PSYCHOMETRIC ANALYSIS AND VALIDATION

MASTER YOUR SUPERPOWERS

PERSONALITY ASSESSMENT

Introduction

This report presents an investigation of the psychometric properties of the Master Your Superpowers Personality Assessment. Currently, this instrument consists of a 50-item inventory composed of 5 scales: Water, Wood, Fire, Earth and Metal. Each of these scales is composed of 10 items. All items consist of a statement (e.g., “I feel vulnerable when others know too much about what’s going on in my life.”), to which respondents are asked to indicate their degree of agreement, using a 5-point Likert scale.

There are no reversed items in the scale. For each respondent, scale scores are calculated by summing responses to the corresponding items. A person is attributed a type based on the largest score obtained.

Methods

The dataset analyzed consisted of the responses of 199 adults. No data was missing in the analysis. All analyses were performed using the statistical programming language R, by a consultant with a PhD in Psychology and several published peer-reviewed articles involving psychometric methods.

Reliability Analysis

To assess reliability, we used various methods. First, we computed internal consistency using Cronbach's alpha (Cronbach, 1951), its most popular measure. Internal consistency is a common measure of reliability based on the correlation between items in a scale (i.e., it estimates the extent to which items in a scale produce scores that are statistically coherent with one another). A value of 0.7 or higher is usually considered acceptable. In addition, even though no decision was made at this stage to delete an item, we computed the alpha that would be obtained should a given item be deleted ("alpha if deleted"), which is a common way of identifying problematic items. These were computed using the "psych" package for R (Revelle, 2024).

In addition to Cronbach's alpha, we used another popular measure of unidimensionality: McDonald's omega (Dunn et al., 2014; McDonald, 2000). Unlike Cronbach's alpha, which is based on the assumption of equal factor loadings among items (i.e., essential tau-equivalence), McDonald's omega is a more flexible and robust indicator that assumes the congeneric model, in which items can have different loadings. Because the congeneric model is often more realistic than the essentially tau-equivalent model, McDonald's omega provides a more accurate estimate of a scale's internal reliability. Here as well, we used a threshold of .7. Similar to Cronbach's alpha, we also computed the omega that would be obtained should a given item be deleted ("omega if deleted"). These were also computed using "psych" (Revelle, 2024).

We should finally note that internal reliability is also a concept that exists in the item response theory (IRT) tradition. Since we provide an IRT analysis of the scales later, reliability was also estimated as part of it, and is presented in the relevant section.

Factor Analysis

In order to investigate the factor structure of the different scales, we used exploratory factor analysis (EFA), using the "psych" package for R (Revelle, 2024), and using ggplot2 (Wickham, 2009) to produce scree plots.

We used maximum likelihood estimation to extract factors. To decide on a parsimonious and appropriate number of factors to retain, we used a combination of popular techniques. We used Kaiser's K1 criterion (Kaiser, 1960), which consists in retaining factors that have eigenvalues (which represent how much of the total item variance is explained by a factor) above 1. We also used the scree test (Cattell, 1966), which consists in a visual inspection of the scree plot where we try to separate meaningful and spurious factors by detecting sharp drops in explained variance. We also used parallel analysis (Horn, 1965), a technique which consists in randomly simulating factors to use as a comparison against the factors extracted via factor analysis (factors were considered here non-spurious when their eigenvalues were above the 95% confidence intervals of the parallel analysis simulated factors). 100 simulations were used for parallel analysis. The interpretability of the factors (decided on using factor loadings after a direct oblimin rotation) was also considered in the decision to retain factors.

Because it was anticipated that each scale was essentially unidimensional, we concluded to satisfactory structural validity when a one-factor solution appeared to optimally represent the data. Once a factorial solution was decided on, we re-estimated an EFA with the number of factors to extract and inspected the factor loadings to identify potentially problematic items. A threshold of .3 (corresponding to a moderate loading) was used to detect potentially problematic items. To note, the item response theory analysis used also produces factor loadings (i.e., discrimination parameters), which are presented in the IRT section.

Item Response Theory Analysis

Item response theory (IRT) is a statistical modeling approach used to analyze the relationship between item responses and latent traits. In IRT models for Likert scale data, we obtain the probability that a person endorses each response category for each item. An item response theory analysis was performed for each scale using the R package "mirt" (Chalmers, 2012). The packages "ggplot2" (Wickham, 2009) and "jrt" (Myszkowski, 2021) were used to produce plots.

For each scale, we used various popular ordinal unidimensional models, including the graded response model (GRM; Samejima, 1969), the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992). To note the Rating Scale Model (Andrich, 1978) was also considered at first, but it requires all response categories to be observed for all items, which was not the case, and thus was discarded. Although the differences between the models are extensively explained elsewhere (e.g., De Ayala, 2022; Nering & Ostini, 2010), we shall note that the models make slightly different assumptions. Notably, the PCM assumes all items to have the same discrimination (i.e., same loading), while the GPCM and GRM do not. The GPCM and GRM both allow items to vary on the response category structure (i.e., the distance between the response categories), discrimination (i.e., the strength of the relation between the latent trait measured and the item response) and location (i.e., how easy an item is to agree to). They are however distinct from one another in that the GPCM predicts the probability to choose between adjacent response categories (here, 2 vs. 1, 3 vs. 2, 4 vs. 3 and 5 vs. 4), while the GRM predicts the probability to choose a response category or above (here, 2 or above, 3 or above, 4 or above, and 5).

Since the models tested are not all nested, using likelihood ratio tests to compare them was not possible, and therefore we used information criteria – more specifically, the Akaike information criterion (AIC; Akaike, 1978) and the Bayesian information criterion (BIC; Schwarz, 1978) – to decide on which model to retain.

After having selected the best fitting model, we assessed its fit by looking at the item fit statistics of the model. Item fit statistics yield significance tests that indicate if an item if significantly misfitted by the model. We used the signed chi-squared (S-X2) statistic (Orlando & Thissen, 2000), which is one of the most commonly used measures of item fit, and the default in "mirt". The p-values were Bonferroni-adjusted in order to decide on overall model fit from the multiple item fit statistics. P-values all above .05 were used as an indicator that the IRT model used fitted the data sufficiently well. Since the IRT models used were all unidimensional, this was also interpreted as a sign of good structural validity.

Item response functions were inspected using item response category curves plots. In addition, we computed for each scale the empirical reliability (i.e., the average reliability observed in the sample) and expected reliability (i.e., the average reliability expected for a standard normal distribution of latent person locations), using a threshold of .7 to decide on sufficient reliability. Finally, since the instrument is currently used with sum scores, we compared factor scores from the model and sum scores using a Pearson correlation coefficient, expecting a strong positive correlation.

Results

Reliability Analysis

Internal consistency was overall satisfactory, with Cronbach's alpha estimates of .71 for Water, .74 for Wood, .83 for Fire, .70 for Earth, and .72 for Metal.

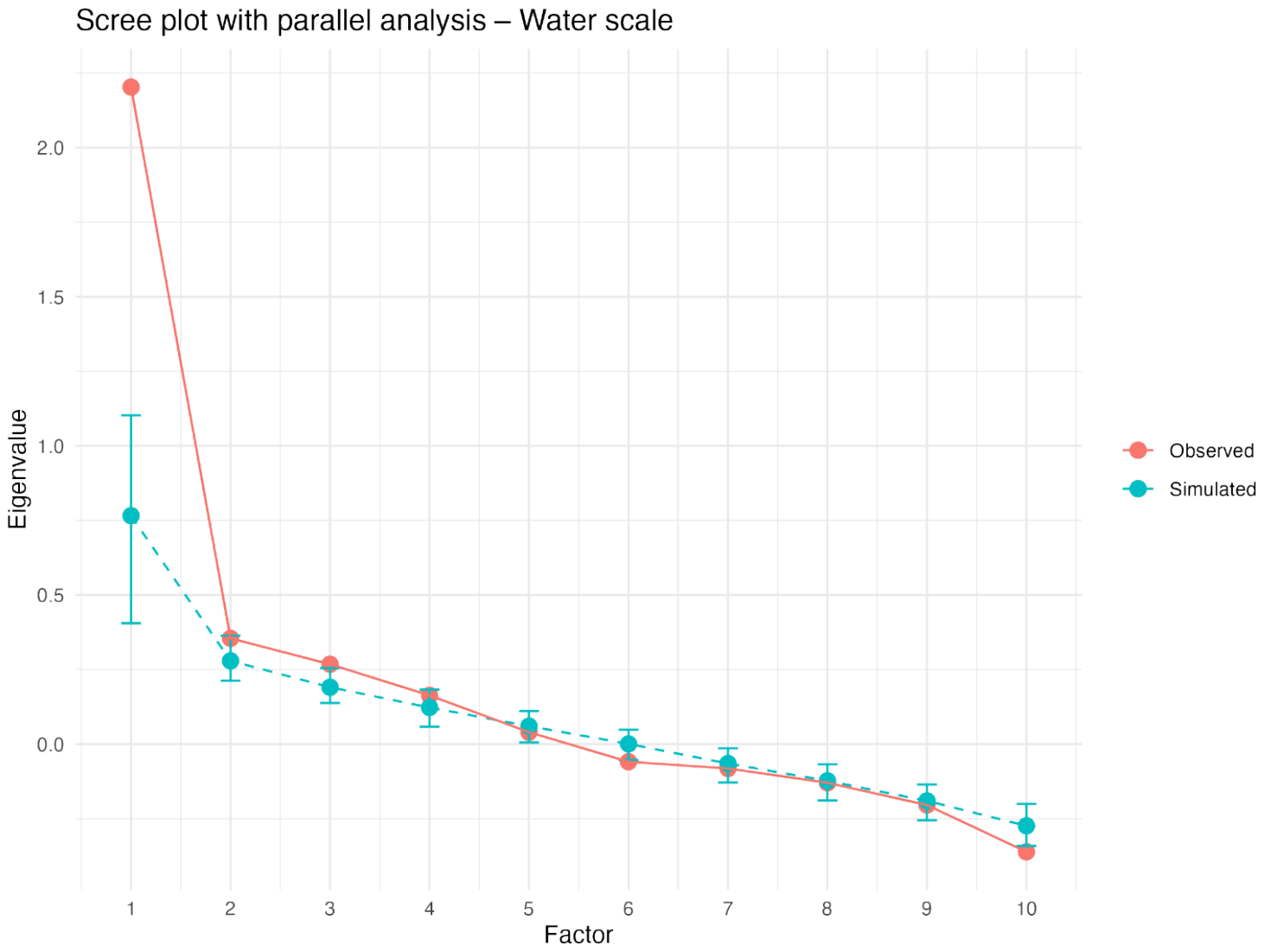
Regarding unidimensionality (investigated using McDonald's omega), similar results were observed, as we found satisfactory McDonald's omega estimates for Water (.81), Wood (.80), Fire (.88), Earth (.78) and Metal (.80).

Overall, we can conclude from this reliability analysis that the scales present satisfactory reliability.

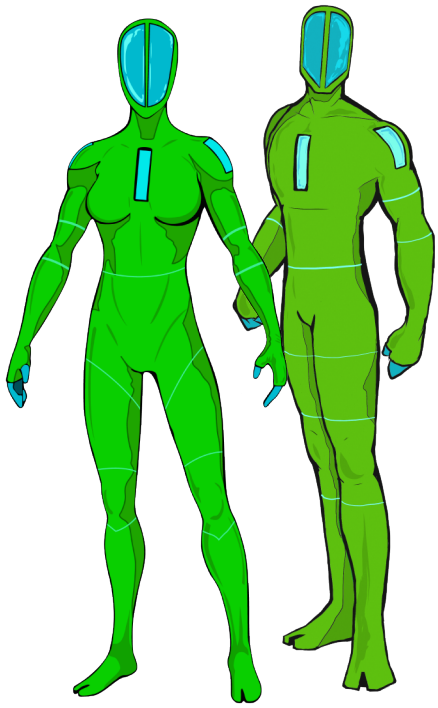
Factor Analysis

WATER

For the Water scale, the exploratory factor analysis indicated to retain 1 factor per Kaiser's K1 criterion, per Cattell's scree test (scree plot below) and per the parallel analysis. As a consequence, we concluded to a unidimensional structure for this scale. In line with previously however, the 3 items (2, 7 and 10) that were identified as problematic had negative loadings, which suggests that they should be replaced or removed.

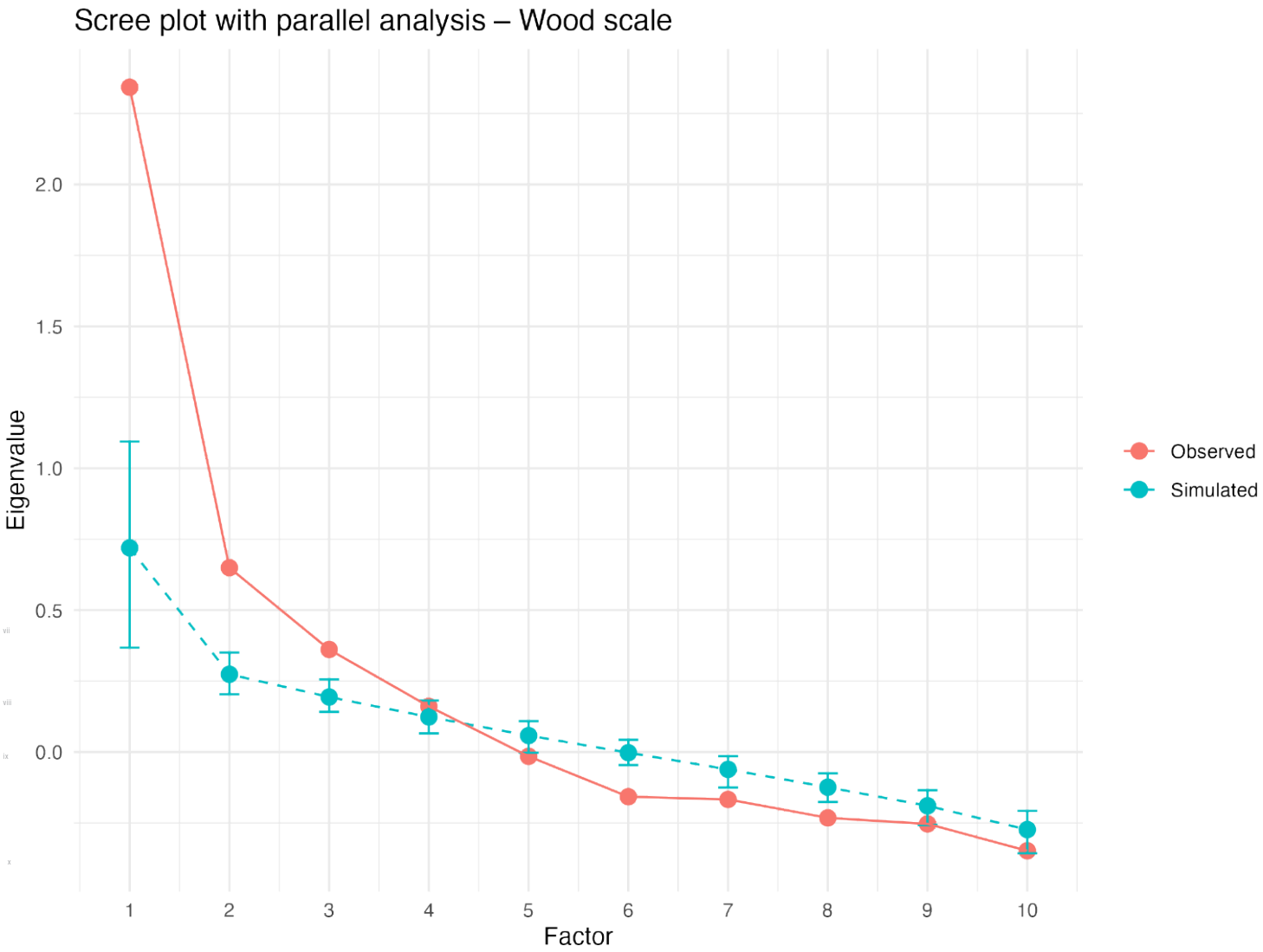


Factor Analysis



WOOD

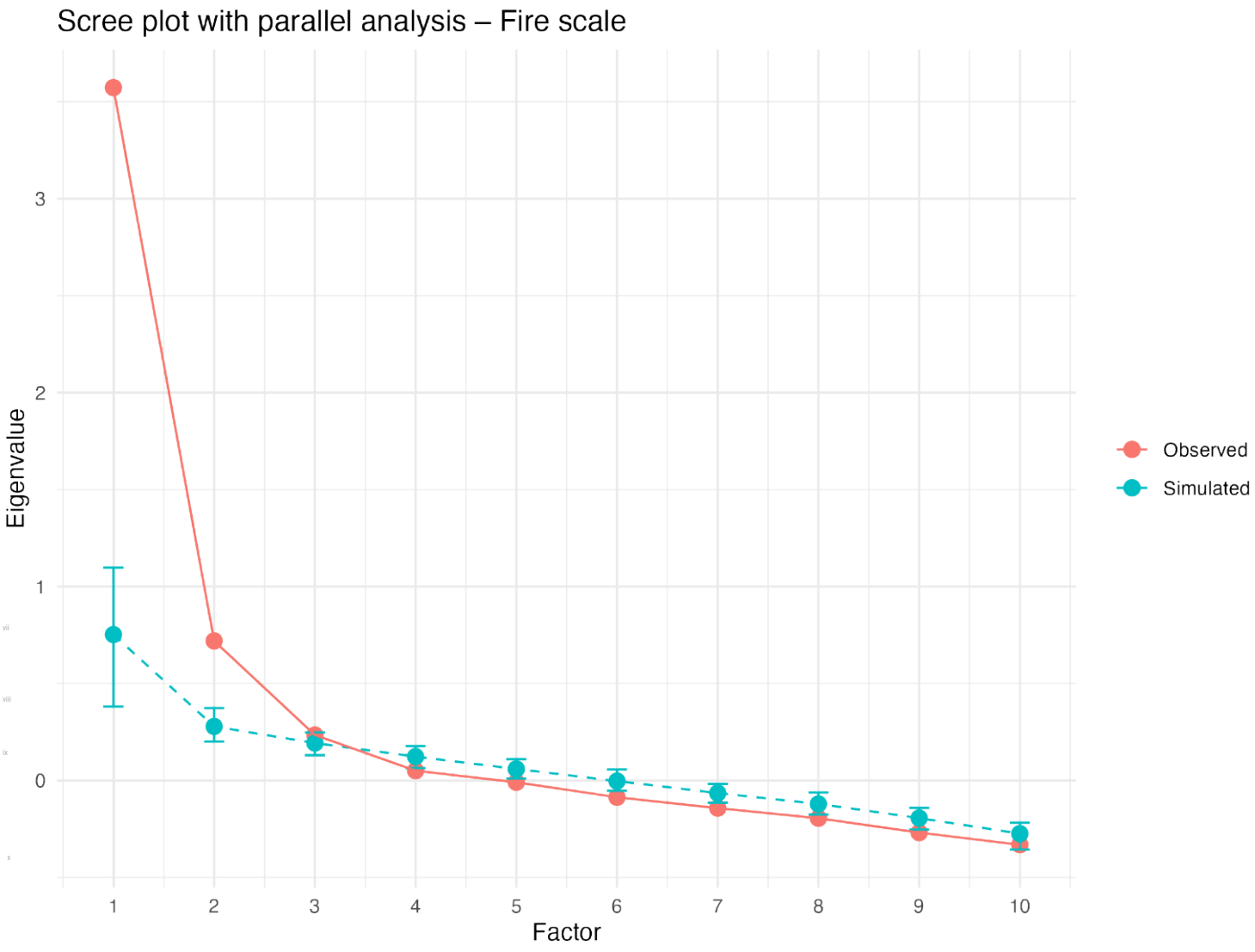
For the Wood scale, the exploratory factor analysis indicated to retain 1 factor per Kaiser's K1 criterion and per Cattell's scree test (scree plot below), while the parallel analysis suggested to retain 3 factors. The 3-factor solution was not clearly interpretable, so we concluded that the scale was essentially unidimensional. In the 1-factor solution, all items had positive loadings that were at least moderate (>.3).

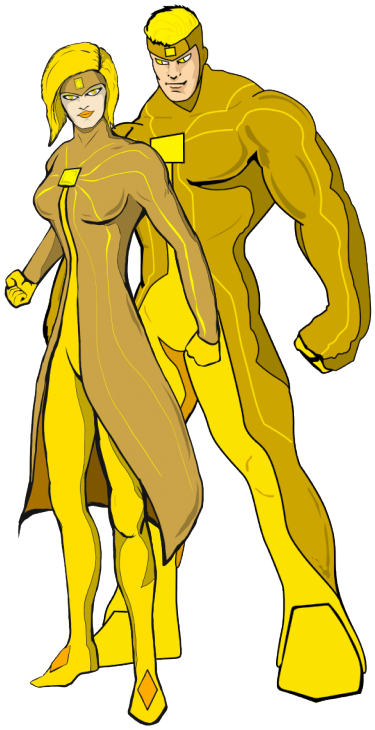


Factor Analysis

FIRE

For the Fire scale, the exploratory factor analysis indicated to retain 1 factor per Kaiser's K1 criterion and per Cattell's scree test (scree plot below), while the parallel analysis suggested to retain 3 factors. The 3-factor solution was not clearly interpretable, so we concluded that the scale was essentially unidimensional. In the 1-factor solution, all items had positive loadings that were at least moderate (>.3), except the last item (loading = .22).

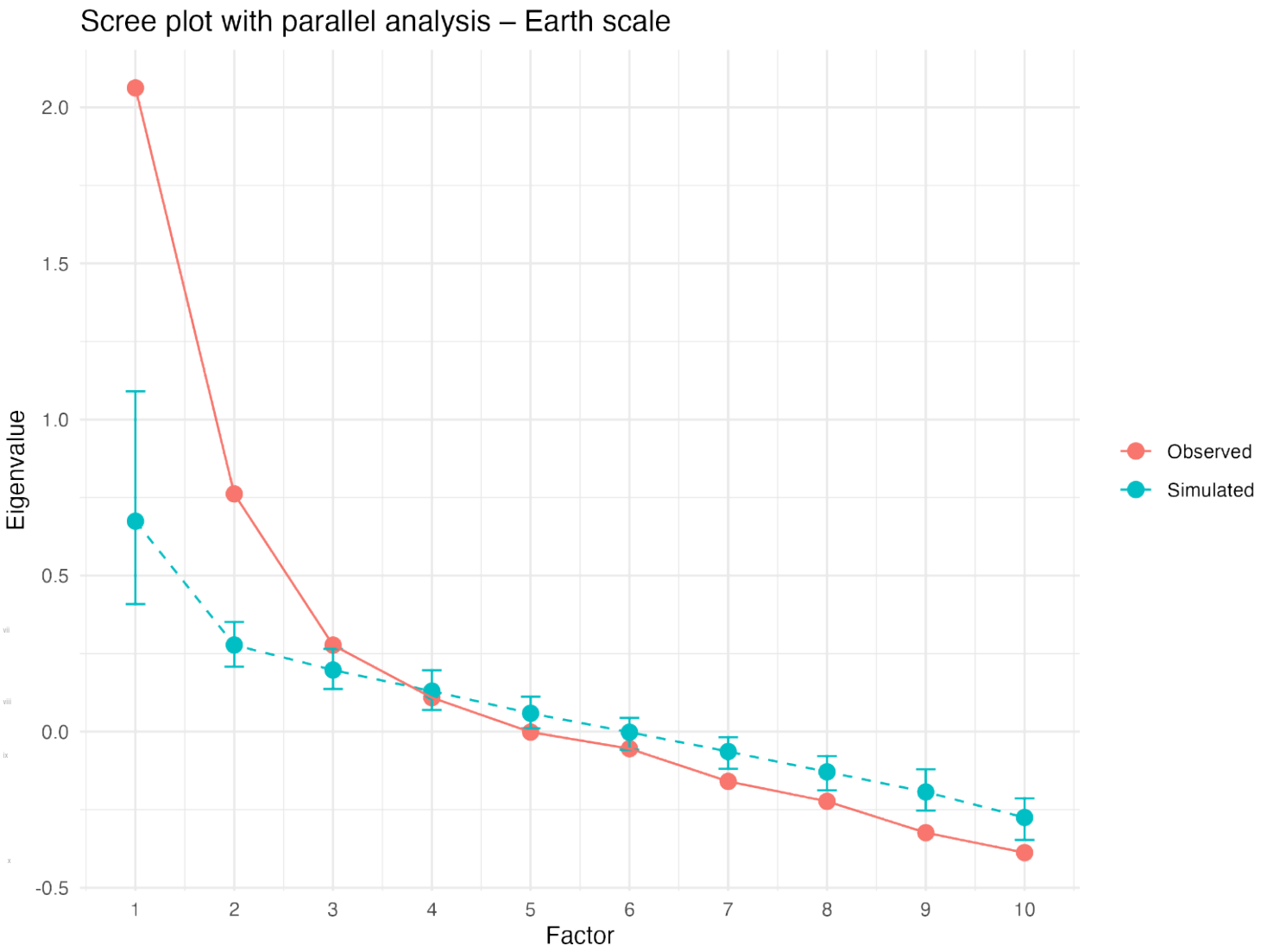


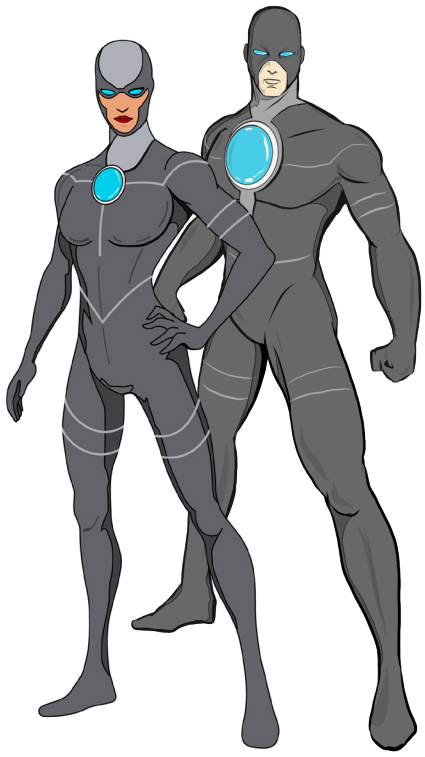


Factor Analysis

EARTH

For the Earth scale, the exploratory factor analysis indicated to retain 1 factor per Kaiser's K1 criterion and per Cattell's scree test (scree plot below), while the parallel analysis suggested to retain 3 factors. The 3-factor solution was not clearly interpretable, so we concluded that the scale was essentially unidimensional. In the 1-factor solution, all items had positive loadings that were at least moderate (>.3), except for item 7 (loading = .29).

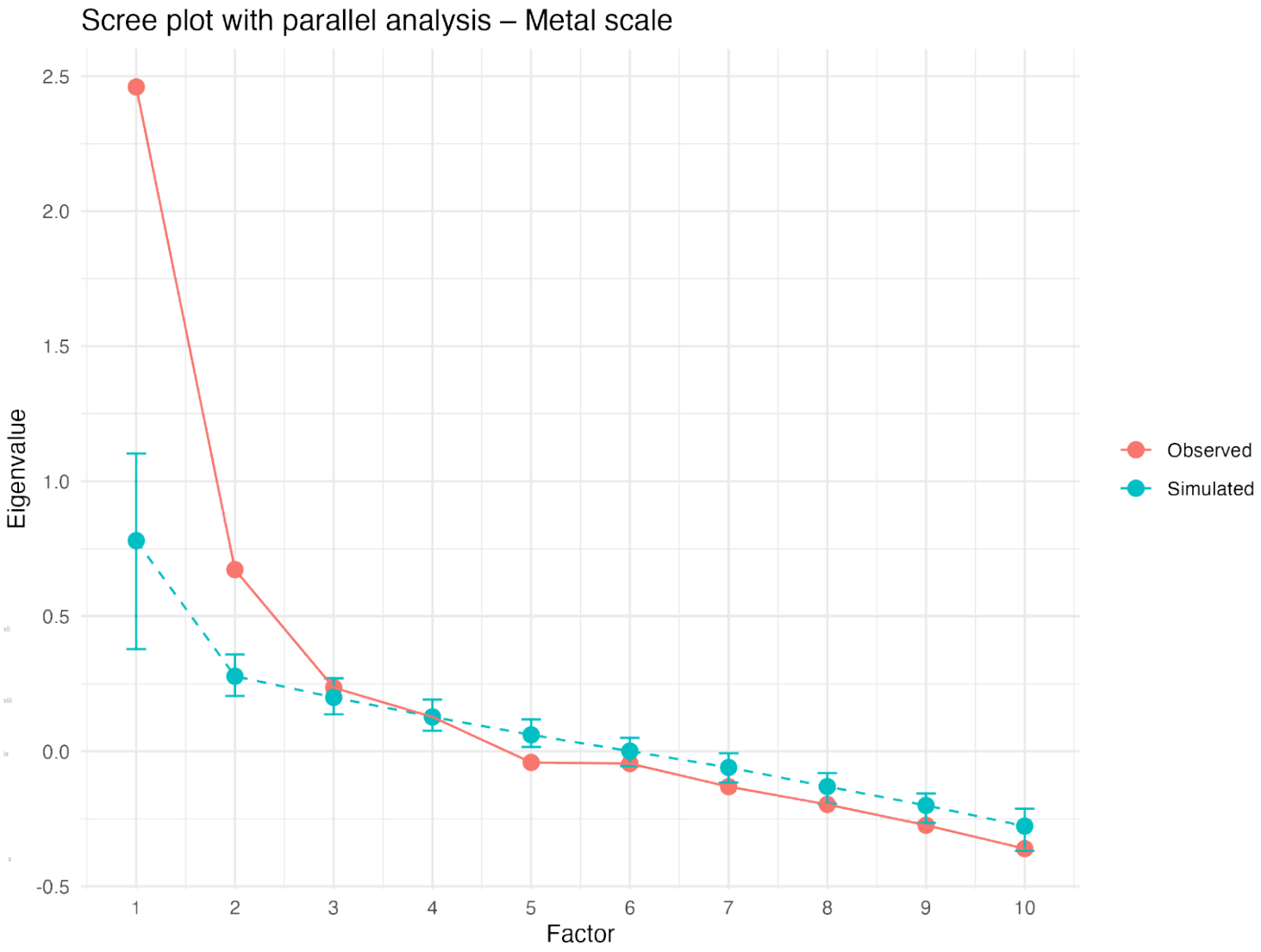




Factor Analysis

METAL

For the Metal scale, the exploratory factor analysis indicated to retain 1 factor per Kaiser's K1 criterion and per Cattell's scree test (scree plot below), while the parallel analysis suggested to retain 2 factors. The 2-factor solution was not clearly interpretable, so we concluded that the scale was essentially unidimensional. In the 1-factor solution, all items had positive loadings that were at least moderate (>.3), except for items 4 (loading = .29) and 8 (loading = .02). We would recommend replacing or removing item 8.

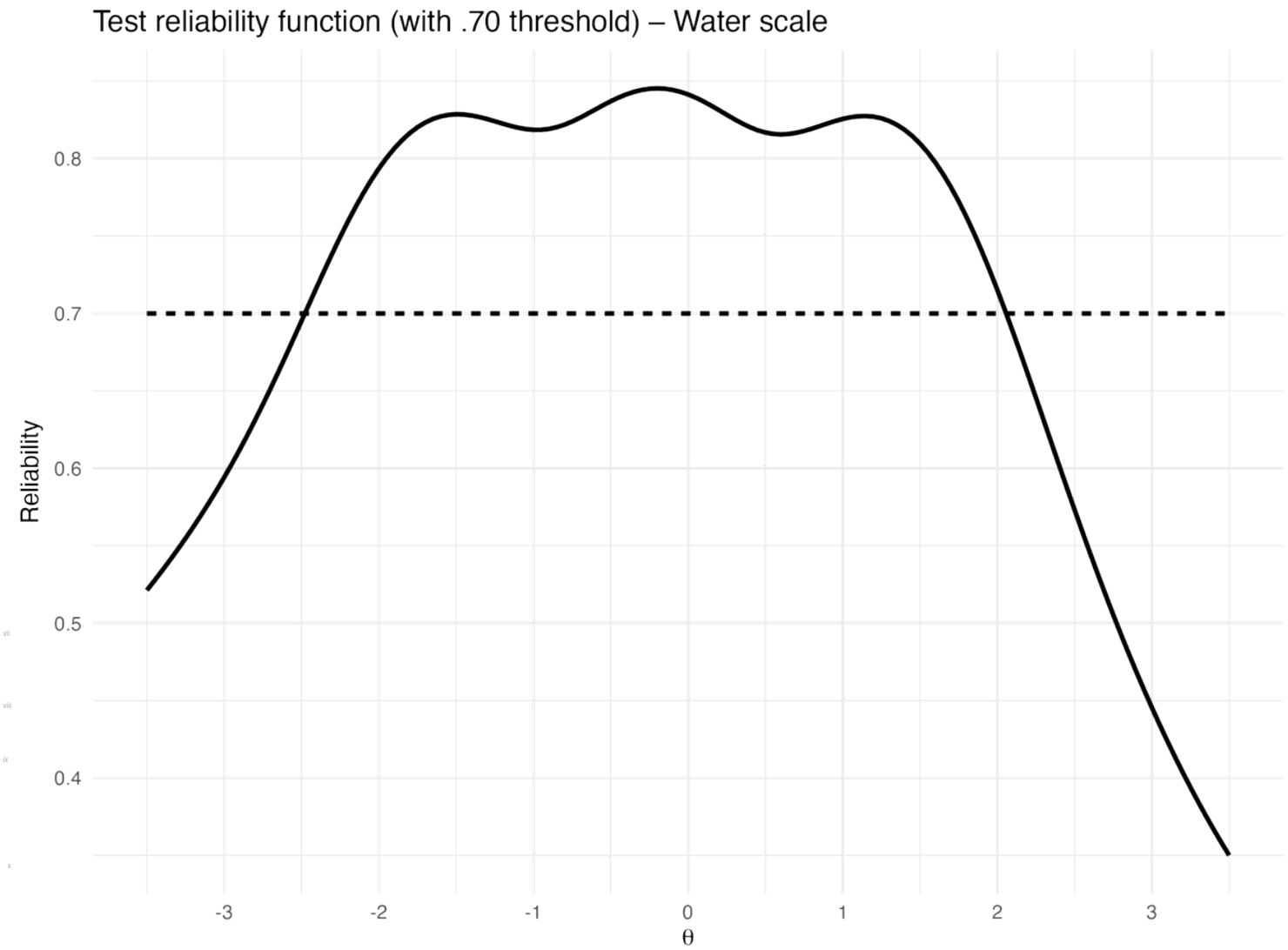


Item Response Theory Analysis

WATER

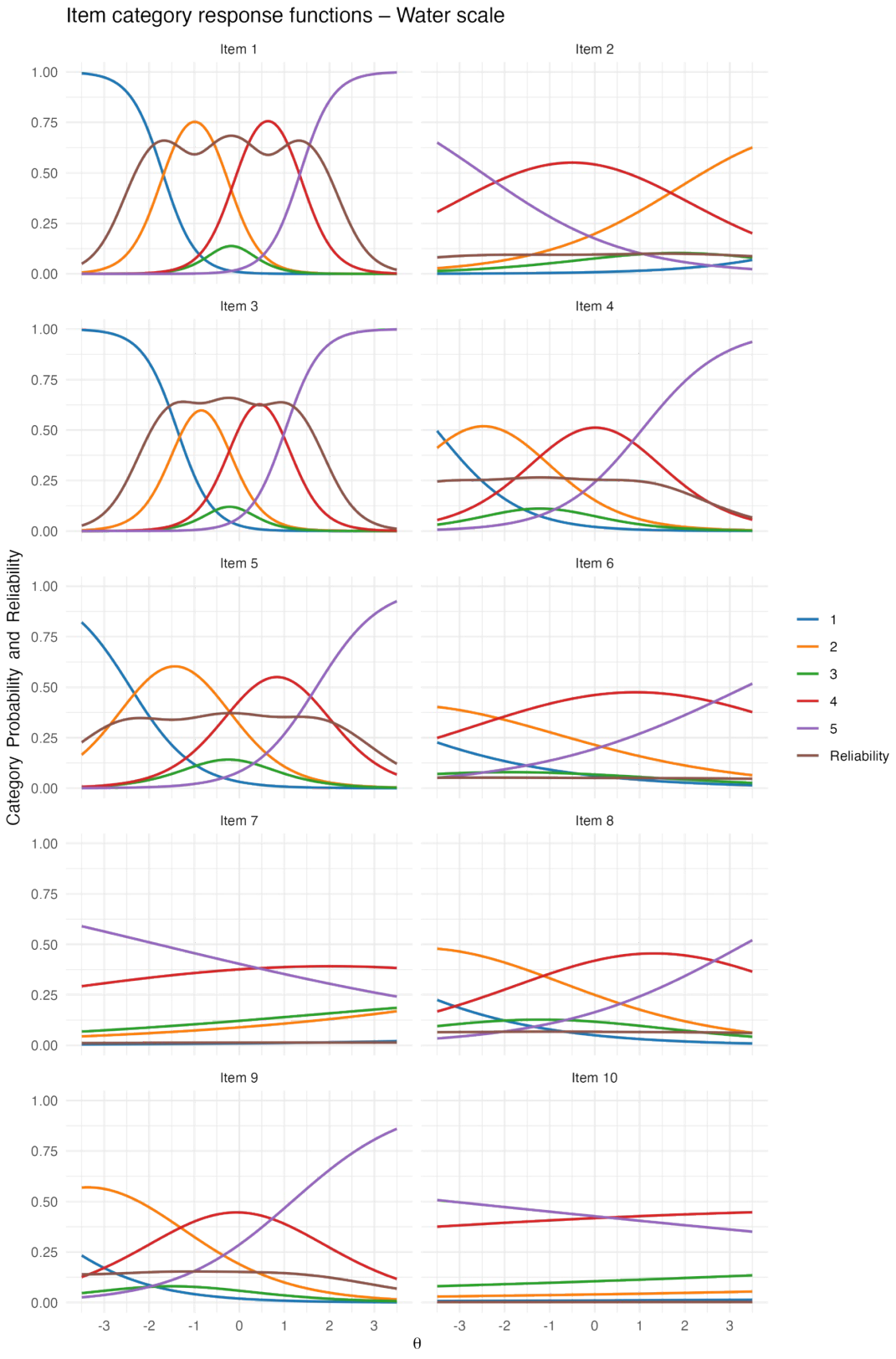
For the Water scale, the graded response model (GRM) outperformed the partial credit and generalized partial credit model according to both the AIC and BIC. Consequently, we here interpreted the GRM. Overall, the model showed good fit, with only 1 item (item 4) significantly misfitted (Bonferroni-corrected p-value = .048) – interestingly, this is not one of the items that was previously flagged as problematic. This suggests good structural validity, since the test was hypothetically unidimensional, and the model tested, which had good fit, was also unidimensional.

The empirical (i.e. observed) reliability of the scores per the GRM was .82, while the expected reliability for a standard normal prior distribution was .82, indicating satisfactory reliability. The reliability function is presented below, and indicates that the scale had good reliability at most levels of the latent trait.



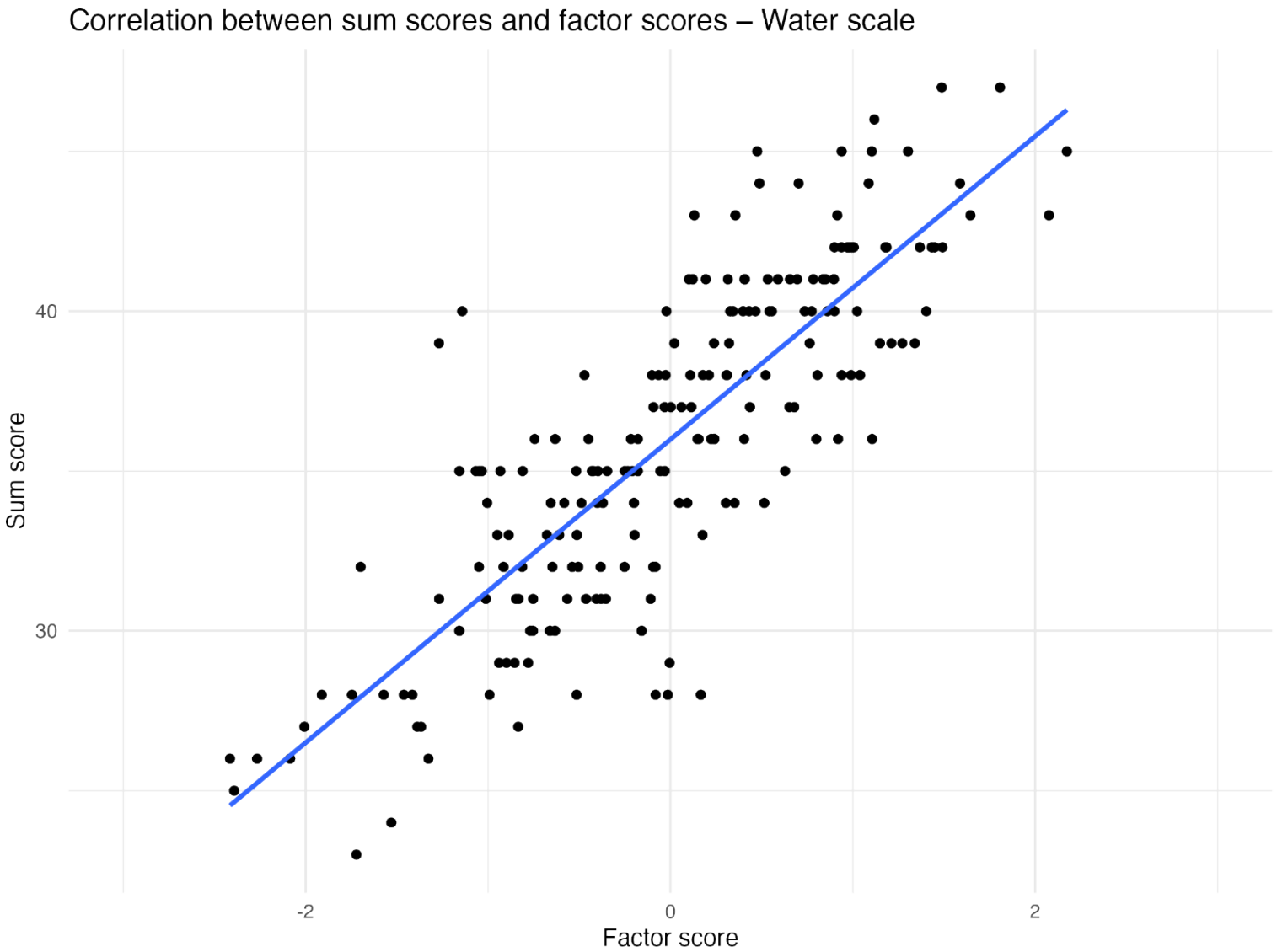
WATER

IRT loadings were similar to those observed in the EFA. Notably, the 3 previously identified items (2, 7, and 10) had negative loadings. The other loadings ranged between .24 and .85, which is satisfactory. Item category response curves, along with item reliability functions, are presented below.



WATER

Finally, the IRT factor scores correlated at .83 with the sum scores (see scatterplot below), indicating that sum scores can be used as good proxies (this will likely increase if items with negative loadings are removed or replaced in the future)

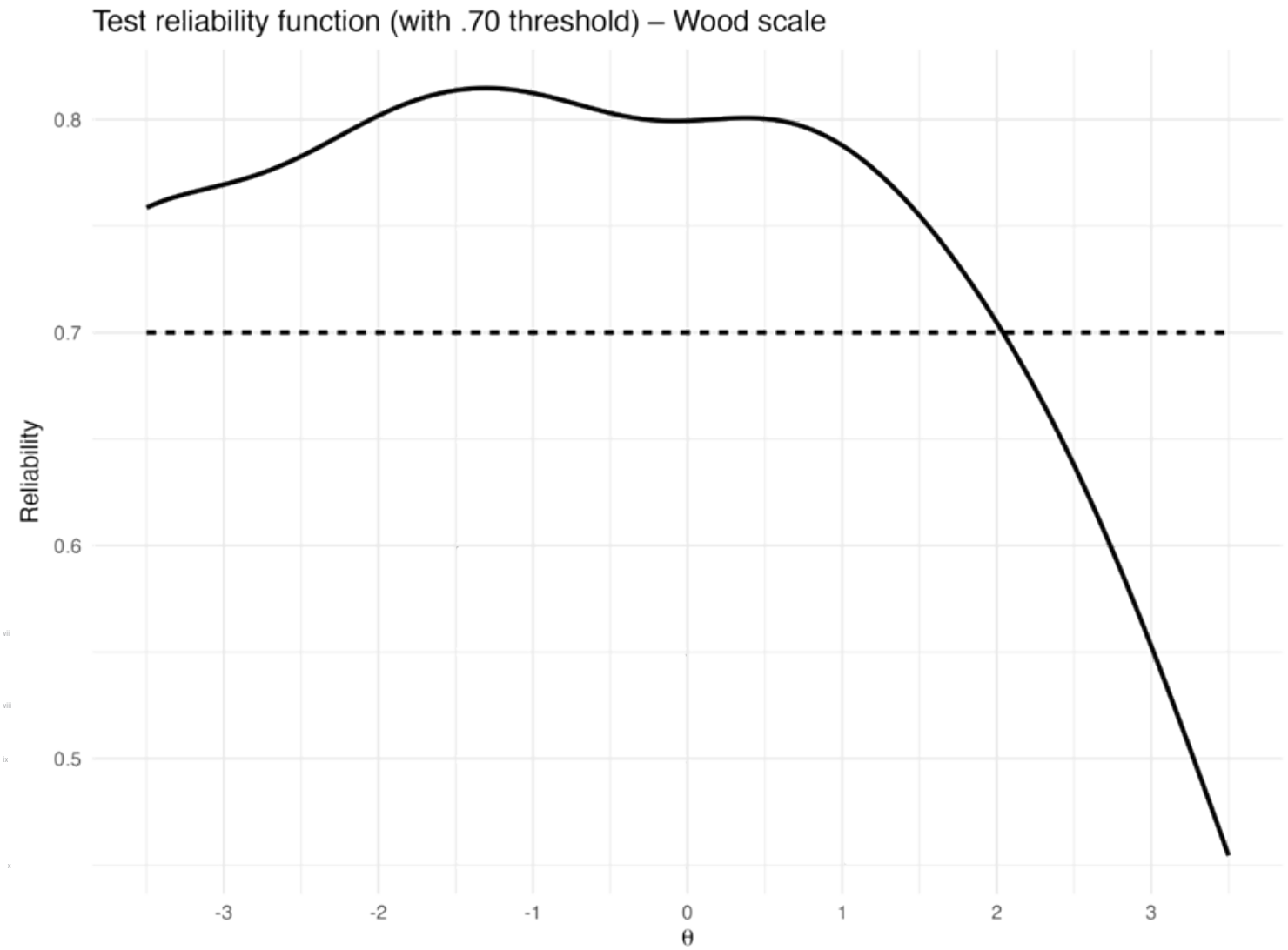


Item Response Theory Analysis

WOOD

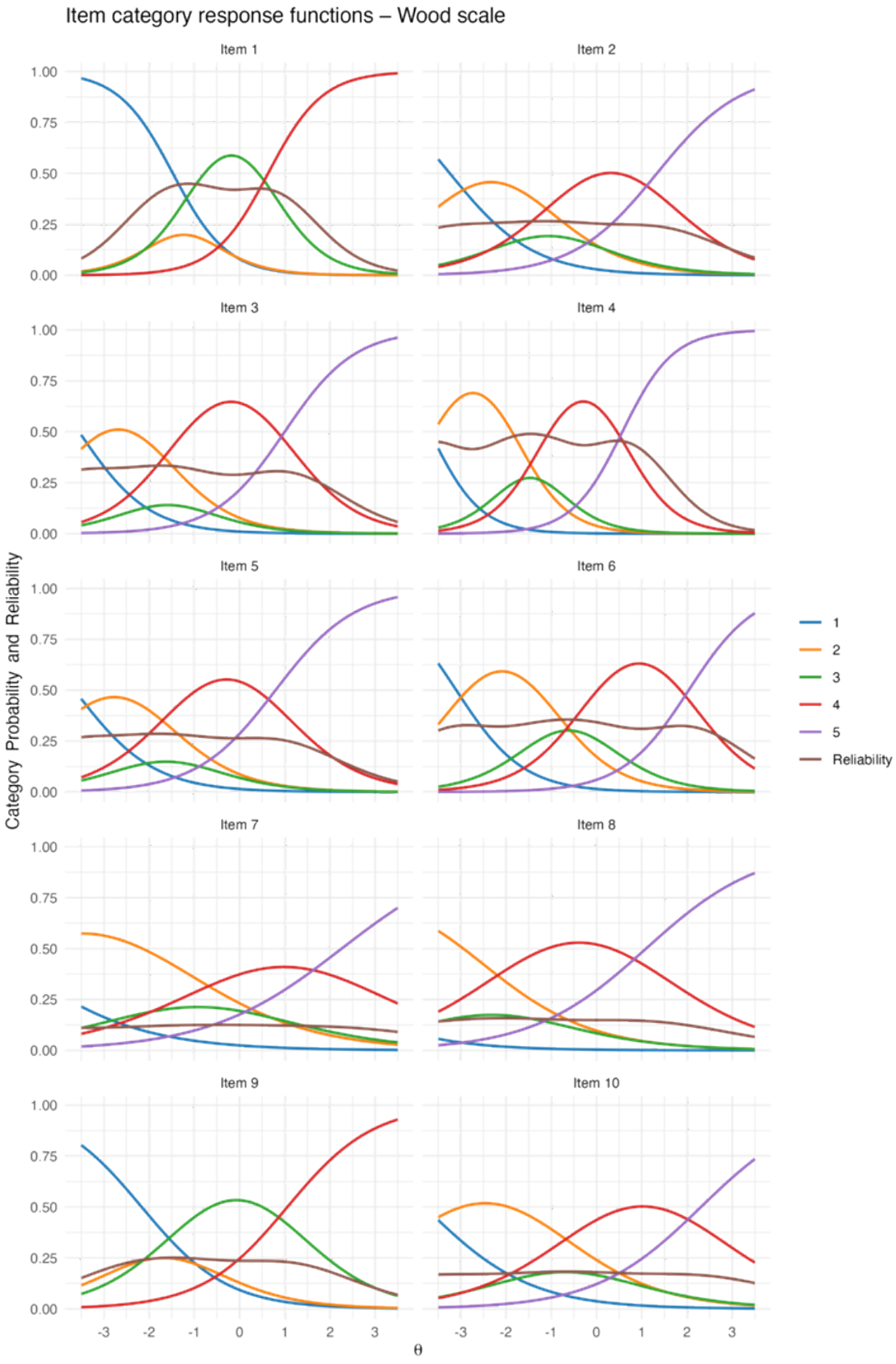
For the Wood scale, the graded response model (GRM) outperformed the partial credit and generalized partial credit model according to both the AIC and BIC. Consequently, we here interpreted the GRM. Overall, the model showed good fit, with no item significantly misfitted per Bonferroni-corrected p-values. This suggests good structural validity, since the test was hypothetically unidimensional, and the model tested, which had good fit, was also unidimensional.

The empirical (i.e. observed) reliability of the scores per the GRM was .79, while the expected reliability for a standard normal prior distribution was .79, indicating satisfactory reliability. The reliability function is presented below and indicates that the scale had good reliability at most levels of the latent trait.



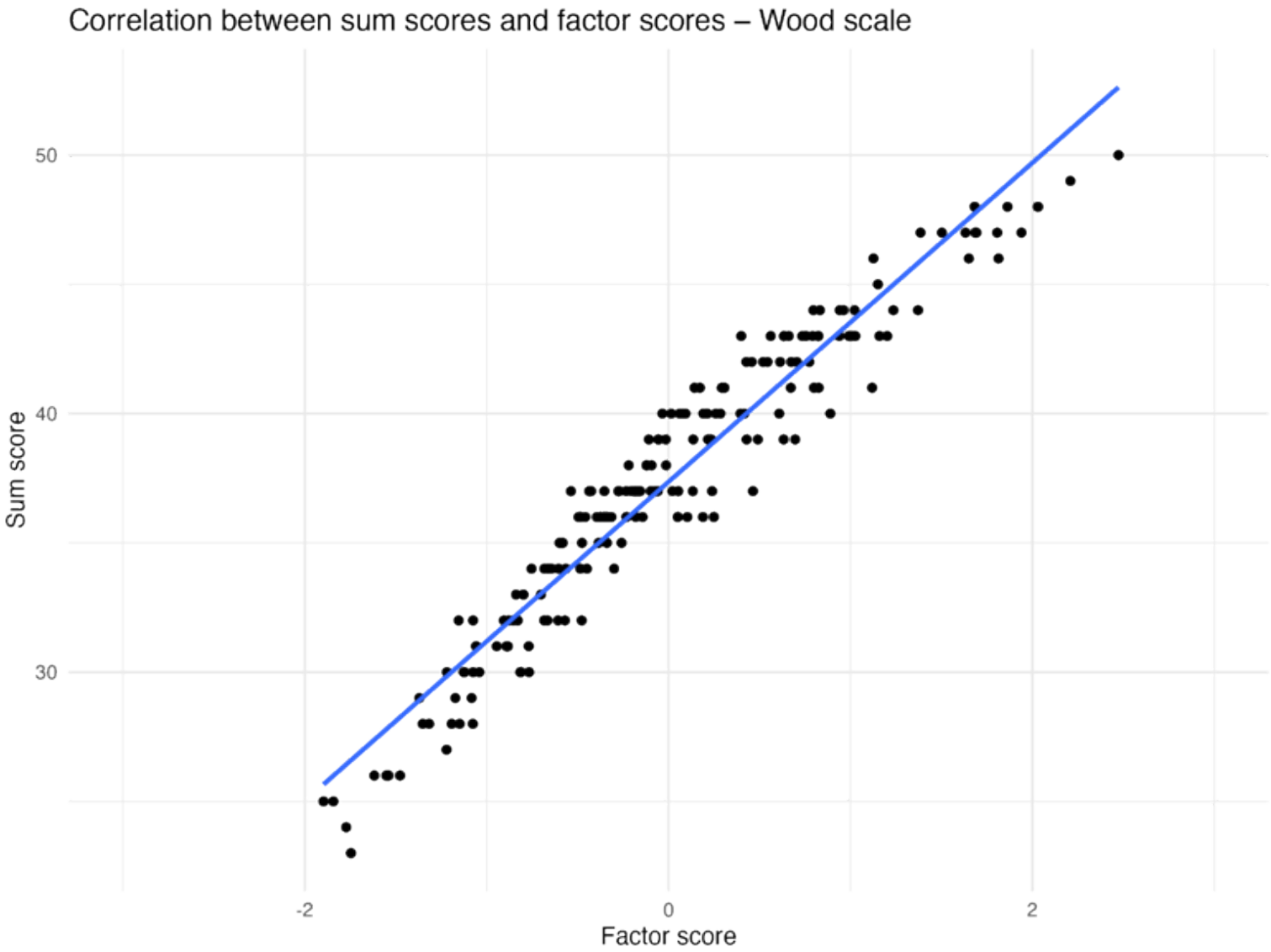
WOOD

IRT loadings were comparable to those observed in the EFA. With all loadings above .3. Item category response curves, along with item reliability functions, are presented below.



WOOD

Finally, the IRT factor scores correlated at .96 with the sum scores (see scatterplot below), indicating that sum scores can be used as good proxies.

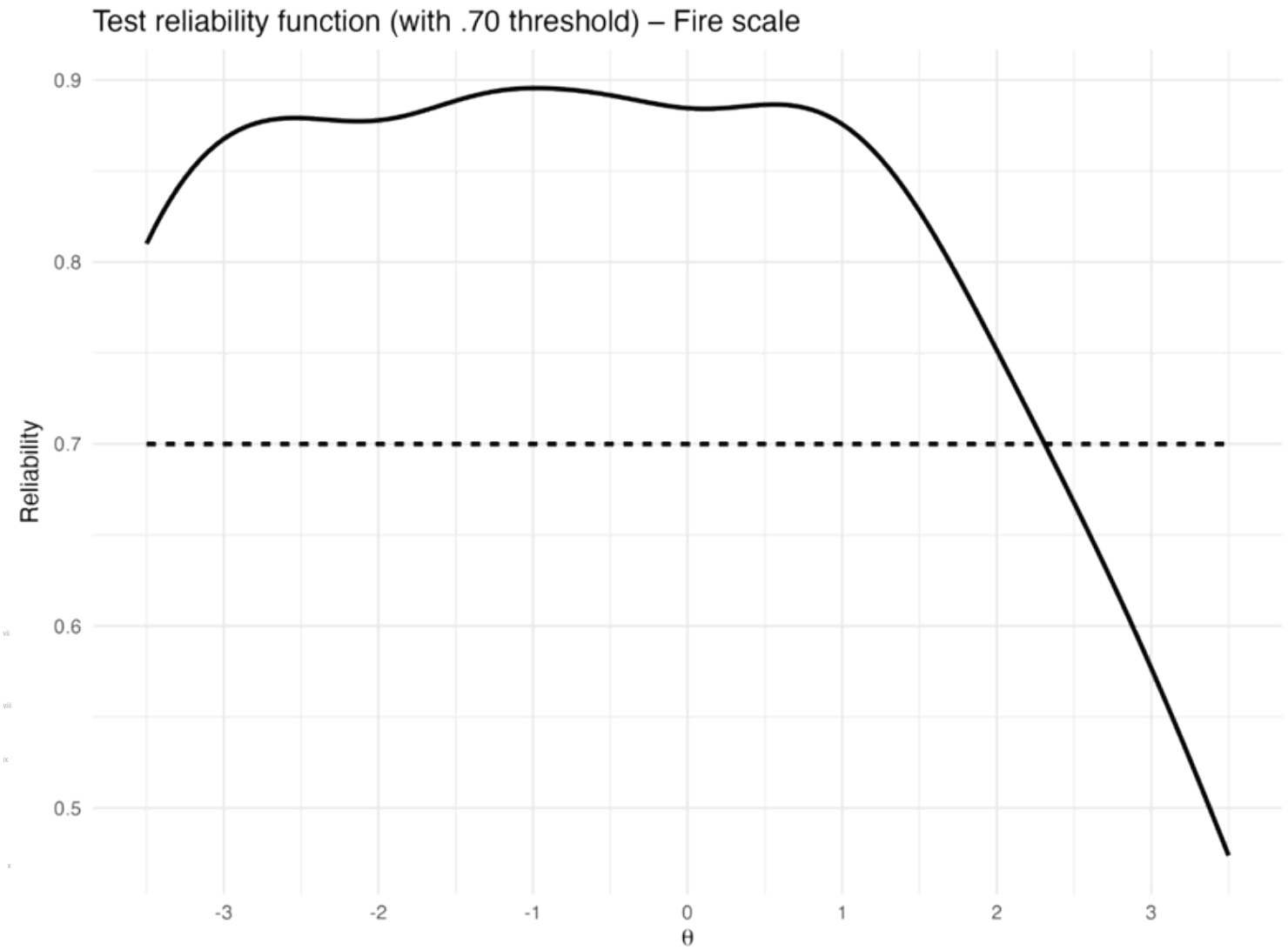


Item Response Theory Analysis

FIRE

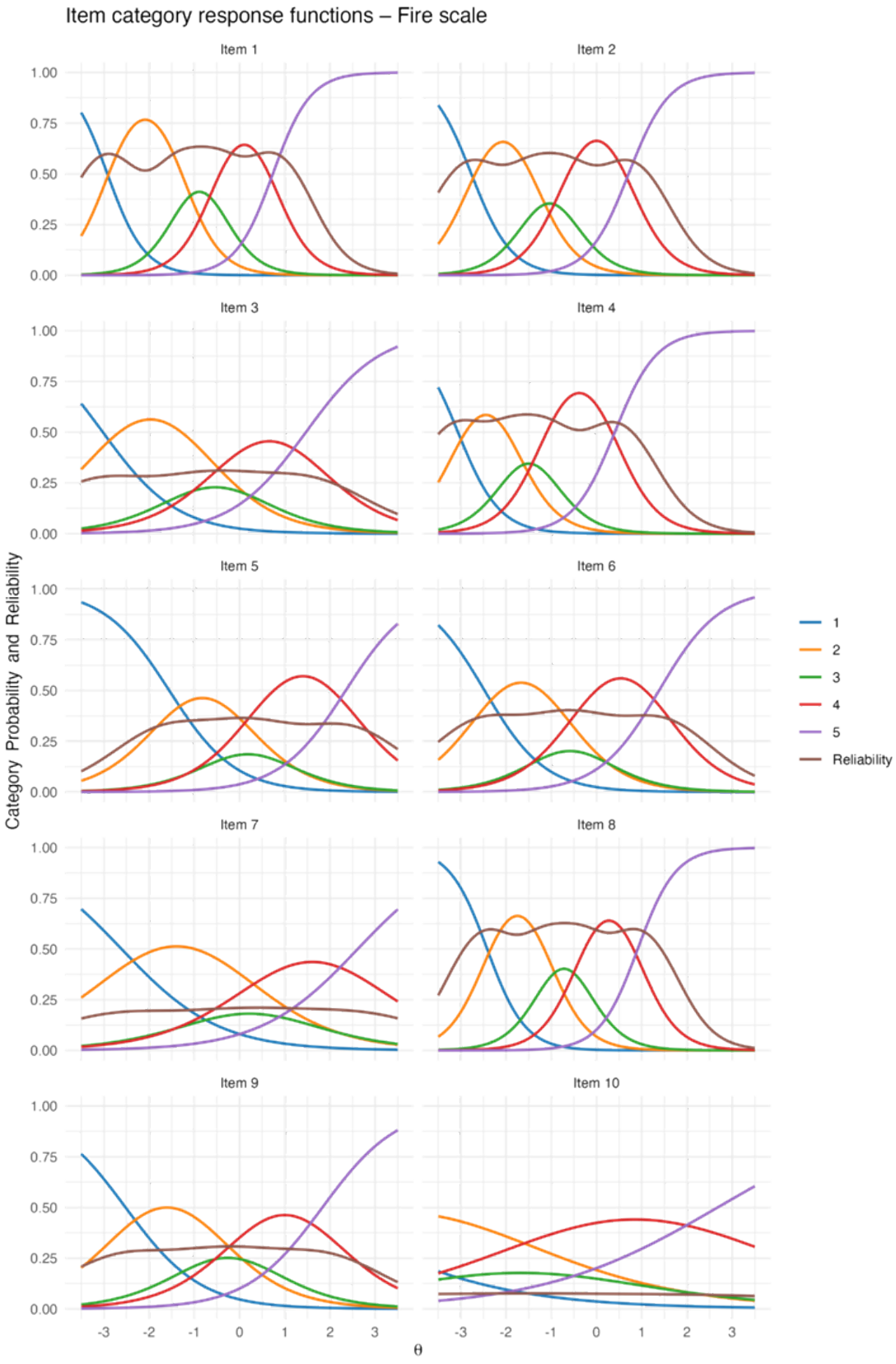
For the Fire scale, the graded response model (GRM) outperformed the partial credit and generalized partial credit model according to both the AIC and BIC. Consequently, we here interpreted the GRM. Overall, the model showed good fit, with no item significantly misfitted per Bonferroni-corrected p-values. This suggests good structural validity, since the test was hypothetically unidimensional, and the model tested, which had good fit, was also unidimensional.

The empirical (i.e. observed) reliability of the scores per the GRM was .88, while the expected reliability for a standard normal prior distribution was .88, indicating satisfactory reliability. The reliability function is presented below, and indicates that the scale had good reliability at most levels of the latent trait.



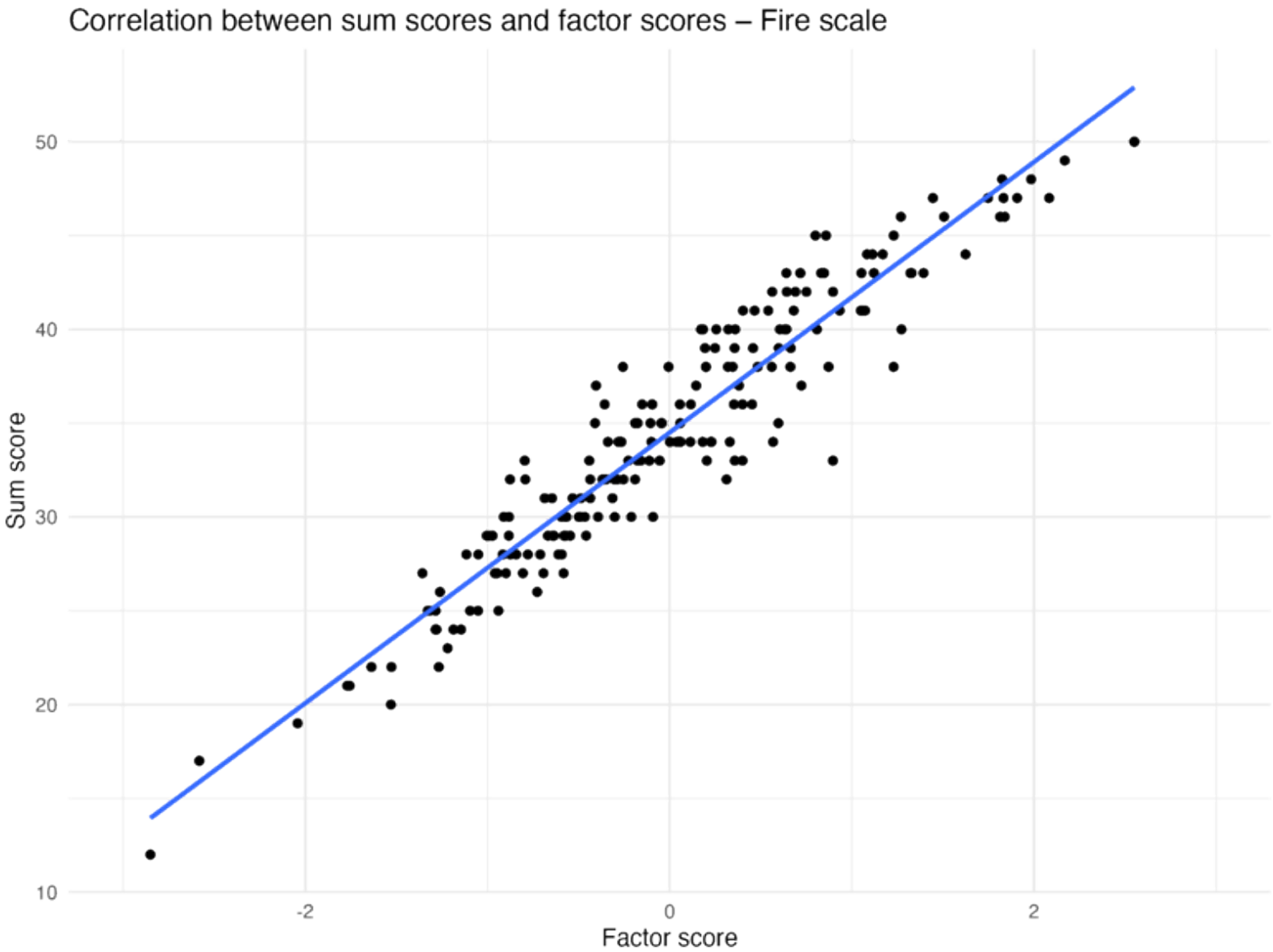
FIRE

IRT loadings were similar to those observed in the EFA. All loadings were above .3, except for the last item (loading = .29), which is satisfactory. Item category response curves, along with item reliability functions, are presented below.



FIRE

Finally, the IRT factor scores correlated at .95 with the sum scores (see scatterplot below), indicating that sum scores can be used as good proxies.

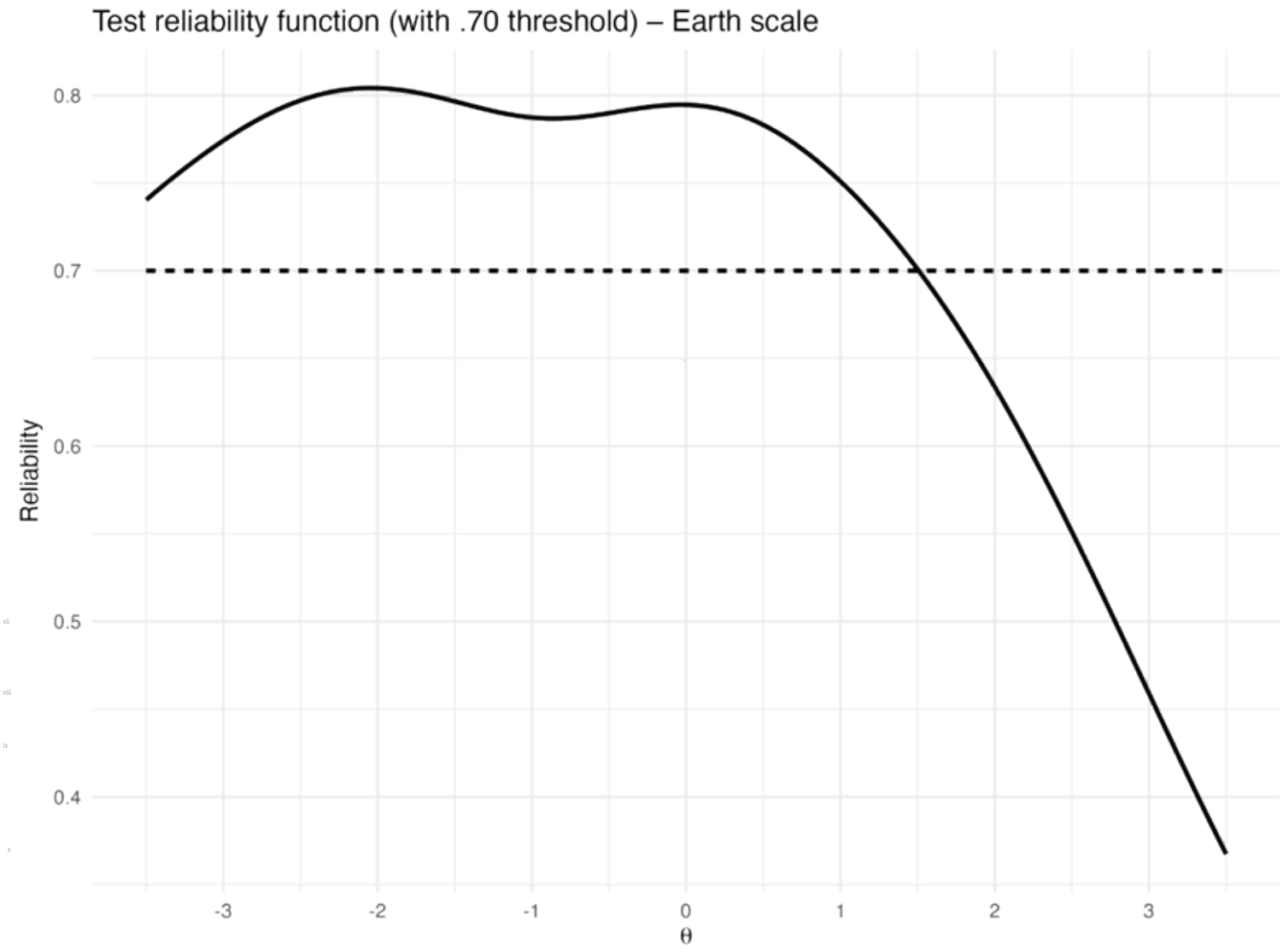


Item Response Theory Analysis

EARTH

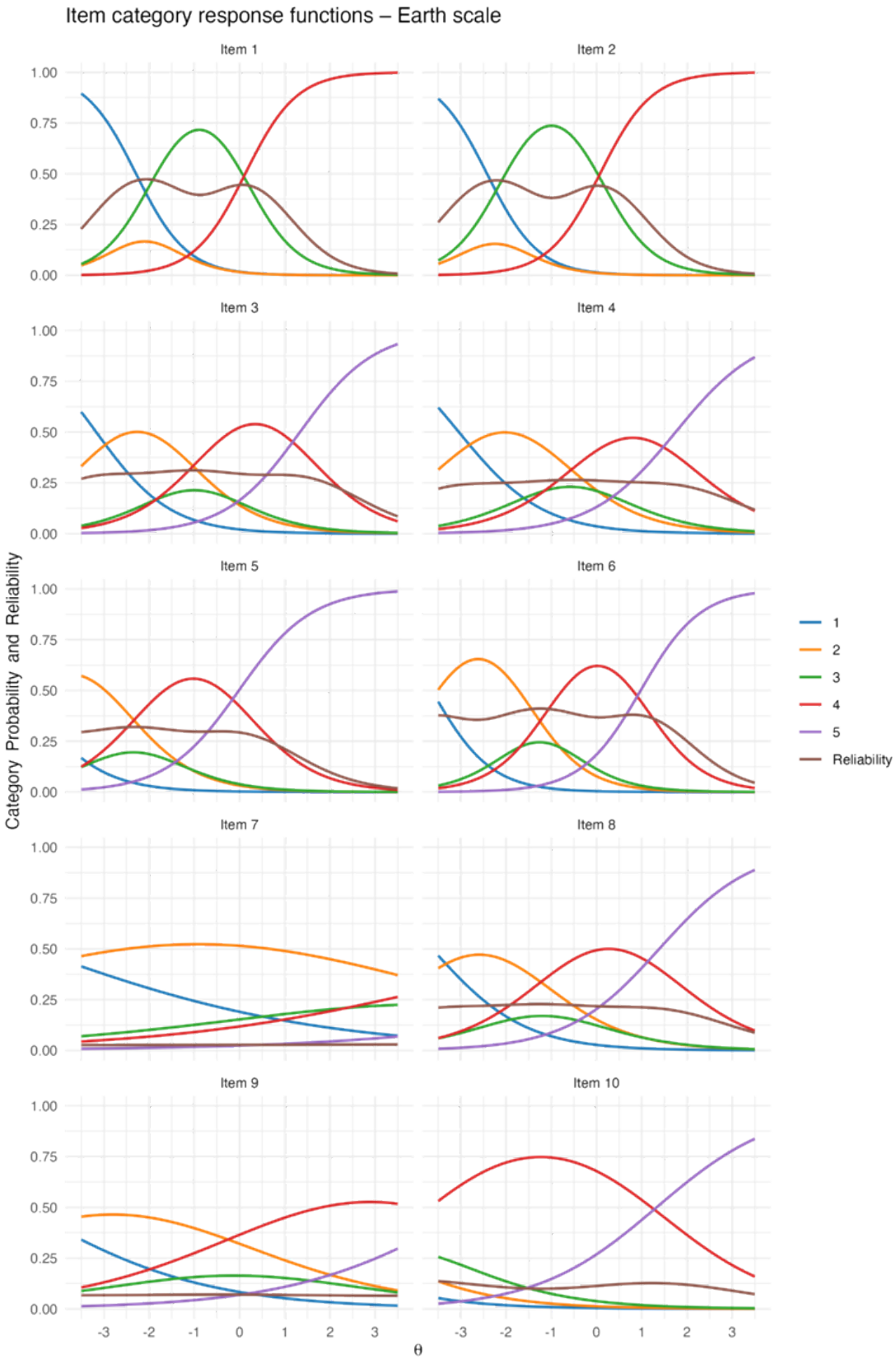
For the Earth scale, the graded response model (GRM) outperformed the partial credit and generalized partial credit model according to both the AIC and BIC. Consequently, we here interpreted the GRM. Overall, the model showed good fit, with no item significantly misfitted per Bonferroni-corrected p-values. This suggests good structural validity, since the test was hypothetically unidimensional, and the model tested, which had good fit, was also unidimensional.

The empirical (i.e. observed) reliability of the scores per the GRM was .77, while the expected reliability for a standard normal prior distribution was .77, indicating satisfactory reliability. The reliability function is presented below, and indicates that the scale had good reliability at most levels of the latent trait.



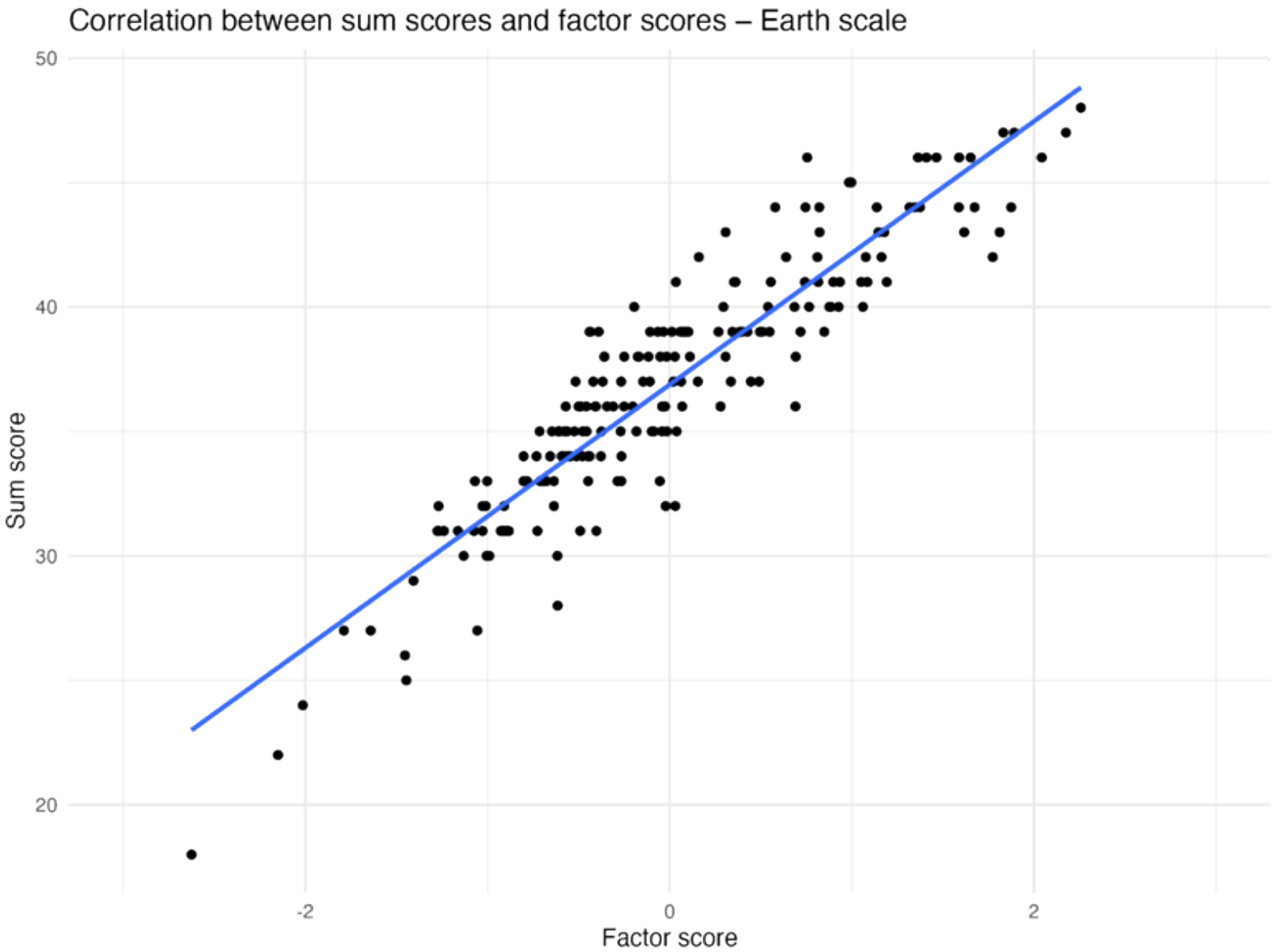
EARTH

IRT loadings were similar to those observed in the EFA. All loadings were above .3, except for items 7 and 9 (with loadings respectively of .18 and .28), which is satisfactory. Item category response curves, along with item reliability functions, are presented below.



EARTH

Finally, the IRT factor scores correlated at .92 with the sum scores (see scatterplot below), indicating that sum scores can be used as good proxies.

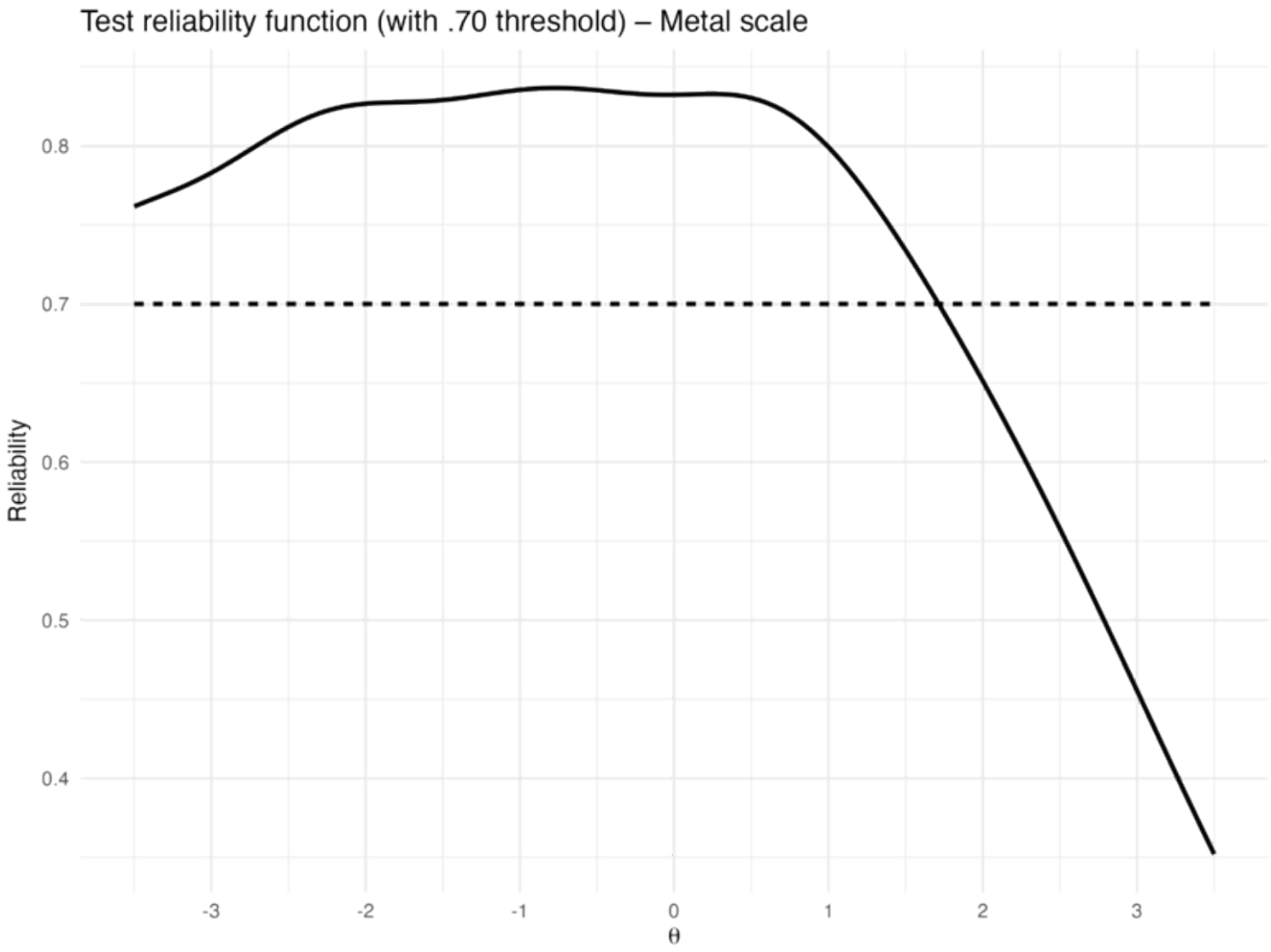


Item Response Theory Analysis

METAL

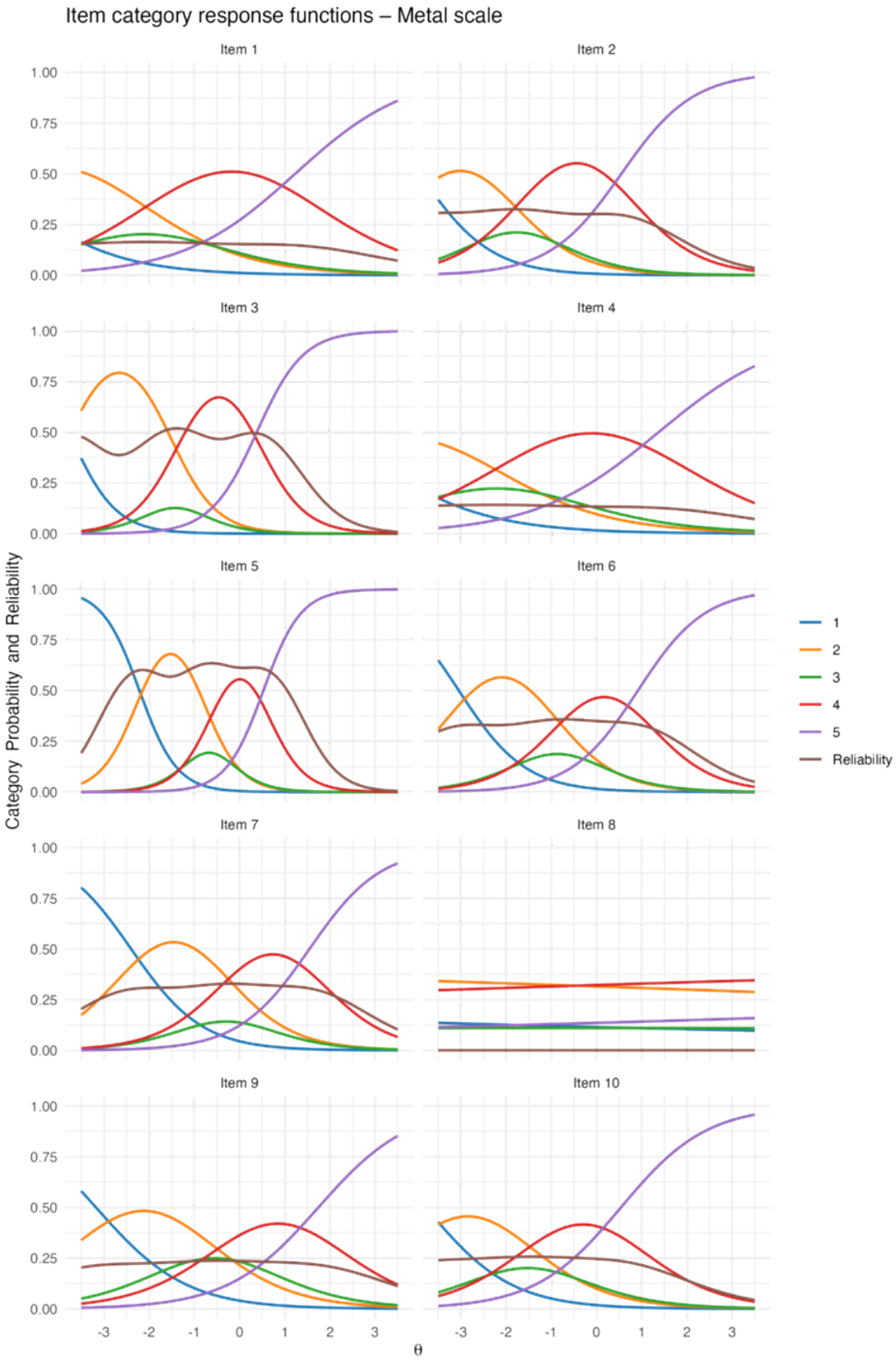
For the Metal scale, the graded response model (GRM) outperformed the partial credit and generalized partial credit model according to both the AIC and BIC. Consequently, we here interpreted the GRM. Overall, the model showed good fit, with no item significantly misfitted per Bonferroni-corrected p-values. This suggests good structural validity, since the test was hypothetically unidimensional, and the model tested, which had good fit, was also unidimensional.

The empirical (i.e. observed) reliability of the scores per the GRM was .81, while the expected reliability for a standard normal prior distribution was .81, indicating satisfactory reliability. The reliability function is presented below and indicates that the scale had good reliability at most levels of the latent trait.



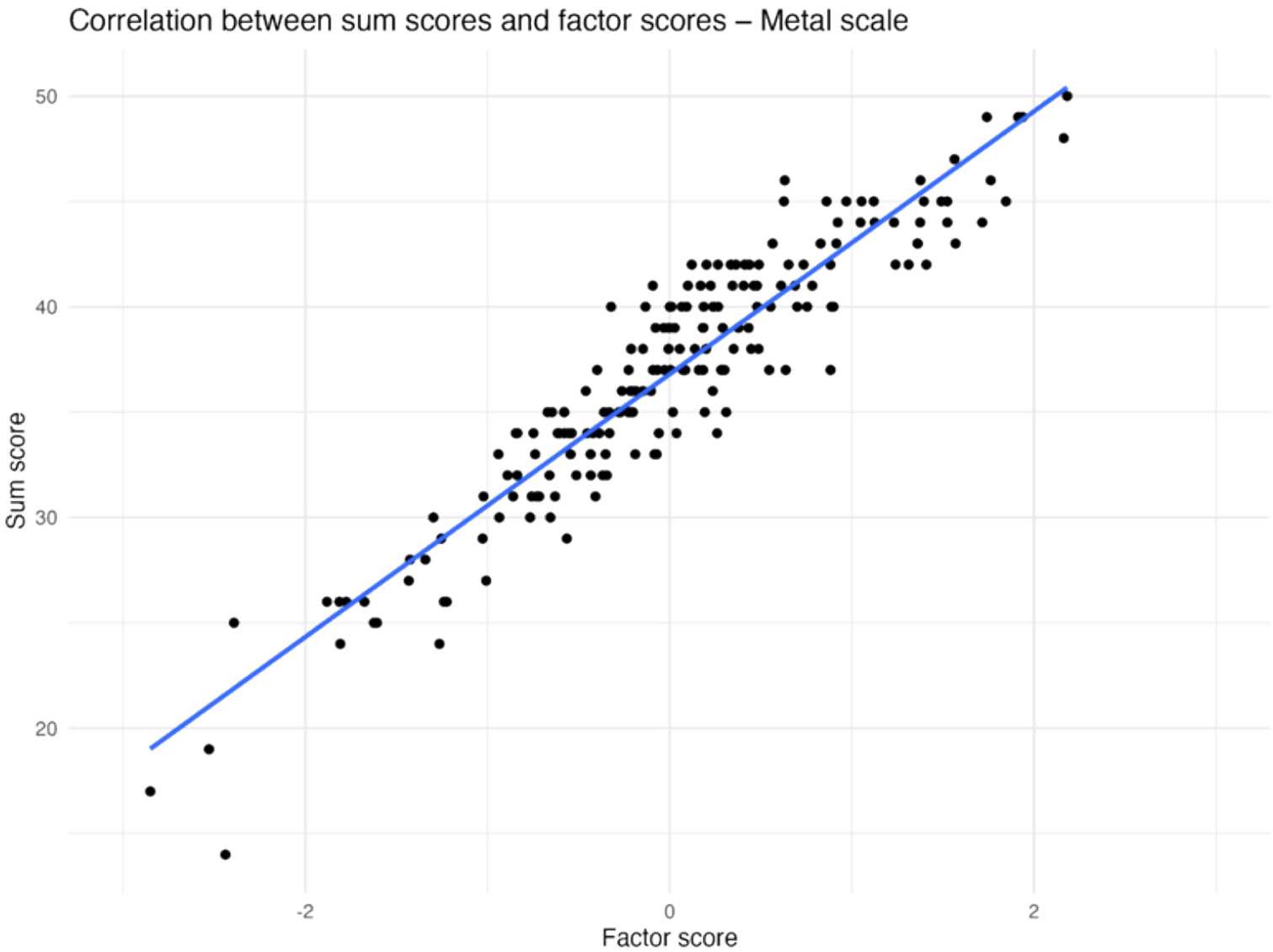
METAL

IRT loadings were similar to those observed in the EFA. All loadings were above .3, except for item 8 (loading = .03). In concordance with what was observed in the EFA, we would therefore recommend removing this item. Item category response curves, along with item reliability functions, are presented below.



METAL

Finally, the IRT factor scores correlated at .83 with the sum scores (see scatterplot below), indicating that sum scores can be used as good proxies.

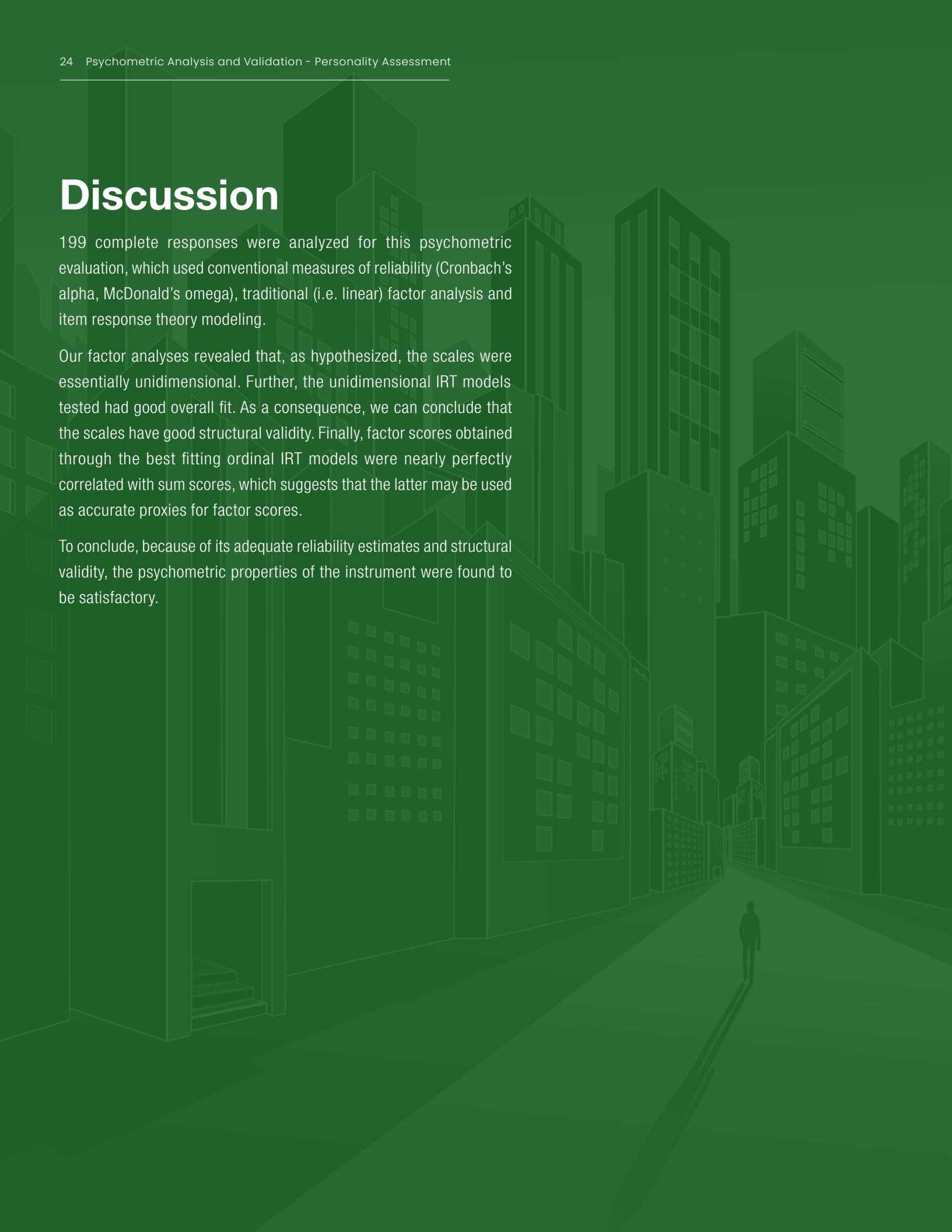


Discussion

199 complete responses were analyzed for this psychometric evaluation, which used conventional measures of reliability (Cronbach's alpha, McDonald's omega), traditional (i.e. linear) factor analysis and item response theory modeling.

Our factor analyses revealed that, as hypothesized, the scales were essentially unidimensional. Further, the unidimensional IRT models tested had good overall fit. As a consequence, we can conclude that the scales have good structural validity. Finally, factor scores obtained through the best fitting ordinal IRT models were nearly perfectly correlated with sum scores, which suggests that the latter may be used as accurate proxies for factor scores.

To conclude, because of its adequate reliability estimates and structural validity, the psychometric properties of the instrument were found to be satisfactory.



References

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics, 30(1), 9–14.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573. https://doi.org/10.1007/BF02293814

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R Environment. Journal of Statistical Software, 48(1), 1–29. https://doi.org/10.18637/jss.v048.i06

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

De Ayala, R. J. (2022). The theory and practice of item response theory (Second edition). The Guilford Press.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105(3), 399–412. https://doi.org/10.1111/bjop.12046

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185. https://doi.org/10.1007/BF02289447

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 141–151. https://doi.org/10.1177/001316446002000116

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149–174. https://doi.org/10.1007/BF02296272

McDonald, R. P. (2000). Test Theory: A Unified Treatment: 1st (first) Edition. Taylor & Francis, Inc.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. ETS Research Report Series, 1992(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Myszkowski, N. (2021). Development of the R library “jrt”: Automated item response theory procedures for judgment data and their application with the consensual assessment technique. Psychology of Aesthetics, Creativity, and the Arts, 15(3), 426–438. https://doi.org/10.1037/aca0000287

Nering, M. L., & Ostini, R. (Eds.). (2010). Handbook of polytomous item response theory models (1 edition). Routledge.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24(1), 50–64. https://doi.org/10.1177/01466216000241003

Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research [Manual]. Northwestern University. https://CRAN.R-project.org/package=psych

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(1), 1–97. https://doi.org/10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464.

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis. Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4