# Getting Started with the Graded Response Model (GRM): A gentle introduction and tutorial in R

Rizqy Amelia Zein1,2 and Hanif Akhtar3,4

1Department of Psychology, Universitas Airlangga

2Department of Psychology, Ludwig-Maximilians-Universität

3Doctoral School of Psychology, ELTE Eötvös Loránd University

4Faculty of Psychology, Universitas Muhammadiyah Malang

# Author Note

Rizqy Amelia Zein https://orcid.org/0000-0001-7840-0299

Hanif Akhtar https://orcid.org/0000-0002-1388-7347

We declare we have no competing interests. The first author receives a PhD scholarship from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD).

Correspondence concerning this article should be addressed to Rizqy Amelia Zein, Department of Psychology, Universitas Airlangga, Jalan Airlangga 4-6, Surabaya, Jawa Timur, Indonesia 60286, Email: amelia.zein@psikologi.unair.ac.id

# Abstract

This tutorial paper introduces the Graded Response Model (GRM), a tool for testing measurement precision under the Item Response Theory (IRT) paradigm. Addressing common problems of measurement imprecision and lack of construct validity, the tutorial guides researchers through a one-dimensional GRM analysis in the R environment, using psych, mirt, and ggmirt packages. GRM is specifically designed to examine the psychometric properties of psychological scales with polytomous items. The tutorial illustrates the procedure using data from the Open Psychometrics Database on the Right-Wing Authoritarianism (RWA) scale, outlining the theoretical underpinnings of GRM, and steps for data preparation, model fitting, interpretation of results, and dealing with common issues and anomalies that may typically arise in the process.

*Keywords*: graded response model, item response theory, mirt, ggmirt, psych, R

# Getting Started with the Graded Response Model (GRM): A gentle introduction and tutorial in R

A foundation of scientific research is a measurement process, which starts from defining the construct being measured, determining measurement assumptions, and inspecting its measurement validity and precision. To illustrate this process, consider a researcher designs a study aiming to identify the association between authoritarian disposition and prejudice toward a social group. As the first step, they must clearly define the variables they are interested in and determine a strategy to measure those variables accurately. Ensuring that the measurement strategy precisely measure what it is intended to measure often poses a challenge to applied psychology researchers because, most of the time, we have no direct access to observe most psychological constructs. Instead, we rely on statistical techniques, which assume that non-observable psychological constructs are “out there,” or later we call this variable a *latent construct* or a *manifest variable*. More sharply, we create materials, such as scale items, stimulus, tasks, observation check-lists, etc. that we *assume* to represent these underlying latent construct (or constructs) being measured.

Relatedly, as most psychological constructs are not directly observable, the questions on whether a scale actually measures what it is intended to measure (i.e., construct validity, see Cronbach and Meehl (1955)) becomes critical. Construct validity, as argued by Cronbach and Meehl (1955), can be sustained by linking test scores to other constructs in which researchers theoretically assume to be related with the construct of interest. Further, Cronbach and Meehl (1955) introduce the notion of “nomological network,” which refers to a theoretical framework connecting the measured construct with other (theoretically) related constructs and relevant observations. Therefore, validity of the test depends on the behavior of the construct of interest: whether or not it behaves as specified in the theory within this network. To scrutinize test validity, researchers correlate the test score with some criterion (i.e., constructs or future/current behaviors theoretically related to the measured construct), and this is a common and widespread practices in psychology (Zumbo & Chan, 2014).

However, validity theory has evolved over time and been expanded since. Rather than a property of a specific test, scholars argue that validity of a scale refers to questions about its consequential aspects (Kane, 2013, 2016; Messick, 1995). In this sense, a test is deemed valid when researchers can demonstrate evidence for articulating a logical argument (Kane, 2013, 2016) in defending test interpretations and its practical use (Messick, 1995). Consequently, scholars argue that scale validity is an ongoing process in which researchers should continually and incrementally gather (and report) the evidence for validity as an important complement to ensure the credibility of their findings (Flake et al., 2017; Messick, 1995; Zumbo & Chan, 2014).

Yet in practice, researchers pay far less attention to the validity (and reliability) of their measures. According to a study examining articles published in a flagship journal in the field of social and personality psychology shows that only a little more than half (53%) of sample articles cited validity evidence from previous validation studies but 19% of those articles had adapted or modified the measures in various fashions so that the evidence for validity may not extend to the modified scales (Flake et al., 2017). A larger study tapping into fifteen commonly used measures in social and personality psychology (e.g., Big Five Inventory, belief in just world scale, need for cognitive closure scale, etc.) even paints a grimmer picture (Hussey & Hughes, 2020). While majority of the tested scales (88%) purportedly had good validity, a more exhaustive examination (i.e., internal consistency, test-retest reliability, factor structure, and invariance) demonstrated that only 4% of these scales actually possess good validity (Hussey & Hughes, 2020). Furthermore, applied psychological researchers seem to be less aware of assumptions underlying their chosen method for evaluating scale reliability. For example, while Cronbach’s is always almost reported (Flake, 2021) and continuously popular (McNeish, 2018), it is unclear that researchers are aware of its underlying assumptions (that true scores and item variances should be equal and that item residuals should be uncorrelated, i.e., equivalence), and thus, makes its usefulness extremely limited (Sijtsma, 2008).

With this in mind, it is not surprising that researchers’ lack of attention to their measurement quality raises questions about replicability of their findings (Flake & Fried, 2020; Lilienfeld & Strother, 2020). In this sense, a serious doubt about the validity of the measures used in a study can affect the validity of the findings, and the decisions that leads to this doubt is called *questionable measurement practice* (QMP, see Flake and Fried (2020). The difficulty to reach a consensus about the best way to measure a certain construct is also a major downstream consequence of the lack of validity reporting and may become a major obstacle to establish cumulative psychological science (Elson et al., 2023).

Therefore, it is very important to equip applied psychological researchers with practical steps to systematically evaluate their measurement validity and precision. By expanding their measurement evaluation toolkit, applied psychological researchers can ensure the quality of their findings and enable them to contribute to cumulative science.

In this article, we aim to gently introduce a graded response model (GRM), a family of item response model,

A meta-communicative cue here—

# A Brief Overview of Item Response Theory

* Contrasting the concept of validity in IRT and CTT
* Contrasting the concept of measurement precision in IRT and CTT Inserting Table 1

# Graded Response Model

Jelasin konsep dasarnya dan plot-plotnya apa aja bedanya sama inserting Figure 1

## Assumptions

Jelasin dua asumsi penting: unidimensionality dan local independence.

# Disclosure and Data Availability Statements

To maximise reproducibility of our analysis, we wrote the article as a Quarto (.qmd) document, where we integrate the R codes used in the analysis as well as its outputs. We also include the complete R script for the example we used as a supplementary document. The Quarto file (and its corresponding .docx and .pdf output) and R script are publicly available on [a Github repository](https://github.com/rameliaz/grm-tutorial-paper). To demonstrate the use of GRM, we use a dataset obtained from [Open-Source Psychometrics Project](http://openpsychometrics.org/_rawdata/), which is publicly available.

# An Illustrative Example of Graded Response Model: The Right Wing Authoritarianism (RWA) Scale

To demonstrate the procedure of running a graded response model, we

## A Brief Overview of the Altemeyer’s RWA Scale

## Step 1: Preparation

install.packages("tidyverse", "psych", "devtools", "mirt", "caret", dependencies=TRUE)  
devtools::install\_github("masurp/ggmirt") # remote installation through GitHub repository

Blabla

library(ggmirt); library(tidyverse); library(psych); library(mirt); library(caret)

Blabla

ds <- read.csv("data/data.csv")

Blabla

str(ds)

Blabla

rwa <- subset(ds, select = Q1:Q22)  
str(rwa)

Blabla

## Step 2: Inspecting Key Descriptive Statistics

psych::describe(rwa)

As we see in Table 2, Kok ada nilai 0-nya? padahal kan skor minimalnya 1. Coba kita hitung frekuensi nilai 0 di tiap item.

zero <- colSums(rwa == 0) / nrow(rwa) \* 100 # Computing the frequency of "0" in each column.  
print(zero) # The proportion of "0" for each item.

Kita anggap nilai 0 adalah NA dan kita delete semua case dengan nilai NA

rwa <- rwa %>%  
 mutate\_all(~na\_if(., 0)) %>% # Replacing 0 with NA in all columns.  
 drop\_na() # Removing cases with any NA values.

Kalau lihat di Table 2, meannya kok beda2? oh iya karena ada unfavorable items. Ayo kita reverse score dulu ya.

unfav <- c("Q4","Q6","Q8","Q9","Q11","Q13","Q15","Q18","Q20","Q21") # Now we create a vector defining which items will be coded reversely.  
rwa <- rwa %>%   
 mutate(across(all\_of(unfav), ~ 10 - .))# We simply subtract the scores from 9 (the maximum) + 1 to reverse code the unfavorable items.

## Step 3: Examining Dimensionality

irt.fa(rwa, nfactors = 1, fm = "minres")

lalalala

fa.parallel(rwa, nfactors = 1, fm="minres", fa="fa", cor = "poly")

lalalalal

cor <- cor(rwa,method="pearson") # First, creating a (pearson) correlation matrix.  
efa <- fa(rwa, nfactors=1, fm="minres") # Now, running exploratory factor analysis.   
print(efa) # Print the results.

Scree plot

Now lets do parallel analysis

pa <- fa.parallel(rwa, fm="minres", fa="fa")

lalalala

pa$fa.values

Eigenvalues 12.225/0.843

## Step 4: Model Estimation, Parameters, and Fit Statistics

model <- 'rwa = 1-22'

lalalala

fit <- mirt(data=rwa, 1, model=model, itemtype="graded", SE=T, verbose=F)

llalalala

coefs <- coef(fit, IRTpars=T, simplify=T) # Storing model parameters in a data frame.  
print(coefs) # Yielding model parameters: item discriminations (a) and threshold (b).

lalalala

summary(fit)

bla bla lhalalala

M2(fit, type="C2")

lhalalala

According to our analysis, the model does not fit the data well (*M*(209) = 1.980251^{4}, *p* = , *RMSEA* = 0.098, )

item.fit <- itemfit(fit)

## Step 5: Model Residuals

ld <- residuals(fit, type = "LD") # Running local dependency statistics  
up <- which(upper.tri(ld), arr.ind = T) # Extracting values only on the upper side of the diagonal.  
lar <- up[ld[up] > 0.2 | ld[up] < -0.2, ] # Defining unusually large residuals (>0.2).

lhalalala

for (i in 1:nrow(lar)) {  
 row <- lar[i, 1]  
 col <- lar[i, 2]  
 value <- ld[row, col]  
 cat(sprintf("A large residual correlation is found between item %d and item %d: %f\n", row, col, value))  
} # Now we detect the problematic pairs.

lhalalala

q3 <- residuals(fit, type = "Q3") # Running Yen's Q3 statistics  
findCorrelation(q3, cutoff = 0.2, verbose = T) # Detecting problematic correlation pairs.

## Step 6: IRT Plots

Now spit out the

tracePlot(fit, facet=T, title = "Category Probability Functions of RWA Scale") + labs(color="Response Options")

lhalalala

itemInfoPlot(fit, facet=T, title = "Item Information Curves of the RWA Scale")

lhalalala

testInfoPlot(fit, title="Test Information Curve of the RWA Scale")

## Step 7: Computing Reliability

theta\_se <- fscores(fit, full.scores.SE = T) # Extracting the estimated theta score of each participant.  
e\_rel <- empirical\_rxx(theta\_se) # Then use the estimated theta to calculate empirical reliability.

lhalalala

m\_rel <- marginal\_rxx(fit)

lhalalala

omega(rwa)

# Conclusions

# References

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Test. *Psychological Bulletin*, *52*(4), 281–302.

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren’t toothbrushes. *Commun Psychol*, *1*(1), 1–4. <https://doi.org/10.1038/s44271-023-00026-9>

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, *56*(2), 132–141. <https://doi.org/gj2hw7>

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. <https://doi.org/10.1177/2515245920952393>

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*. <https://doi.org/10.1177/1948550617693063>

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166–184. <https://doi.org/10.1177/2515245919882903>

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. <https://doi.org/10.1111/jedm.12000>

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 309–311. <https://doi.org/10.1080/0969594X.2016.1156645>

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 281–288. <https://doi.org/10.1037/cap0000236>

McNeish, D. (2018). Thanks coefficient alpha, we’ll take it from here. *Psychological Methods*, *23*(3), 412–433. <https://doi.org/10.1037/met0000144>

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, 9.

Sijtsma, K. (2008). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach’s Alpha. *Psychometrika*, *74*(1), 107. <https://doi.org/10.1007/s11336-008-9101-0>

Zumbo, B. D., & Chan, E. K. H. (2014). Setting the Stage for Validity and Validation in Social, Behavioral, and Health Sciences: Trends in Validation Practices. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54, pp. 3–6). Springer International Publishing. <https://doi.org/10.1007/978-3-319-07794-9>

Table 1

Comparison Between Common IRT Models

| Model | Key Characteristics | Data Type | Response Options |
| --- | --- | --- | --- |
| 1-PL Model (Rasch Model) | 1. Estimates only item difficulties (a). 2. Assumes that all items have the same discrimination parameters (b). 3. Item and person parameters are independent. | Dichotomous | Correct/Incorrect (0/1) |
| 2-PL Model | 1. Estimates item difficulties (a) and item discriminations (b). 2. Less stringent than 1-PL model since it allows item discrimination parameters (b) to vary. | Dichotomous | Correct/Incorrect (0/1) |
| 3-PL Model | 1. Estimates item difficulties (a), discriminations (b), and pseudo-guessing parameters (c). 2. Appropriate for modeling a test data with multiple responses (e.g., multiple-choice tests), and thus, guessing might influence participants’ responses. | Dichotomous | Correct/Incorrect (0/1) |
| Graded Response Model (GRM) | 1. Appropriate for modeling ordinal data with more than two response categories (i.e., Likert-style). 2. Estimates a discrimination parameter (a) and multiple threshold parameters (b) per item. | Polytomous | Ordered Categories |
| Partial Credit Model (PCM) | 1. An extension of the 1-PL (Rasch) model for polytomous items. 2. Estimates thresholds between adjacent categories but assumes equal discrimination across items. | Polytomous | Ordered Categories |
| Generalized Partial Credit Model (GPCM) | Extending PCM to allow differential discrimination parameters across items. | Polytomous | Ordered Categories |
| Nominal Response Model (NRM) | 1. Appropriate for modeling categorical responses with no order. 2. Estimates discrimination parameters (a) and multiple category-specific parameters (b). | Categorical | Unordered Categories |

Figure 1

An Item Probability Function from a GRM Model

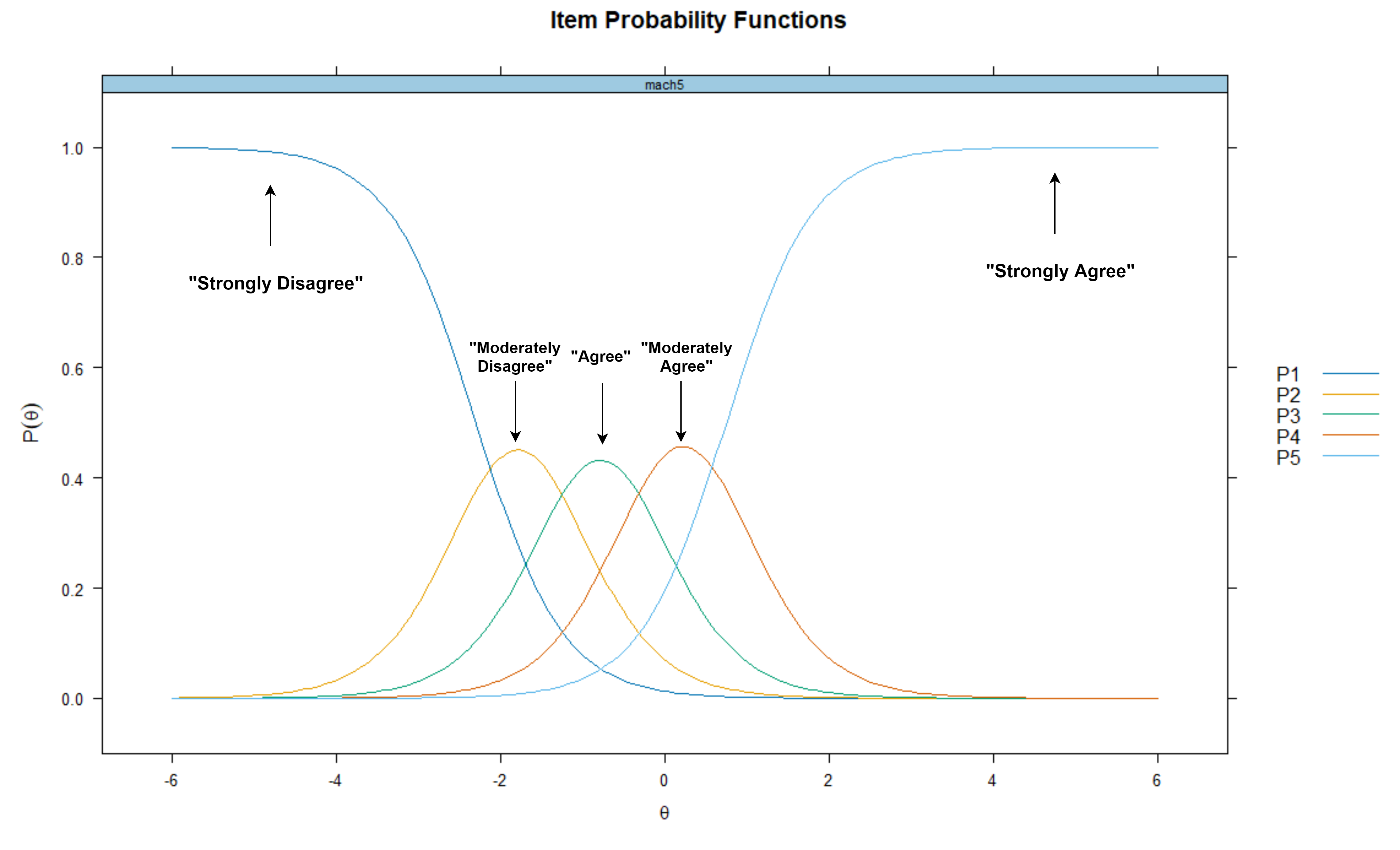


Table 2

Descriptive Statistics of RWA Scale

| Item | Mean | SD | Minimum | Maximum | Range | n |
| --- | --- | --- | --- | --- | --- | --- |
| Q1 | 3.38 | 2.19 | 0 | 9 | 9 | 9,881 |
| Q2 | 2.11 | 2.08 | 0 | 9 | 9 | 9,881 |
| Q3 | 2.72 | 2.54 | 0 | 9 | 9 | 9,881 |
| Q4 | 7.54 | 2.41 | 0 | 9 | 9 | 9,881 |
| Q5 | 2.95 | 2.12 | 0 | 9 | 9 | 9,881 |
| Q6 | 7.45 | 2.29 | 0 | 9 | 9 | 9,881 |
| Q7 | 2.87 | 2.54 | 0 | 9 | 9 | 9,881 |
| Q8 | 6.67 | 2.42 | 0 | 9 | 9 | 9,881 |
| Q9 | 7.51 | 2.03 | 0 | 9 | 9 | 9,881 |
| Q10 | 3.21 | 2.64 | 0 | 9 | 9 | 9,881 |
| Q11 | 7.65 | 2.06 | 0 | 9 | 9 | 9,881 |
| Q12 | 3.61 | 2.47 | 0 | 9 | 9 | 9,881 |
| Q13 | 7.04 | 2.48 | 0 | 9 | 9 | 9,881 |
| Q14 | 3.55 | 2.74 | 0 | 9 | 9 | 9,881 |
| Q15 | 6.84 | 2.32 | 0 | 9 | 9 | 9,881 |
| Q16 | 2.14 | 2.09 | 0 | 9 | 9 | 9,881 |
| Q17 | 3.00 | 2.55 | 0 | 9 | 9 | 9,881 |
| Q18 | 7.86 | 2.11 | 0 | 9 | 9 | 9,881 |
| Q19 | 2.81 | 2.31 | 0 | 9 | 9 | 9,881 |
| Q20 | 7.54 | 2.06 | 0 | 9 | 9 | 9,881 |
| Q21 | 6.27 | 2.68 | 0 | 9 | 9 | 9,881 |
| Q22 | 2.90 | 2.49 | 0 | 9 | 9 | 9,881 |

*Note*. SD = Standard Deviation

Table 3

Descriptive Statistics of RWA Scale

| Item | Mean | SD | Minimum | Maximum | Range | n |
| --- | --- | --- | --- | --- | --- | --- |
| Q1 | 3.38 | 2.19 | 1 | 9 | 8 | 9,680 |
| Q2 | 2.10 | 2.08 | 1 | 9 | 8 | 9,680 |
| Q3 | 2.73 | 2.54 | 1 | 9 | 8 | 9,680 |
| Q4 | 2.43 | 2.38 | 1 | 9 | 8 | 9,680 |
| Q5 | 2.96 | 2.12 | 1 | 9 | 8 | 9,680 |
| Q6 | 2.54 | 2.28 | 1 | 9 | 8 | 9,680 |
| Q7 | 2.87 | 2.54 | 1 | 9 | 8 | 9,680 |
| Q8 | 3.31 | 2.41 | 1 | 9 | 8 | 9,680 |
| Q9 | 2.47 | 1.99 | 1 | 9 | 8 | 9,680 |
| Q10 | 3.20 | 2.63 | 1 | 9 | 8 | 9,680 |
| Q11 | 2.33 | 2.03 | 1 | 9 | 8 | 9,680 |
| Q12 | 3.61 | 2.46 | 1 | 9 | 8 | 9,680 |
| Q13 | 2.95 | 2.47 | 1 | 9 | 8 | 9,680 |
| Q14 | 3.55 | 2.74 | 1 | 9 | 8 | 9,680 |
| Q15 | 3.15 | 2.31 | 1 | 9 | 8 | 9,680 |
| Q16 | 2.14 | 2.09 | 1 | 9 | 8 | 9,680 |
| Q17 | 3.00 | 2.55 | 1 | 9 | 8 | 9,680 |
| Q18 | 2.13 | 2.09 | 1 | 9 | 8 | 9,680 |
| Q19 | 2.80 | 2.31 | 1 | 9 | 8 | 9,680 |
| Q20 | 2.44 | 2.04 | 1 | 9 | 8 | 9,680 |
| Q21 | 3.72 | 2.67 | 1 | 9 | 8 | 9,680 |
| Q22 | 2.90 | 2.50 | 1 | 9 | 8 | 9,680 |

*Note*. Descriptive Statistics After Reversing Unfavorable Items and removing cases with NA. SD = Standard Deviation.

Figure 2

Scree Plot

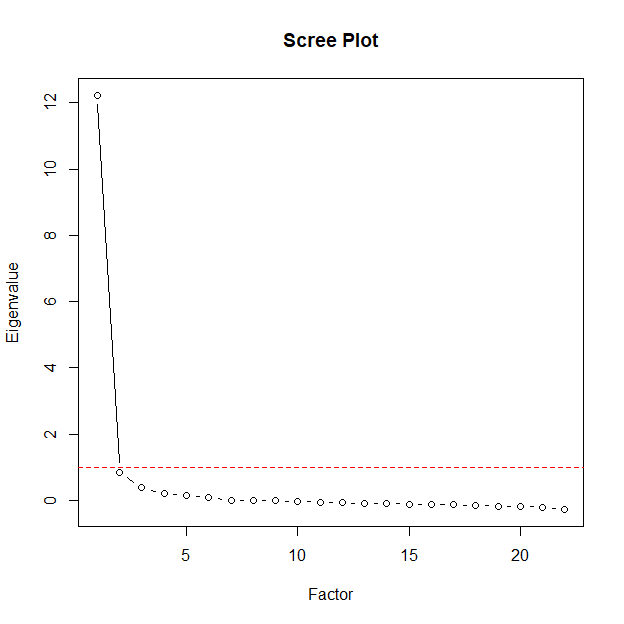


Figure 3

Parallel Analysis

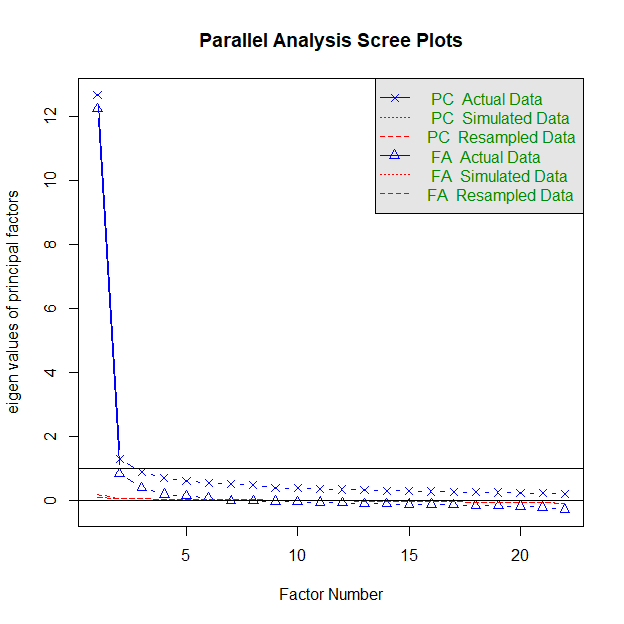


Table 4

Item Parameters

| Item | α | β1 | β2 | β3 | β4 | β5 | β6 | β7 | β8 | λ | h2 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Q1 | 1.57 | -1.08 | -0.23 | 0.48 | 0.78 | 1.20 | 1.67 | 2.41 | 3.19 | 0.68 | 0.46 |
| Q2 | 2.45 | 0.47 | 0.94 | 1.19 | 1.33 | 1.58 | 1.86 | 2.13 | 2.47 | 0.82 | 0.67 |
| Q3 | 3.04 | 0.03 | 0.50 | 0.74 | 0.87 | 1.04 | 1.30 | 1.54 | 1.85 | 0.87 | 0.76 |
| Q4 | 2.50 | 0.22 | 0.74 | 0.98 | 1.08 | 1.30 | 1.53 | 1.82 | 2.10 | 0.83 | 0.68 |
| Q5 | 1.92 | -0.62 | 0.13 | 0.67 | 1.01 | 1.38 | 1.82 | 2.30 | 2.90 | 0.75 | 0.56 |
| Q6 | 1.91 | 0.05 | 0.62 | 0.98 | 1.16 | 1.53 | 1.81 | 2.13 | 2.52 | 0.75 | 0.56 |
| Q7 | 3.32 | -0.14 | 0.38 | 0.66 | 0.83 | 0.97 | 1.23 | 1.51 | 1.82 | 0.89 | 0.79 |
| Q8 | 1.58 | -0.74 | -0.06 | 0.47 | 0.70 | 1.28 | 1.64 | 2.04 | 2.44 | 0.68 | 0.46 |
| Q9 | 1.89 | -0.23 | 0.54 | 1.07 | 1.54 | 1.83 | 2.08 | 2.37 | 2.74 | 0.74 | 0.55 |
| Q10 | 3.30 | -0.30 | 0.20 | 0.47 | 0.63 | 0.82 | 1.13 | 1.40 | 1.75 | 0.89 | 0.79 |
| Q11 | 2.03 | 0.04 | 0.69 | 1.13 | 1.42 | 1.71 | 1.99 | 2.35 | 2.74 | 0.77 | 0.59 |
| Q12 | 2.61 | -0.77 | -0.20 | 0.25 | 0.46 | 0.79 | 1.19 | 1.60 | 2.02 | 0.84 | 0.70 |
| Q13 | 2.60 | -0.31 | 0.29 | 0.65 | 0.91 | 1.16 | 1.35 | 1.58 | 1.89 | 0.84 | 0.70 |
| Q14 | 2.30 | -0.46 | 0.03 | 0.32 | 0.47 | 0.75 | 1.09 | 1.42 | 1.79 | 0.80 | 0.65 |
| Q15 | 2.32 | -0.62 | 0.04 | 0.50 | 0.87 | 1.19 | 1.44 | 1.77 | 2.21 | 0.81 | 0.65 |
| Q16 | 2.59 | 0.43 | 0.85 | 1.13 | 1.30 | 1.54 | 1.82 | 2.11 | 2.41 | 0.84 | 0.70 |
| Q17 | 2.56 | -0.15 | 0.35 | 0.60 | 0.77 | 1.00 | 1.35 | 1.66 | 2.00 | 0.83 | 0.69 |
| Q18 | 2.32 | 0.40 | 0.98 | 1.26 | 1.44 | 1.61 | 1.86 | 2.14 | 2.45 | 0.81 | 0.65 |
| Q19 | 3.14 | -0.24 | 0.32 | 0.67 | 0.90 | 1.12 | 1.48 | 1.78 | 2.11 | 0.88 | 0.77 |
| Q20 | 1.75 | -0.14 | 0.61 | 1.14 | 1.52 | 1.74 | 2.07 | 2.46 | 2.89 | 0.72 | 0.51 |
| Q21 | 2.31 | -0.70 | -0.19 | 0.17 | 0.43 | 0.85 | 1.09 | 1.32 | 1.61 | 0.81 | 0.65 |
| Q22 | 3.07 | -0.16 | 0.35 | 0.63 | 0.82 | 1.02 | 1.32 | 1.59 | 1.88 | 0.87 | 0.76 |

*Note*. λ = Standardized Factor Loadings, h2 = Commonality, α = Discrimination, β1.4 = Response specific difficulty parameters (item threshold).

Table 5

Item Fit Statistics

| Item | χ2 | df | RMSEA | p |
| --- | --- | --- | --- | --- |
| Q1 | 1,187.23 | 957.00 | 0.00 | 0.00 |
| Q2 | 917.58 | 915.00 | 0.00 | 0.47 |
| Q3 | 983.47 | 847.00 | 0.00 | 0.00 |
| Q4 | 978.46 | 930.00 | 0.00 | 0.13 |
| Q5 | 1,072.31 | 957.00 | 0.00 | 0.01 |
| Q6 | 1,310.80 | 1,041.00 | 0.01 | 0.00 |
| Q7 | 972.91 | 780.00 | 0.01 | 0.00 |
| Q8 | 1,191.41 | 1,041.00 | 0.00 | 0.00 |
| Q9 | 1,254.66 | 981.00 | 0.01 | 0.00 |
| Q10 | 981.24 | 770.00 | 0.01 | 0.00 |
| Q11 | 1,206.09 | 988.00 | 0.00 | 0.00 |
| Q12 | 842.01 | 811.00 | 0.00 | 0.22 |
| Q13 | 1,088.69 | 904.00 | 0.00 | 0.00 |
| Q14 | 1,031.66 | 903.00 | 0.00 | 0.00 |
| Q15 | 1,152.22 | 932.00 | 0.00 | 0.00 |
| Q16 | 997.76 | 891.00 | 0.00 | 0.01 |
| Q17 | 1,142.47 | 906.00 | 0.01 | 0.00 |
| Q18 | 1,182.62 | 936.00 | 0.01 | 0.00 |
| Q19 | 912.71 | 793.00 | 0.00 | 0.00 |
| Q20 | 1,262.10 | 1,046.00 | 0.00 | 0.00 |
| Q21 | 1,031.48 | 878.00 | 0.00 | 0.00 |
| Q22 | 1,122.72 | 832.00 | 0.01 | 0.00 |

*Note*. Scaled χ2 Statistics. RMSEA = root mean square error of approximation, CFI = comparative fit index, TLI = Tucker-Lewis index, SRMR = Standardized Root Mean Square Residual.

Figure 4

Item Probability Functions of RWA Scale

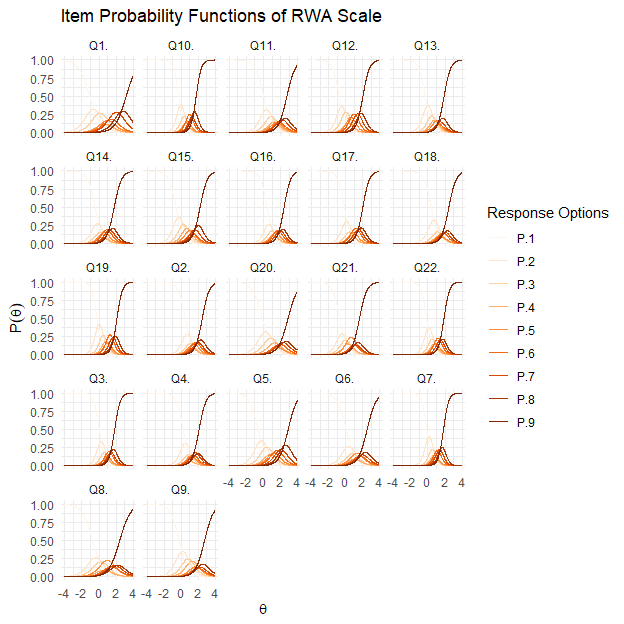


Figure 5

Item Information Curves of the RWA Scale

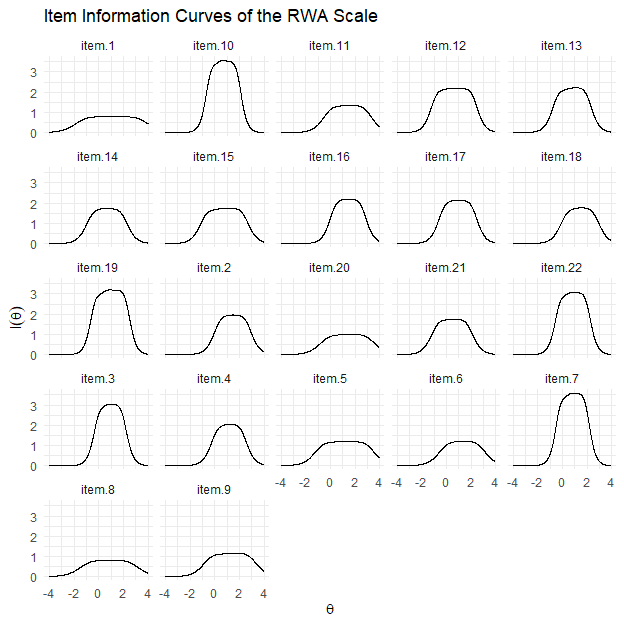


Figure 6

Test Information Curve of the RWA Scale

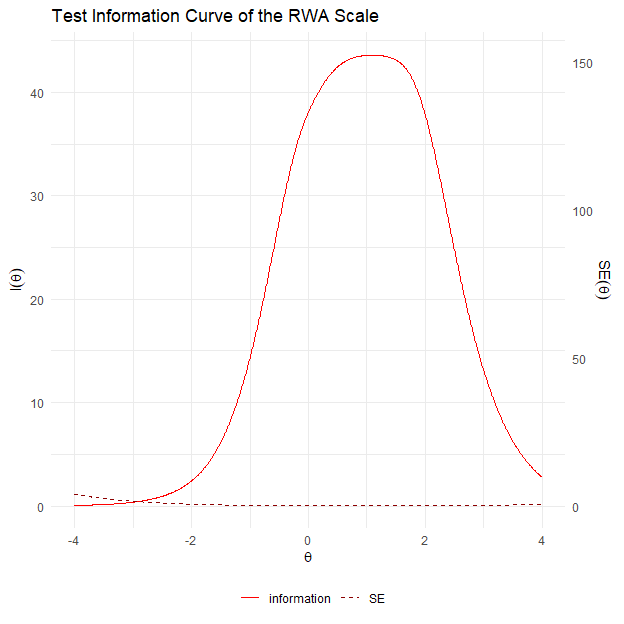


Figure 7

Reliability of the RWA Scale Given to the θ Level

