Do Items Order? Analysis with Shift-Scale IRT Models

Item ordering refers to the statement that if one item is harder than another for one person, then it is harder for all people. Whether item ordering holds or not is a psychological statement because it describes how people may qualitatively vary. Yet, modern item response theory (IRT) makes an a priori commitment to item ordering, such as in the Rasch model where items have to order, or, conversely, in the 2PL model, where items cannot order. Needed is an IRT model where item ordering or its violation is a function of the data rather than an a priori commitment. We develop a two-parameter shifted-exponential model for this purpose. We show how item ordering may be assessed and discuss computational issues with shift-scale IRT models.

Keywords: Item Ordering, 2PL, Item response theory

Item-response models make only limited assumptions about the psychological processes underlying performance. For instance, there are no assumptions about memory, attention, or the processes that result in differences among individuals. And rarely are there accounts about why people exhibit specific responses for certain items (though see, e.g., De Boeck & Wilson, 2004). Instead, models are specified so that they have desirable measurement, statistical, and computational properties. For example, the Rasch model has the desirable properties that the estimate of performance depends only on the number of correct responses to a set of items, and that the relative performance of two individuals is invariant across items (e.g., Baker & Kim, 2004; Wright, 1977). In our view, this lack of psychological content and emphasis on technical properties in item-response models is desirable. It allows the models to be used broadly across a large number of domains, say from education to cognition to psychopathology, in a fairly unified way, and in turn, keeps test theory timely and topical.

Although item-response models are designed to be content-free, there is one element of psychological content that seems unavoidable in current modeling. That element is *invariant item ordering* (Sijtsma & Hemker, 2000)—if Item A is harder than Item B for one person, then it is at least as hard as Item B for all people. Item ordering and its violation may be seen in Figure 1. If two items order, then the item response curve of one dominates (is always greater than or equal to) the response curve of the other. Items A and B order, with Item A being easier than Item B. Item C does not order with the other two items. For example, Item C is harder than Item B for low-ability individuals but it is easier than Item B for high-ability individuals. Therefore, Items B and C are not consistently easier or harder than one another. A violation of item ordering occurs when item-response curves cross.

Item ordering is a psychological statement because it puts constraints on how individuals may differ. One-parameter IRT models, such as the Rasch model, make the bold prediction that all items order. Two-parameter IRT models, such as 2PL, predict that two items order only if they have the same discriminability. Moreover, if all items in a set order, then 2PL reduces to 1PL. Whereas 1PL forces all items to order, 2PL forces at least some items to not order. By choosing between 1PL and 2PL, the analyst makes a content choice that has psychological implications. Either the analyst forces item orderings or forces their violation.

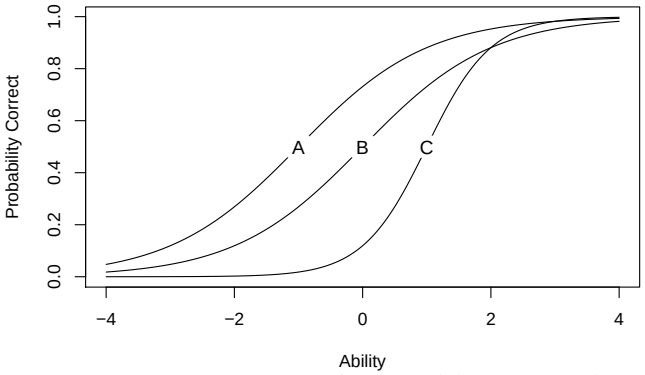

Figure 1. Item response curves from Model 2PL. Items have varying location and scale.

Most modern psychometric modeling is as rich as the 2PL in that items have at a minimum unique ability and discriminability parameters with continuous support. And, consequently, in most psychometric applications, it is implicitly stipulated that all items do not order. There is a minority of psychometricians that use the Rasch model (Wright, 1977). Yet, the Rasch model makes the constraining assumption that items are translations of one another, and this assumption may not provide for an adequate statistical description of the data. For further distinction between statistical arguments and philosophical arguments for and against the Rasch model, see Andrich (2004).

We think the current situation, where researchers must choose between enforcing item ordering or enforcing violations of it, is undesirable. Rather, we think psychometric models should be sufficiently flexible to enforce or violate item orderings

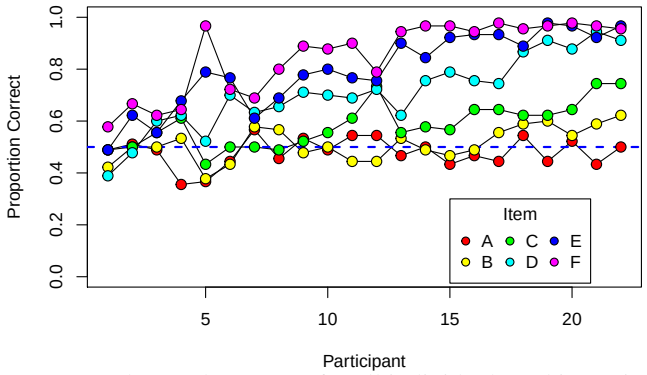

Figure 2. Observed accuracy for all individuals and items in masked digit naming. Items A through F are presentations of 16.7 ms, 25.0ms, 41.7ms, 58.3 ms, 75 ms, and 100 ms. An accuracy value of .5 serves as an at-chance baseline. Data are taken from Morey, Pratte, and Rouder (2018), Figure 7.

dependent on the data. In this paper, we explore alternatives to the logistic link to attain this desirable property.

Insights From Psychophysics

We find it helpful to consider psychophysics as guidance in designing psychometric models. To model psychophysical data, one typically considers each stimulus as an item. For example, Morey, Rouder, and Speckman (2008) asked participants to identify digits that were briefly flashed and subsequently masked. The duration that the digit was presented was varied in six levels from 16.7 ms to 100 ms. We treat each of these six levels as an item, and, in this application, we are fortunate to have 90 replicates for each participant-by-item combination. Because there are replicates, it is possible to visualize response proportions without modeling assumptions. Figure 2 shows these proportions for all individuals and items. The pattern is informative. Here, item ordering seems plausible inasmuch as the small violations may be due to sample noise. Moreover, the item-response curves seemingly are not shifts (or location translations) of one another, and there is a change of discriminability across the items (Item A is far less discriminable than Item F). Hence, we need a model that may account for item orderings while not forcing item-response curves to be shifts of one another.

In this application, it seems highly plausible from a theoretical perspective that the six items order. All people respond at least as well to digits presented for the longer durations than for the shorter durations. Indeed, it is hard to conceive of an individual that has better true performance to digits presented say for 16.7 ms than for 41.7 ms. Morey, Rouder, and Speckman (2009) account for these data with a novel psychometric model that uses a half-normal link and participant discriminability parameters. We develop a more conventional model with item discriminability here.

Item ordering strikes us as plausible in many contexts. Haaf and Rouder (2017) consider the question for response time data in priming and Stroop tasks in cognition. They conclude that response times for congruent and incongruent items in priming and Stroop tasks order for all people.

Yet, item ordering cannot be a universal property, and it is always possible to design items to violate it. Take, for example a depression inventory. Suppose we have a set of Rasch items that span the range of a latent one-dimensional depression construct. We can always add an item, say a general knowledge question, that is far less sensitive for depression yet is located in the middle of the range. In this sense, we can always force item ordering violations in the data. But the question is not whether we can force violations, it is whether item orderings occur in well-articulated domains where items are designed to measure a single latent construct.

Shift-Scale Links and the 2PE Model

To develop our approach, we first highlight the structural and distributional assumptions in 2PL. The structural property is:

$$p_{ij} = G\left(\frac{\theta_j - \alpha_i}{\beta_i}\right),$$

where p_{ij} is the probability of a correct response for the jth individual on the ith item, $i = 1, \ldots, I$, $j = 1, \ldots, J$. The parameter θ_j is the ability of the jth individual; the parameters α_i and β_i are respectively the location and scale of the ith item. The function G, the link, is monotonically increasing.

In 2PL, the distributional assumption is that G is the logistic link given by $G(x) = 1/(1 - \exp(-x))$. In the model, the parameter α_i describes the center of the link and parameter β_i describes the scale. Figure 1 shows three items. The centers are denoted by the letter, and all three items have different centers. Item A and B have the same scale but different centers; Item C has a larger scale. Evident in the figure is the violation of item ordering; for example, Item C is easier than Item B for low ability individuals but harder than Item B for high ability individuals. This violation occurs whenever two items have different scale parameters β_i.

The main insight here is that location-scale models with unbounded support, such as 2PL, lead to the critical flaw where scale changes necessarily violate item ordering. Shift-scale models, where the location parameter serves to shift the bound on support, do not have this problem. The shift-scale link is the CDF of a distribution with support on the positive half line. In this report, we specify G as the exponential link, $G(x) = 1 - \exp(-x)$ for $x > 0$, though other links are possible. With this link:

$$p_{ij} = \begin{cases} 1 - \exp\left(-\frac{\theta_j - \alpha_i}{\beta_i}\right), & \theta_j \geq \alpha_i, \\ 0, & \alpha_i > \theta_j. \end{cases}$$

The model is analogous in some regards to 2-PL and parameters α and β play the same role as locating and scaling the

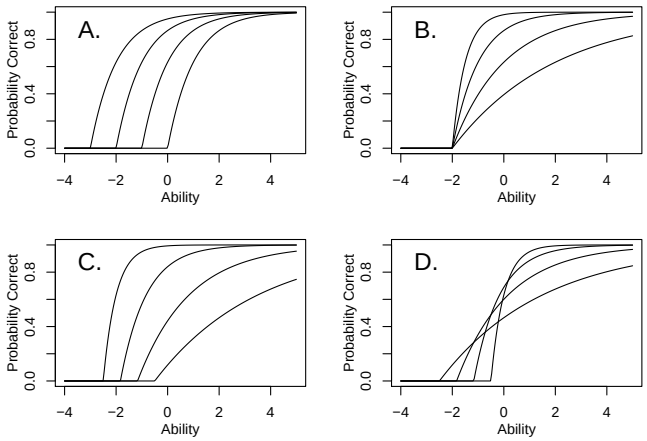

Figure 3. Item response curves from the shifted exponential model. A. Changes in shift with constant scale preserve item orderings. B. Changes in scale with constant shift preserve item orderings. This property is not true for 2-PL. C. Changes in shift and scale preserve item ordering so long as scale increases with shift. D. When shift increases and scale decreases, item ordering is violated.

distribution. Figure 3 shows a few sets of item response curves from the shifted exponential model. In Panel A, the items differ only in shift, and item response curves are translations of one another. Item ordering is evident, and the model is analogous to the Rasch model, albeit with a different link. In Panel B, the items differ only in scale, and again item ordering is evident. This pattern contrasts with 2PL because in 2PL scale changes necessitate violations of item orderings. Panel C shows item orderings, and these orderings are achieved by having shift and scale increase for each successive item. Panel D shows how violations of item orderings come about in the model—here, increases in shift correspond to decreases in scale. Figure 3 shows the desirable flexibility of the 2PE model with respect to item ordering. Item ordering occurs for two items i_1 and i_2 if $\alpha_{i_1} > \alpha_{i_2} \iff \beta_{i_1} \geq \beta_{i_2}$.

We refer to the parameter α as a shift parameter. Shift parameters are location parameters that also define the bound of support for a distribution. The key property of the 2PE model is the shift parameter, and in this parameterization, scale increases make items harder for all people. Although shift-scale models have the potential to account for item orderings and their violations, there are technical drawbacks. These models are not as statistically or computationally convenient as models where the bound of support is not a parameter. We highlight these statistical difficulties as they are consequential in application.

Model Specification and Analysis

The shifted exponential model may be analyzed in a Bayesian framework. Needed are priors for the collections of $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. We follow the convention of fixing the mean and variance of people's abilities:

$$\theta_j \overset{iid}{\sim} \text{Normal}(0, 1).$$

The priors for the item parameters are

$$\alpha_i \overset{iid}{\sim} \text{Normal}(c_1, c_2),$$
$$\beta_i \overset{iid}{\sim} \text{Inverse Gamma}(c_3, c_4),$$

where $(c_1, \ldots, c_4)$ are prior parameters that need to be set before data analysis. We use the following settings in application: $c_1 = -1$, $c_2 = 10^2$, $c_3 = c_4 = 1$, and the resulting priors are quite broad and only weakly informative.

The resulting joint posterior for the parameters is:

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{Y}) \propto \prod_i \prod_j [1 - \exp(-(\theta_j - \alpha_i)/\beta_i)]^{Y_{ij}} [\exp(-(\theta_j - \alpha_i)/\beta_i)]^{1-Y_{ij}} \times \prod_j \phi(\theta_j) \times \prod_i \phi((\alpha_i - c_1)/c_2^{1/2}) \times \prod_i f(\beta_j; c_3, c_4), \quad (1)$$

where ϕ is the normal density and f is the density on an inverse gamma with shape c_3 and scale c_4.

We initially derived conditional posteriors and implemented a MCMC chain with Metropolis Hasting sampling of each parameter. Yet, we were unable to achieve good mixing with this approach. The problem is obvious; the shift and scale parameters are too interdependent. A large increase in shift and a corresponding decrease in scale results in likelihoods that change little. One way of obtaining chains that mix well is to use an alternative sampler that better handles ridges in the likelihood. We chose the `Stan` package (Stan Development Team, 2018), which uses Hamiltonian sampling and is ideal for this application. The `R` code is integrated into this manuscript and can be found at https://github.com/PerceptionAndCognitionLab/irt-2pe.

Applications

To assess whether item ordering holds in common IRT applications, we reanalyzed a few readily-available sets from the `mirt` (Chalmers, 2012) and `sirt` (Robitzsch, 2016) packages as well as for the Morey et al. data previously presented.

LSAT-6

The LSAT-6 data set, from Thissen (1982), consists of the responses of one-thousand individuals to five dichotomously scored items from the Law School Admissions Test, Section 6. To examine mixing, we ran a single chain for 800 iterations, the first 200 of which served as a burn-in. The quality of mixing was assessed by inspection of chains and by the effective samples statistic. Figure 4 shows the trace plots, and each plot is for an item. The two traces are for the two item parameter

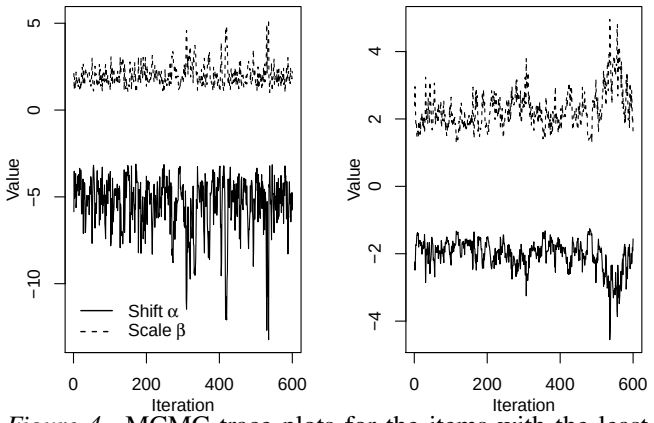

Figure 4. MCMC trace plots for the items with the least autocorrelated (left) and most autocorrelated (right) item parameters.

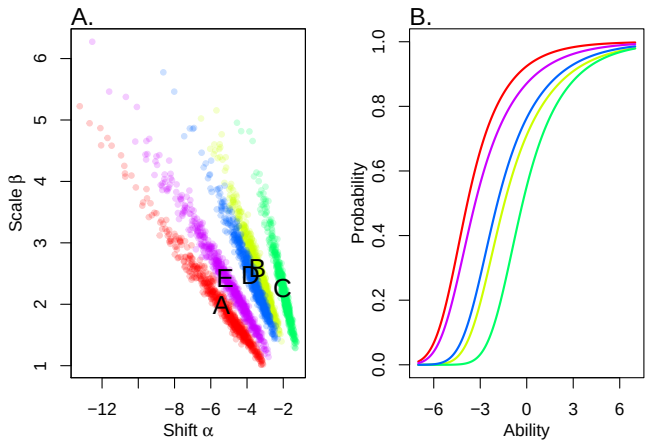

Figure 5. Results for the LSAT6 data set. A. Posterior samples from the five distributions. B. Item response curves derived from posterior means.

shift (α_i) and scale (β_i). The left plot is for the item that mixed best; the right plot is for the one that mixed worst. As can be seen, mixing, while not particularly good, is sufficient. The effective sample sizes (Ripley, 1979) at worst are 1/10 the nominal values, and obtaining hundreds or thousands of samples is not particularly time consuming. Hence, posterior distributions may be estimated to a reasonable degree of confidence.

Figure 5A shows the bivariate posterior samples for all item parameters. There are five clouds, with one cloud for each item in the set. Here we see that item ordering may hold. Overall, it seems that Item A has the smallest shift and scale, and the remaining items increase in shift while maintaining about the same scale. The bivariate posterior mean for each item is located by the letter, and the corresponding item-response curves from these posterior means are shown in Figure 5B. As can be seen, the item ordering is plausible over a reasonable range of abilities.

Although the 2PE model is helpful for assessing item ordering, it has difficult statistical properties. The bivariate posterior distribution is highly correlated, and this correlation reflects long ridges in the likelihood. Such ridges necessitate care in evaluating posterior distributions as parameters are only weakly identifiable.

Astute readers may inquire about alternative parameterizations of the 2PE model that may lead to less correlated posteriors. For example, we can define an alternative *center-scale* parameterization of the 2PE model by using the mean, $\mu = \alpha + \beta$, as a parameter. The 2PE model may then be expressed as

$$p_{ij} = \begin{cases} 1 - \exp\left(-\left[1 + \frac{\theta_j - \mu_i}{\beta_i}\right]\right), & \theta_j \geq \mu_i - \beta_i, \\ 0, & \theta_j < \mu_i - \beta_i. \end{cases}$$

The inequality constraint for ordering items i_1 and i_2 is $\beta_{i_1} > \beta_{i_2} \iff \mu_{i_1} - \beta_{i_1} \geq \mu_{i_2} - \beta_{i_2}$.

We have explored this parameterization, and the joint posteriors of μ and β have the same degree of ridge-like concentration (though the orientation of these ridges is rotated 45 degrees). Therefore, there is no gain to the alternative parameterization. The statistical difficulties do not reflect the parameterization, but are inherent in the shifted-exponential link itself.

LSAT7

The LSAT-7 data set, from Bock and Lieberman (1970), consists of the response of one-thousand individuals to five dichotomously scored items from the Law School Admissions Test, Section 7. Figure 6A shows the bivariate posterior samples for all item parameters. There are five clouds, with one for each item in the set. Here we see a clear violation of item ordering. Consider Items E and C; Item E has a smaller shift but larger scale than Item C. The violation of ordering is also prominent in Figure 6B, which shows the derived item response curves at posterior mean values.

Masked Digit Identification

One domain in which it is highly plausible that item ordering holds is psychophysics. Previously, we treated stimulus duration as an item variable, and it seems inconceivable that anyone is truly better at identifying stimuli presented at shorter than at longer durations. We reanalyzed the accuracy data in Figure 2 from Morey et al. (2008) with a modified 2PE model:

$$Y_{ij} \sim \text{Binomial}(p_{ij}, N_{ij}),$$
$$p_{ij} = \frac{1}{2} + \frac{1}{2}\left(1 - \exp\left[-\frac{\theta_i - \alpha_i}{\beta_i}\right]\right).$$

The main modification here is a baseline accuracy of .5 rather than zero. In Morey et al's experiment, half the digits were less-than-five and half were greater-than-five. Individuals had

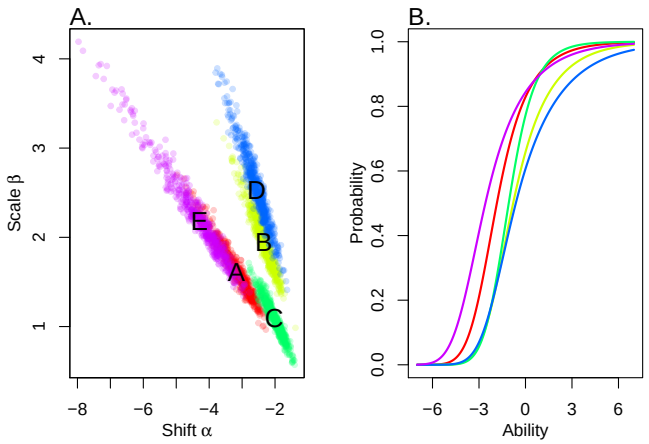

Figure 6. Results for the LSAT7 data set. A. Posterior samples from the five distributions. B. Item response curves derived from posterior means.

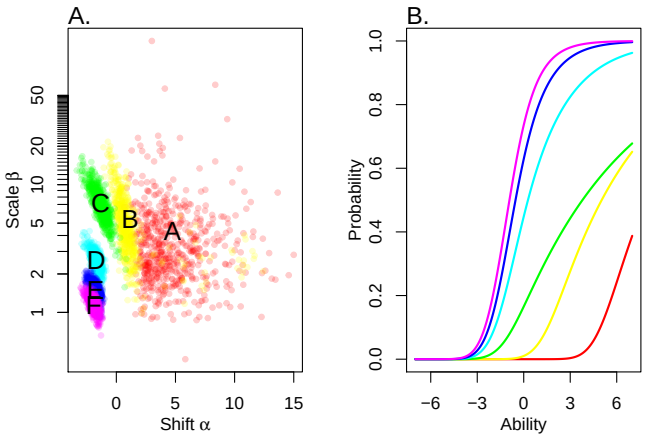

Figure 7. Results for the digit identification data set. A. Posterior samples from the six presentation durations. The y-axis is shown in log scale. B. Item response curves derived from posterior means.

to choose among these two alternatives. With this experimental setup, individuals unable to identify the stimulus at all have a 50% chance of correct response.

The results of the analysis are shown in Figure 7, and as can be seen, the pattern is nonmonotonic. The easiest four items, Items C through F, order strongly and are notably different in scale. The hardest two items, Items A and B, however, order in reverse. The reason for this reverse ordering is more of a statistical issue than any statement about the items. Items A and B correspond to the shortest stimulus durations, and inspection of Figure 2 reveals that it is highly likely that none of the individuals could identify any of the targets above baseline. In this case, there is no information from the performance data other than the shift is quite high. The scale, in particular, reflects only the prior settings, which are lower in value than the data-driven value for Item C. As the posterior on scale reflects the prior to a greater and greater degree, that is, for harder items, the scale estimates may actually decrease in value depending on the prior settings. We have confirmed this excessive dependence for hard items by trying various values of priors, and they do have the described effect. Further, the attenuated correlation between shift and scale for hard items is also explained by the same lack of data dependency. As the items become increasingly difficult and the posterior reflects more the prior, the independence in the priors becomes more evident in the posterior. Because the likelihood is so diffuse, it no longer influences the posterior.

This behavior is not unique to the 2PE model and poses no special limitation. In any two-parameter IRT model, including 2PL, if all participants perform near baseline on an item, it is very difficult to estimate all parameter values other than to know the item difficulty is quite high. In any IRT analysis, all items that are estimable must admit some heterogeneity in performance.

General Discussion

Standard IRT models such as the Rasch model and 2PL make strong substantive commitments about item ordering. In the Rasch model, all items must order; in 2PL all items must not order. These commitments, in our view, are not generally applicable and must be assessed on a case by case basis. For example, in psychophysics, where items index levels of physical strength, it is reasonable to expect item ordering. Such an ordering would be in violation of 2PL. While most psychometricians do not use stimuli that vary on a single, physical dimension, the point remains that psychometric models *should* apply to this case as an important boundary. Absent from data, it is difficult to justify a commitment to item ordering or to its violation.

We propose a class of IRT links that are flexible with regard to item ordering. These links allow for item ordering or violations of item ordering depending on the data. The links use bounded support, and the parameter of support, the shift, along with the scale, form the critical item parameters. Although we develop a shifted exponential link, other choices have the same flexibility. Examples include a shifted lognormal with fixed shape or a uniform. These links are "you can have your cake and eat it" links in that one retains the flexibility of a two-parameter model without the strong commitment to violations of item ordering.

The downside of the current proposal is statistical. These models are not very computationally convenient. The main issue is the weak identifiability of item parameters which is manifest in long ridges in the likelihood. In this case, the issue is a trade-off between shift and scale. Logistic, probit, and t-linked models do not have this issue because the support is fixed across items. Although we find the hybrid Monte Carlo routines in `Stan` to be excellent for this application, the

problem of weak identifiability remains. Localizing parameters, and assessing whether item ordering holds is not as of yet straightforward.

We ask psychometricians pay heed to a basic, almost minimalist element of psychological content, that is, whether items in a context vary consistently across all people. More pragmatically, we believe domains should be classified as admitting item orderings or violating item orderings. And when they admit a natural ordering, the claim that the items measure a unidimensional latent meaningful psychological construct is all the more stronger.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, I7–I16.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 1–29. Retrieved from https://www.jstatsoft.org/article/view/v048i06

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer-Verlag.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2009). A truncated-probit item response model for estimating psychophysical thresholds. *Psychometrika*, *74*, 603–618.

Ripley, B. D. (1979). Tests ofrandomness' for spatial point patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 368–374.

Robitzsch, A. (2016). *Sirt: Supplementary item response theory models*. Retrieved from https://CRAN.R-project.org/package=sirt

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of irt models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25*(4), 391–415.

Stan Development Team. (2018). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Wright, B. D. (1977). Solving measurement problems with the rasch model. *Journal of Educational Measurement*, *14*(2), 97–116.