large sparse tables with small expected counts that threaten the accuracy of the test (García-Pérez, 1994; García-Pérez & Núñez-Antón, 2004). Small expected counts are nonetheless unavoidable and compromise a comparative analysis of calibrations so that item misfit under some calibration does not necessarily speak of the inadequacy of the corresponding assumption about category order. Nevertheless, if the estimated disorder of suspect items is authentic, one should expect to find consistent significant misfit for those items under the constrained calibration but not under the unconstrained calibration. The presence (or lack thereof) of such differential patterns of misfit across calibrations will be demonstrated later, which attest to the utility of this criterion for an assessment of category order in suspect items.

An alternative measure that is free of statistical artifacts is the value of twice the negative log-likelihood ($-2 \log L$) of the data given the estimated parameters, computed for each item as

$$-2 \log L_j = -2 \sum_{i=1}^N \log(p_{R_{ij}}(\hat{\theta}_i)). \quad (6)$$

This statistic is deprived of an inferential framework but it allows a dependable comparison of the appropriateness of item parameters from alternative calibrations. If the estimated disorder of suspect items is authentic, $-2\log L$ should be meaningfully smaller for those items with unconstrained parameter estimates; if the estimated disorder is spurious, $-2\log L$ should be nearly identical whether computed with constrained or unconstrained parameter estimates.

Finally, standard errors for estimated item parameters (which are immediately available from the diagonal elements in the Hessian matrix at the last iteration) are not useful in a comparison of calibrations and they will not be reported here. For the record, they did not differ meaningfully or systematically across calibrations for suspect or nonsuspect items.

Simulation I: Large Samples, Ordered Categories, and Ordered ORF Crossings

Besides comparing calibrations, an important goal of this simulation was to evaluate our software. This includes not only checking its ability to recover the generating parameters but also comparing its unconstrained estimates with those obtained with MULTILog. Large samples of respondents were used and true item parameters were created so that categories are ordered as preassumed (i.e., $a_k \leq a_{k+1}$ in all items). The study involved numbers J of items between 20 and 50, numbers K of response categories between 3 and 7, and sample sizes N between 1,000 and 5,000 respondents. Each simulation was also repeated several times with different initial seeds for the random number generators. Results were thoroughly analogous across the board and only results for one of the simulations with $N = 2,000$, $J = 25$, and $K = 5$ are presented.

True θ s were randomly drawn from a unit-normal distribution and a parameters were generated to satisfy the conditions in Equation (2). For each item, a_1 was randomly drawn from a uniform distribution on $[-5, -0.5]$, consecutive a_k (for $2 \leq k < K$) were then randomly drawn from uniform distributions bounded as described in Equation (2), and a_K was then computed as indicated also in Equation (2). The c parameters were indirectly generated via conditions on the crossings τ_k (for $1 \leq k < K$) between the ORFs for consecutive categories, given by

$$\tau_k = \frac{c_{k+1} - c_k}{a_k - a_{k+1}}. \quad (7)$$

A vector of $K - 1$ ordered crossings was generated for each item by drawing random numbers from a uniform distribution on $[-2.5, 2.5]$ and sorting them in ascending order so that $\tau_k \leq \tau_{k+1}$. The c parameters were then computed by solving the system of Equation (7) using the item τ s and a s. Note that category order is only determined by the order of the a s, whether the τ s are ordered or disordered (see Figure 1 for

examples of disordered crossings with categories that are ordered or disordered). Then, with parameters generated as in this simulation, ordered *as* ensure that categories are ordered as preassumed whereas ordered *ts* ensure that they are sequentially modal or super-modal as defined by Linacre (2004): Category *k* is modal in the range of θ for which $p_k(\theta) > p_m(\theta)$ for all *m*, if any such range exists; otherwise category *k* is nonmodal. Category *k* is super-modal if it is modal by the preceding definition and $p_k(\theta) > 0.5$ at some point within that range. The ORFs of sample items will be shown below.

Responses were then generated by a *K*-bin multinomial process with a probability vector whose values were computed using the respondent's θ and the item ORFs. The resultant data array was then submitted to constrained calibration with our software and to unconstrained calibration with our software and with MULTILOG.

Figure 2 shows scatterplots of item parameter estimates against their true values for constrained (top row) and unconstrained (center row) calibration with our software and also for unconstrained calibration with MULTILOG (bottom row). Data points are tightly packed around the diagonal in all panels, indicating that all parameters were accurately recovered.

When samples of respondents are sufficiently large to permit accurate recovery, enforcing the order constraint on the *as* (which holds here by design) seems unnecessary. In other words, unconstrained estimates of *a* parameters should turn up ordered just as their true counterparts are. Yet unconstrained calibration returned disordered estimates for Items 3, 6, 7, 9, 10, 12, and 13, both with our software and with MULTILOG. Seven items with disorderly estimated categories represent 28% of the items on the test, a large fraction indeed. Estimated disorder for so many items could be taken to mean that a sample of 2,000 respondents is still insufficient to recover the true parameters precisely, but it definitely indicates that disorderly estimated categories are a common outcome not necessarily reflecting an underlying reality, and an outcome that may occur even when sample size is unquestionably large.

Figure 3 plots true and estimated ORFs for Item 1 (a nonsuspect item by unconstrained calibration) and for Items 3 and 9 (suspect items). For the three items, estimated ORFs are virtually identical to their true ORFs, whether estimation is order-constrained (second row in Figure 3) or unconstrained (third and bottom rows). Note also that unconstrained calibration produces virtually the same ORFs with our software and with MULTILOG despite minor differences in parameter values.

The true ORFs for Item 1 clearly display the defining item characteristics in this simulation: Categories are ordered as preassumed and become progressively modal as θ increases. Large separation between consecutive *as* in this item makes all categories discriminative. Hence, responses to this item are sufficiently informative for unconstrained calibrations to capture the actual order of categories. The true ORFs for Item 3 also reflect ordered categories with progressive modality, but the smaller distance between some of the consecutive *as* produces nondiscriminative intermediate Categories 2 to 4 whose order is understandably misconstrued in unconstrained calibrations; the same holds for the nondiscriminative Categories 3 and 4 in Item 9.

Figure 2. Scatterplots of parameter estimates versus true values in Simulation I under the order-constrained (top row), unconstrained (center row), and MULTILOG (bottom row) calibrations. Columns show a s (red symbols) and c s (blue symbols) for each category.

The presence of nondiscriminative categories might be seen as the cause of estimated disorder and something that demands collapsing, but it will be seen later that estimated disorder also occurs with discriminative categories.

When unconstrained calibration returns ordered category parameters (as for Item 1 in Figure 3), enforcing the order constraint should not result in different parameter estimates if all other estimation options remain identical. Parameter estimates listed on the right of the second and third panels down the first column of Figure 3 confirm this. Small differences will later be seen to occur in other cases but such eventuality is easily understood: Imposing inequalities forces the optimization algorithm to explore the parameter space differently, something that may prevent it from arriving at the exact same solution in both cases.

It is also interesting to note that constrained calibration does not seem to affect parameter estimates other than those which turned up disordered in the unconstrained calibration. For instance, in Item 3 (center column in Figure 3), unconstrained calibration rendered $a_3 = -0.15 > a_4 = -0.20$ whereas constrained calibration rendered $a_3 = a_4 = -0.18$ with no change in the value of any other parameter. This is only because estimated disorder in this illustration with large sample sizes occurs for non-discriminative categories whose respective true a s are similarly valued. It will later be seen that alternative calibrations may affect all parameters, and also that disordered parameters in the unconstrained calibration are not always estimated to have the same value in a constrained calibration.

Figure 3. True option response functions (ORFs) for three items in Simulation 1 (top row) and estimated ORFs under the order-constrained (second row), unconstrained (third row), and MULTLOG (bottom row) calibrations. The three calibrations return ordered categories for Item 1 (left column); yet unconstrained calibrations return disordered categories for Items 3 (center column) and 9 (right column) without producing meaningfully different ORFs.

Given the accurate parameter recovery, true and estimated θ s should only differ because of sampling error. Also, estimated θ s under both calibrations should be virtually identical if estimated disorder is inconsequential. All this is confirmed in Figure 4.

Results presented thus far make clear that estimated disorder in unconstrained calibration is spurious, but it is also worth noting that fit indices corroborate this conclusion. Fit statistics with $G = 15$ were significant at $\alpha = .05$ only for Item 8 in the unconstrained calibration and for Items 2 and 8 in the constrained calibration. Both calibrations returned ordered parameter estimates for these two items. Thus, neither item is suspect of disordered categories and there is no room for wondering whether ordered or disordered categories fit these items better. On the other hand, values of $-2\log L$ ranged across items between 1,762 and 2,906 in both calibrations and the largest absolute difference between calibrations for any individual item was 0.2. Again, both types of calibration offer the same account of the data by this criterion.

In sum, Simulation 1 proves the adequacy of our software for NRM parameter estimation. Also, constrained estimation shows that disordered estimates in unconstrained calibrations can be meaningless: They do not reflect any underlying truth and the account that disordered parameters provide is no better than an account with ordered parameters. The conclusion does not require foreknowledge of the true item

Figure 4. Scatterplots of trait estimates and true values in Simulation 1. Trait estimates are equally related to true traits whether obtained using order-constrained (a) or unconstrained (b) item parameters estimated with our software, obviously because both types of trait estimate are virtually identical (c). Unconstrained trait estimates from our software are also virtually identical to those obtained with MULTILOG (d).

parameters: A decision on whether disordered categories must be claimed rests only on a comparison of the outcomes of unconstrained and constrained calibrations. This simulation illustrated spuriously estimated disorder in an admittedly narrow scenario involving items with nondiscriminative categories and large samples. Subsequent simulations demonstrate the same problem in alternative scenarios.

Simulation 2: Small Samples, Ordered Categories, and Ordered ORF Crossings

This simulation assessed whether small samples are any more likely to render disorderly estimated categories, or whether they hamper the identification of spurious disorder. Simulation 2 is thus identical to Simulation 1 with only 200 respondents.

Not unexpectedly, scatterplots like those in Figure 2 revealed less accurate estimation in all cases. Unconstrained calibration with our software or with MULTILOG returned disorderly estimated *as* for the same eight items (3, 5, 6, 7, 10, 12, 13, and 18), only one more than in Simulation 1. Then, spuriously disordered categories are no more prevalent with small samples. Interestingly, Item 9 was suspect in Simulation 1 but not in Simulation 2. This may come as a surprise but it only makes clear that large samples do no protect against disordered parameter estimates in unconstrained calibrations. The complex multidimensional parameter space seems to permit alternative and equally fitting solutions with ordered or disordered *as* almost independently of sample size.

True and estimated ORFs for three items in this simulation are shown in Figure 5. Items 1 (left column) and 3 (center column) were also in Figure 3 and a comparison confirms that estimated ORFs are not as close to true ORFs when parameters are estimated with fewer data. Item 3 is suspect here and Item 1 is not, just as in Simulation 1. Estimated disorder affects the three nondiscriminative intermediate categories of Item 3, although they are purported to be completely reversed in this case. In suspect

Figure 5. True option response functions (ORFs) for three items in Simulation 2 (top row) and estimated ORFs under the order-constrained (second row), unconstrained (third row), and MULTLOG (bottom row) calibrations. The three calibrations return ordered estimates with virtually identical ORFs for Item 1 (left column); yet unconstrained calibrations return disordered categories for Items 3 (center column) and 12 (right column) and produce ORFs that differ meaningfully from those arising from a constrained calibration.

Item 12 (right column of Figure 5), Categories 3 to 5 are nondiscriminative and unconstrained parameter estimates indicate that the middle category is the highest.

A comparison of Figures 3 and 5 reveals effects of sample size on parameter estimates in calibrations of either type. For Item 1 (left column in Figures 3 and 5), whose categories were orderly estimated in both calibrations at both sample sizes, estimated ORFs do not differ across calibrations and are only closer to the true ORFs with the larger sample size. Yet for items whose categories are disorderly estimated in unconstrained calibrations (e.g., Item 3; center column in Figures 3 and 5), small samples make estimated ORFs differ markedly across types of calibration (compare the second and third panels down the center column of Figure 5), differences that were not found with larger samples (see the second and third panels down the center column of Figure 3). In these cases, constrained ORFs are also closer than unconstrained ORFs to the true ORFs (see the right column of Figure 5 for another example).

Scatterplots like those in Figure 4 revealed the exact same pattern, with virtually identical trait estimates from constrained or unconstrained item parameters. Fit statistics with $G = 5$ were in this case significant for Items 2 and 10 under both calibrations. Of these, only Item 10 was suspect of disordered categories by the unconstrained calibration, but significant statistics in both calibrations fails to

provide the extraordinary evidence required to claim disordered categories in it. On the other hand, values of $-2\log L$ ranged across items between 161 and 287 in both calibrations and the largest absolute difference between calibrations was 5.26. As for the suspect Item 10, the difference was 0.72 (286.82 vs. 286.10).

In sum, small samples are not more likely to give rise to disorderly estimated categories for items with ordered categories, and identification of spurious disorder is not hampered. With small samples, estimated ORFs differ more across calibrations for suspect items but, as in Simulation 1, disorder does not account for the data better.

Simulation 3: Large Samples, Ordered Categories, Extreme ORF Crossings

The preceding simulations used items whose ORFs crossed within the central range of a unit-normal distribution of θ . In this simulation, crossings covered broader ranges of θ across items so that they cooperatively allow person discriminations at all ranges of θ . Simulation 3 is linked to one of the studies in which estimated disorder was reported (Murray et al., 2016) and, thus, uses $K = 3$ with $N = 1,000$. Twenty-four items were used so that exactly one third of them had all their crossings in each of three broad ranges of θ : $\theta < -2$ (Items 1-8), $-2 < \theta < 2$ (Items 9-16), and $\theta > 2$ (Items 17-24). Parameter generation proceeded as before except that the (random) location of crossings was bounded in each group as just described.

Figure 6 shows results for one item in each group. When true crossings occur in the low range of θ (see Item 8 in Figure 6), unconstrained calibration returns disordered Categories 1 and 2; this outcome occurred for three of the eight items in the first group. Analogously, when true crossings occur in the high range of θ (see Item 24 in Figure 6), unconstrained calibration returns disordered Categories 2 and 3, an outcome that also occurred for three of the eight items in this group. In contrast, central crossings resulted in orderly estimated categories for all items (see Item 15 in Figure 6). As in previous results also with large samples, constrained (second row in Figure 6) or unconstrained (third and bottom rows) ORFs did not differ meaningfully, whether or not categories were orderly estimated in both cases. Trait estimates were also virtually identical across calibrations. Finally, fit statistics with $G = 15$ were significant for three (in the constrained calibration) or four (in the unconstrained) items in the second group, which were not suspect items by the unconstrained calibration.

Two aspects of these results are worth discussing. The first one is that estimated disorder was associated with J-shaped response distributions in which one of the extreme categories is massively endorsed. All items in the first and third groups produced these distributions but disorder was estimated for only three items in each group, suggesting that minor differences in the counts of responses across rarely endorsed categories play some role in the estimation of disorder. These features agree with the incomplete information provided by Murray et al. (2016) for three-category items in which an unconstrained NRM calibration portrayed the intermediate category to be highest or lowest on some items. They reported that intermediate categories had

Figure 6. True option response functions (ORFs) for three items in Simulation 3 (top row) and estimated ORFs under the order-constrained (second row), unconstrained (third row), and MULTLOG (bottom row) calibrations. With the extreme locations of crossings in Items 8 (left column) and 24 (right column), unconstrained calibrations return disordered categories; with the central location of crossings in Item 15 (center column) categories are orderly estimated.

low or very low endorsement rates in such items and estimated ORFs for one of them had their crossings at low θ s, indicating that Category 3 was massively endorsed in this item (see their figure 5). Additional details were not provided for this or other items regarding whether the response distributions were J-shaped only (or mainly) in the items whose categories turned up disordered in the NRM calibration.

The second aspect relates to the fact that disorderly estimated categories occurred only for items whose true ORF crossings were placed far out, and for which estimated crossings were placed accordingly. Such items are common in patient-reported outcome scales (see, e.g., Sharp et al., 2012). Nonetheless, the relevant issue is the location of the ORFs relative to the trait distribution of the respondents. In empirical studies with sample trait distributions that do not spread far out (in one or both directions), items with crossings at less extreme locations will be affected by this problem. This will become apparent in the empirical examples presented later.

Simulation 4: Large Samples, Mixture of Ordered and Disordered Categories

The preceding simulations investigated whether encountering disorderly estimated categories in an unconstrained calibration is sufficient evidence to claim disordered

categories. All results concur in indicating that it is not: A constrained calibration offers essentially the same picture and accounts for the data equally well with ordered categories. The immediate next question, addressed now, is whether constrained calibration will also provide a fitting account when items have truly disordered categories.

This simulation also used 24 three-category items and a sample of 1,000 respondents. Item parameters were generated as before, but a parameters were rearranged to render disordered categories. Four groups of six items were thus created: one in which categories remained ordered, one in which Categories 1 and 2 were flipped, one in which Categories 2 and 3 were flipped, and one in which Categories 1 and 3 were flipped. With $K = 3$, the latter group implies completely reversed categories, as might occur when responses to reverse-worded items are not coded in reverse. Crossings were created ordered before the a s were rearranged and, then, the crossings between factually consecutive categories ended up disordered for some items.

Parameters of items in the first group were estimated in their true order in both calibrations and graphical results for these items are omitted. All items in the three other groups were estimated to have disordered categories in unconstrained calibrations. Figure 7 shows results for an item in each of those suspect groups. Clearly, constrained calibration (third row in Figure 7) fails to recover the true ORFs (second row) whereas unconstrained calibration (fourth and bottom rows) recovers them adequately. Without knowledge of the true ORFs, five lines of evidence linked to our criteria suggest that category disorder is authentic here.

The first one comes from a comparison of ORFs across calibrations. When categories are reversed (Item 19 in Figure 7), constrained ORFs are entirely flat and only their height differs to reflect the observed proportion of responses in each category (top row in Figure 7), whereas unconstrained ORFs give a different picture with reversed categories and good-looking ORFs. The evidence of reversed categories is overwhelming: If the item content were unrelated to the trait as the constrained ORFs suggest, unconstrained ORFs would have been identically flat. On the other hand, when disorder is only partial and affects Categories 1 and 2 (Item 8 in Figure 7) or Categories 2 and 3 (Item 15), constrained and unconstrained ORFs differ only for the suspect categories (i.e., red and green curves in the left column of Figure 7 and blue and green curves in the center column) and only at the lower or upper range of θ . In these ranges, constrained ORFs are flat whereas unconstrained ORFs are not. Analogous patterns of differences were sometimes observed in previous simulations (see, e.g., Item 12 in Figure 5) but it seems that the differences do not occur here at outer and inconsequential ranges of θ .

The second line of evidence comes from the response distribution (top row in Figure 7) and relates also to whether constrained and unconstrained ORFs differ at inconsequential ranges of θ . If this were the case, estimated disorder would occur for rarely endorsed categories due to minor differences in the small counts of responses in such categories, but there is no sign of this in the response distributions for Items 8 and 15 in Figure 7 (or other items in their groups).

Figure 7. Response distributions for three items with disordered categories in Simulation 4 (top row), true option response functions (ORFs; second row) and estimated ORFs under the order-constrained (third row), unconstrained (fourth row), and MULTILOG (bottom row) calibrations. In Item 8 (left column), Categories 1 and 2 are disordered; in Item 15 (center column), Categories 2 and 3 are disordered; in Item 19 (right column), categories are reversed.

The third line of evidence comes from a comparison of estimated θ s from constrained and unconstrained item parameter estimates. Figure 4c showed that trait estimates were virtually identical in both cases in Simulation 1. Figure 8a shows that this was also true under the analogous conditions of Simulation 3 (with $K = 3$ instead). In contrast, large differences between constrained and unconstrained ORFs in the conditions of Simulation 4 produce a substantial scatter (Figure 8b), even when trait levels are estimated without the six items whose categories were reversed (Figure 8c). This result provides additional evidence that the differences between constrained and unconstrained calibrations are not inconsequential.

The fourth line of evidence comes from item fit statistics. With $G = 15$, fit statistics were significant in the constrained calibration for the 18 suspect items (i.e., all items in the last three groups) and also for the nonsuspect Item 6 (from the first group). In contrast, only the fit statistic for Item 13 (from the third group) was significant in the unconstrained calibration. Massive significance in the constrained

Figure 8. Scatterplot of constrained versus unconstrained estimates of trait levels in Simulation 3 (a), in Simulation 4 using all items in the test (b), and in Simulation 4 without using the six items whose constrained option response functions turned up flat (c).

calibration along with lack of significance in the unconstrained calibration for suspect items seems too systematic to reflect Type I errors or statistical artifacts due to small expected frequencies. Statistical evidence of misfit is dependable when it consistently favors unconstrained calibration for suspect items.

The fifth line of evidence comes from the values of $-2\log L$ in each calibration, shown in Figure 9. The six nonsuspect items (with orderly estimated categories in both calibrations) are plotted as green circles and they all lie near the diagonal, paralleling the distinctive equality described in preceding simulations. Suspect items (estimated to have disordered categories by the unconstrained calibration) are plotted in other colors. Clearly, $-2\log L$ is meaningfully smaller with unconstrained estimates for those items. Some colored symbols are perhaps sufficiently close to the diagonal to consider that constrained parameters (i.e., ordered categories) account for the data reasonably well too. These are indeed the items in the second and third groups, whose categories were created to be only partially disordered. In addition, the red or blue circles closer to the diagonal come from the items whose ORF crossings occurred at the farther locations, where even true disorder is arguably inconsequential. Altogether, converging evidence from the five sources justifies the claim of disordered categories in the last three groups of items along with ordered categories for items in the first group. The criteria thus successfully identify items with truly disordered categories. Needless to say, consideration of these sources of evidence is crucial in empirical studies in which the true item parameters and ORFs are unknown and, hence, the question of whether or not their categories are ordered is under blindfolded investigation.

Simulation 4 was repeated with analogous results with mixtures of items in which disorder was not as prevalent as it was here. In sum, Simulation 4 and the variants just mentioned prove that items with truly disordered categories will not pass unnoticed in a comparison of constrained and unconstrained calibrations. Although none of the

Figure 9. Scatterplot of constrained versus unconstrained values of $-2\log L$ for the items in Simulation 4. Color denotes group membership as indicated in the inset. Items in Group 1 were not suspect of disordered categories by previous analyses.

lines of evidence is conclusive individually (except when categories are estimated to be completely reversed), their joint occurrence is a dependable indicator of authentic disorder.

Empirical Example 1

An application of our procedure is presented here for the analysis of category ordering in an empirical test whose items obviously have unknown ORFs. Data come from the administration of a 37-item inventory to 262 respondents. The inventory was developed as a classroom exercise to create an instrument that measures interest in soccer and it consisted of six-category Likert-type items. Omissions occurred that did not seem relevant: There were 45 omitted responses altogether (13 coming from a single respondent) which were scattered across 27 items, and the largest number of omissions on any individual item was 3. Some items were worded in reverse and responses were accordingly coded to prevent the trivial outcome of reversed categories in those items. Calibrations were obtained as before.

Unconstrained calibration with our software returned disordered categories for 17 items (46% of the items) and calibration with *MULTILOG* returned disordered categories for the exact same items and one more. Figure 10 shows response distributions and estimated ORFs for three items. Item 22 (left column of Figure 10) is representative of the 20 items for which both types of calibration returned ordered categories and virtually identical ORFs. All categories are modal in this item; other items shared this characteristic but still others had nonmodal categories. It is also apparent that *MULTILOG* estimates the ORFs to be slightly displaced leftward, something that occurred for all items. This only indicates a different representation of the latent continuum, which did not manifest in simulations because of the unit-normal trait distributions. The implied transformation of the trait scale will be discussed later.

Figure 10. Response distributions for three items in empirical Example 1 (top row) and estimated option response functions under the constrained (second row), unconstrained (third row), and MULTLOG (bottom row) calibrations.

For suspect Item 27 (center column in Figure 10), ORFs differ across calibrations only at $\theta > 1.5$. Unconstrained calibration purports that categories are ordered as 1–2–5–3–6–4, which is difficult to interpret with category labels that express increasing levels of agreement. The J-shaped distribution in the top panel indicates that Categories 4 to 6 were, respectively, endorsed by only 5, 4, and 3 respondents and it is hardly contentious that the ORFs for these categories were not estimated accurately. An analogous outcome arose for suspect Item 29 (right column in Figure 10), where unconstrained calibrations returned partial disorder affecting Categories 5 and 6. These are also the only categories whose ORFs differed across calibrations and only for $\theta > 2$. A J-shaped distribution with nine endorsements of Category 5 and eight endorsements of Category 6 seems insufficient to claim that these categories are reversed.

With $G = 5$, significant fit statistics occurred for 8 items in the unconstrained calibration and for 10 items in the constrained calibration, which included the exact same 8 items and two more. Of the 17 items suspect of disordered categories, 5 were in both lists, 2 were only in the list for constrained calibration, and the remaining 10 were not in any of the lists. Although fit statistics are not entirely dependable, there is nothing in these results indicating that suspect items are better described with unconstrained parameters. A comparison of $-2\log L$ across calibrations (Figure 11a) reveals

Figure 11. (a) Scatterplot of constrained versus unconstrained values of $-2\log L$ for the items in empirical example 1. Open symbols depict items above suspicion of having disordered categories; solid symbols depict items suspect of having disordered categories. (b) Scatterplot of constrained versus unconstrained estimates of trait levels obtained with our software. (c) Scatterplot of unconstrained estimates obtained with our software versus multiLOG.

that suspect items (solid symbols) lie on the identity line just as nonsuspect items (open symbols) do.

A scatterplot of traits estimated with constrained and unconstrained item parameters (Figure 11b) reveals virtual identity, further supporting the notion of spuriously estimated disorder in suspect items. Figure 11c shows a tight relation between traits estimated with unconstrained item parameters obtained with our software and with MULTILOG. The displacement of the string of symbols about 0.8 units down from the identity line is consistent with the leftward shift in the estimated ORFs (compare the two bottom rows in Figure 10). This only reveals different representations of the trait continuum in our software and in MULTILOG.

In sum, although an unconstrained NRM calibration returned disordered categories for 17 of the 37 items in this inventory, there are no signs in the analyses subsequently conducted that might justify the claim that those items have disordered categories.

Empirical Example 2

This example is analogous to the previous one and uses data downloaded from the archives of the International Personality Item Pool (<http://ipip.ori.org>). Data consist of responses from 1,071 individuals to the *Humor Styles Questionnaire* (HSQ; Martin, Puhlik-Doris, Larsen, Gray, & Weir, 2003), a 32-item inventory with eight 5-category Likert-type items contributing to each of four dimensions of humor use (affiliative, self-enhancing, aggressive, and self-defeating). The data were retrieved from http://personality-testing.info/_rawdata/HSQ.zip. Eleven items were reverse worded (see table 1 in Martin et al., 2003) and responses were recoded before analysis. The data were calibrated separately for each scale so that there are only eight items in each individual calibration. There were also a relatively large number of

Figure 12. Response distributions for three items in empirical Example 2 (top row) and estimated option response functions under the constrained (center row) and unconstrained (bottom row) calibrations.

omissions which affected the size of the sample of respondents for each scale and also the number of available responses to each individual item on each scale.

Results of the MULTILOG calibration were equivalent to those of unconstrained calibration with our software and they will not be reported. Figure 12 shows results for three items in the affiliative humor scale, five of whose eight items (62.5%) were estimated to have disordered categories in the unconstrained calibration. Note the J-shaped distributions for Items 21 and 25 (top row in Figure 12), with only 12 and 15 responses, respectively, in Category 1, and only 29 and 32 responses in Category 2. Accordingly, constrained (center row) and unconstrained (bottom row) ORFs differed for these categories (yellow and red curves) but only at $\theta < -2$; constrained and unconstrained ORFs for Category 3 (green curves) also differed in that range on Item 25. By the unconstrained calibration, the estimated order of categories is 2–3–1–4–5 for both items: The intermediate Categories 2 and 3 appear to be lower than Category 1. The three other suspect items had analogous characteristics. The right column of Figure 12 shows results for a nonsuspect item in which the response distribution had only some skew. ORFs and parameters did not differ meaningfully across calibrations. Skewed response distributions and similar ORFs across calibrations occurred for the two other nonsuspect items.

Fit statistics were significant for the same six items in both calibrations. At the same time, a scatterplot of $-2\log L$ across calibrations placed all items virtually on the diagonal as in Figure 11a earlier. A scatterplot of estimated traits also showed data points virtually on the diagonal, as in Figure 11b. No evidence of true disorder is found in these results.

As for the three other HSQ scales, both calibrations rendered ordered categories with nearly identical parameter estimates and no further analyses were conducted for lack of suspect items. Analyses were also conducted on two other data sets. One of them consisted of the first 1,000 responses (out of 47,974) to the 10 four-category Likert-type items of the *Rosenberg Self-Esteem Scale* (retrieved from http://personality-testing.info/_rawdata/RSE.zip). Both calibrations rendered nearly identical parameter estimates with ordered categories and no further analyses were conducted for lack of suspect items. The other data set consisted of the 1,005 responses to the 40 five-category Likert-type items of the *IPIP Assertiveness, Social Confidence, Adventurousness, and Dominance scales* (with 10 items per scale; retrieved from http://personality-testing.info/_rawdata/AS+SC+AD+DO.zip). Unconstrained calibration estimated disorder for three of the items on the assertiveness scale and for eight of the items on the adventurousness scale. Differences between constrained and unconstrained ORFs in suspect items were identical to those described for Examples 1 and 2: J-shaped distributions often yielded different ORFs but only at extreme ranges of θ for the suspect categories and with no accompanying signs that disorder accounts for the data better.

Some of these analyses were repeated under conditions that mimic an inadvertent failure to recode reverse-worded items. The anomaly was instated on numbers of items that ranged from one to half the number of items on the scale. The results were identical irrespective of the number of “miskeyed” items: The ORFs of items not affected by this problem turned up nearly identical in both calibrations; for the remaining items, constrained calibration rendered flat ORFs whereas unconstrained calibration returned appropriate ORFs with reverse-ordered category parameters (as expected from the results of Simulation 4) and with all the accompanying signs that disorder is authentic (although, in fact, due to a coding error).

Discussion

Justification for the use of the NRM to assess the order of response categories rests on the assumptions that (1) successive category parameters will turn up ordered if the response categories are indeed ordered as preassumed and (2) if this is not the case, disordered estimates will indicate the order of categories. The fact that many items appear to have inexplicably disordered categories in empirical applications suggests that this strategy is not as rigorous and dependable as one might like to think. Simulation results presented here show that items with ordered categories are often estimated to have disordered categories by the NRM, even with large samples. This outcome arises in (at least) two distinct situations. One is when items include two or more nondiscriminative categories whose ORFs superimpose (see, e.g., Items 3 and 12 in Figure 5), producing response distributions in which minor irregularities in the counts of responses across the nondiscriminative categories determine how their order will be estimated. The other situation affects items with discriminative categories at extreme ranges of θ (see, e.g., Items 8 and 24 in Figure 6), producing J-

shaped response distributions in which irregularities in response counts across rarely endorsed categories also determine how their order will be estimated. The analysis of empirical data sets confirmed that estimated category disorder is associated with these response distributions, although unavailability of the true ORFs prevents linking observed response distributions to true item characteristics.

This article has also proposed a robust test of category order based on a comparison of the outcomes of order-constrained and unconstrained NRM calibrations. If categories are truly ordered, both calibrations should account for the data equally well even when unconstrained calibration estimates categories to be disordered on some items; if categories are instead truly disordered, unconstrained calibration should account for the data meaningfully better than order-constrained calibration. Several criteria were proposed to determine whether or not both calibrations account for the data equally well and hence, whether category disorder should be regarded authentic or spurious. The use of such criteria was illustrated with empirical data and in simulations with items that had ordered or disordered categories.

In simulations with items whose categories were ordered, constrained and unconstrained calibrations accounted for the data equally well and rendered sets of ORFs that did not differ in any meaningful respect. In simulations with mixtures of items with ordered and disordered categories, differences between calibrations occurred and application of the criteria could tell apart items with ordered or with disordered categories. The analysis of empirical data yielded results analogous to those observed in simulations with ordered categories. Of course, signs of disordered categories might be found in other data sets.

The following sections discuss aspects pertaining to the problem addressed in this article, the proposed solution, and its implications.

Can the Potential for Spurious Disorder Be Eliminated Beforehand?

Although not only and not always in these circumstances, estimated disorder definitely occurs with J-shaped response distributions. Low endorsement rates in some categories are known to affect the recovery of NRM item parameters and Preston et al. (2015) recommended collapsing categories as needed to ensure counts per category in excess of 10% of the size of the sample, something that can be actioned before NRM calibration. We confirmed the utility of this strategy by recoding the six categories in Example 1 as 1–1–2–2–3–3 and the five categories in Example 2 as 1–1–2–3–4 before NRM analysis. Unconstrained calibration estimated disordered categories for a single item in the first case (compared with 17 out of 37 when the original six categories were used) and for no item in the second (compared with 5 out of 8 with the original categories). Although this strategy seems to eliminate the problem of disorderly estimated categories, its applicability is limited for reasons discussed next.

First, collapsing seems appropriate for categories whose labels are hard to discriminate by the respondents, but it is not clear that rarely endorsed categories reveal

this problem. For instance, outer categories in patient-reported outcome scales may read “never” and “very rarely” at the lower end and “almost daily” and “daily” at the upper end. Such labels are unmistakable and can surely be discriminated by respondents, but the item may refer to a symptom whose complete absence or constant presence is only experienced by a very small fraction of the target population. Collapsing extreme categories on one or both sides in the only interest of NRM parameter estimation gives up the ability to tell apart respondents at different locations within extreme ranges of θ , something that may be practically relevant (see, e.g., Sharp et al., 2012).

Second, collapsing is not an option for three-category Likert-type items with the usual response options (i.e., negative, unsure, and positive) and this should never be attempted at the encounter of a J-shaped response distribution. Even if unconstrained NRM calibration returns disordered categories in such cases, the middle category still cannot be combined with any of the others. Out of 158 three-category items on the *16 Personality Factor Questionnaire Version 5*, Murray et al. (2016) reported estimated disorder for 22 items in a sample of 10,261 respondents and for 50 items in another sample of 1,216 respondents. When estimated disorder is so prevalent and collapsing is not an option, alternative action is necessary.

In sum, although collapsing can prevent disorderly estimated categories in an NRM calibration, an alternative approach that tests for disorder and estimates ordered category parameters (if warranted) is needed when collapsing is either unfeasible or unadvisable.

Can Category Order Be Assessed Differently?

As mentioned in the “Introduction” section, a comparison of order-constrained and unconstrained calibrations of the same data could arguably have been conducted differently. For instance, an order-constrained calibration may instead be obtained with one of the available item response theory models for ordered responses. All divide-by-total models for ordered responses are variants of the NRM that incorporate additional assumptions (Thissen & Steinberg, 1986) and, hence, they may also be suitable for our purpose. The single variant that apparently adds only the assumption of ordered categories is the generalized partial credit model (GPCM; Muraki, 1992, 1997); the remaining variants explicitly add other assumptions that are irrelevant in an assessment of the empirical order of categories. But the GPCM includes an extra assumption that makes it unsuitable for our purposes, as discussed next.

To make the relation with the NRM directly apparent, consider a reparameterization of the GPCM in which the ORF for category k is expressed as

$$p_k(\theta) = \frac{\exp[(k-1)\alpha\theta - \gamma_k]}{\sum_{m=1}^K \exp[(m-1)\alpha\theta - \gamma_m]}, \quad (8)$$

subject to the identifiability constraint $\gamma_1 = 0$. A K -category item is thus described with only K parameters under the GPCM, compared with $2(K-1)$ parameters under

the NRM. The GPCM parameters are a single α and $K-1$ γ s. GPCM parameters can be converted into NRM-like form by making $a_k = (k-1)\alpha$ and $c_k = -\gamma_k$, which shows that the order constraint is embedded into the GPCM via an equispaced integer scoring function that makes $a_1 = 0$ and $a_{k+1} = a_k + \alpha$ so that $a_k < a_{k+1}$ by construction. From Equation (7), the crossings of consecutive ORFs thus occur at $\tau_k = (\gamma_{k+1} - \gamma_k)/\alpha$ and these crossings (referred to as thresholds) are the category parameters in the GPCM, in which

$$\frac{p_k(\theta)}{p_{k-1}(\theta) + p_k(\theta)} = \frac{1}{1 + \exp[-\alpha(\theta - \tau_{k-1})]}$$

for $1 < k \leq K$.

The identifiability constraint in the GPCM can also be converted into NRM-like form by simply subtracting the average value of the putative a parameters from each of them and similarly for the putative c parameters, but this is inconsequential.

Then, fitting the GPCM amounts to fitting the NRM under an order constraint that also requires a parameters to increase with k in constant steps of size α . Using the GPCM for our purposes here would thus compare an unconstrained NRM calibration (with no constraint on the separation between consecutive a s) with an order-constrained calibration in which the a s are additionally constrained to be equispaced. Spacing and category order are confounded in such a comparison of order-constrained (GPCM) and unconstrained (NRM) calibrations. The equispacing assumption is also the reason that NRM and GPCM calibrations of the same data render markedly different ORFs when NRM calibration estimates nonequispaced a s (see, e.g., figure 11 in García-Pérez, 2017).

The alternative approach proposed here is free of this difficulty, as it isolates the order constraint for a comparison of calibrations with and without it. When unconstrained NRM calibration estimates categories to be ordered, item parameters and ORFs are directly seen to be virtually identical to those obtained in an order-constrained NRM calibration of the same data, whether or not consecutive a s happen to be equispaced. In contrast, when unconstrained NRM calibration results in disorderly estimated categories (again, with or without equal spacing between consecutive a s), order-constrained NRM calibration renders one of two outcomes: Either very similar ORFs except (at most) over inconsequential ranges of θ for the categories suspect of disorder (which is indicative of spuriously estimated disorder) or flat ORFs over the entire θ continuum or a large part of it (which is indicative of true disorder by proof of the absence of a fitting solution with ordered category parameters).

It is evident that we have not considered that thresholds (in the GPCM and Rasch-type models for ordered responses) or crossings (τ s in the NRM) inform of category order. The reasons are discussed at length in García-Pérez (2017) but a brief presentation may be worth giving here. Recall that all order-constrained variants of the NRM (i.e., the GPCM and the various Rasch-type models for ordered responses) share the property that $\frac{p_k(\theta)}{p_{k-1}(\theta) + p_k(\theta)}$ is a monotone increasing function of θ for $1 < k \leq K$, which implies that category $k-1$ precedes category k in their preassumed order. Thus, the

preassumed order is structurally embedded in the models by ensuring that $a_k < a_{k+1}$ in an NRM-like reparameterization, as shown above for the GPCM. Disorderly estimated thresholds cannot contradict this structural assumption and, then, the order of estimated thresholds cannot inform of the order of categories. Instead, ordered thresholds (or ordered crossings in the NRM) inform that all categories are modal within some range of θ in the sense discussed by Linacre (2004) whereas disordered thresholds (or crossings) inform that one or more categories are nonmodal in the same sense (see, e.g., figure 4 in García-Pérez, 2017).

Implications for Item Scoring

Thissen, Cai, and Bock (2010) proposed an alternative parameterization of the NRM in which the ORFs in Equation (1) are expressed as

$$p_k(\theta) = \frac{\exp[a_k^s a^* \theta + c_k]}{\sum_{m=1}^K \exp[a_m^s a^* \theta + c_m]}. \quad (9)$$

Thus, c parameters remain unaltered but the original set of a parameters are expressed as the product of an overall slope parameter a^* (common to all categories) and category parameters a_k^s referred to as the scoring function. The original zero-sum identifiability constraint is also inconsequentially replaced with the constraints $a_1^s = c_1 = 0$ and $a_K^s = K - 1$. Note that a_k^s in Equation (9) replaces the factor $k - 1$ in Equation (8) and, thus, this new parameterization generalizes the equispaced scoring function in the GPCM (with scores ranging from 0 to $K - 1$ in integer steps) to arbitrarily spaced scores also in the range from 0 to $K - 1$, in principle.

The identifiability constraints assume that Category 1 is the first and category K is the last, but parameter estimates may still say otherwise because an order restriction is not imposed on the scoring function. If categories are estimated to be disordered in such circumstances, the estimated a_k^s will be less than zero or greater than $K - 1$ for one or more k . If out-of-range scores are unacceptable, Thissen et al. (2010) suggested using some method to identify the first and last categories (e.g., by fitting first the unrestricted NRM to identify category order) and then placing the corresponding identifiability constraints on them so that the scoring function ranges from 0 to $K - 1$.

Under their new parameterization of the NRM, the order of categories is still taken from the face value of the estimated a^s s, which may be spuriously disordered as shown in this article. Consider Item 25 on the HSQ (center column of Figure 12), where categories were disorderly estimated in the unconstrained calibration. On reparameterization under Equation (9), the estimated parameters yield the scoring function $\mathbf{a}^s = [0.00, -3.17, -2.05, 1.00, 4.00]$ with $a^* = 0.94$ if out-of-range scores are allowed, or $\mathbf{a}^s = [1.77, 0.00, 0.63, 2.33, 4.00]$ also with $a^* = 0.94$ if out-of-range scores are not allowed. With item parameters from the equally plausible and more realistic order-constrained calibration, the resultant scoring function is instead $\mathbf{a}^s = [0.00, 0.00, 0.25, 1.97, 4.00]$ with $a^* = 1.35$. The differences are not negligible and will be larger at the overall scale level on consideration that five of the eight items

on the HSQ affiliative humor scale turned up with disorderly estimated categories in the unconstrained calibration. Then, a comparison of order-constrained and unconstrained NRM calibrations to assess the empirical order of categories will also be useful for determining appropriate scoring functions for the items.

Conclusion

Application of a dependable method to assess category order is useful for that precise purpose but it is also crucial when item scoring is to be tailored to the estimated order of categories, as in Murray et al. (2016) or Sharp et al. (2012; see also Thissen & Cai, 2016). Given the suspiciously high frequency with which conventional NRM calibration estimates categories to be disordered, a comparison of the outcomes of order-constrained and unconstrained calibrations is needed for informed and dependable decisions on the order of categories. Both types of calibration were implemented and tested with simulated and empirical data in this article. The results show that spurious disorder (in unconstrained calibrations) can easily be detected whereas authentic disorder is also easily identified in the inability of constrained calibration to provide a fitting account with ordered categories. Order-constrained parameter estimates should be used unless a comparison of calibrations reveals that unconstrained estimates suit suspect items meaningfully better. The code developed here to implement order-constrained and unconstrained NRM calibrations is slightly inefficient but an optimized version will be made available in due course.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Grant No. PSI2015-67162-P (Ministerio de Economía y Competitividad, Spain).

References

- Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, 72, 547-573. doi:10.1177/0013164411432166
- Andrich, D. A. (2004). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 167-200). Maple Grove, MN: JAM.
- Andrich, D. (2013a). An expanded derivation of the threshold structure of the polytomous Rasch model that dispels any "threshold disorder controversy". *Educational and Psychological Measurement*, 73, 78-124. doi:10.1177/0013164412450877

- Andrich, D. (2013b). The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical order of categories. *Educational and Psychological Measurement*, 73, 553-580. doi:10.1177/0013164413477107
- Andrich, D., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59-70). New York, NY: Waxmann.
- Annoni, P., Weziak-Bialowolska, D., & Farhan, H. (2013). Measuring the impact of the Web: Rasch modelling for survey evaluation. *Journal of Applied Statistics*, 40, 1831-1851. doi:10.1080/02664763.2013.796351
- Ashley, L., Smith, A. B., Keding, A., Jones, H., Velikova, G., & Wright, P. (2013). Psychometric evaluation of the Revised Illness Perception Questionnaire (IPQ-R) in cancer patients: Confirmatory factor analysis and Rasch Analysis. *Journal of Psychosomatic Research*, 75, 556-562. doi:10.1016/j.jpsychores.2013.08.005
- Bee, P., Gibbons, C., Callaghan, P., Fraser, C., & Lovell, K. (2016). Evaluating and quantifying user and carer involvement in mental health care planning (EQUIP): Co-development of a new patient-reported outcome measure. *PLoS One*, 11, e0149973. doi:10.1371/journal.pone.0149973
- 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
- Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York, NY: Springer.
- Bokhary, K. A., Suttle, C., Alotaibi, A. G., Stapleton, F., & Boon, M. Y. (2013). Development and validation of the 21-item Children's Vision for Living Scale (CVLS) by Rasch analysis. *Clinical and Experimental Optometry*, 96, 566-576. doi:10.1111/cxo.12055
- Bourke, M., Wallace, L., Greskamp, M., & Tormoehlen, L., (2015). Improving objective measurement in nursing research: Rasch model analysis and diagnostics of the Nursing Students' Clinical Stress Scale. *Journal of Nursing Measurement*, 23, E1-E15. doi:10.1891/1061-3749.23.1.E1
- Brogårdh, C., Lexell, J., & Lundgren-Nilsson, Å. (2013). Construct validity of a new rating scale for the self-reported impairments in persons with late effects of polio. *Physical Medicine & Rehabilitation*, 5, 176-181. doi:10.1016/j.pmrj.2012.07.007
- das Nair, R., Moreton, B. J., & Lincoln, N. B. (2011). Rasch analysis of the Nottingham Extended Activities of Daily Living Scale. *Journal of Rehabilitation Medicine*, 43, 944-950. doi:10.2340/16501977-0858
- Dougherty, B. E., Nichols, J. J., & Nichols, K. K. (2011). Rasch analysis of the Ocular Surface Disease Index (OSDI). *Investigative Ophthalmology and Visual Science*, 52, 8630-8635. doi:10.1167/iovs.11-8027
- du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.
- García-Pérez, M. A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *British Journal of Mathematical and Statistical Psychology*, 47, 247-282. doi:10.1111/j.2044-8317.1994.tb01037.x
- García-Pérez, M. A. (2017). An analysis of (dis)ordered categories, thresholds, and crossings in difference and divide-by-total IRT models for ordered responses. *Spanish Journal of Psychology*, 20, e10. doi:10.1017/sjp.2017.11

- García-Pérez, M. A., & Núñez-Antón, V. (2004). Small-sample comparisons for goodness-of-fit statistics in one-way multinomial with composite hypotheses. *Journal of Applied Statistics*, 31, 161-181. doi:10.1080/0266476032000148849
- González-Romá, V., & Espejo, B. (2003). Testing the middle response categories «Not sure», «In between» and «?» in polytomous items. *Psicothema*, 15, 278-284.
- Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*, 49, 146-160. doi:10.1037/a0027899
- Hernández, A., Espejo, B., & González-Romá, V. (2006). The functioning of central categories *Middle Level* and *Sometimes* in graded response scales: Does the label matter? *Psicothema*, 18, 300-306.
- Kim, S.-H., Bao, Y., Horan, E., Kim, M., & Cohen, A. S. (2015). Gauss-Hermite quadrature in marginal maximum likelihood estimation of item parameters. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014* (pp. 43-58). New York, NY: Springer.
- Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions*, 13, 675.
- Linacre, J. M. (2004). Rasch model estimation: Further topics. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 48-72). Maple Grove, MN: JAM.
- Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, 37, 48-75. doi:10.1016/S0092-6566(02)00534-2
- Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, 40, 261-279. doi:10.1207/s15327906mbr4002_5
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. doi:10.1177/014662169201600206
- Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York, NY: Springer.
- Murray, A. L., Booth, T., & Molenaar, D. (2016). When middle really means “top” or “bottom”: An analysis of the 16PF5 using Bock’s nominal response model. *Journal of Personality Assessment*, 98, 319-331. doi:10.1080/00223891.2015.1095197
- Oluboyede, Y., & Smith, A. B. (2013). Evidence of a unidimensional 15-item version of the CASP-19 using a Rasch model approach. *Quality of Life Research*, 22, 2429-2433. doi:10.1007/s11136-013-0367-z
- Peterson, J., Zhang, M., Wang, S., Pak, S., Lee, W.-C., & Kolen, M. J. (2014, December). A comparison of several item response theory software programs with implications for equating mixed-format tests. In M. J. Kolen & W.-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 3, pp. 123-159). CASMA Monograph Number 2.3. Iowa City, IA: CASMA, the University of Iowa. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.4829&rep=rep1&type=pdf>
- 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
- Preston, K. S. J., Parral, S. N., Gottfried, A. W., Oliver, P. H., Gottfried, A. E., Ibrahim, S. M., & Delany, D. (2015). Applying the nominal response model within a longitudinal framework to construct the Positive Family Relationships Scale. *Educational and Psychological Measurement*, 75, 901-930. doi:10.1177/0013164414568717
- Preston, K. S. J., & Reise, S. P. (2014). Detecting faulty within-item category functioning with the nominal response model. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modelling: Applications to typical performance assessment* (pp. 386-405). London, England: Routledge.
- Salzberger, T. (2015). The validity of polytomous items in the Rasch model—The role of statistical evidence of the threshold order. *Psychological Test and Assessment Modeling*, 57, 377-395.
- Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph No. 18*. Richmond, VA: Psychometric Society. Retrieved from <https://www.psychometric-society.org/sites/default/files/pdf/MN18.pdf>
- Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, 23, 17-35. doi:10.2333/bhmk.23.17
- Sharp, C., Steinberg, L., Yaroslavsky, I., Hofmeyr, A., Dellis, A., Ross, D., & Kincaid, H. (2012). An item response theory analysis of the Problem Gambling Severity Index. *Assessment*, 19, 167-175. doi:10.1177/1073191111418296
- Thissen, D., & Cai, L. (2016). Nominal categories model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1. Models* (pp. 51-73). Boca Raton, FL: CRC Press.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43-75). New York, NY: Routledge.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577. doi:10.1007/BF02295596
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26, 161-176. doi:10.1111/j.1745-3984.1989.tb00326.x
- Wang, Y.-C., Deutscher, D., Yen, S.-C., Werneke, M. W., & Mioduski, J. E. (2014). The self-report Fecal Incontinence and Constipation Questionnaire in patients with pelvic-floor dysfunction seeking outpatient rehabilitation. *Physical Therapy*, 94, 273-288. doi:10.2522/ptj.20130062
- Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment*, 21, 765-774. doi:10.1177/1073191114530775
- Wetzel, E., Hell, B., & Pässler, K. (2012). Comparison of different test construction strategies in the development of a gender fair interest inventory using verbs. *Journal of Career Assessment*, 20, 88-104. doi:10.1177/1069072711417166