TPM

SHOULD WE CONSTRAIN THE THRESHOLDS IN FACTOR ANALYSIS OF RATING SCALE RESPONSES?

GREGOR SOČAN
UNIVERSITY OF LJUBLJANA, SLOVENIA

In factor analysis of ordinal variables, category thresholds determine the values of a hypothetical latent response at which a transition to a higher response option occurs. In general, category thresholds are estimated as free parameters. In the Rasch measurement theory framework, the rating-scale model has been proposed that prescribes equal sets of distances between category thresholds for all items. As a factor analytic parallel, I propose a model with a constrained threshold structure (FACTS). The application of the model is illustrated with a real data example. A simulation study showed that the thresholds are estimated more accurately with the FACTS model than with the standard unconstrained model for different sample sizes, test lengths, and number of response categories. In addition, the likelihood ratio test generally showed good power in comparing the two models. Because the FACTS model performs well and provides a meaningful interpretation of category thresholds, it may be used routinely in factor analysis of categorical item responses.

Keywords: Rating scales; Threshold; Categorical data; Factor analysis; Simulation.

Correspondence concerning this article should be addressed to Gregor Sočan, Department of Psychology, University of Ljubljana, Aškerčeva c. 2, SI-1000 Ljubljana, Slovenia. Email: gregor.socan@ff.uni-lj.si

(Self-)rating scales are regularly used in the measurement of personality traits and other typical-response assessments. Usually, participants are asked to choose among a limited number of discrete and ordered response categories. The categories may cover the continuum between complete agreement and complete disagreement or between a very high frequency and a very low frequency. Sometimes only the meaning of the two extreme categories is defined (especially when the number of categories is high), and sometimes each category is given its verbal label. The item scores (i.e., the ratings, reversed if necessary) are regularly subjected to factor analysis, either to examine the dimensionality of the items or to construct a structural equation model. In such analyses, the factor analysis model for ordered categorical variables seems to be preferable. Applying the standard linear model of factor analysis on categorical data violates the basic model assumptions, resulting in less accurate loading estimates and incorrect standard errors and model fit statistics (e.g., Li, 2016; Rhemtulla et al., 2012).

Factor analysis for categorical indicators, as developed by Muthén (1978, 1984), is based on the introduction of intermediate variables, usually referred to as latent responses (Muthén, 1984) or underlying variables (Bartholomew & Knott, 1999). For the 1-factor model, this approach can be summarized as follows. Let *y* be a categorical item score with *C* ordered categories. The relationship between the latent trait and the latent response is modeled by the ordinary linear model:

$$y_i^* = \nu + \lambda_y \xi_i + \epsilon_{iy*} , \quad (1)$$

where y_i^* stands for the latent response of person *i*, ν is the factor intercept, λ_y is the factor loading of item latent response y_i^*, ξ_i is the latent trait value of person *i*, and ϵ_{iy*} is the residual.

The latent responses and the manifest responses are related through a set of $C - 1$ thresholds (τ_1, τ_2, …, $\tau_{C\text{-}1}$,):

$$y_i = \begin{cases} 0, \text{if } y_i^* \leq \tau_1 \\ 1, \text{if } \tau_1 < y_i^* \leq \tau_2 \\ \vdots \\ C - 1, \text{if } y_i^* > \tau_{C-1} \end{cases} \quad (2)$$

The scale of the latent response is arbitrary; therefore, the intercept can be set to 0. In the usual delta parameterization (Muthén & Asparouhov, 2002), the latent response is standardized. The latent response is assumed to be normally distributed, and thresholds are estimated from the univariate marginal frequencies. Factor loadings are usually estimated by applying a diagonally weighted least squares estimator to the matrix of polychoric correlations between items (Muthén & Muthén, 1998-2017; Rosseel, 2024).

It should be noted that the latent response already contains the measurement error. In the context of personality assessment, the latent response can be understood as the level of item endorsement at the time of response. The *k*-th threshold can then be viewed as the highest level of endorsement, at which the participant still prefers category $k - 1$ to category k (cf. Muthén & Asparouhov, 2002, Section 2.2). The translation from the latent response to the manifest response is deterministic; more specifically, it is a case of surjective mapping. For example, if the latent response is less than τ_1, the participant will certainly choose the lowest category. The correlation between test scores on two occasions will be less than 1 because the same value of the latent trait ξ_i will result in different values of the latent response y_i^* on different testing occasions, as a result of the addition of the random error ϵ_{iy*}. Item intercorrelations are further reduced by the rounding error introduced by the discretization of the continuous latent response: all values of the latent response between two consecutive thresholds are mapped to the same discrete observed response (for a discussion of the effects of the rounding error see Schneeweiss et al., 2010).

In factor analyses of psychological scales, thresholds often receive less attention than loadings. However, thresholds should play an important role in psychometric analyses because they reflect the difficulty (or "endorsability") of the items. For instance, consider an item with four categories and thresholds of –2, –1, and 0. If the latent responses are normally distributed, participants with above-average endorsement will choose the highest category and only 2.3% of participants with the lowest endorsement will choose the lowest category (because in the standard normal distribution, the mean is 0 and the probability of a *z* score equal or lower than –2 is 2.3%). The item can therefore be described as "easy" in the sense that a relatively low level of latent response is sufficient for a relatively high manifest response. If the factor loading is positive, this implication also holds with respect to the latent trait value, but only in a probabilistic sense: because the latent response also contains random error, a relatively low factor value may suffice for a high response and vice versa.

Thresholds are assumed to be the same for all participants. Although it might seem reasonable to allow for variation in the mean structure across participants — for example, due to individual differences in response styles — such a model would be overparameterized unless some stringent restrictions and assumptions were introduced (see Maydeu-Olivares & Coffman, 2006, for a proposed solution based on the ordinary factor analysis model, and Falk & Cai, 2016, for a more general IRT-based treatment). At the same time, the default solutions generated by state-of-the-art software packages such as Mplus (Muthén & Muthén, 1998-2017) and lavaan (Rosseel, 2012, 2024) do not impose any constraints on the threshold structure across items. This makes sense in cases where response categories have different meanings for different items. However, if the same set of anchors is used for all items, it is reasonable to assume that the categories are used in the same way for all items. If participants respond consistently, the choice of some response category (for

example, *strongly agree*) reflects the same level of item endorsement regardless of the item content. This implies that the distance between thresholds for each unique pair of categories (such as *strongly agree* and *agree*) should be the same for all items. Notable differences in the threshold structure indicate an item × rating-scale interaction, which may be difficult to interpret in a meaningful way.

The idea of constraining the threshold structure is not new in psychometrics. In Rasch measurement theory, the rating-scale model (RSM) has been proposed (Andrich, 1978, 2016) as a restricted version of a polytomous Rasch model:

$$p(y_i = t) = \frac{\exp \sum_{k=0}^{t} \left(\theta_i - (b_y + \tau_k)\right)}{\sum_{h=0}^{C-1} \exp \sum_{k=0}^{h} \left(\theta_i - (b_y + \tau_k)\right)} \tag{3}$$

In the above equation, y_i is the item score, θ_i represents the position on the latent trait for person i, b_y is the location parameter for item y, τ_k is the threshold deviation for category k ($k = 0, 1, \ldots C - 1$), and t is the value of the chosen category. Because RSM is a member of the Rasch family, no item discrimination parameter is included. Each item is characterized only by a single location parameter ("difficulty"). Additionally, the $C - 1$ category thresholds $b_y + \tau_k$ can be calculated, which determines the points at which adjacent categories are equally likely to be endorsed. A single set of τ parameters is estimated; therefore, the threshold structure is the same for all items.

A Factor-Analytic Model with a Constrained Threshold Structure

In this paper, I propose a factor-analytic analog of the RSM. The factor analysis with a constrained threshold structure (FACTS) model is defined by Equations 1 and 2 with the following additional constraints. Let us define a set of $C - 2$ threshold differences:

$$\begin{aligned} &\tau_2 - \tau_1 = \delta_1 \\ &\vdots \\ &\tau_{C-2} - \tau_1 = \delta_{C-2} \end{aligned} \tag{4}$$

Each of the differences $\delta_1, \ldots, \delta_{C-2}$ is constrained to be equal across all items. From the substantive viewpoint, this means that the differences between response categories are the same for all items. On the other hand, the difficulty of the items, reflected in the average value of the thresholds for an item, may still vary across items. Figure 1 illustrates the rationale of the model. The thresholds for three 4-category items are marked with triangles. The items vary in difficulty (item C is the most difficult and item B is the easiest), but the differences between the thresholds (δ) are the same. We can also see that the second threshold is closer to the first threshold than to the third one, that is $(\delta_2 - \delta_1) > \delta_1$, which indicates that the lower response categories are closer to each other than the upper categories.

The FACTS model can be readily fitted using state-of-the-art structural equation software that provides the DWLS estimation and can impose equality constraints on the thresholds. In addition, it may be interesting to test whether the FACTS model fits significantly worse than the unconstrained model. In lavaan, this can be accomplished with the scaled χ^2 test (Satorra, 2000). The score (Lagrange multiplier) test can be used to investigate whether the constraints should be released for items with the greatest differences between the constrained and unconstrained thresholds (post hoc testing should preferably be supplemented by a Holm-Bonferroni correction). An R script that automatically generates the code for the FACTS model and runs the analyses in lavaan, and an illustrative Mplus script are available as electronic supplements.

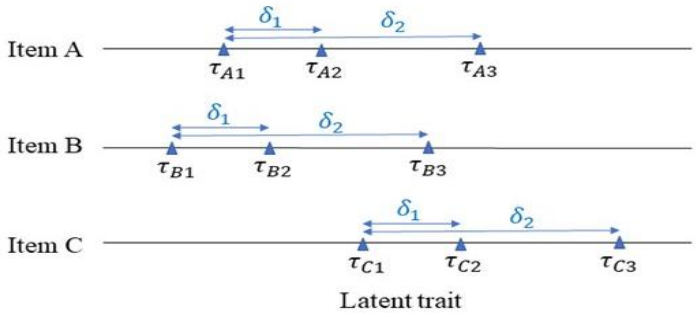

FIGURE 1
An ideal case of the constrained threshold structure for three items with four response categories

The FACTS model has two major advantages over the standard unconstrained model. First, the number of parameters is smaller. Given p items with C categories, we freely estimate $p \times (C - 1)$ parameters in the unconstrained model. Because of the equality constraints, the FACTS model has $(p - 1)(C - 2)$ more degrees of freedom than the unconstrained model. In addition, the thresholds in the unconstrained model are based on univariate marginals. If the extreme categories are endorsed by a small fraction of participants, the corresponding thresholds will not be estimated accurately. On the other hand, in the FACTS model, all estimates are based on all responses because of the difference constraints. Thus, the estimates from FACTS should be more stable than the unconstrained estimates. Second, the interpretation of the thresholds is simpler in the FACTS model because the relative positions of the thresholds are the same for all items. The FACTS model is more intuitive because it assumes that the category anchors are understood equally for all items.

It should be noted that the interpretation of thresholds in the FACTS model and in the rating scale model is different (see Bartholomew & Knott, 1999, and Muthén & Asparouhov, 2002, for a discussion of the relationship between the factor-analytic and item-response approaches). In RSM, thresholds operate probabilistically (they denote points of equal probability), while in FACTS they operate deterministically. A consequence is that — unlike RSM thresholds — the FACTS thresholds can never be reversed, that is, they are always ordered. Disordered FACTS thresholds would imply paradoxical consequences where a single latent response value would be simultaneously mapped to different observed response values (cf. Equation 2).

AN ILLUSTRATION USING REAL DATA

I will illustrate the proposed procedure using the stress scale of the Depression Anxiety Stress Scales-DASS-21 (Lovibond & Lovibond, 1995). For each of the seven items, respondents rate how much the statements applied to them in the past week using a 4-point severity/frequency scale (0 = *not at all*; 1 = *to some degree* or *some of the time*; 2 = *to a considerable degree* or *a good part of the time*; 3 = *very much* or *most of the time*). The Slovenian version of the scale was administered to a sample of 431 adults (Kavčič et al., 2023).

Despite the significant test statistic, the approximate model fit for the FACTS model was good: $\chi^2(26) = 59.25$, $p < .001$; CFI = .99; RMSEA = .05; SRMR = .03. The model fit was not significantly worse compared with the unconstrained 1-factor model: $\chi^2(12) = 18.08$, $p = .113$. Both sets of thresholds are shown in Figure 2. In most cases, the constrained thresholds were very close to the unconstrained thresholds. The difference was slightly larger for the third threshold of items S6 and S8 ($\tau_{\text{FACTS}} - \tau_{\text{unconstrained}} = -0.13$ and

0.17, respectively), but none of the differences were significant according to the score test with the Holm-Bonferroni correction. Thus, we can conclude that the same threshold structure can be applied to all items. The delta parameters were $\delta_1 = 1.15$ and $\delta_2 = 2.02$, which means that the difference between the first two thresholds ($\tau_2 - \tau_1 = \delta_1 = 1.15$) was slightly larger than the difference between the second and third ($\tau_3 - \tau_2 = \delta_2 - \delta_1 = 0.86$). Thus, it appears that the lower response categories are further apart than the higher ones. This is not surprising, because the difference between the total absence and some presence is fundamentally greater than the difference between the different levels of presence of a phenomenon. The items can be considered psychometrically difficult because all thresholds were relatively high, implying that a high level of latent response is required to select the middle or high response categories.

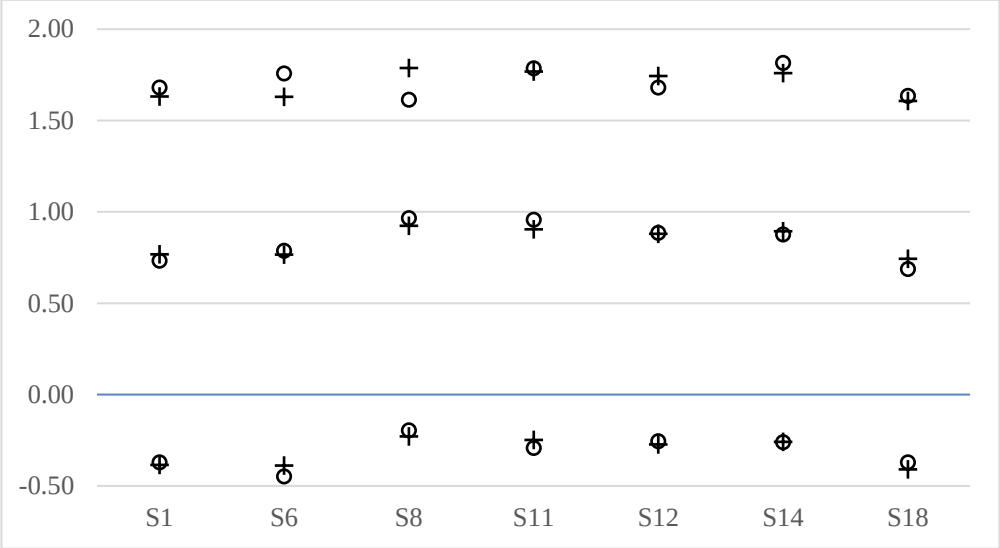

FIGURE 2
Unconstrained and constrained thresholds for the DASS-21 stress scale items
Note. The constrained thresholds are indicated by crosses, the unconstrained ones by circles.

EMPIRICAL EVALUATION OF THE FACTS MODEL

The usefulness of the FACTS model depends on the empirical behavior of the threshold estimates. It is particularly important to check whether the use of the model notably improves the accuracy of the threshold estimates and whether it is possible to detect a misfit with a satisfactory power. I conducted two simulation studies to evaluate the performance of the model.

First, I evaluated the accuracy of the FACTS estimates compared with the unconstrained estimates. I expected that, provided that the FACTS model was the correct population model, the constrained estimates would be more accurate (i.e., deviate less from the population values) than the unconstrained estimates. Because the parameters δ in Equation 4 are based on all the data, the thresholds should be less affected by the sampling error than the unconstrained thresholds, which are based on the cumulative proportions of responses within a single item. As subhypotheses, I also hypothesized that the improvement in accuracy would be larger in the following situations:

(a) Smaller sample size. In very large samples, the unconstrained thresholds are close to the population values, so constraining them has little effect.

(b) Longer tests. With a larger number of items, parameters δ are based on more data and are therefore more accurately determined. In contrast, the accuracy of unconstrained parameters should not depend on the test length.

(c) Thresholds with large absolute values. Because the unconstrained estimates for such thresholds are based on proportions that are relatively close to 0 or 1, respectively, they are particularly susceptible to sampling error.

Second, the statistical properties of the χ^2 difference test for testing the fit of the FACTS model were evaluated against the unconstrained model. As mentioned earlier, the difference test as implemented in popular software such as the lavaan package (Rosseel, 2024) is not optimal. Therefore, it was important to investigate the power to detect differences in threshold structure across items to prevent constraining the thresholds when this does not correspond to the actual threshold structure.

Simulation 1: Evaluation of the Accuracy of the FACTS Estimates

In the first simulation, the following conditions were manipulated:

(1) sample size (n = 200, 500, or 1000 persons),
(2) number of response categories (C = 3, 4, or 5 categories),
(3) test length (p = 10 or 20 items),
(4) size of standardized factor loadings (λ = .30 or .60).

Note that the comparisons related to thresholds with different absolute values — see point (c) above — did not require additional manipulations, because all threshold values were used in each condition.

For each of the 36 conditions, 1000 sample data matrices were created as follows. In the first step, sample latent trait values were obtained using a pseudo-random number generator. Then, the latent responses were calculated according to Equation 1 and finally converted into observed responses using Equation 2. The latent trait and residuals were normally distributed. The variances of the latent responses were set to 1. Threshold values are given in the online Appendix A.[1] All computations were performed with the R 4.0.4 program (R Core Team, 2021).

The primary measure of the accuracy of the threshold estimates was the mean absolute deviation (MAD) from the population value, which was calculated as:

$$\mathrm{MAD} = \frac{\sum_{s=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{C-1}|\hat{\tau}_{sjk}-\tau_{jk}|}{np(C-1)}, \tag{5}$$

where $\hat{\tau}_{sjk}$ is the empirical estimate of the k-th threshold for item p, estimated in sample s, and τ_{jk} is the true (population) value of this threshold. In relation to Hypothesis (c), MAD was calculated separately for thresholds with the same absolute values:

$$\mathrm{MAD_A} = \frac{\sum_{s=1}^{n}\sum_{j=1}^{p}\sum_{a=1}^{A}|\hat{\tau}_{sja}-\tau_{ja}|}{npA}, \tag{6}$$

where A was the number of thresholds with the same absolute value pertaining to an item. In our simulations, A was either 1 or 2 (see threshold matrices in the online Appendix A). Additionally, the absolute bias of the estimates was calculated as the difference between the mean sample estimate and the true value of the threshold:

$$\mathrm{bias}_{\tau_k} = \frac{\sum_{s=1}^{n}\sum_{j=1}^{p}\hat{\tau}_{sjk}}{np} - \tau_{jk} \tag{7}$$

While MAD is a measure of accuracy (i.e., how close to the true value the estimates fall, on average), bias tells us whether the sample values tend to be systematically too high or too low, respectively. Although bias is typically reported in the relative form (absolute bias divided by the true value), this would not make sense in our case, because the values would depend on the scaling of latent response and would approach infinity if the true value approached zero. Nevertheless, the absolute bias values can be interpreted by taking into account the fact that the latent response is standardized and thus one unit corresponds to the standard deviation.

In 26 out of 36 combinations of conditions, the MAD value for the constrained estimates was lower than for the unconstrained estimates *in all samples*. For the remaining combinations of conditions, the smallest percentage of samples for which the constrained estimates were more accurate was 98.8%.

Table 1 shows the results of the mixed model ANOVA, in which sample size, number of categories, test length, and size of factor loadings were treated as between-subjects factors and the threshold structure model (constrained vs. unconstrained) was treated as a repeated measures factor. The dependent variable was the absolute deviation from the true threshold value (H_0 thus stated that MAD was the same in all conditions). Effects are sorted by the value of the generalized eta-squared statistic η^2_G (Bakeman, 2005; Olejnik & Algina, 2003). To save space, only significant effects are presented; the complete table is available in the online Appendix B. For our purposes, only the effects that include the model are of interest. In addition to the large overall effect of the model, interactions with sample size, number of categories, size of factor loadings, and test length were also statistically significant. It should be noted, however, that only the interaction with sample size exceeded the medium effect size cut-off suggested by Cohen (1988), and the effect sizes for the remaining interactions were less than .010 (only the effect of the interaction with the number of categories was close to the small size cutoff).

TABLE 1
ANOVA table for the effects on the mean absolute deviation

Effect	df_1	F	p	η^2_G	e.s.
n	2	61902.03	< .001	.746	Large
model	1	268853.99	< .001	.521	Large
n : model	2	18060.40	< .001	.128	Medium
C	2	425.27	< .001	.020	Small
C : model	2	811.45	< .001	.007	
n : C	4	33.33	< .001	.003	
λ : model	1	242.31	< .001	.001	
p : model	1	196.80	< .001	.001	
n : C : model	4	46.33	< .001	.001	
p	1	31.07	< .001	.001	
λ	1	26.60	< .001	.001	
n : λ : model	2	10.06	< .001	.000	
n : p : model	2	5.19	= .006	.000	

Note. df_2 = 35964. n = sample size; C = number of categories; p = test length; λ = factor loading; η^2_G = generalized effect size; e.s. = interpretation of the effect size according to Cohen (1988, p. 283).

Figure 3 illustrates the effects of sample size and number of categories on the accuracy of threshold estimates. In all conditions, the mean absolute deviation was smaller for the constrained estimates than for the unconstrained estimates. The difference between the MAD values decreased with sample size. Constrained estimates were slightly more accurate in the conditions with more categories. The unconstrained estimates were the least accurate for 4-category items, which may be attributed to the fact that in this condition the average absolute value in the threshold matrix was slightly larger ($|\bar{\tau}| = 1.1$) than in the 3-category and 5-category conditions ($|\bar{\tau}| = 1.0$), respectively (cf. the online Appendix A and the results in the next paragraph). Test length, on the other hand, had a negligible (although statistically significant) interaction with the model: for the 10-item test, the MAD values were $MAD_{UC} = 0.070$ and $MAD_{CTS} = 0.047$, while for the 20-item test, these values were $MAD_{UC} = 0.070$ and $MAD_{CTS} = 0.046$.

Figure 4 shows that the accuracy of the unconstrained estimates was particularly low for very high or very low estimates (note that the absolute population value of the threshold is on the x-axis). This was to be expected, because these estimates are based on proportions close to 0 or 1, respectively. A small change in such a proportion (due to sampling error) leads to a relatively large change in the corresponding quantile and, consequently, in the threshold estimate. On the other hand, the mean absolute deviations remained about the same for all thresholds in the constrained estimation: due to constraints, these estimates depend on all data, making all threshold estimates approximately equally accurate.

I further computed absolute bias for each threshold value (from –2 to 2) in each of the 36 experimental conditions. The bias of the constrained thresholds ranged from –0.006 to 0.010, and the bias of the unconstrained thresholds ranged from –0.041 to 0.047. The average bias across all conditions was less than .0001 for both models. Although these values are negligible from a practical viewpoint, I still conducted ANOVA (with model and threshold value as repeated-measures factors, and sample size, loading size, test length, and number of categories as between-subjects factors). While several effects were significant, only the model × threshold interaction had a small effect size (η^2_G = .013). Unconstrained threshold estimates tended to be closer to zero, especially in smaller samples, and constrained thresholds showed a very slight tendency to be farther from zero than the true values. Full results are presented in the online Appendix C.

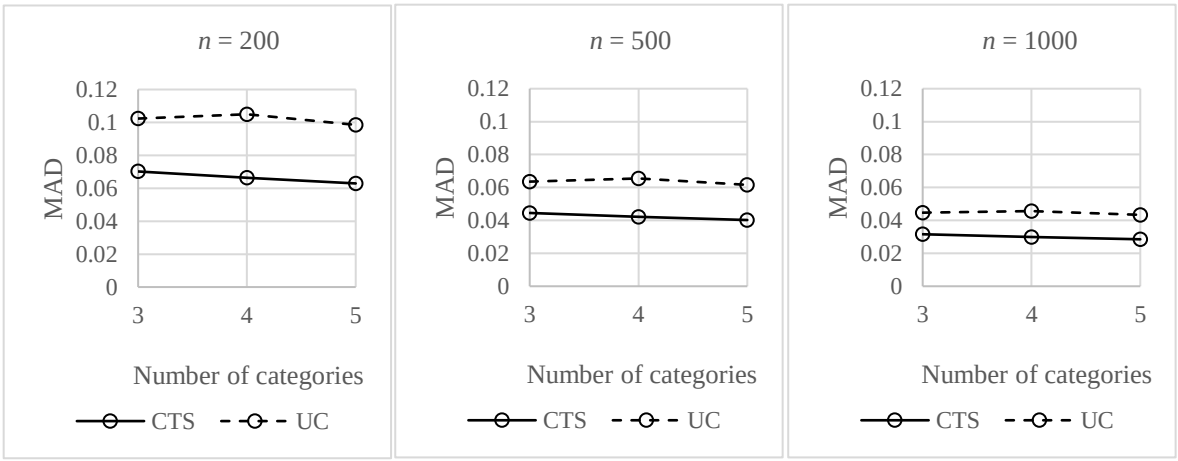

FIGURE 3
Mean absolute deviation for various levels of sample size and number of categories
Note. UC = unconstrained estimates; CTS = constrained threshold structure.

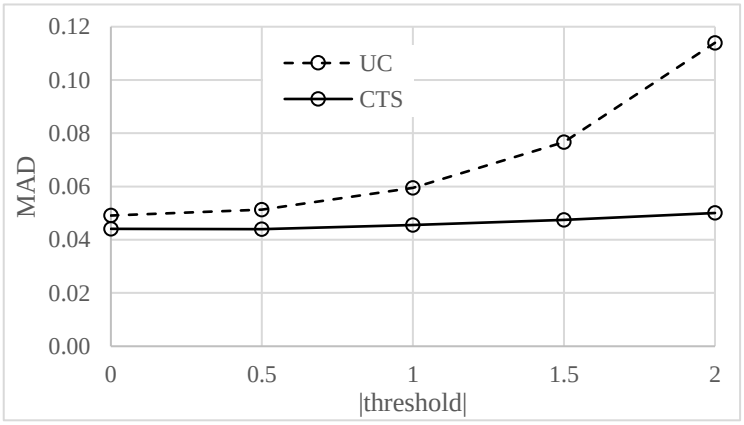

FIGURE 4
Mean absolute deviation in relation to the absolute threshold value
Note. UC = unconstrained estimates; CTS = constrained threshold structure.

Simulation 2: The Power of the Model Fit Test

This simulation aimed to estimate the power of the χ^2 difference test for testing the fit of the FACTS model against the unconstrained model. It seems reasonable to postulate that the degree of misfit is higher when there is a larger proportion of misfitting items (i.e., items with a deviating pattern of threshold differences) and when the threshold differences are larger. In the second simulation, the following conditions were manipulated:

1. proportion of misfitting items (10 or 20%),
2. deviation of threshold differences for misfitting items from common threshold differences (d = 0.20 or 0.50; these values were chosen to reflect the commonly accepted criteria for a small and medium value of the Cohen's d, respectively),
3. sample size (n = 200, 500, or 1000 persons),
4. test length (p = 10 or 20 items),
5. factor loadings structure:
   a) unidimensional structure, all loadings equal .30,
   b) unidimensional structure, all loadings equal .60,
   c) 2-dimensional structure: first-factor loadings equal .60, second factor loadings equal ± .30.

Data were analyzed according to the 1-factor model. Condition (c), therefore, represents a departure from unidimensionality (mean sample fit indices for the unconstrained model were RMSEA = .10 and CFI = .87).

For each of the 72 conditions, 3000 sample data matrices were generated. The number of response categories was set at 4. The remaining settings were the same as for the first simulation. The thresholds for the misfitting items are given in the online Appendix A, and the factor loading matrix for Condition (c) is in the online Appendix D.

Table 2 shows the power under different conditions. The power was higher when the sample size was larger and when the test consisted of more items. When factor loadings were very small (λ = .30) and when factor structure notably departed from unidimensionality, the power was somewhat lower compared to the unidimensional structure with high loadings. When the threshold structure of the misfitting items notably deviated from the common structure (d = 0.50), the power was generally very high. When the factor loadings

were high (.60), the power was above 80% even with a sample size of 200 and a single misfitting item (in a test with 10 items). On the other hand, if we want to detect even very small deviations from the threshold structure (d = 0.20), the sample size should be quite large and the proportion of misfitting items should be higher (about 20% or more) to achieve high power; higher factor loadings and a good fit of a 1-factor model are also beneficial in this case.

TABLE 2
Power to detect the misfit of the FACTS model

			λ = .30		λ = .60		λ_1 = .60, λ_2 = ± .30	
%mf	*n*	*p*	*d* = 0.20	*d* = 0.50	*d* = 0.20	*d* = 0.20	*d* = 0.20	*d* = 0.50
10	200	10	8	65	10	81	7	68
		20	8	86	13	98	8	88
	500	10	23	100	29	100	24	100
		20	35	100	50	100	37	100
	1000	10	56	100	66	100	57	100
		20	78	100	93	100	80	100
20	200	10	12	95	18	99	13	96
		20	17	100	32	100	16	100
	500	10	46	100	53	100	51	100
		20	69	100	91	100	71	100
	1000	10	88	100	91	100	89	100
		20	99	100	100	100	99	100

Note. λ = size of standardized factor loadings; %mf = percentage of misfitting items; *n* = sample size; *p* = test length; *d* = threshold deviation for misfitting items.

DISCUSSION

Compared with the unconstrained factor analysis model, the constrained threshold structure (FACTS) is attractive because of the smaller number of free parameters and simpler interpretation. Given that the differences between the thresholds are the same for all items, the difficulty of the items is easier to assess. The FACTS model can also facilitate the analysis of measurement invariance. If the hypothesis of equal thresholds in all groups is rejected, the search for partial invariance can be complicated because of the large number of thresholds. Within the framework of the FACTS model, the researcher can first allow parameters δ to differ between groups. For example, the perceived differences between category labels may be different in different language versions, while the item content remains the same.

The results of the simulations show good empirical behavior of the FACTS model. The model provides for a more accurate estimation of thresholds, especially for relatively small samples and for thresholds with large absolute values. The constrained estimates are also almost unbiased, although the bias is generally very small for the unconstrained thresholds, as well. The reported results on accuracy and bias pertain to the situation when the data are perfectly unidimensional in the population. I also ran a smaller simulation based on data that departed from unidimensionality. The results were almost identical to those reported here and are available in the online Appendix D.

The power of the standard likelihood ratio test to detect the misfit of the FACTS model compared with the unconstrained model also appears to be good, at least when the threshold deviations are not very small. As expected, the power is also slightly reduced by model error, introduced by either large residual variances or large departures from unidimensional structures.

It is surprising that the idea of constraining the threshold structure has not yet received attention in psychometric practice. One reason for this may be that the factor analytic approach has traditionally focused primarily on factor loadings, whereas in the item response theory approach, and especially in Rasch measurement theory, item difficulty is considered at least as important as item discrimination.

The main limitation remains the assumption of normally distributed latent responses. As has been pointed out (e.g., Robitzsch, 2020), its validity should not be taken for granted; serious deviations may be detrimental not only to the estimation of thresholds but also to the estimation of loadings (Foldnes & Grønneberg, 2020). It is less clear what practical conditions lead to a notable nonnormality of latent responses. One can assume that the latent errors (ϵ_{iy*} in Equation 1) are unbounded, continuous, and influenced by many factors, implying a normal distribution. The shape of the distribution of the latent trait, on the other hand, can sometimes be inferred from theory: for example, in the general population, the distribution of extraversion is expected to be closer to the normal distribution than the distribution of psychoticism. Because the latent response is the weighted sum of the latent trait and the error, normally distributed errors would attenuate the effect of a possibly nonnormal distribution of the latent trait, especially if the factor loadings are not very high: when the factor loading is lower than about .71, the latent response is more affected by error than by the latent trait (because the communality is $.71^2 = .50$ and the uniqueness is $1 - .71^2 = .50$). In any case, the hypothesis of multivariate normality of the latent responses can be tested (Foldnes & Grønneberg, 2020; Maydeu-Olivares, 2006).

Moreover, the FACTS model seems to be useful only for (self-)ratings in personality and similar domains. Clearly, the number of categories should be at least three and equal for all items for the FACTS model to be applicable. Even if an ability test or an educational test consists of items with the same rating categories (e.g., incorrect, partially correct, correct), the relative positions of the thresholds depend on the relative difficulty of the steps needed to solve the task. These step difficulties vary across items, therefore constraining the threshold structure is not warranted. Considering these limitations, I propose to use the FACTS model as the default factor analysis model when item responses are categorical ratings, and all items are rated using the same response scale.

Note

1. All online Appendices, the R code for fitting and testing the FACTS model, and a sample Mplus code are available at https://osf.io/fe6mu/?view_only=c74c4016256847788d29a6121908852d

Fundings

The author acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P5-0062).

Acknowledgements

I wish to thank Tina Kavčič for providing the raw data on DASS-21.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-574. https://doi.org/10.1007/BF02293814

Andrich, D. (2016). Rasch rating-scale model. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume one: Models* (pp. 75-94). CRC Press.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379-384. https://doi.org/10.3758/BF03192707

Bartholomew, B. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). Hodder Arnold.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328-347. https://doi.org/10.1037/met0000059

Foldnes, N., & Grønneberg, S. (2020). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 525-543. https://doi.org/10.1080/10705511.2019.1673168

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Kavčič, T., Zager Kocjan, G., & Dolenc, P. (2023). Measurement invariance of the cd-risc-10 across gender, age, and education: A study with Slovenian adults. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *42*(3), 1727-1737. https://doi.org/10.1007/s12144-021-01564-3

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335-343. https://doi.org/10.1016/0005-7967(94)00075-u

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57-77. https://doi.org /10.1007/s11336-005-0773-4

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344-362. https://doi.org/10.1037/1082-989X.11.4.344

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551-560. https://doi.org/10.1007/BF02293813

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132. https://doi.org/10.1007/BF02294210

Muthén, B. O., & Asparouhov, T. (2002, December 9). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus.* Muthén & Muthén. https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén..

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434-447. https://doi.org/10.1037/1082-989X.8.4.434

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under sub-optimal conditions. *Psychological Methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.589965

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2024). *The lavaan tutorial*. https://lavaan.ugent.be/tutorial.pdf

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* [*A tribute for Heinz Neudecker*] (pp. 233-247). Kluwer Academic Publishers.

Schneeweiss, H., Komlos, J., & Ahmad, A.S. (2010). Symmetric and asymmetric rounding: A review and some new results. *AStA Advances in Statistical Analysis*, *94*, 247-271. https://doi.org/10.1007/s10182-010-0125-2