# 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

The foundation of scientific research is the measurement process, which involves defining the construct to be measured, determining measurement assumptions, and evaluating measurement validity and precision. For instance, a researcher studying the role of authoritarian personality in explaining prejudice must first clearly define these variables, and then, develop a measurement strategy to operationalize the variables. This is challenging in applied psychology, as most psychological constructs are not directly observable. To ensure a precise measurement, researchers rely on statistical techniques, assuming that these non-observable traits are “out there,” or later we name this concept a *latent trait* or a *manifest variable*. Researchers then create tools, like scales, stimuli, tasks, observation checklists, etc., which are assumed to represent these these latent traits.

In psychological research, where variables are often not directly observable, construct validity, which refers to a question on whether a scale measures what it is intended to measure (Cronbach & Meehl, 1955), becomes crucial. Construct validity, as argued by Cronbach and Meehl (1955), can be scrutinized by linking test scores to other related constructs and observations. Test validity relies on whether the construct behaves as theory predicts within a “nomological network” (Cronbach & Meehl, 1955). In practice, researchers establish validity by correlating the test score with external criterion i.e., other theoretically related traits or behavior (Zumbo & Chan, 2014).

Over time, validity theory has evolved and has been expanded. Validity is now seen not as a property of a specific test, but rather lies in the consequential aspects of the test use (Kane, 2013, 2016; Messick, 1995). In this sense, a test is deemed valid when researchers can demonstrate evidence for formulating a logical argument (Kane, 2013, 2016), such as causal inference (Rutkowski & Delandshere, 2016), used to defend test interpretations and its practical use (Messick, 1995). This makes construct validity an ongoing process where researchers are advised to continuously gather and report evidence for validity of the measures they use in their research (Flake et al., 2017; Messick, 1995; Zumbo & Chan, 2014).

Yet in practice, researchers tend to overlook to the quality of their measures. A study reviewing articles published in a social and personality psychology journal found that only a little more than half (53%) of sample articles cited validity evidence from previous validation studies. Nineteen percent of those articles, nonetheless, had adapted or modified the measures in many different ways so that the evidence for validity may not extend to these modified versions (Flake et al., 2017). Additionally, a larger study examining 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 the majority of the tested scales (88%) purportedly had good validity, a more exhaustive examination through internal consistency, test-retest reliability, factor structure, and invariance tests revealed that only 4% of these scales possess good measurement validity (Hussey & Hughes, 2020). Furthermore, applied psychological researchers seem to overlook the assumptions underlying their chosen method for evaluating reliability. For example, while Cronbach’s is always almost reported (Flake, 2021) and exceptionally popular (McNeish, 2018), it is unclear whether 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, limits its usefulness (Sijtsma, 2008).

With this in mind, this oversight in measurement quality can raise reasonable concerns about the replicability of research findings because measurement contributes to the (in)validity of the findings (Flake & Fried, 2020; Lilienfeld & Strother, 2020). Measurement oversight can be identified in decisions that researchers make, which potentially compromise the validity of the measures used in their studies, a phenomenon known as “*questionable measurement practices*” (Flake & Fried, 2020). Further, the lack of evidence for measurement validity contributes to the difficulty in reaching a consensus about the best way to measure a construct. In the long run, it becomes a major obstacle to establishing cumulative psychological science (Elson et al., 2023). Given these circumstances, it is critical to provide applied psychological researchers with a set of tools and practical steps to assist them in evaluating measurement validity so that they can improve their measurement practices, and thus, the credibility of their findings.

In this article, we aim to gently introduce the graded response model (GRM), which is a family of item response models, specifically aimed to assess the measurement precision of a polytomous (Likert-style) scale. We start this paper by briefly explaining the basics of Item Response Theory (IRT), the theoretical foundation of GRM, as well as contrasting its differences with the widely known Classical True Score Theory (CTT). In this paper, we aim to keep our tutorial concise so we do not cover all IRT concepts in detail. Interested readers are referred to available didactic texts on IRT, such as Embretson and Reise (2000) for the introductory level, and Baker and Seock-Ho (2017) and de Ayala (2022) for a more technical and comprehensive overview.

Further, we briefly explain what GRM is and how a GRM model is specified. In this part, we highlight how model and item parameters are specified and interpreted as well as visualizations (i.e., plots) derived from a GRM analysis. We also introduce three key assumptions underlying a GRM and how to test these assumptions. Subsequently, we provide an illustrative example of testing measurement precision using GRM, and in this part, we show how to practically implement GRM analysis and interpret its results. We complement the illustrative example with code in the open-source programming language R (R Core Team, 2023) so that the readers can implement and reproduce the example presented in this paper. Readers with some experience with R and RStudio can easily follow this tutorial, but for those who are not yet familiar with R, we refer them to excellent, easy-to-follow materials developed by Navarro (2018). At last, we present some concluding remarks to encourage researchers to implement a GRM analysis as a part of their routine.

# A Brief Overview of Item Response Theory

To scrutinize their measurement quality, psychological researchers have relied upon two foundational principles underlying psychological testing, i.e., CTT and IRT, to decompose observed score into its *deterministic* and *random* or stochastic elements (Zumbo, 2006). On the one hand, CTT decomposes observed score into *true* score and its *error* components. In practice, researchers implement this principle by conducting Structural Equation Modeling (SEM), Exploratory Factor Analysis (EFA), or Confirmatory Factor Analysis (CFA). Specifying the measurement part of SEM, or employing EFA and CFA models, allows for the decomposition of the covariance matrix of a given data into latent factors, which represent true score or actual ability or trait level, and error terms (i.e., residuals or unique variances). Consequently, the relationship between observed score and true score is *linear*, as true score is essentially a linear transformation of observed score (Embretson & Reise, 2000). IRT, on the other hand, models this relationship in a *probabilistic* rather than a linear fashion. More sharply, an IRT model assumes a probabilistic relationship between the observed score and the latent trait being measured (), and in particular, this probability takes into account the properties of the items, such as item discrimination, difficulty, and, if necessary, guessing (Baker & Seock-Ho, 2017).

The most critical distinction between CTT and IRT is how they conceptualize measurement error. In CTT, standard deviation of the observed score (*s*) is used to calculate standard error of measurement. Since measurement error relies on *s*, CTT assumes that measurement error is sample-dependent and constant for all individuals in the sample regardless of their trait level. This shortcoming restricts the utility of CTT, especially since psychologists often need to interpret individual scores, not just evaluate the test as a whole. Additionally, since standard error of measurement is also a function of reliability, which is typically denoted as a correlation between two parallel tests, reliability is presumed to be equal across different levels of a trait or ability. This is unrealistic since researchers frequently encounter situations where a test is excessively difficult for a group of low-performing participants, providing little information beyond indicating that their trait or ability level is significantly lower than what the test can measure.

IRT offers an elegant solution to this issue by enabling the calculation of standard errors for each individual (*SE*), and test reliability can be inferred from the average of *SE* across a sample of participants (Embretson & Reise, 2000; Lang & Tay, 2021). This approach allows for the estimation of reliability at varying levels of trait or ability (i.e., test information function - TIF). By doing this, when a test is overly challenging for low-performing individuals or too easy for high-performing individuals, IRT analysis can help identify the levels of trait or ability where the test is most reliable. A concrete example of this is a study scrutinizing the reliability of the Short Dark Tetrad (SD4) scale from an IRT perspective (Blötner & Beisemann, 2022). According to this study, the sadism subscale of the SD4 scale is the most reliable for measuring individuals with average to high levels ( < < 2.5 *SD*) of sadism but suboptimal for measuring those with low levels of sadism (Blötner & Beisemann, 2022). Therefore, IRT models offer a more informative and nuanced perspective, proving useful for researchers who wish to closely examine the performance of their measures.

Further, IRT allows researchers to examine the performance of a specific item (hence “item” response theory) by specifying the relationship between item score and given one-, two-, or three-parameter. One-parameter logistic (1PL) or Rasch model, the simplest variant of the IRT models, accounts only item difficulty (*b*) while presuming equal item discrimination (*a*) across items. Practically, this model slightly overlaps with CTT in the sense that they assume that all items carry equal informative value thus the estimated of a 1PL/Rasch model is identical to the sum score of CTT (Lang & Tay, 2021; Stemler & Naples, 2021). This notion actually stems from a philosophical principle of *specific objectivity*, which requires that comparisons between measurements remain independent of both the item and individual characteristics (Rasch, 1977). Put simply, Rasch model assumes that the latent trait () should remain unaffected by specific items used in the test. For example, the difference in depression levels between two individuals should be always the same regardless of the scale used to measure their depression levels. While this principle reasonably enforces objectivity in measurement practices, it demands a strong theory delineating the construct and strict requirements of data-model fit while both assumptions rarely hold in a real-world scenario, especially in psychological science.

Furthermore, the two-parameter logistic (2PL) model fills this gap by allowing item ability to differentiate individuals with varying levels of (i.e., item discrimination parameter - *a*) to differ. Additionally, in some contexts, researchers may suspect that a part of the probabilistic relationship between observed score and is explained by guessing thus three-parameter logistic (3PL) model incorporates guessing parameter (*c*). The field of IRT has rapidly developed to include various models suited to specific contexts, such as handling ordinal responses (Muraki, 1992; Samejima, 1997), categorical responses (Thissen et al., 2013), or assessing multidimensional traits simultaneously (Bock & Aitkin, 1981; Chalmers, 2012). It is important to note that 1PL, 2PL, and 3PL models are only applicable to binary or dichotomous data (e.g., true/false response), and the focus of this article is nonetheless to show the utility of IRT for fitting the ordered (Likert-style) responses. We briefly summarize the features of the most frequently employed IRT models in Table 1.

# Graded Response Model

The GRM is a family of IRT models specifically designed to analyze ordered polytomous (Likert-style) data (Samejima, 1997, 2010, 2016). The fundamental idea of GRM is to extend the logic of simpler dichotomous IRT models to polytomous scales by applying a probabilistic function to each response category. In dichotomous 2PL IRT models, the probability of answering a correct response is modeled as a function of , item difficulty, and item discrimination. GRM extends this idea by modeling the probability of picking a certain response category or higher on an item (i.e., *step function*) given to item difficulty and discrimination. For each response category (e.g., answering “agree” on a five-point Likert scale), GRM calculates the cumulative probability of a participant answering “agree” or above, given to their . Therefore, the notion of item difficulty in dichotomous IRT models is extended to the step function (i.e., item threshold) in order to handle ordered polytomous data.

To illustrate the step function, consider a scale measuring sadistic personality (e.g., “*watching a fist-fight excites me*,” etc.) with response categories ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Imagine these categories are hierarchically ordered from the lowest category (*strongly disagree*) to the highest (*strongly agree*), like a staircase. GRM aims to estimate the level of sadistic personality () required for “stepping” from one response category to another (e.g., *moderately agree* to *strongly agree*). Therefore, a GRM model of a five-point Likert scale calculates four item threshold parameters (*b*): the location of level where individuals are equally likely to respond 1 or 2 (*b*1), 2 or 3 (*b*2), 3 or 4 (*b*3), and 4 or 5 (*b*4). Each threshold is exactly the point where a participant is equally likely to respond to either of the adjacent response categories. For example, if *b*1 = -1.23 for the item “*watching a fist-fight excites me*”, it implies that participants with sadism level 1.23 below the mean are equally likely to either answer *strongly disagree* or *moderately disagree*.

Let’s take a closer look at Figure 1, which shows an example of an item probability function (IPF) from an item of a scale with five response categories. In general, IPF describes a relationship between (x-axis) and the probability of endorsing a response category (y-axis - *P*). Figure 1 consists of five category probability curves, each of which represents the probability of endorsing a response category given . Item threshold is exactly the location of where two adjacent category probability curves cross each other.

## Assumptions

Before running an analysis, researchers need to be aware of assumption underlying a GRM model. The first key assumption is that is monotonically correlated with the probability of endorsing a response category. In other words, individuals with higher levels of have a higher probability of endorsing response category that represents greater intensity of the traits (Hambleton et al., 2010). For instance, people with higher tendency of sadistic personality have a greater chance of answering “*moderately agree*” or “*strongly agree*” on a sadism scale compared to those with lower levels of sadism.

The second key assumption pertains to unidimensionality of . This implies that the item pool tested using a GRM analysis should represent only one latent trait (). Although it is indeed possible to account for multiple in a IRT model (multidimensional IRT, see Bock et al. (1988), we limit the scope of our tutorial to the unidimensional GRM. To test whether an item pool represents only one , researchers can apply a parallel analysis (Guo & Choi, 2023) or a factor analysis (Hambleton & Rovinelli, 1986), before specifying their GRM models.

Relatedly, a GRM model assumes that the item pool are locally independent, as the third assumption. Local independence suggests that participants’ responses to one item do not influence their response to another. It is important to note that evidence for unidimensionality does not always warrant local independence. As an eyeball example - consider a trust in science scale containing these two items: “*I trust scientists working in the natural science*” and “*most biologists are trustworthy*.” While both items might measure the same latent trait (i.e., trust in science), those who trust scientists working in the natural science are very likely to also trust biologists. To inspect whether items are locally independent, researchers can examine the relationships between item responses after accounting for using residual correlation analysis (Chen & Thissen, 1997).

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

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

To illustrate the application of GRM for analyzing measurement precision of an ordered polytomous scale, we provide an example of a GRM analysis of the RWA scale data (Altemeyer, 1981, 2006). Altemeyer (1981) defines RWA as a personality propensity to blindly abide established authorities (i.e., authoritarian submission), to act aggressively toward individuals who are perceived to be punished by these established authorities (i.e., authoritarian aggression), and to uphold traditional values promoted by the authorities (i.e., conventionalism). Individuals with higher levels of RWA tend to view the world as a dangerous place (Duckitt & Sibley, 2017), and thus, to ensure stability and safety, they are motivated to preserve social order by deferring to those whom they perceive as legal, social, or moral authorities; for instance, the government, religious institutions, or political or religious or military leaders (Saunders & Ngo, 2017). The RWA scale developed by Altemeyer (1981) remains as one of the most important measures in the field of social and political psychology (Saunders & Ngo, 2017), and has helped researchers to conceptualize the role of RWA in shaping various social psychological outcomes, such as prejudice, political behavior, and various antisocial behaviors Akrami and Ekehammar (2006); (Osborne et al., 2023; Sibley et al., 2012).

In this paper, we use the latest version of the RWA scale (Altemeyer, 2006), which consists of 22 items that participants are asked to respond their agreement to those items on a nine-point scale, ranging from “*strongly disagree*” (-4) to “*strongly agree*” (+4). While the RWA scale consists of three sub-dimensions, i.e., submissiveness, aggression, and conventionalism, the RWA scale is (theoretically) unidimensional because (Altemeyer, 2006) shows that the sub-dimensions are strongly intercorrelated. While some evidence shows otherwise (Duckitt & Bizumic, 2013), for the purpose of this article, we assume that the RWA scale is unidimensional.

# Disclosure and Data Availability Statements

To maximize the reproducibility of our illustrative example, 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 annotated R script file (.R) for the example we used as a supplementary document, which we highly recommend to use for a didactic purpose. The Quarto file (and its corresponding .docx and .pdf output) and R script file are publicly available on [a Github repository](https://github.com/rameliaz/grm-tutorial-paper). The dataset we used in this tutorial paper was obtained from [Open-Source Psychometrics Project](http://openpsychometrics.org/_rawdata/), which is publicly accessible.

## 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 discrimination (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

m\_rel <- marginal\_rxx(fit)

lhalalala

conRelPlot(fit, title="Reliability of the RWA Scale Given to the θ Level")

lalalala

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

omega(rwa)

# Conclusions

Since construct validation is an ongoing process (Messick, 1995), applied researchers can contribute to the cumulative science by reporting evidence for measurement validity (Flora & Flake, 2017), and this can be done by implementing and reporting a GRM analysis.

# References

Akrami, N., & Ekehammar, B. (2006). Right-Wing Authoritarianism and Social Dominance Orientation. *Journal of Individual Differences*, *27*(3), 117–126. <https://doi.org/10.1027/1614-0001.27.3.117>

Altemeyer, B. (1981). *Right-wing authoritarianism*. The University of Manitoba Press.

Altemeyer, B. (2006). *The Authoritarians*.

Baker, F. B., & Seock-Ho, K. (2017). *The basics of item response theory using R*. Springer Science+Business Media.

Blötner, C., & Beisemann, M. (2022). The Dark Triad is dead, long live the Dark Triad: An item-response theoretical examination of the Short Dark Tetrad. *Personality and Individual Differences*, *199*, 111858. <https://doi.org/10.1016/j.paid.2022.111858>

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. <https://doi.org/10.1007/BF02293801>

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement*, *12*, 261–280. <https://doi.org/10.1177/014662168801200305>

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29. <https://doi.org/10.18637/jss.v048.i06>

Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. <https://doi.org/10.3102/10769986022003265>

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

de Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory*. The Guilford Press.

Duckitt, J., & Bizumic, B. (2013). Multidimensionality of Right-Wing Authoritarian Attitudes: Authoritarianism-Conservatism-Traditionalism: Authoritarianism-Conservatism-Traditionalism. *Political Psychology*, *34*(6), 841–862. <https://doi.org/10.1111/pops.12022>

Duckitt, J., & Sibley, C. G. (2017). The dual process motivational model of ideology and prejudice. In *The Cambridge handbook of the psychology of prejudice* (pp. 188–221). Cambridge University Press. <https://doi.org/10.1017/9781316161579.009>

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>

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. L. Erlbaum Associates.

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>

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *49*(2), 78–88. <https://doi.org/10.1037/cbs0000069>

Guo, W., & Choi, Y.-J. (2023). Assessing Dimensionality of IRT Models Using Traditional and Revised Parallel Analyses. *Educational and Psychological Measurement*, *83*(3), 609–629. <https://doi.org/10.1177/00131644221111838>

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement*, *10*(3), 287–302. <https://doi.org/10.1177/014662168601000307>

Hambleton, R. K., Van Der Linden, W. J., & Wells, C. S. (2010). IRT models for the Analysis of Polytomously Scored Data: Brief and Selected History of Model Building Advances. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21–42). Routledge.

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>

Lang, J. W. B., & Tay, L. (2021). The Science and Practice of Item Response Theory in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 311–338. <https://doi.org/10.1146/annurev-orgpsych-012420-061705>

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.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. <https://doi.org/10.1177/014662169201600206>

Navarro, D. J. (2018). R for Psychological Science. In *UNSW Computational Cognitive Science*. https://psyr.djnavarro.net/.

Osborne, D., Costello, T. H., Duckitt, J., & Sibley, C. G. (2023). The psychological causes and societal consequences of authoritarianism. *Nat Rev Psychol*, *2*(4), 220–232. <https://doi.org/10.1038/s44159-023-00161-4>

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rasch, G. (1977). On Specific Objectivity an Attempt at Formalizing the Request for Generality and validity of Scientific Statements. *Danish Yearbook of Philosophy*, *14*(1), 58–94. <https://doi.org/10.1163/24689300-01401006>

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*, *4*(1), 6. <https://doi.org/10.1186/s40536-016-0019-1>

Samejima, F. (1997). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). Springer. <https://doi.org/10.1007/978-1-4757-2691-6_5>

Samejima, F. (2010). The General Graded Response Model. In *Handbook of Polytomous Item Response Theory Models*. Routledge.

Samejima, F. (2016). Graded Response Models. In *Handbook of Item Response Theory*. Chapman and Hall/CRC.

Saunders, B. A., & Ngo, J. (2017). The Right-Wing Authoritarianism Scale. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–4). Springer International Publishing. <https://doi.org/10.1007/978-3-319-28099-8_1262-1>

Sibley, C. G., Osborne, D., & Duckitt, J. (2012). Personality and political orientation: Meta-analysis and test of a Threat-Constraint Model. *Journal of Research in Personality*, *46*(6), 664–677. <https://doi.org/10.1016/j.jrp.2012.08.002>

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>

Stemler, S., & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research, and Evaluation*, *26*(1). <https://doi.org/10.7275/v2gd-4441>

Thissen, D., Cai, L., & Bock, R. D. (2013). The Nominal Categories Item Response Model. In *Handbook of Polytomous Item Response Theory Models*. Routledge. <https://doi.org/10.4324/9780203861264.ch3>

Zumbo, B. D. (2006). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 45–79). Elsevier. <https://doi.org/10.1016/S0169-7161(06)26003-6>

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 (*b*). 2. Assumes that all items have equal discrimination parameters (*a*). 3. Item and person parameters are independent. | Dichotomous | Correct/Incorrect (0/1) |
| 2-PL Model | 1. Estimates item difficulties (*b*) and item discrimination (*a*). 2. Less stringent than 1-PL model since it allows item discrimination parameters (*a*) to vary. | Dichotomous | Correct/Incorrect (0/1) |
| 3-PL Model | 1. Estimates item difficulties (*b*), discrimination (*a*), and pseudo-guessing parameter (*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

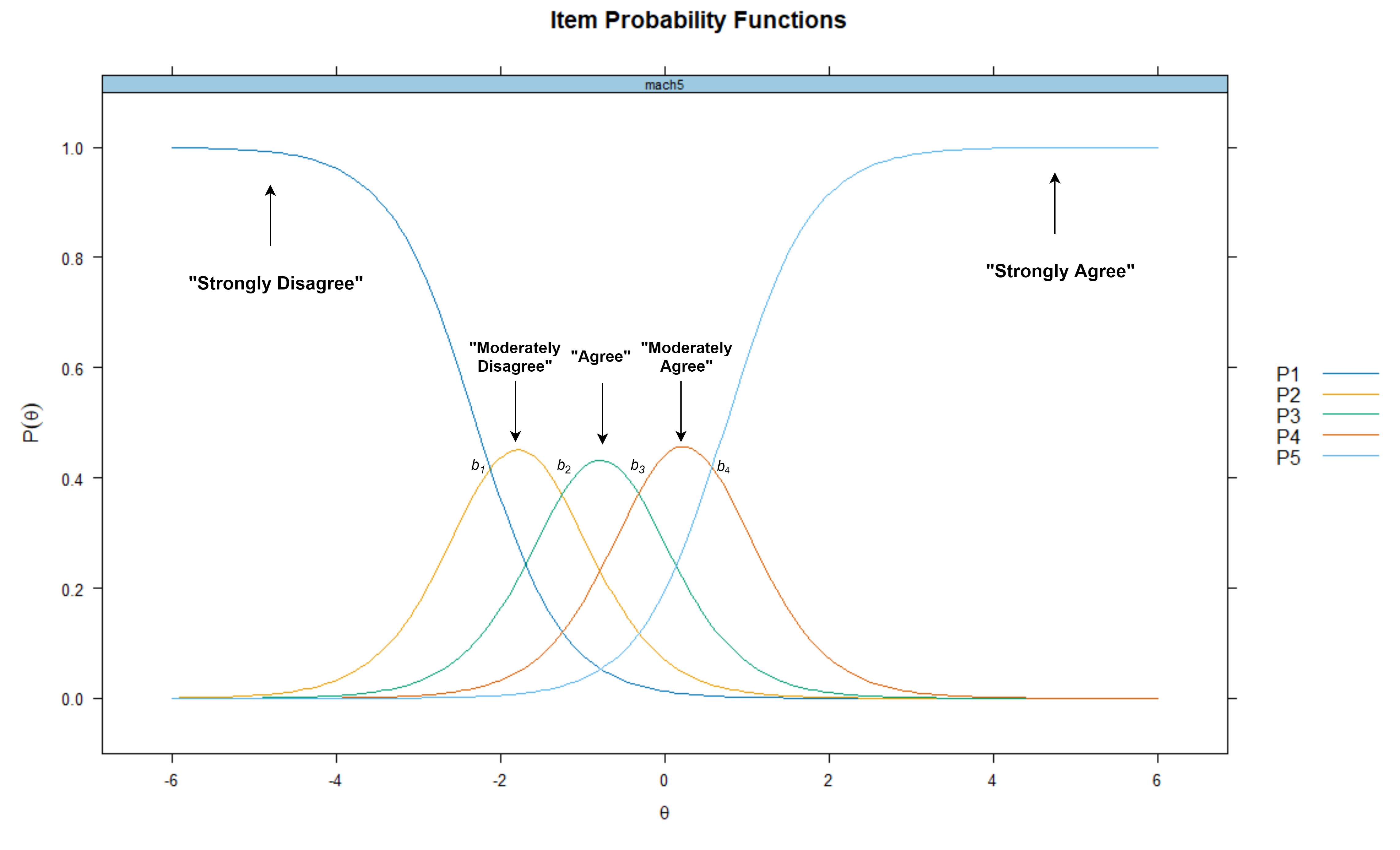


Table 2

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

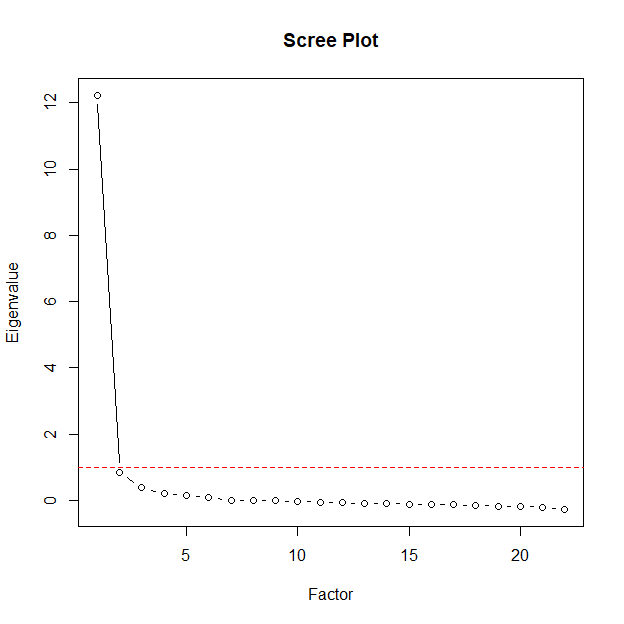


Figure 3

Parallel Analysis

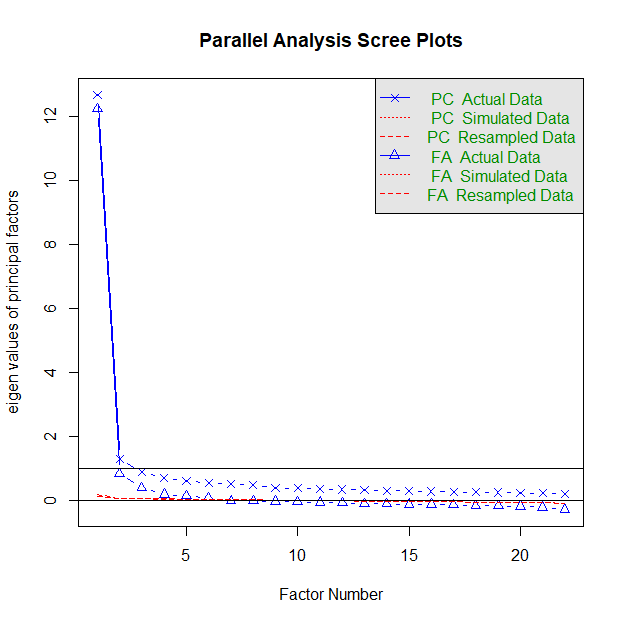


Table 4

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

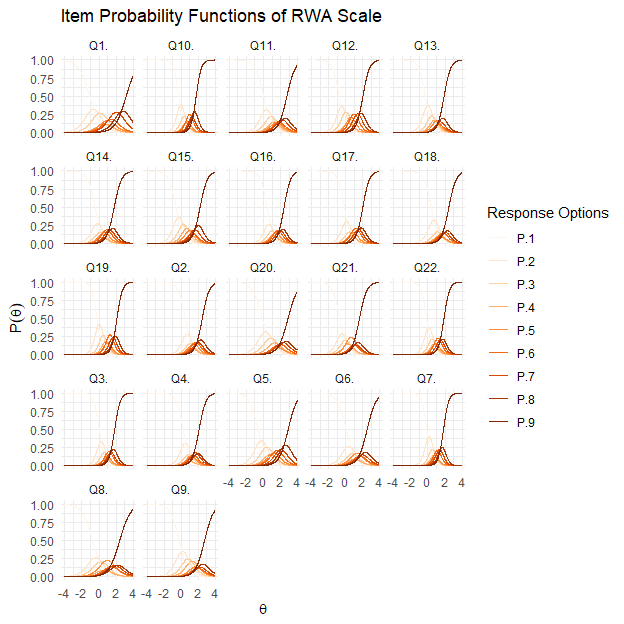


Figure 5

Item Information Curves of the RWA Scale

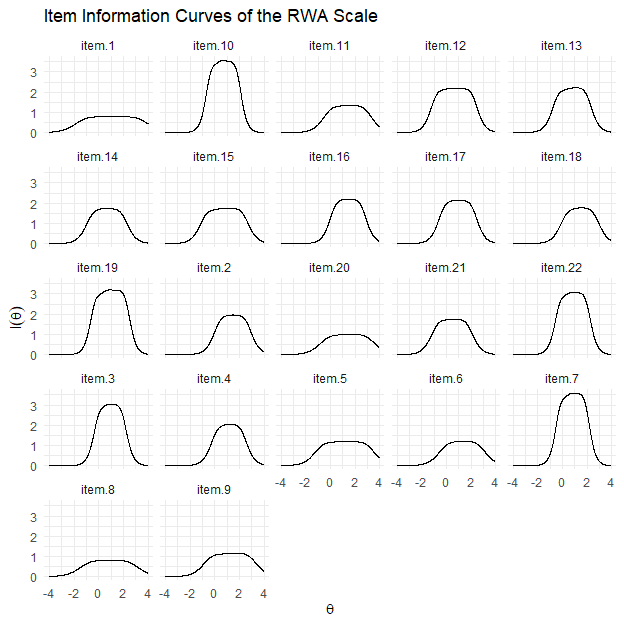


Figure 6

Test Information Curve of the RWA Scale

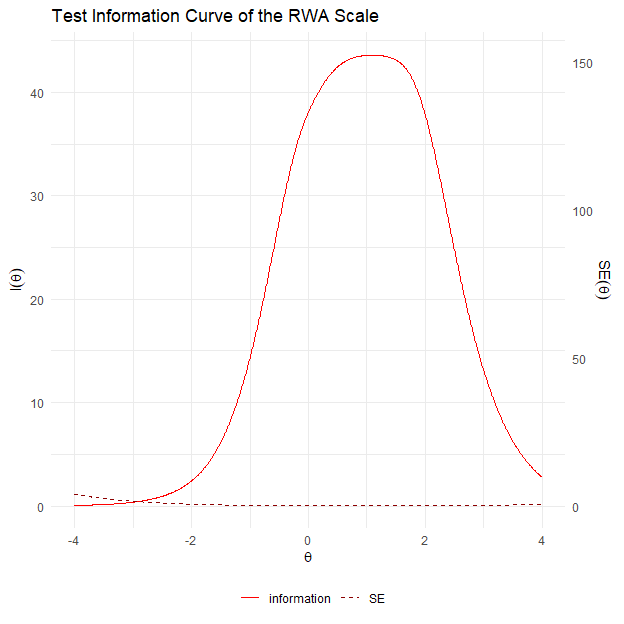


Figure 7

Reliability of the RWA Scale Given to the θ Level

