Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

From the Roots: Likelihood Ratio DIF testing for IRTrees

Anne Thissen-Roe* and Amy Li

Harver, Dallas, TX, United States
*Corresponding author. Email: anne.thissen-roe@harver.com

Abstract

Assessments of personality and similar constructs often use items in a rating scale format, such as Likert items, with flexible or inferred scale anchors. Intrinsically multidimensional polytomous IRT models, including IRTrees, can disentangle construct-relevant response behavior from scale use behavior or response sets. We propose a set of general principles for applying IRT likelihood ratio DIF testing to whole IRTrees, and demonstrate a real application to an operational employee selection personality instrument.

Keywords: DIF, IRTrees, Likert items, response sets

Assessments of personality and similar constructs often use items in a rating scale format, such as Likert items, with flexible or inferred scale anchors. When we measure a construct using this type of item, responses are affected by both the intended construct and also individual differences in scale use: response sets. In recent years, our field has developed intrinsically multidimensional polytomous IRT models for this purpose. These include IRTrees, which model multiple sequential processes with latent construct and response set traits (De Boeck and Partchev, 2012). These models improve both data fit and the modeling of construct relationships, with applications ranging from basic behavioral research to employee selection.

DIF testing methods have not kept up with the new item models, limiting our options in evaluating the content these models fit. If a set of items is affected by response sets, but we use a DIF test based on a construct-only model, we are left unsure if our DIF findings are really DIF, or a symptom of model misfit. Jin, Wu and Chen evaluated non-IRT DIF tests (Logistic Regression and Odds Ratio) applied to binary pseudo-items that constitute an IRTree (Jin et al., 2018). Here, we propose a set of general principles for applying IRT likelihood ratio DIF testing to whole IRTrees, even those that are not binary-decomposable, and demonstrate a real application to an operational employee selection personality instrument using six-point dual-anchored response scales.

1. Background: Extreme Response Set

Personality items, as well as other subjective self-report items, often have several ordered response options, but with flexible or inferred scale anchors, as shown in Figure 1. (This is in contrast to, for example, task frequency surveys, with highly specific response options such as: "Once a year," "Once a month," "Once a week," "Every day," "Several times a day.") Due to clear semantic ordering, responses are assumed monotonically related to a single latent construct dimension, even when that dimension represents a composite of personality traits.

These items are vulnerable to response sets, including acquiescent response set, extreme response set, and midpoint response set. Responses influenced by sets violate the classical assumption that scores have a linear relationship with a construct, and the modern assumption that ordered response options have a monotonic progression according to a construct. A model such as the Graded Response Model

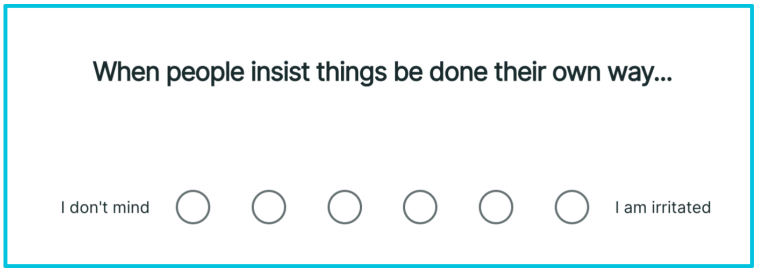

Figure 1. A personality item with four inferred intermediate scale anchors, from Harver's Personality Print assessment.

(Samejima, 1969), which depends on that assumption, may fit data poorly when response sets are in effect. Another symptom is excess correlation between scales that should be well-differentiated.

Applying an unordered model such as the Nominal Response Model (Bock, 1972) can reveal such violations through improved fit statistics and diagnostic plots, such as the plot shown in Figure 4. Note that both ends of the scale are at the same end of theta! This item has "folded" (Thissen-Roe and Thissen, 2013). Extreme response set has overriden the semantic construct, placing the middle responses at the left, and the ends of the scale at the right.

Researchers in our field have come up with a number of models that can handle extreme response set. One is the Multidimensional Nominal Response Model (Bolt and Johnson, 2009), which can handle a broad range of intrinsic multidimensionality in polytomous item responses. The Proportional Thresholds Model (Johnson, 2003) is more specific to the case of extreme response set in ordered polytomous items.

Finally, there are IRTrees (De Boeck and Partchev, 2012). IRTrees model multiple sequential processes underlying a response, which may depend on different latent traits for constructs and/or response sets, in a decision tree structure that ends in response option nodes or "leaves." This modeling of internal processes can provide enough structure to disentangle response set effects from constructs of interest, and thereby provide significantly cleaner construct level estimates for a variety of purposes.

For items with five or six ordered response options, Böckenholt considered three-process models, with as many as six possible orders of sequential pseudoitems for midpoint (M), extremity (E) and agreement (A) (Böckenholt, 2012, 2017). Figure 2 shows one of these orders, the Midpoint-Agreement-Extremity Model (MAE). This sequence was Böckenholt's preference for classical Likert items with one response (as opposed to a prompted sequence of responses that itself involves a tree). Each pseudoitem is modeled using a unidimensional IRT model; Böckenholt used the one-parameter probit model (Birnbaum, 1968), while some later researchers preferred the two-parameter logistic model (Alarcon and Lee, 2022). The probability of responding at a particular response option leaf is the product of the probabilities given by the pseudoitem responses leading to that leaf. (This works because the pseudoitem probabilities are modeled as independent; any correlations between underlying traits or response sets are modeled as a feature of the population, existing across all items.)

All six sequences of pseudoitems are valid models for a six-option item; a researcher can pick the one that makes sense as a theory of the process. That said, unless distinct item parameters are used for branch decision processes at the same level (such as left and right branch Agreement), not all of the models result in different predictions for the data; the proper place of agreement in the sequence can be reasoned about but not empirically tested.

Another IRTree option is the Two-Decision Model (TDM), shown in Figure 3 (Thissen-Roe and Thissen, 2013). The TDM calls for only two processes, agreement and a combined midpoint-extremity process. As in the three-process model, agreement is permitted to correlate at the population level with extremity, but is modeled as independent within the item. Because the correlations between person parameters for each tree level are estimated as parameters, a three-process AME model might

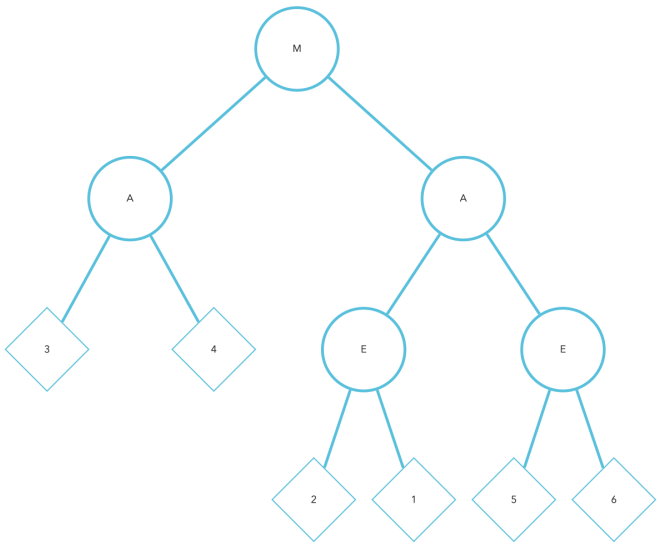

Figure 2. The Midpoint-Agreement-Extremity Model for items with six ordered response options (Böckenholt, 2017). According to this model, a person responding to such an item first decides whether to express an opinion or use a near-middle response (M pseudoitem, top), then proceeds to decide which end of the scale to select (A pseudoitems, second level). Finally, if an opinion is being expressed, the person decides whether to use the most extreme response options (E pseudoitems, third level on the right-hand branch).

be considered to devolve into a two-process model if the midpoint and extremity processes approach collinearity. It is beyond the scope of this chapter to determine which process model is more correct. Let us treat them both as plausible.

In addition to this difference, and using the Graded Response Model for some nodes with more than two branches, the Two-Decision Model is an unorthodox IRTree in that it has a cross-level parameter ν that incorporates root-theta agreement information into branch pseudoitems. This parameter allows for some people, some of the time, to be using a Likert scale the way it was intended, such that extreme values actually indicate more extreme agreement or disagreement. The TDM is therefore not fully decomposable into unidimensional latent pseudoitems.

The response option probabilities for a six-point item under the TDM are given as follows. First, the Agreement node follows a simple two-parameter logistic model:

$$p_a = \frac{1}{1 + e^{-(c_a + a_a \theta_a)}}. \tag{1}$$

The second-level nodes are modified by the cross-level parameter ν, combined with the logit term of the Agreement pseudoitem. For convenience we will name this combination ν_e:

$$\nu_e = \nu(c_a + a_a \theta_a). \tag{2}$$

Then, the second-level nodes produce trace lines for response options 1 through 6 as follows:

$$T_1 = (1 - p_a)\frac{1}{1 + e^{-(c_{e2} + a_e(\theta_e - \nu_e))}}. \tag{3}$$

$$T_2 = (1 - p_a)\frac{1}{1 + e^{-(c_{e1} + a_e(\theta_e - \nu_e))}} - T_1. \tag{4}$$

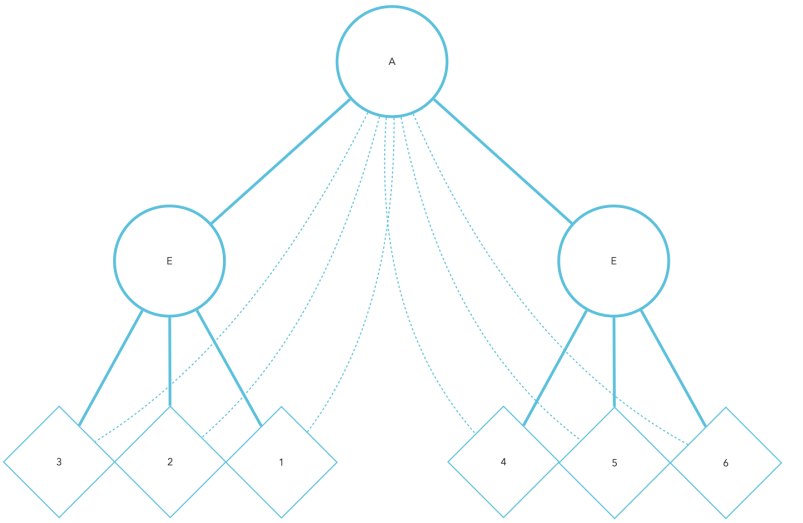

Figure 3. The Two-Decision Model (Thissen-Roe and Thissen, 2013) for six ordered response options. At the first node (Agreement; A), a person decides which end of the scale to prefer, a binary decision because no absolute middle response option is available with an even number of options. On each branch, a second decision is made regarding how much emphasis to use, a single midpoint-extremity process (Extremity; E). The level of emphasis is influenced by the decisiveness of the Agreement node result, which is to say, persons near the item's agreement threshold, who might consider responding at each end of the scale, may use the middle response options to indicate their ambivalence, while persons without ambivalence may prefer the extreme response options for that reason. This gives an item-specific level of interaction between the first- and second-level nodes.

$$T_3 = (1 - p_a)(1 - \frac{1}{1 + e^{-(c_{e1} + a_e(\theta_e - \nu_e))}}). \tag{5}$$

$$T_4 = p_a(1 - \frac{1}{1 + e^{-(c_{e1} + a_e(\theta_e + \nu_e))}}). \tag{6}$$

$$T_5 = p_a \frac{1}{1 + e^{-(c_{e1} + a_e(\theta_e + \nu_e))}} - T_6. \tag{7}$$

$$T_6 = p_a \frac{1}{1 + e^{-(c_{e2} + a_e(\theta_e + \nu_e))}}. \tag{8}$$

What if one wants to check for DIF in these items? A plot of NRM trace lines might make it obvious that DIF exists, but not how it occurred. It would be more helpful to examine DIF in the IRTree model. We have both non-IRT and IRT DIF methods for one-step non-tree items, and non-IRT DIF methods for IRTrees. Let us introduce IRT DIF for IRTrees.

2. Likelihood Ratio DIF testing for IRTrees: Principles and Examples

Here, we apply the IRT Likelihood Ratio DIF testing method to IRTrees. Consider the common practice of testing one candidate item at a time (Thissen, 2001; Thissen, 2025), with reference to the scale composed of that item and an anchor set of items, or by all other candidate items if no anchor has been established. Differences between group means and variances should be allowed; this can be

done by fixing the reference group mean to zero and variance to one for identification, and freely estimating the focal group mean and variance as parameters. We believe that it is reasonable to hold the correlations (not covariances) between traits and response sets equal across groups. Within such a structure, we propose a set of general principles for testing each item, working through the tree from roots to branches, as follows:

1. Begin with all parameters at all model levels unconstrained across groups; they take on group-specific values for reference and focal groups.
   (a) Conduct an overall test for all parameters free vs. all parameters constrained.
2. Test the component models of the tree sequentially from the root node (first branching) to the end nodes.
   (a) For each component model, conduct an overall test for all parameters free vs. all parameters constrained, with all parameters of previous branching models constrained and subsequent branching models free on both sides of the test.
   (b) If the overall test reveals significant DIF at a model level, proceed to individual parameter tests.
3. The general form of IRTrees does not require that the left and right branches lead to models with the same parameters, or same structure. If there are dissimilar models at a branching level of the tree, and that level is reached in the test sequence, test all models at that level (i.e., do not apply a stopping rule between models).
4. Within each component model, if there is an established sequence for ordering parameter tests, follow it.
   (a) Parameter test sequences were previoiusly established for the two-parameter logistic model, three-parameter logistic model, and Graded Response Model (Thissen, 2001).
   (b) If there is no established sequence, adhere as closely to precedent as possible.

Note that conducting all steps for many items can result in a large number of statistical tests. One can apply methods to control the false discovery rate, or, as is our preference, split the available sample and conduct multiple replications of the whole study.

As an example, we apply the principles to the three-process Midpoint-Agreement-Extremity Model (MAE; Figure 2), wherein we have used the two-parameter logistic model for each pseudoitem, with slope and intercept parameters (a_m, c_m) for the Midpoint pseudoitem, symmetrically (a_a, c_a) for both Agreement pseudoitems, and symmetrically (a_e, c_e) for both Extremity pseudoitems:

1. Test all parameters free versus all equal; this is the overall test for the whole item.
2. Midpoint Pseudoitem tests:
   (a) Overall test.
   (b) Test all parameters free versus (a_m) constrained.
   (c) Test (a_m) constrained versus (a_m, c_m) constrained.
3. Agreement Pseudoitem tests:
   (a) Overall test.
   (b) Test (a_m, c_m) constrained versus (a_m, c_m, a_a) constrained.
   (c) Test (a_m, c_m, a_a) constrained versus (a_m, c_m, a_a, c_a) constrained.
4. Extremity Pseudoitem tests:
   (a) Overall test.
   (b) Test (a_m, c_m, a_a, c_a) constrained versus (a_m, c_m, a_a, c_a, a_e) constrained.
   (c) Test (a_m, c_m, a_a, c_a, a_e) constrained versus all parameters equal.

We can also apply the principles to the Two-Decision Model (TDM; Figure 3). However, in addition to using the Graded Response Model for some nodes, the Two-Decision Model has a cross-level parameter ν that incorporates root-theta agreement information into branch pseudoitems. This allows for some people, some of the time, to be using a Likert scale the way it was intended, such that extreme values actually indicate more extreme agreement or disagreement. There is no established sequence that includes such a parameter, but DIF findings are most interpretable when ν is tested at the beginning of the branch extremity nodes, before a and c. This results in a sequence of steps:

1. Test all parameters free versus all equal.
2. Agreement Pseudoitem tests:
   (a) Overall test.
   (b) Test all parameters free versus (a_a) constrained.
   (c) Test (a_a) constrained versus (a_a, c_a) constrained.
3. Extremity Pseudoitem tests:
   (a) Overall test.
   (b) Test (a_a, c_a) constrained versus (a_a, c_a, ν) constrained.
   (c) Test (a_a, c_a, ν) constrained versus (a_a, c_a, ν, a_e) constrained.
   (d) Test (a_a, c_a, ν, a_e) constrained versus all parameters equal.

3. An Applied Example

Personality Print is a personality assessment offered by Harver for employee selection. It has five factors and, in its short form, eleven narrow scales. Four of its five factors are approximately Big Five aligned; the fifth is only modestly correlated due to its focus on a work context. It uses unidimensional pairwise preference items, with a six point rating scale and no intermediate anchor labels, and is vulnerable to response sets. It is used globally, but most candidates test in English, Dutch, and other European languages. Candidates choose their preferred language from a list, and in some cases choose a localization within the language (e.g., Portuguese versus Brazilian Portuguese; French as spoken in France, Belgium or Canada).

Thus, our example comes from an off-label use of DIF testing. Rather than compare demographic groups, we compare users who tested in American English (N = 24,596) with users who tested in British English (N = 2,834) over the same period of time, part of an effort to ensure we haven't introduced any issues in localizing the assessment. (For the purpose of this demonstration, only a single replicate is presented.) The majority of users selecting American English are candidates for jobs in the United States, and the majority of users selecting British English are candidates for European jobs. However, English is spoken in many places, candidates self-select, and candidates may apply to jobs from distant points. Therefore, the American English sample, and to a lesser degree the

Table 1. Sequence of tests for an example item, under TDM

Test	Δdf	λ_{LR}	α = 0.05	significant?
Overall	6	50.33	12.59	yes
a_a	1	0.07	3.84	no
c_a	1	21.53	3.84	yes
ν	1	1.10	3.84	no
a_e	1	0.21	3.84	no
c_e	2	27.43	5.99	yes

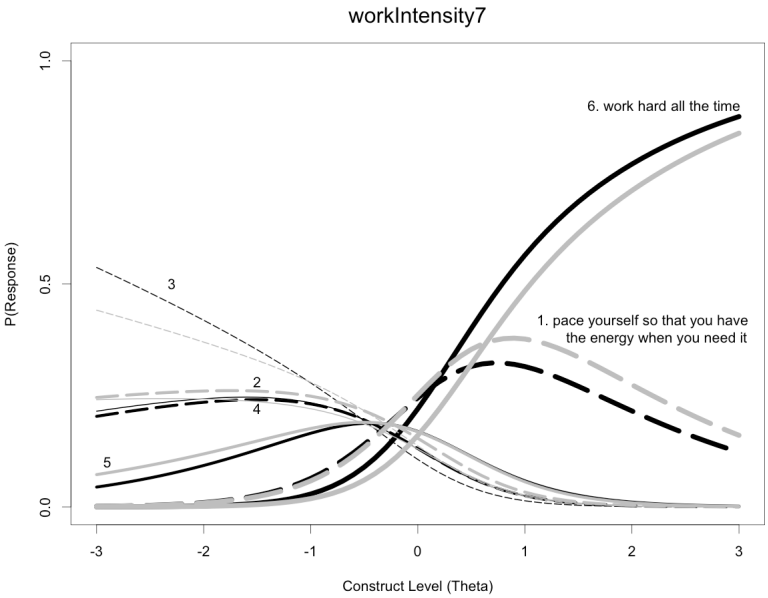

Figure 4. Nominal response model trace lines for a personality item with ordered response options, showing "item folding" characteristic of un-modeled extreme response set, and DIF between two languages. Trace lines for American English appear in black; trace lines for British English appear in grey. Dashed lines are for the three response options associated with the left stem, "Pace yourself...", while solid lines are associated with the right stem, "Work hard all the time." The thickest lines are for the extreme options and the thinnest lines for the near-midpoint options. In this representation, it is clear that there are differences between the response patterns of the two groups of users, but less than clear what cognitive processes give rise to the differences.

British English sample, includes respondents from around the world. It is not accurate to describe the samples as simply "American" and "British."

We would expect items to behave the same in the two versions of English, and need not be concerned with translation errors such as accidental negation. The item shown in Figure 4 doesn't even have any spelling differences, and yet, its folded trace lines under the Nominal Response Model visibly differ. The same words can be interpreted differently based on a larger linguistic and cultural context, and these differences may affect response option use as well as relative agreement with the item's two stems. We can apply DIF analysis with IRTree models to understand the differences and break them down in terms of component processes.

In order to disentangle the effects of the stems and the response options, we analyzed the data for DIF using the TDM. We analyzed each scale separately so that we were working with unidimensional personality constructs. Prior practice for modeling multidimensional personality batteries uses universal response set traits across all item content, as the response sets are associated with the shared item structure rather than the personality construct measured (Thissen-Roe and Thissen, 2013). By contrast, examining the items of each scale separately implies that each of the eleven scales has its own associated extremity and (in MAE) midpoint traits. There is no good theoretical reason to expect different response set behavior between scales. In practice, however, response set traits are sufficiently unidimensional across item content, and achieve high reliability so fast, that we obtain eleven highly correlated estimates of each response set, consistent with one underlying individual difference.

The likelihood ratio DIF procedure allows for inferred mean differences between groups, and

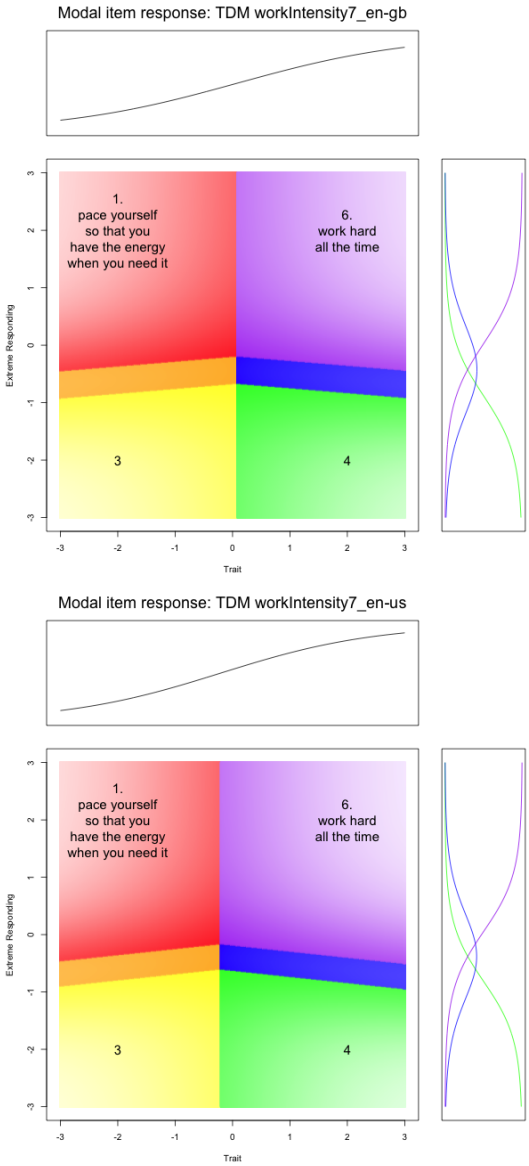

Figure 5. The Two-Decision Model shown in two dimensions for the same Work Intensity item as in Figure 4. Distinct item models are shown for British English responses (top) and American English responses (bottom). Levels of agreement (for the designated right-hand stem, "Work hard all the time") are higher among American English users; the difference between intercept parameters is about 0.21 and is visible as a slight rightward shift of the midline in the upper graph. American English users also show a greater preference for the extreme response options and less for the near-middle options. These intercept differences are about 0.07 and 0.05 respectively, and while they are significant at this sample size, they are harder to spot.

on some of the scales such differences were in evidence. For example, on Assertiveness under the Two-Decision Model, our British English candidates scored half a standard unit higher on average than their American English counterparts. In contrast, the Work Intensity scale doesn't have a large mean difference.

On our example item, we found statistically significant DIF in the overall test (see Table 1),

which stemmed mainly from the intercept parameters. Figure 5 shows a plot of dominant regions for each response option under TDM, in two dimensions: horizontal for agreement and vertical for midpoint up to extremity. The vertical divide between top (British English) and bottom (American English) plots shows the difference in relative preference for the two stems ("agreement" in a pairwise preference item). American English candidates are more likely to endorse "work hard all the time." British English candidates are more likely to choose the "pace yourself" end of the scale, although it is still a minority opinion expressed among job applicants.

How does all of this help us as test developers? The clear difference in stem preferences between self-selected localization groups, when disentangled from scale use elements, speaks to a difference in interpretation or context that is relevant to the specific item. If translation were involved, this type of result would point us to examining the translation of the stems rather than the response options; because it is not, our next steps are to examine what the reasons for the difference might be, and decide whether to leave the item alone because it is reflecting a meaningful and useful difference, declare the item problematic and replace it, or adjust item parameters according to selected localization.

4. Conclusion

In this chapter, we have proposed a set of principles for applying likelihood ratio DIF tests in sequence to individual parameters of IRTree models, and explored the application of those principles to two particular IRTree models, using a specific personality item as an example. We hope that researchers working with other IRTrees or in other contexts will find the principles and examples helpful, and we wish all of our readers the best of luck in their applications.

Acknowledgement

We would like to thank our teammates at Harver for their kind support of our research project, and in particular Ben Porr and Sean Noble. We are also grateful to Hao Wu for helpful comments on a previous draft of this work.

Funding Statement This research was supported by Harver B.V..

Competing Interests None.

References

Alarcon, G. M., & Lee, M. A. (2022). The relationship of insufficient effort responding and response styles: An online experiment. *Frontiers in Psychology*, *12*, 1–15.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

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

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17:4*, 665–678.

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*, 69–83.

Bolt, D., & Johnson, T. (2009). Applications of a MIRT model to self-report measures: Addressing score bias and DIF due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28.

Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2018). Adopting the multi-process approach to detect differential item functioning in likert scales. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 307–318). Springer.

Johnson, T. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*, 563–583.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, No. 18*.

Thissen, D. (2001). *IRTLRDIF v2.0b—software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. L. L. Thurstone Psychometric Laboratory, University of North Carolina. Chapel Hill, NC.

Thissen, D. (2025). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research*, *60(2)*, 211–235. https://doi.org/10.1080/00273171.2024.2396148

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38:5*, 522–547.