# 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 unidimensional 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 (Likert-style) 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 some common issues 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. To illustrate this process, consider 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 for applied psychology researchers, 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,” later, we name this concept a *latent trait* or a *manifest variable*. Researchers then create tools, such as scales, stimuli, tasks, and observation checklists, which are assumed to represent these latent traits.

In psychological research, where variables are often not directly observable, construct validity, which refers to the question of whether a scale measures what it is intended to measure (Cronbach & Meehl, 1955), becomes crucial. Construct validity can be scrutinized by linking test scores to other related constructs and observations (Cronbach & Meehl, 1955). 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 scores with external criteria i.e., other theoretically related traits or behavior (Zumbo & Chan, 2014).

Validity theory has evolved and has been expanded from time to time and validity is now seen not as a property of a specific test, but rather lies in the consequential aspects of 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 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., the 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) as the most popular reliability coefficient (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), which limits its usefulness (Sijtsma, 2008).

In this sense, measurement oversight can raise legitimate concerns about the replicability of research findings because measurement directly contributes to the (in)validity of the findings (Flake & Fried, 2020; Lilienfeld & Strother, 2020). To better identify how these oversights manifest themselves in the research process, we should pay more attention to decisions that researchers make that potentially compromise the validity of the measures used in their studies, a phenomenon known as “*questionable measurement practices*” (Flake & Fried, 2020). Furthermore, the lack of evidence for measurement validity contributes to the difficulty of reaching a consensus on the best way to measure a construct. In the long run, this is a major obstacle to the establishment of 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 help them assess measurement validity so that they can improve their measurement practices, and thus, the credibility of their findings.

The purpose of this article is to provide a brief introduction to the Graded Response Model (GRM), a family of item response models specifically designed to assess the measurement precision of a polytomous (Likert-style) scale. We begin this paper by briefly explaining the basics of Item Response Theory (IRT), and 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 it 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. We then 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 in 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 scores 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 dataset into latent factors, which represent true score or trait level, and error terms (i.e., residuals or unique variances). Consequently, the relationship between observed score and true score is *linear*, since 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 (), in particular, this probability takes into account the properties of the items, such as item discrimination, difficulty, and, when appropriate, 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 limits the utility of CTT, especially since psychologists often need to interpret individual scores, rather than 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 assumed to be the same across different levels of a trait or ability. This approach is unrealistic since researchers often encounter situations where a test is overly difficult for a group of low-performing participants, providing little information beyond an indication 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 so, when a test is overly challenging for low-performing individuals or too easy for high-performing individuals, IRT can help identify the levels of trait or ability at which the test is most reliable. A concrete example of this is a study that examined 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 is suboptimal for measuring individuals with low levels of sadism (Blötner & Beisemann, 2022). Therefore, IRT models offer a more informative and nuanced perspective that proves 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-parameters. The one-parameter logistic (1PL) or Rasch model, the simplest variant of the IRT models, accounts for only item difficulty (*b*) while assuming equal item discrimination (*a*) across items. In practice, this model overlaps slightly with CTT in the sense that they assume that all items have equal informative value, so that 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 measures 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 always be the same regardless of the scale used to measure their depression levels. While this principle reasonably enforces objectivity in measurement practices, it requires a strong theory delineating the construct and strict requirements of data-model fit while both of these assumptions rarely hold in a real-world scenario, especially in psychological science.

Furthermore, the two-parameter logistic (2PL) model fills this gap by allowing for differences in item ability to differentiate individuals with varying levels of (i.e., item discrimination parameter - *a*). Additionally, in some contexts, researchers may suspect that some parts 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 evolved 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; R. P. 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 responses), and the focus of this article is nonetheless to demonstrate the utility of IRT for fitting the ordered (Likert-style) responses. We briefly summarize the features of the most commonly used 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) 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*.

Figure 1 shows an example of an item probability function (IPF) from an item on 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 the assumptions 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 a 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 the 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 an IRT model (multidimensional IRT, see Bock et al. (1988), we limit the scope of our tutorial to the unidimensional GRM. A unidimensional IRT model can still be robustly applied to multidimensional data if multiple latent traits are moderately intercorrelated or when there is a strong general factor (*g*) underlying the data (Reise et al., 2015). 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 is 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 The Trust in Science Scale containing these two items: “*I trust scientists working in 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 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 by 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 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 & 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 in which participants are asked to indicate their agreement with 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 included 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 outputs) 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: Data Preparation

Before running the analysis, we need to install the necessary R packages, which are, *mirt* (P. Chalmers et al., 2023; R. P. Chalmers, 2012), *psych* (Revelle, 2023), *ggmirt* (Masur, 2023), *tidyverse* (Wickham et al., 2019), *caret* (Kuhn, 2024), and *devtools* (Wickham et al., 2022), with these following commands:

install.packages("tidyverse", # for data wrangling  
 "psych", # for descriptive statistics and unidimensionality test  
 "devtools", # for installing R package that is not available on CRAN  
 "mirt", # for conducting the main GRM analysis  
 "caret", # for helping us detecting large residuals correlation (local independence test)  
 dependencies=TRUE)  
devtools::install\_github("masurp/ggmirt") # remote ggmirt installation through GitHub repository

The above command also automatically requests to install dependency packages. Note that we install *ggmirt* directly from its GitHub repository because *ggmirt* is not yet available in the Comprehensive R Archive Network (CRAN). Next, we should activate the R packages by typing a command as follows:

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

After the packages are activated, we need to import the dataset into our R environment. The dataset (data.csv) and codebook (codebook.txt) are available in data folder in our repository. Readers can download them from our repository and then simply import the dataset with the following command:

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

However, since the dataset is openly accessible on the Open Psychometrics webpage, readers can also directly import the dataset from the Open Psychometrics webpage to their R environment. For interested readers, we show how to do this in detail in our annotated R script file.

Next, it is important to examine the structure of the data frame (ds) to get a glimpse of its structure. By typing str() command, readers can examine the variables’ name, number of cases, types of variables (integer or character), and the values of each variable. Then readers can match the information provided in their R console to the codebook.

str(ds)

According to the information provided in the codebook, the responses to the RWA scale are denoted in Q1 to Q22, each representing the score of each item of the RWA scale. Readers can also see the information of Q1 to Q22 variables in their R console: these variables are integer and have values ranging from 1 to 9. In the codebook, there is no information regarding what these values (i.e., 1-9) mean. However, we could reasonably assume that participants who chose “*strongly disagree*” (-4) were scored 1, and those who opted for “*strongly agree*” (+4) were scored 9.

Please note that we are only interested in the responses of the RWA scale, so we should subset the dataset to contain only the variables we are interested in, and then, examine the structure again. To this end, readers can type these commands:

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

Here, we have a new data frame namely rwa, which contains only the responses to the RWA scale (Q1:Q22). In the data frame rwa, we have 22 variables with 9,881 participants. Before proceeding to the analysis stage, we recommend readers to inspect descriptive statistics, which we explain the procedure in the next sub-section.

## Step 2: Inspecting Key Descriptive Statistics

To efficiently examine descriptive statistics, readers can make use of describe() function from *psych* package (Revelle, 2023) by typing a simple command:

describe(rwa)

This command summarizes key descriptive statistics of a data frame, and the summary is provided in Table 2. As we can see in Table 2, all the variables have a minimum value of 0, while the minimum score should be 1. Again, there is no information provided in the codebook regarding what 0 actually represents. Therefore, before deciding what we should do next, it is better to find out how many 0 we have in the data frame. To this end, we can calculate the percentage of 0 in each item by running these commands:

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

After executing these commands, readers can view the percentage of 0 in each items in their R console, which overall are very small, ranging from 0.09% to 0.32% of the total cases. Since the percentages are very small, 0 is likely a code for a missing value. It is important to note that *mirt* package uses the Expected-Maximization (EM) algorithm to estimate model parameters, and the EM algorithm can handle missing values (R. P. Chalmers, 2012). If we decide to eliminate cases with missing values, we still have a sizable number of cases, nonetheless. Therefore, to simplify our analysis, we only keep cases with complete responses by executing these commands:

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

If readers prefer to keep cases with missing values, especially when the sample size is low, readers can exclude drop\_na() command.

Interestingly, if we look at Table 2 again, the mean scores of the items seem to differ drastically. Some items have high mean scores, ranging from 6 to 7, but some of those are comparatively low. The RWA scale indeed has some unfavorable items but, again, there is no information in the codebook whether these unfavorable items have been reversely coded. However, if we look closely at ten items with high mean ( > 6), these are the items that are supposed to be unfavorable. Therefore, we can reasonably assume that unfavorable items have not been reversed in the original dataset, and we should do so by running these commands:

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.

First, we define a vector namely unfav to identify which items should be reversed, and then we use mutate() function to reverse the score. This is simply done by subtracting the values of the unfavorable items from the maximum score plus one. We can examine descriptive statistics of our modified data frame by running describe() function again, and as shown in Table 3, the mean score across items does not drastically vary and has minimum value of 1.

## Step 3: Examining Dimensionality

Next, we show the readers the ways of testing unidimensionality. To test the number of latent factors of an item pool, De Ayala and Hertzog (1991) suggest three alternatives, which are multidimensional scaling (MDS), EFA, and CFA. In this tutorial, we demonstrate the use of EFA, and *psych* package provides a very efficient way to do this by calling this command:

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

irt.fa function is to run an EFA with a polychoric correlation matrix as an input, not a Pearson’s correlation matrix, because our data is ordinal. The option fm = "minres" implies that we choose minimum residual as a factoring method due to its effectiveness in handling non-normally distributed data (Harman & Jones, 1966). Alternatively, readers can also run a parallel analysis with a polychoric correlation matrix by calling the following command:

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

A parallel analysis compares the eigenvalues from the data with those yielded from a randomly generated dataset with the same sample size to determine the number of factors one should retain in EFA (Guo & Choi, 2023). The codes indicate that we tell R to run a parallel analysis on the data frame rwa, with one factor. The option fa="fa" implies that we want to show the eigenvalues for an EFA, while cor = "poly" means that we use a polychoric correlation matrix as an input.

The call to both commands above produces the following error: Error in polychoric(x): You have more than8 categories for your items, polychoric is probably not needed. Since we have more than eight response categories, we should instead run an EFA with Pearson’s correlation matrix as an input. Therefore, we first need to create a Pearson’s correlation matrix from our data frame, and then, run an EFA by calling these following commands:

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

After calling the commands, readers can find Proportion Var in the output, which reflects the amount of variance explained by a latent factor. The variance proportion of our EFA model is 0.56, which means that 56% of the total variance is explained by a latent factor. Since the latent factor can account for more than 50% of the variance, we can reasonably assume that the latent factor underlying our data is unidimensional. To better visualize the Eigenvalues of the EFA, we can draw a scree plot using the following commands:

plot(efa$values, type = "b", main = "Scree Plot", xlab = "Factor", ylab = "Eigenvalue") # Scree plot  
abline(h = 1, col = "red", lty = 2) # Add new line to eigenvalue = 1.

The first line of code is a function to plot the Eigenvalues (efa$values, x-axis) given the number of latent factors (y-axis). The option type = "b" implies that we ask R to display the Eigenvalues exact points and a line connecting these points along the number of factors. The second line of code aims to display a horizontal red line (i.e., a reference line) to see which factors have eigenvalues greater than one, which is a common criterion for factor retention. Figure 2 shows that there is a clear break or an “elbow” that displays the location where the Eigenvalues start to level off. This implies that factors beyond the first may not be meaningful and could be only noise. Therefore, the scree plot strengthens our assumption that the item pool is, indeed, unidimensional.

To scrutinize the unidimensionality assumption, we can also run a parallel analysis by running the following command:

pa <- fa.parallel(rwa, nfactors = 1, fm = "minres", fa = "fa")

Readers should find a plot similar to Figure 3 appear right after executing this code as well as a message, which says: Parallel analysis suggests that the number of factors = 6 and the number of components = NA. This means that the parallel analysis suggests that there are six factors underlying our item pool, not one. However, according to Figure 3, we can see clearly that the Eigenvalues from our actual data fall steeply after the first factor. To paint a clearer picture, we can examine the Eigenvalues of each latent factor estimated by the parallel analysis by running the following command:

pa$fa.values

The output shows that the Eigenvalue for the first factor is much larger than the the adjacent factor (12.225/0.843), further supporting the evidence for unidimensionality.

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

In this step, we start to specify a vector representing our model specification, such as:

model <- 'rwa = 1-22'

This code implies that we want to estimate a model with namely rwa, and the items representing the are columns 1 to 22 of the data frame that we are going to use in the analysis. Then, we can continue to model fitting by typing the following command:

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

The code above implies that we are fitting an IRT model using data from a data frame namely rwa (data = rwa). The number 1 indicates that we presume that the model contains only one , i.e., unidimensional. We also set the model specification to model = model, which we have previously determined. We set itemtype to "graded" because we want to run a GRM model. The SE option is set to TRUE because we ask R to also estimate the standard error. At last, we set the verbose option to FALSE to avoid R displaying unnecessary information during the model fitting process, so the mirt() function can run quietly.

Apart from this tutorial, readers might encounter four common error messages after running mirt() function: (a) EM cycles terminated after 500 iterations, (b) Could not invert information matrix; model may not be empirically identified, (c) The following items have only one response category and cannot be estimated, and (d) Error in [<-(\*tmp\*, i, nms[[i]], value = items.old[[i]]) : subscript out of bounds. Error messages (a) and (b) correspond to the convergence issue, while (c) and (d) are issues related to the data structure.

*mirt* uses the EM algorithm as a default option. In short, the EM algorithm estimates model parameters in two steps: the expectation (E) and the maximization (M) steps. In the E step, the algorithm estimates the expected value of the , and in the M step, it improves model parameters to maximize the likelihood of the observed data given the expected value of . The process continues until model parameters are converged in stable values that fit the data. If the error message (a) appears, it means that the package cannot estimate a stable solution for model parameters after 500 iterations. As a solution, readers can simply increase the maximum number of iterations to, for instance, 1000 cycles by adding this option technical = list(NCYCLES = 10000) to the mirt() function.

The solution to the error message (b) is less straightforward than (a) because it indicates that the model cannot be identified at all, which can have several causes. It could be due to an error in specifying assumptions about the number of latent traits or item properties, or it could be due to insufficient sample size. Therefore, when error message (b) appears, it is important to take a step back and think twice about the model assumptions, the fit between the model and the underlying theory, and whether the sample size is large enough to estimate the model properly.

The error message (c) and (d) are easier to solve because both are related to the data structure. The former means that certain items have no variance, consisting only one response category. The solution for this issue is to exclude these items from model estimations. The latter indicates a mismatch between the vector of model specification and the data frame, for instance, specifying a vector with ten variables, while in fact, the data frame has only nine columns. To solve this issue, readers are suggested to check again whether the specified vector representing the model specification matches the structure of the data frame.

Next, we can ask the package to calculate item discrimination and thresholds by calling these functions below:

coefs <- coef(fit, IRTpars = TRUE, simplify = TRUE) # Saving model parameters (item discrimination (a) and threshold (b)) in a data frame.  
print(coefs) # Calling the data frame.  
summary(fit) # Displaying factor loadings and commonality.

The first line of code indicates that we are calling a coef() function to calculate item parameters from fit model, and then, store them in a new data frame called coefs. The IRTpars and simplify option is set to TRUE because we want to display the traditional IRT parameters (*a* and *b*) in a simple, readable format without showing the standard errors.

The output from this procedure is presented in Table 4, which shows item discrimination (*a*) of each item and item thresholds (*b1-b8*) of each response category within an item. To interpret *a*, readers can use a rule of thumb suggested by Baker and Seock-Ho (2017), which categorizes item discrimination exceeding 0.00, 0.35, 0.65, 1.35, and 1.70, as very low, low, moderate, high, and very high. As we see in Table 4, item discrimination parameters of the RWA scale range from high to very high (1.57 (Q1) to 3.32 (Q7)), suggesting efficient differentiation among individuals at different RWA levels. All threshold parameters show a consistent increase, implying that those with higher RWA levels tend towards a higher response categories. However, *b1* and *b2* for all items are very close to zero, which indicates that individuals with the RWA level around the mean are likely to opt for “*strongly disagree*” or stepping to its adjacent response category. This might suggest that the RWA scale is less sensitive to individuals with a low or very low RWA level.

The third line of code is a function to calculate a factor loading () and commonality (h2) of each item, which are also provided in Table 4. Before estimating a graded response model, *mirt* ran an EFA, and now as we can see in the console, we are looking at the EFA results. The results are slightly different from our previous EFA analysis, because *mirt* runs EFA using a quasi-polychoric correlation matrix, while the one we ran earlier to test unidimensionality used a Pearson’s correlation matrix as an input. However, most importantly, we see that all items are significantly loaded to one factor, and the factor now substantially accounts for 65.1% of the variance in the data, which strengthens our assumption that the RWA scale is unidimensional.

Next, we can estimate the model goodness-of-fit (GOF) statistics (*M*2) by calling this command below:

M2(fit, type = "C2")

We ask the package to calculate the scaled *M*2 statistics (Maydeu-Olivares & Joe, 2006) and other GOF indices, such as Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Standardized Root-Mean-Square Residual (SRMSR) by running this code above. Here, we set the type option to "C2" because it is suitable for computing *M*2 statistics in polytomous models. The interpretation of GOF statistics is similar to a general guide for interpreting GOF in CFA or SEM models (Schermelleh-Engel et al., 2003).

According to our analysis, the model overall does not fit the data well (*M*2(209) = 1.980251^{4}, *p* < .005, *RMSEA* = 0.098, *SRMSR* = 0.055, TLI = 0.969, CFI = 0.972), and this may be caused by local dependency between the items. We will explore this issue further in the next part.

Next, we calculate item-level fit statistics, i.e., the scaled 2 statistics, by running the following function:

item.fit <- itemfit(fit) # Estimating item fit statistics and saving them in a data frame.  
print(item.fit) # Calling the data frame.

The output of this function is provided in Table 5, which shows fit statistics for an individual item. All items have very good (low) *RMSEA* values but normally, we want the *p* values of the scaled 2 to be nonsignificant, implying that there is no discrepancy between the observed response pattern with what the model predicts. In Table 5, only three items here are nonsignificant (Q2, Q4, and Q12), which strengthens our suspicion that the RWA items might not be locally independent.

## Step 5: Model Residuals

To ensure that the items are locally independent, *mirt* offers several alternatives to examine the behavior of residuals. First, we demonstrate how to run Local Dependency (LD) 2 statistics (Chen & Thissen, 1997), which looks at the covariance between item pairs after accounting for . To that end, readers can run two lines of codes as follows:

ld <- residuals(fit, type = "LD") # Running local dependency (LD) statistics.  
up <- which(upper.tri(ld), arr.ind = TRUE) # Creating a new matrix containing values only on the upper side of the diagonal (correlations).

The first line of code aims to extract residuals of fit model and store them as a matrix, namely ld. The type option is set to "LD" because we ask the package to calculate LD statistics of the fit model. After running the first line, readers may view a matrix containing residual covariance (in the lower diagonal) and correlations (in the upper diagonal). Correlations are more interpretable, so we focus on the upper side of the diagonal. Therefore, the second line of code is to create a new matrix containing only correlation values in the upper diagonal.

We are only interested in large correlations between a pair of items because they indicate that responses to one item depend on responses to another item beyond what is explained by . It is important to note that there is no consensus on a rule of thumb for the cutoff value of residual correlations because LD statistics are contingent on the parameters of a specific dataset (Christensen et al., 2017). For the purpose of this tutorial, we set |0.2| as a paired correlation value that we want to look closely at. To do so, we create a vector to define the value of residual correlations that we are interested in:

lar <- up[ld[up] > 0.2 | ld[up] < -0.2, ] # A vector defining unusually large residuals ( > |0.2|).

We can then use the new vector lar to ask R to tell us the item pairs with residual correlations above |0.2| by running this function:

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.

After running the function above, readers may encounter an error message: Error in lar[i, 1] : subscript out of bounds. This message indicates that R does not detect any residual correlation above |0.2|. To paint a clearer picture, we may need to use another way to examine model residuals. In the next part, we demonstrate the use of Yen’s *Q*3 statistics (Yen, 1984), which is considered more powerful for detecting underlying local dependency (Chen & Thissen, 1997). To run Yen’s *Q*3, readers can type and run this command:

q3 <- residuals(fit, type = "Q3") # Running Yen's Q3 statistics.

This code asks R to run residuals() function using "Q3" as the type, and then, keep the value as a new matrix, namely q3. Again, we are only interested in item pairwise correlations above |0.2|, and to help us flag problematic pairs, we can make use of findCorrelation() function from *caret* package (Kuhn, 2024) as follows:

findCorrelation(q3, cutoff = 0.2, verbose = T) # Flagging problematic item pairwise correlations.

After running the code, readers may see in their R console that residual correlations between item Q3, Q4, Q5, Q7, Q11, Q13, Q14, Q18, Q19, and Q21 are above |0.2|. Interestingly, all of those items, except for Q4, have nonsignificant *S*-2 statistics (see Table 5) as well, which tells us that we have to look at these items closely and then decide whether these items have shared characteristics beyond what is explained by the model.

## Step 6: IRT Plots

*mirt* offers features to visualize model parameters but the options for customizing these plots are rather limited. *ggmirt* package (Masur, 2023) fills this gap by combining *mirt* and *ggplot2* functions. In this tutorial, we demonstrate the use of *ggmirt* to visualize Item Probability Function (IPF), Item Information Function (IIF), Test Information Function (TIF), and conditional probability plot.

To visualize IPF, readers can type and run this function:

tracePlot(fit, title = "Item Probability Functions of RWA Scale") + labs(color="Response Categories")

The labs(color) function is to assign different colors to each curve representing a response category. The output of this function is Figure 4. In Figure 4, we see that as increases, respondents are more likely to choose higher response categories. The gradual transition from one response option to the next across the spectrum of values indicates that the items are capturing incremental increases in the RWA level. However, all IPFs of RWA items seem to be significantly overlapping and tend to peak on a value close to or higher than the mean. This implies that RWA items are more sensitive to differentiate participants with high levels of RWA.

To evaluate the performance of each item in measuring the RWA trait, we can visualize the amount of information explained by each item with this simple line of code:

itemInfoPlot(fit, facet = TRUE, title = "Item Information Functions of the RWA Scale")

The facet option is set to TRUE so that R displays IIF for each item. The output for this function is Figure 5, where the x-axis represents the range of , and the y-axis indicates the amount of information provided by each item. Peaks that are higher and narrower indicate items that are very informative for specific levels of . In our case, most items provide the highest amount of information near the center to the right side of the distribution, indicating they are most useful for individuals with average to high levels of RWA.

It is also possible to evaluate the overall performance of the RWA scale by plotting a TIF, which can be done by running the following code:

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

The output of this code is Figure 6, and according to this, the RWA scale as a whole is informative for measuring a group of individuals with a wide range of RWA levels, i.e., between -2*SD* to +4*SD*. However, the RWA scale is not optimal for measuring individuals with RWA levels below or beyond this range.

## Step 7: Computing Reliability

The heart of an IRT analysis is to estimate the precision of the measure, which can vary given different levels of . *mirt* provides two strategies to compute reliability that differ in their assumption regarding the distribution of . First, readers may calculate marginal reliability by running this simple command:

marginal\_rxx(fit)

The output of the above code is 0.948, which indicates that the overall RWA scale is reliable, assuming that the underlying distribution follows the Gaussian or normal distribution (R. P. Chalmers, 2019). Since reliability can vary across different levels of , readers can also visualize the scale’s reliability given the levels of by running the function of *ggmirt* package, as follows:

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

The output of this code is Figure 7, which shows that the RWA scale can measure individuals with levels between -2*SD* and +4*SD* with sufficient reliability, i.e., *r*xx 0.75.

Second, *mirt* can also calculate the RWA scale overall reliability using predicted by the model. To compute empirical reliability, one needs to calculate the estimated value of each participant first, and then, calculate reliability based on these values. To do this, readers can run these two lines of codes:

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

The fscores() function is used to compute the estimated value for each participant predicted by the fit model. The full.scores.SE is set to TRUE because the package uses the value and standard error of each participant (*SE*) to compute the empirical reliability. We store the estimated and its *SE* in a new data frame, namely theta\_se, and then run empirical\_rxx() function to calculate empirical reliability. The output of these codes is 0.953, which is very close to marginal reliability we estimated before. Notably, while marginal and empirical reliability aim to quantify the overall reliability of the scale, they base the calculations on different assumptions. Marginal reliability assumes that is normally distributed, and empirical reliability uses the distribution predicted by the model, which still contains measurement error but is more realistic. Note that our model has a poor fit, perhaps due to local dependence, as we showed in steps 4 and 5. In practice, we recommend that readers address model-fitting issues first before interpreting reliability or drawing substantive conclusions based on these estimates.

To paint a more comprehensive picture, we can compare IRT-based reliability calculations with unweighted sum-score-based reliability, such as Cronbach’s and McDonald’s . To do this, we can use omega() function provided by *psych* package (Revelle, 2023), which can be done by running this command:

omega(rwa)

The output of omega() function includes several information but for the purpose of this tutorial, we only need to focus on Cronbach’s and McDonald’s h. Both Cronbach’s ( = 0.964) and McDonald’s (h = 0.85) shows that the RWA scale is overall internally consistent.

# Discussion

Construct validation is an ongoing process (Messick, 1995), so reporting evidence of validity should always be an important part of the routine. In practice, however, researchers are less interested in examining the performance of the measures they use in their study, making it difficult to assess the integrity of their research findings (Flake et al., 2017; Hussey & Hughes, 2020). In this tutorial, we demonstrate the applicability of GRM analysis as part of the IRT family to help applied psychology researchers assess their measurement quality. We provide a non-technical guide to implementing a GRM analysis through a simple 7-step process using a real, openly available dataset. To maximize the effectiveness of this tutorial, we show how to perform a GRM analysis using R (R Core Team, 2023), an open source statistical software, and make all materials publicly available. We also supplement our tutorial with some tips for troubleshooting common problems when running the GRM analysis.

We begin the tutorial with a theoretical overview of IRT, which underlies GRM, so that readers can relate the practical steps to the theory behind the analysis. Since IRT is less popular than CTT, we also briefly explain how IRT differs from CTT in its assumptions. However, we want to emphasize that our goal here is not to argue for the superiority of CTT over IRT, since the choice of analytic tool depends largely on the specific research question at hand. It is important to highlight that researchers should be aware of the merits and limitations of their chosen methods. Therefore, we argue that researchers should always justify why they choose a particular method over many available alternatives.

We agree with the call for psychological researchers to shift their focus from improving their methodological practices to advancing the process of theory building (Eronen & Bringmann, 2021; Grahek et al., 2021; Proulx & Morey, 2021). Therefore, theory building should precede the creation of a psychological measure. It is important to note, however, that construct validation is a critical step in the iterative process of theorizing a psychological phenomenon (Grahek et al., 2021). Therefore, we encourage applied psychology researchers to report evidence of measurement quality to ensure the validity of their findings and contribute to cumulative psychological science (Flora & Flake, 2017). One way to do this is to conduct and report a GRM analysis of their measures. We hope that this tutorial can be seen as a small step towards the larger goal of building robust psychological science.

# Conclusion

Applied psychology researchers strive to advance psychological science by producing robust and replicable findings. To do so, they should begin by carefully defining the construct, developing an appropriate research design, and, most importantly, creating a measure that accurately captures the phenomenon of interest. In this tutorial, we provide a non-technical guide that we hope will help applied psychology researchers critically evaluate and improve the quality of their measurement instruments. By using a real, openly available dataset, we provide a realistic scenario for evaluating the performance of a psychological scale. By making the materials publicly available, this tutorial paper facilitates applied psychology researchers to easily implement the GRM in their own research project.

# 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, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2023). *Mirt: Multidimensional Item Response Theory*.

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>

Chalmers, R. P. (2019). Answer to "Difference between empirical and marginal reliability of an IRT model". In *Cross Validated*.

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>

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen’s Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas*, *41*(3), 178–194. <https://doi.org/10.1177/0146621616677520>

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.

De Ayala, R. J., & Hertzog, M. A. (1991). The Assessment of Dimensionality for Use in Item Response Theory. *Multivariate Behavioral Research*, *26*(4), 765–792. <https://doi.org/10.1207/s15327906mbr2604_9>

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.

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspect Psychol Sci*, *16*(4), 779–788. <https://doi.org/10.1177/1745691620970586>

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>

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a Psychological Theory: Integrating Construct-Validation and Computational-Modeling Methods to Advance Theorizing. *Perspect Psychol Sci*, 1745691620966794. <https://doi.org/ghtczb>

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.

Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, *31*(3), 351–368. <https://doi.org/10.1007/BF02289468>

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>

Kuhn, M. (2024). *Caret: Classification and Regression Training*.

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>

Masur, P. K. (2023). *Ggmirt*.

Maydeu-Olivares, A., & Joe, H. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. <https://doi.org/10.1007/s11336-005-1295-9>

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>

Proulx, T., & Morey, R. D. (2021). Beyond Statistical Ritual: Theory in Psychological Science. *Perspect Psychol Sci*, *16*(4), 671–681. <https://doi.org/10.1177/17456916211017098>

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>

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the Impact of Multidimensionality on Unidimensional Item Response Theory Model Parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 13–40). Routledge.

Revelle, W. (2023). *Psych: Procedures for Psychological, Psychometric, and Personality Research*.

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>

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

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>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

Wickham, H., Hester, J., Chang, W., Bryan, J., & RStudio. (2022). *Devtools: Tools to Make Developing R Packages Easier*.

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(2), 125–145. <https://doi.org/10.1177/014662168400800201>

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

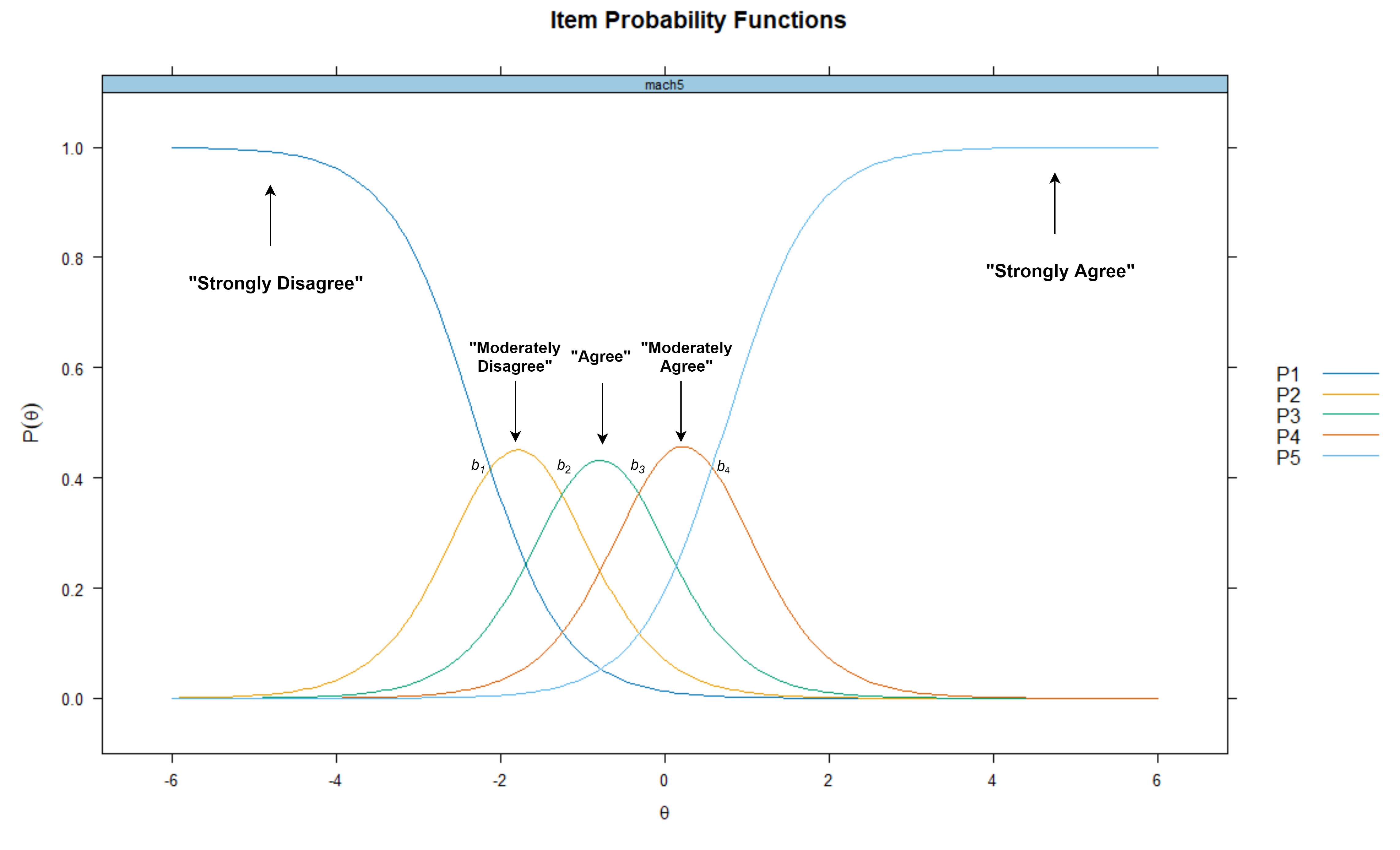


Table 2

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

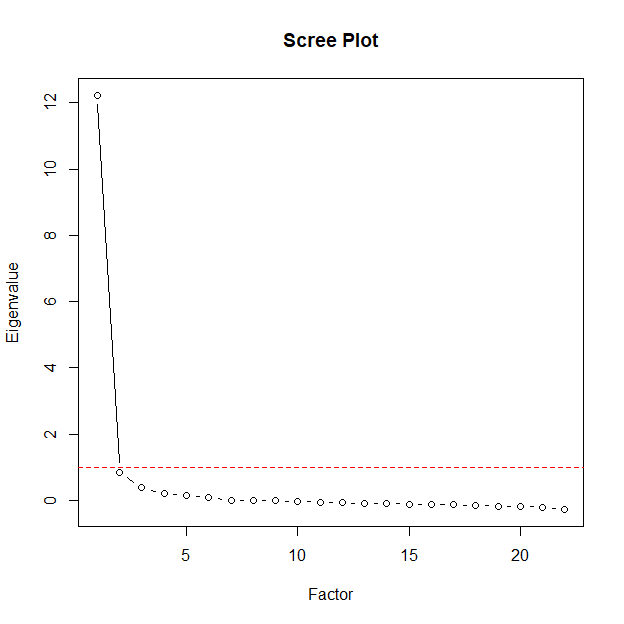


Figure 3

Parallel Analysis

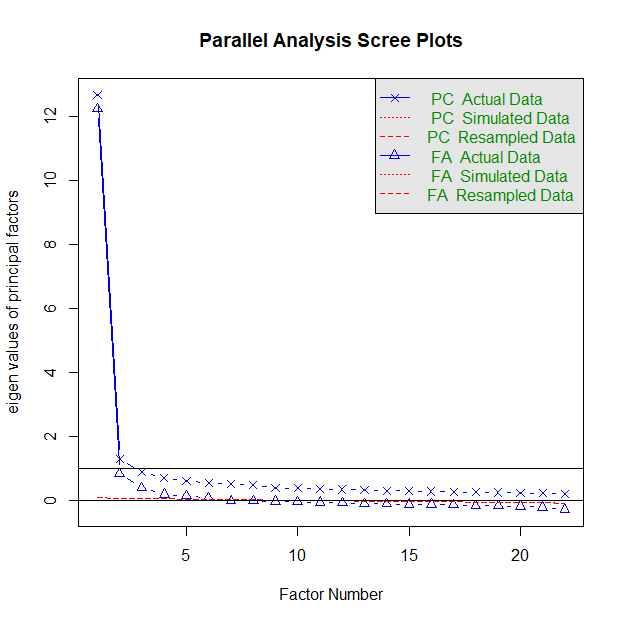


Table 4

Item Parameters

| Item | a | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | 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*. a = Discrimination, b1-b8 = Response specific threshold parameters, λ = Standardized Factor Loadings, h2 = Commonality.

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

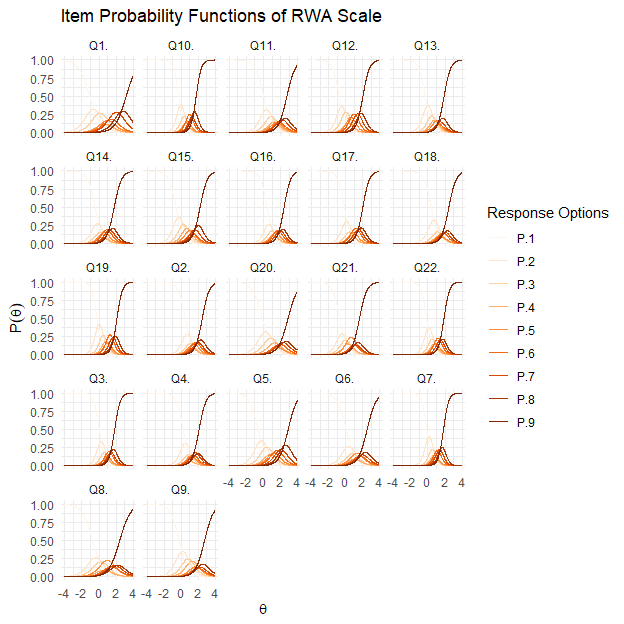


Figure 5

Item Information Curves of the RWA Scale

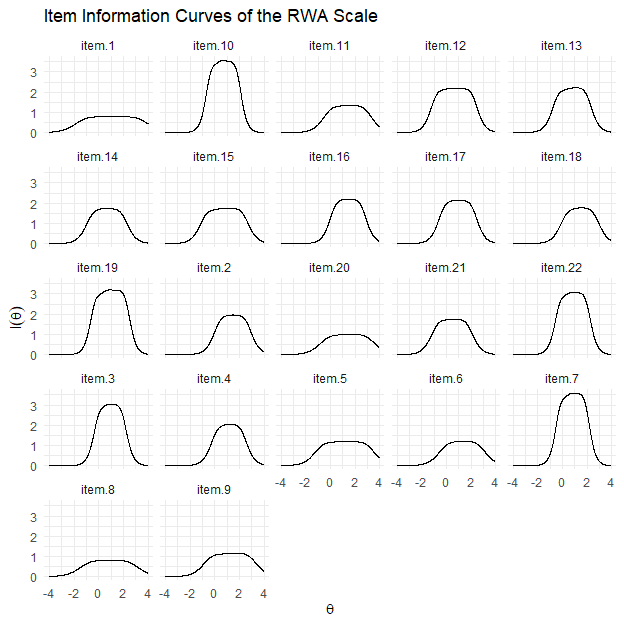


Figure 6

Test Information Curve of the RWA Scale

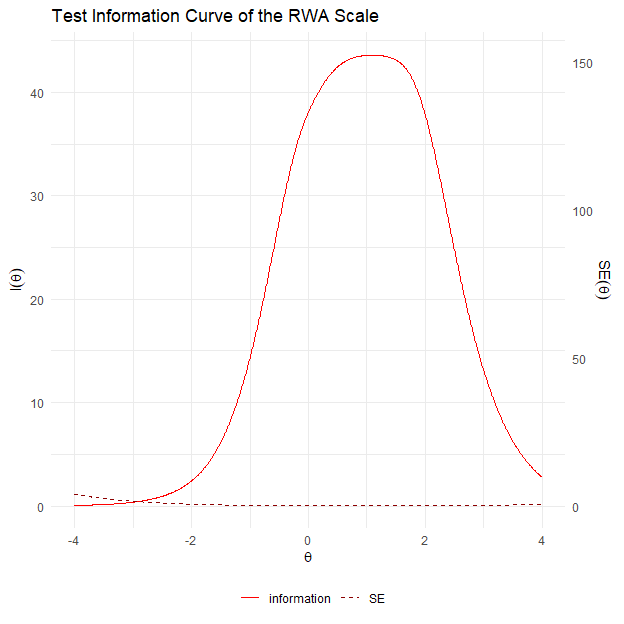


Figure 7

Reliability of the RWA Scale Given to the θ Level

