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

Rizqy Amelia Zein1,2 and Hanif Akhtar3

1Department of Psychology, Ludwig-Maximilians-Universität

2Department of Psychology, Universitas Airlangga

3Faculty 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 that we have no competing interests. The first author receives a PhD scholarship from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD), award number 91803023, and thanks DAAD for its financial support.

Correspondence concerning this article should be addressed to Rizqy Amelia Zein, Department of Psychology, Ludwig-Maximilians-Universität, Leopoldstraße 13, Munich, Germany 80802, Email: [Amelia.Zein@psy.lmu.de](mailto:Amelia.Zein@psy.lmu.de)

# Abstract

This tutorial introduces the Graded Response Model (GRM), a tool for testing measurement precision within the Item Response Theory (IRT) paradigm, which is useful for informing researchers about the item and person properties of their measurement. The tutorial aims to guide applied researchers through a unidimensional GRM analysis in the R environment, using the 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, testing model assumptions, model fitting, plotting item parameters, and interpretation of results.

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

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

To examine the quality of psychological measures, applied researchers implement statistical or formal mathematical models, which always require specific testable assumptions. Classical test theory (CTT) is one of the model-based measurement theories widely used to examine measurement precision, i.e., how accurate a measure is at capturing underlying psychological constructs and minimizing measurement error. While CTT is useful for informing researchers about measurement precision at the sample level, some want to look closely at specific individuals in their sample or items in their scale. In this sense, item response theory (IRT) fills this need by allowing researchers to inspect their measurement quality at the person and item level. IRT models are nonetheless known to be computationally intensive and less practical than CTT-based analysis, so a tutorial with didactic value can potentially help applied researchers to consider IRT analysis as a tool for their research.

The purpose of this tutorial is to provide a brief introduction to the graded response model (GRM), a family of IRT models specifically designed to assess the measurement precision of a polytomous (Likert-style) scale. In this paper, we aim to keep our tutorial concise, so we only briefly introduce the basics of IRT concepts. Interested readers are referred to available didactic texts on IRT, such as Embretson and Reise ([2000](#ref-embretsonItemResponseTheory2000)) for the introductory level, and Baker and Seock-Ho ([2017](#ref-bakerBasicsItemResponse2017)) and De Ayala ([2022](#ref-deayalaTheoryPracticeItem2022)) for a more technical and comprehensive overview.

First, we briefly introduce readers to IRT and explain what GRM is and how the model is specified. In this part, we highlight how model and item parameters are specified and interpreted as well as visualizations derived from a GRM analysis. We also introduce three assumptions underlying a GRM and how to test these assumptions. We then provide an illustrative example of testing measurement precision using a GRM analysis, and in this part, we show how to practically implement a GRM analysis and interpret its results. We complement the illustrative example with code in the open-source programming language R ([R Core Team, 2023](#X9908c3c11d2ce7c45992bd046a78b3587daa4e5)) 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](#ref-navarroPsychologicalScience2018)). Finally, 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 on two theoretical frameworks underlying psychological testing, CTT and IRT, to decompose observed or raw scores into their *deterministic* (i.e., “true” score) and *random* (i.e., measurement error) components ([Zumbo, 2006](#ref-zumboValidityFoundationalIssues2006)). In practice, researchers implementing CTT principles apply structural equation modeling (SEM), exploratory factor analysis (EFA), or confirmatory factor analysis (CFA). Specifying the measurement model of an SEM model, or fitting EFA and CFA models, allows the covariance matrix of a given dataset to be decomposed into latent constructs () representing true scores or actual skill or trait levels, with measurement error (i.e., residuals or unique variances).

In this sense, CTT generally assumes that the relationship between observed score and true score is *linear*, as true score is essentially a linear transformation of observed score ([Embretson & Reise, 2000](#ref-embretsonItemResponseTheory2000)). However, IRT models this relationship differently 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 (). The probabilistic relationship in an IRT model accounts for the statistical properties of the items, such as item discrimination, difficulty, and, if necessary, pseudo-guessing parameter ([Baker & Seock-Ho, 2017](#ref-bakerBasicsItemResponse2017)). IRT methods were originally developed to evaluate traditional aptitude and achievement tests that measure knowledge and skill levels, but the application of IRT has been extended to personality, attitude, and other psychological scales of various formats.

The most critical difference between CTT and IRT is the way each framework conceptualizes measurement error. CTT uses standard deviation of the observed score (*s*) to calculate standard error of measurement. Because measurement error is based on the observed score, CTT assumes that measurement error is sample-dependent and constant for all individuals in the sample, regardless of their level. This shortcoming limits the utility of CTT, especially since psychologists often need to interpret individual scores, not just the test or the sample as a whole. In addition, because the standard error of measurement is also a function of reliability, it is assumed that scale reliability will always be the same across different levels of . This is unrealistic because researchers often encounter situations in which a scale is “too hard” to endorse for a group of “low-performing” (i.e., low ) participants, providing little information beyond an indication that their level is much lower than what the scale can measure.

IRT offers a solution to this issue by enabling the calculation of standard error for each individual (*SE*), and test reliability can be inferred from the average of *SE* across a sample of participants ([Embretson & Reise, 2000](#ref-embretsonItemResponseTheory2000); [Lang & Tay, 2021](#ref-langSciencePracticeItem2021)). This approach allows for the estimation of reliability at varying levels of (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 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](#ref-blotnerDarkTriadDead2022)). 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](#ref-blotnerDarkTriadDead2022)). Therefore, IRT models offer a more informative approach, thus are useful for researchers who wish to closely examine the performance of their measures at the person level.

Further, IRT allows researchers to examine the performance of a specific item by specifying the relationship between item score and given one-, two-, or three-parameter. One-parameter logistic (1PL) or Rasch model accounts only varying 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](#ref-langSciencePracticeItem2021); [Stemler & Naples, 2021](#ref-stemlerRaschMeasurementItem2021)). In this sense, 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, be it Beck Depression Inventory (BDI) or Patient Health Questionnaire-9 (PHQ-9). 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 reality.

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 pseudo-guessing parameter (*c*).

Further, IRT paradigm has rapidly developed to include various models suited to specific contexts, such as handling ordinal responses ([Muraki, 1992](#ref-murakiGeneralizedPartialCredit1992); [Samejima, 1997](#ref-samejimaGradedResponseModel1997)), categorical responses ([Thissen et al., 2013](#ref-thissenNominalCategoriesItem2013)), or assessing multidimensional traits simultaneously ([Bock & Aitkin, 1981](#ref-bockMarginalMaximumLikelihood1981); [Chalmers, 2012](#ref-chalmersMirtMultidimensionalItem2012)). 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 distinctions between 1PL, 2PL, 3PL, and GRM model in {apatb-table1}.

| 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 |

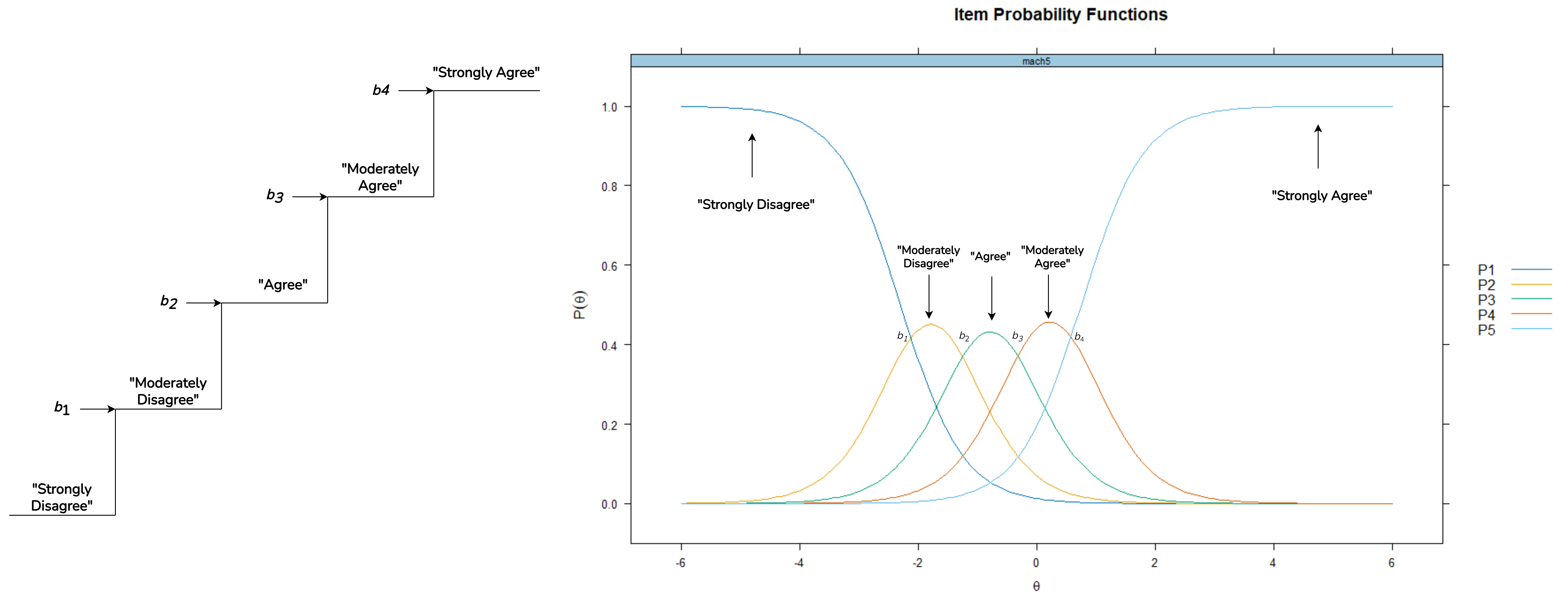
# Graded Response Model

GRM is a family of IRT models specifically designed to analyze ordered polytomous (e.g., Likert-style) data ([Samejima, 1997](#ref-samejimaGradedResponseModel1997), [2010](#ref-samejimaGeneralGradedResponse2010), [2016](#ref-samejimaGradedResponseModels2016)). GRM generally can be applied to a different types of test formats, such as scales measuring behavioral outcomes ([Arias et al., 2016](#X6abcd1fc08bf7c969323376cca841272c783521)), personality and attitude scales ([Rauthmann, 2013](#ref-rauthmannInvestigatingMACHIV2013)), or patient-reported outcome measures ([Normand et al., 2006](#ref-normandGradedResponseModelbased2006)). Situational judgment tests (SJTs), which measure non-cognitive abilities, decision-making, and problem solving in context-specific situations ([Corstjens et al., 2017](#X88412507392387ffab3be0d42015f55cf7770fc)), may also be suitable for GRM analysis when the response categories are ordered.

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 two-parameter logistic (2PL) IRT models, the probability of answering a correct response is modeled as a function of , item difficulty (*b*), and item discrimination (*a*). GRM extends this idea by modeling the probability of selecting certain response category or higher on an item (i.e., *step function*) given to item difficulty (*b*) and discrimination (*a*). 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 - *b*) 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 5 (*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 at which a participant is equally likely to respond to either of the adjacent response categories. Note that the level of the latent trait () that a person has is typically centered around a mean of zero. The mean represents the average level of the trait across all participants in the sample. When we interpret item threshold parameters, such as b1, we are considering how far an individual’s level deviates from this mean. For example, if *b*1 = -1.23 for the item “*watching a fist-fight excites me*”, this means that participants with a sadism level 1.23 below the mean might change their response from *strongly disagree* to *moderately disagree*.

{apafg-fig1} shows an example of an item probability function (IPF) from an item on a scale with five response categories along with the staircase illustration. In general, IPF describes a relationship between (x-axis) and the probability of endorsing a response category (y-axis - *P*). {apafg-fig1} 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. The “Strongly Disagree” and “Strongly Agree” curves are boundary categories in a GRM model, which results in their peaks being at the extreme ends of the . These curves look different than the other because they represent the cumulative probability of selecting the most extreme responses, while intermediate categories peak in the middle and have more defined, bell-shaped curves. The probability curves of boundary categories simply mean that individuals at extreme levels are very likely to select extreme responses.



*Note*. {}

## Assumptions

The first key assumption of a GRM is that is monotonically correlated with the probability of endorsing a response category. More sharply, individuals with higher levels of should have a higher probability of endorsing response category that represents greater intensity of the traits ([Hambleton et al., 2010](#ref-hambletonIRTModelsAnalysis2010)). 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) 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](#X6063c9832c002f4cec709041b07655ee901e5b2)). To test whether an item pool represents one , researchers can apply a parallel analysis, a factor analysis, or a multidimensional scaling (MDS) ([De Ayala & Hertzog, 1991](#X3ee6bae5ef93add7564b14f04f771b6562e9859); [Guo & Choi, 2023](#ref-guoAssessingDimensionalityIRT2023)), 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 affect their responses to another. To test whether items are locally independent, researchers can examine the relationships between item responses after accounting for by performing residual correlation analysis ([Chen & Thissen, 1997](#ref-chenLocalDependenceIndexes1997)). In practice, it is quite difficult to fully satisfy the local dependency assumptions when researchers measure a construct with multi-item measures, so we should always expect some degree of local dependency to be present in the model. That said, some local dependency problems may be insignificant, and there are several strategies to avoid severe model misfit due to local dependency, which we will discuss later.

Note that evidence of unidimensionality does not always warrant local independence. As an eyeball example - consider a scale measuring trust in science that includes: “*I trust scientists working in natural science*” and “*Most biologists are trustworthy*.” While both items may measure the same latent trait (i.e., trust in science), as indicated by moderate or large loading factors, those who trust scientists in the natural sciences are very likely to trust biologists as well. Therefore, in practice, it is likely that these items will still be moderately correlated even after taking into account their loading factors on .

# 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, 2006](#ref-altemeyerAuthoritarians2006)). Altemeyer ([2006](#ref-altemeyerAuthoritarians2006)) defines RWA as a personality tendency to blindly obey 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 authorities (i.e., conventionalism). Individuals with higher levels of RWA tend to view the world as a dangerous place, and thus, to ensure stability and security, they are motivated to maintain social order by deferring to those they perceive as legal, social, or moral authorities, such as the government, religious institutions, or political, religious, or military leaders ([Saunders & Ngo, 2017](#X2ea6ac663606e999b517edf86575b3156e3350a)). The RWA scale remains one of the most important measures in the field of social and political psychology and has helped researchers conceptualize the role of RWA in shaping various social psychological outcomes, such as prejudice, political behavior, and various antisocial behaviors.

In this paper, we use the most recent version of the RWA scale ([Altemeyer, 2006](#ref-altemeyerAuthoritarians2006)), which consists of 22 items in which participants are asked to indicate their agreement with the items on a nine-point scale ranging from “*strongly disagree*” (-4) to “*strongly agree*” (+4), with 0 as the midpoint. While the RWA scale consists of three sub-dimensions, i.e., Submissiveness, Aggression, and Conventionalism, the RWA scale is theoretically assumed to be unidimensional because Altemeyer ([2006](#ref-altemeyerAuthoritarians2006)) shows that these sub-dimensions are strongly intercorrelated. While there is some evidence to the contrary ([Duckitt & Bizumic, 2013](#Xe7e7921bdf8fe3bceba0f362744b385b1b87c20)) and despite the theoretical debates existing in the literature pertaining to the dimensionality of RWA, for the purposes of this article we will assume that the RWA construct is theoretically unidimensional.

# Disclosure and Data Availability Statements

To maximize the reproducibility of our illustrative example, we wrote the article as a Quarto (.qmd) document ([Schneider, 2024](#ref-schneiderWjschneApaquarto2024)), where we integrate the R codes used in the analysis as well as its outputs. We also include a longer, annotated R Markdown file (.Rmd) for the example we used as a supplementary document, which we highly recommend using for a didactic purpose. In the same file, we also include a part where we demonstrate the application of GRM using a simulated dataset, which we do not include in the article, so that readers can compare the example we present here against an ideal scenario. The Quarto file (and its corresponding .docx and .pdf output) and an annotated R Markdown 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 a publicly accessible [Open-Source Psychometrics Project](http://openpsychometrics.org/_rawdata/).

## Step 1: Data Preparation

Before running the analysis, we need to install the necessary R packages, which are, *mirt* ([Chalmers, 2012](#ref-chalmersMirtMultidimensionalItem2012)), *psych* ([Revelle, 2023](#X3df0b348f80d9bf4e12f708b20f66c81afe6902)), *ggmirt* ([Masur, 2023](#ref-masurGgmirt2023)), and *caret* ([Kuhn, 2024](#X43830a7d32074f3481be13b1a7e307f0fdfc671)), and *skimr* ([Waring et al., 2024](#ref-waringSkimrCompactFlexible2024)) 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)  
 "skimr", # for visualizing data distribution in a data frame at glance  
 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); library(skimr)

Now after the packages are activated, we need to import the dataset to our R environment. The cleaned dataset (data.csv) and code book (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:

data <- read.csv("path\_to\_data/data.csv") # Replace "path\_to\_data" with the actual path to your data file

The dataset provided in our repository has been cleaned, with all unfavorable items reversed scored, and is ready for analysis. Since the dataset is openly available on the Open Psychometrics website, readers can also import the dataset directly from the website into their R environment. For interested readers, we show in detail how to perform the import and data cleaning process in our annotated R Markdown file, which is available as supplementary material.

## Step 2: Examining Dimensionality

Next, we show the readers the ways of testing unidimensionality. This step is important so that researchers can decide whether applying unidimensional GRM analysis is the right choice. In doing so, there are three available alternatives, which are multidimensional scaling (MDS), EFA, or CFA ([De Ayala & Hertzog, 1991](#X3ee6bae5ef93add7564b14f04f771b6562e9859)). 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(data, nfactors = 1, fm = "minres")

irt.fa() function is to run an EFA with a polychoric correlation matrix as an input when the data are ordinal, which is often the case in practice, especially when researchers are working with a five-option Likert scale. Since we have more than eight response categories in our current dataset, after running irt.fa() function, an error message, “*polychoric is probably not needed*”, should appear in the console. Therefore, another option would be running an EFA manually with a correlation matrix as an input. To determine which correlation method should be used when constructing the correlation matrix, we need to check whether items are normally distributed by calling this command:

skim(data)

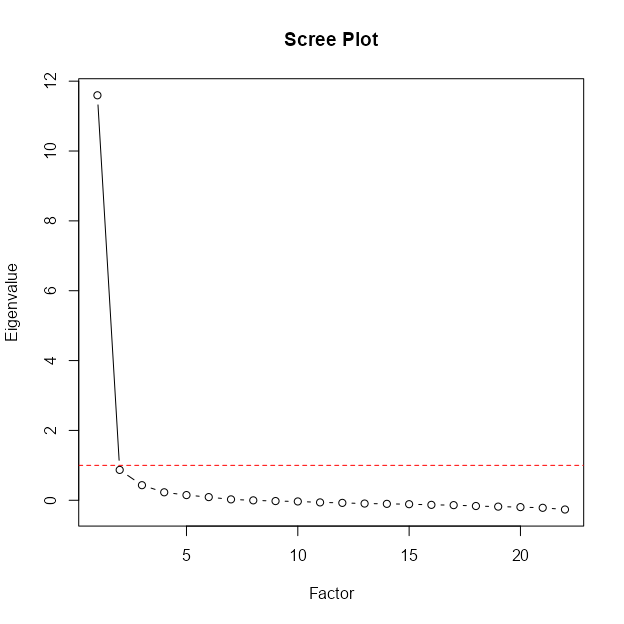
After running this command, readers can see in the console a table summarizing the data frame, including histograms of each item in hist column. Here, readers would find that all items are heavily skewed, so that we should use a Spearman correlation matrix as an input for EFA. In doing so, we need to create a Spearman correlation matrix from our data frame, and then, run an EFA by calling these following commands:

cor <- cor(data, method = "spearman") # First, creating a Spearman correlation matrix.  
efa <- fa(cor, 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 0.53, which means that 53% 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.

{apafg-fig2} shows that there is a distinct break or “elbow” that indicates where the eigenvalues begin to level off below 1. This implies that factors beyond the first may not be meaningful and could be just noise. Therefore, the scree plot supports our assumption that the variance of the item pool is mainly explained by one latent factor.



*Note*. {}

In a case where there is no or unclear theoretical assumption or empirical evidence of dimensionality, such that researchers are unsure of the structure of their item pool, it is suggested that a parallel analysis be performed before proceeding with EFA ([O. L. Liu et al., 2007](#ref-liuInitialInvestigationModified2007)). The purpose of running a parallel analysis is to identify the optimal number of factors underlying the item pool by comparing the eigenvalues of the data with a set of simulated data sets. Parallel analysis can be easily applied by running the following command:

fa.parallel(data, fm="minres", fa="pc")

As suggested by Guo and Choi ([2023](#ref-guoAssessingDimensionalityIRT2023)), here we use principal component analysis (PCA - fa="pc") to estimate the eigenvalues. When researchers have a Likert scale with five or more but less than 8 response categories, it is recommended to perform a parallel analysis with a polychoric correlation matrix ([Guo & Choi, 2023](#ref-guoAssessingDimensionalityIRT2023)), which can be done by adding cor = "poly" to the above command.

After running the code, readers will see in the console that parallel analysis suggests two components underlying the item pool but this is not a clear cut since the eigenvalue of the second component is only slightly above 1 (i.e., 1.303). In this situation, we can run another EFA to be sure by assuming that there are two factors underlying the data, and then compare the results with the single-factor EFA we ran earlier.

efa2 <- fa(cor, nfactors = 2, fm = "minres") # Running EFA assuming that there are two factors.  
print(efa2) # Print the results.

EFA with two factors has slightly more variance explained and lower root mean square residuals (Cumulative Var = 0.57, RMSR = 0.03) than the EFA we ran earlier (Proportion Var = 0.53, RMSR = 0.05), indicating a better model fit. However, the correlation between the first and second factors is large (*r* = 0.79). In this case, it is difficult to definitively conclude that our data show a single factor (i.e., unidimensional), but we will proceed with the unidimensional GRM analysis because the RWA theory underlying the item pool argues for unidimensionality, the one-factor EFA shows acceptable evidence of unidimensionality, and, in general, a one-factor model is easier to interpret. It should be noted, however, that we can expect some local dependency problems with our one-dimensional GRM model later, since the two-factor model is a slightly more accurate representation of our data, according to our EFA results.

When stronger evidence of unidimensionality is needed, especially when testing the precision of a newly developed scale, researchers can also cross-validate the structure of their scale on another sample by running a CFA ([Flora & Flake, 2017](#ref-floraPurposePracticeExploratory2017)), and then decide to run a uni- or multidimensional GRM analysis after obtaining stronger evidence of the number of underlying the item pool. Note that parallel analyses, EFA, and CFA are used to uncover the structure of the item pool and to focus on the snippet of the performance of the measure at the sample level. If researchers are interested in item-level and person-level parameters, then GRM should still be conducted. EFA and GRM can be performed on the same dataset, since the purpose of each analysis is different, but EFA and CFA should be performed on a different sample, with EFA performed at the early phase and CFA performed at a later stage, after obtaining preliminary evidence of the underlying structure ([Flora & Flake, 2017](#ref-floraPurposePracticeExploratory2017)).

## Step 3: Model Estimation, Parameters, and Fit Statistics

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

model <- 'theta = 1-22'

This code implies that we want to estimate a model with one namely theta, which is supported by EFA results we ran earlier. 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 = data, 1, model = model, itemtype = "graded", SE = TRUE, verbose = FALSE)

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 output from this procedure is presented in {apatb-table2}, 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](#ref-bakerBasicsItemResponse2017)), 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 {apatb-table2}, item discrimination parameters of the RWA scale range from high to very high (1.57 - item #1 to 3.32 - item #7), indicating a strong ability of RWA items to differentiate between individuals with different RWA levels. The threshold parameters consistently increase, suggesting that individuals with higher RWA levels are more likely to choose 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 the lowest extreme category (i.e., “*strongly disagree*” or -4) or stepping to its adjacent response category. This might suggest that the RWA scale is less sensitive to individuals with a low and 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 {apatb-table2}. 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 an EFA analysis we ran earlier, because *mirt* runs EFA using a quasi-polychoric correlation matrix, while the one we ran earlier to test unidimensionality used a Spearman 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.

| Item | a | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 | λ | h2 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 001 | 1.57 | -1.08 | -0.23 | 0.48 | 0.78 | 1.20 | 1.67 | 2.41 | 3.19 | 0.68 | 0.46 |
| 002 | 2.45 | 0.47 | 0.94 | 1.19 | 1.33 | 1.58 | 1.86 | 2.13 | 2.47 | 0.82 | 0.67 |
| 003 | 3.04 | 0.03 | 0.50 | 0.74 | 0.87 | 1.04 | 1.30 | 1.54 | 1.85 | 0.87 | 0.76 |
| 004 | 2.50 | 0.22 | 0.74 | 0.98 | 1.08 | 1.30 | 1.53 | 1.82 | 2.10 | 0.83 | 0.68 |
| 005 | 1.92 | -0.62 | 0.13 | 0.67 | 1.01 | 1.38 | 1.82 | 2.30 | 2.90 | 0.75 | 0.56 |
| 006 | 1.91 | 0.05 | 0.62 | 0.98 | 1.16 | 1.53 | 1.81 | 2.13 | 2.52 | 0.75 | 0.56 |
| 007 | 3.32 | -0.14 | 0.38 | 0.66 | 0.83 | 0.97 | 1.23 | 1.51 | 1.82 | 0.89 | 0.79 |
| 008 | 1.58 | -0.74 | -0.06 | 0.47 | 0.70 | 1.28 | 1.64 | 2.04 | 2.44 | 0.68 | 0.46 |
| 009 | 1.89 | -0.23 | 0.54 | 1.07 | 1.54 | 1.83 | 2.08 | 2.37 | 2.74 | 0.74 | 0.55 |
| 010 | 3.30 | -0.30 | 0.20 | 0.47 | 0.63 | 0.82 | 1.13 | 1.40 | 1.75 | 0.89 | 0.79 |
| 011 | 2.03 | 0.04 | 0.69 | 1.13 | 1.42 | 1.71 | 1.99 | 2.35 | 2.74 | 0.77 | 0.59 |
| 012 | 2.61 | -0.77 | -0.20 | 0.25 | 0.46 | 0.79 | 1.19 | 1.60 | 2.02 | 0.84 | 0.70 |
| 013 | 2.60 | -0.31 | 0.29 | 0.65 | 0.91 | 1.16 | 1.35 | 1.58 | 1.89 | 0.84 | 0.70 |
| 014 | 2.30 | -0.46 | 0.03 | 0.32 | 0.47 | 0.75 | 1.09 | 1.42 | 1.79 | 0.80 | 0.65 |
| 015 | 2.32 | -0.62 | 0.04 | 0.50 | 0.87 | 1.19 | 1.44 | 1.77 | 2.21 | 0.81 | 0.65 |
| 016 | 2.59 | 0.43 | 0.85 | 1.13 | 1.30 | 1.54 | 1.82 | 2.11 | 2.41 | 0.84 | 0.70 |
| 017 | 2.56 | -0.15 | 0.35 | 0.60 | 0.77 | 1.00 | 1.35 | 1.66 | 2.00 | 0.83 | 0.69 |
| 018 | 2.32 | 0.40 | 0.98 | 1.26 | 1.44 | 1.61 | 1.86 | 2.14 | 2.45 | 0.81 | 0.65 |
| 019 | 3.14 | -0.24 | 0.32 | 0.67 | 0.90 | 1.12 | 1.48 | 1.78 | 2.11 | 0.88 | 0.77 |
| 020 | 1.75 | -0.14 | 0.61 | 1.14 | 1.52 | 1.74 | 2.07 | 2.46 | 2.89 | 0.72 | 0.51 |
| 021 | 2.31 | -0.70 | -0.19 | 0.17 | 0.43 | 0.85 | 1.09 | 1.32 | 1.61 | 0.81 | 0.65 |
| 022 | 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. There are 8 threshold parameters because each threshold represents the point of transition between adjacent categories, not the number of categories themselves.

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

M2(fit, type = "C2")

*mirt* calculates the *M*2 statistics for model fit ([Maydeu-Olivares & Joe, 2006](#X8a0ee77163ceb391bc971ded1391b8c7e388d5c)) and other goodness-of-fit 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) after running the code above. Here, we set the type option to C2 because it is suitable for computing *M*2 statistics in polytomous models. *M*2 statistic is conceptually similar to the 2 test of overall model fit in SEM ([Maydeu-Olivares & Joe, 2006](#X8a0ee77163ceb391bc971ded1391b8c7e388d5c)), so the interpretation is similar. The rule of thumb for other goodness of fit statistics is also similar to a general guide for interpreting goodness of fit in CFA or SEM models ([Schermelleh-Engel et al., 2003](#X4f674a50e25708f779b591c14889b7d36290a85)), which is *RMSEA* 0.05, *SRMSR* < 0.05, *TLI* and *CFI* 0.97, for a good fit. While 0.05 < *RMSEA* 0.08, 0.05 < *SRMSR* 0.10, 0.97 < *TLI* and *CFI* 0.95, for an acceptable fit.

According to our analysis, the model does not fit the data well (*M*2(209) = 1.980251^{4}, *p* < .001, *RMSEA* = 0.098, *SRMSR* = 0.055, TLI = 0.969, CFI = 0.972), because there is a significant discrepancy between response patterns predicted by the model with the data, as shown by significant *M*2 statistics. When faced with this situation, readers are advised to tentatively reject the model and then identify the sources and magnitude of the misfit ([Kline, 2023](#X841dc8cf9fe2c9aa03f8adf3ba844f5d6f66257); [Y. Liu & Maydeu-Olivares, 2014](#ref-liuIdentifyingSourceMisfit2014)). It is likely that a model fails the 2 test, but the residuals show an insignificant discrepancy between the observed and expected values when we have a very large sample ([Kline, 2023](#X841dc8cf9fe2c9aa03f8adf3ba844f5d6f66257); [Maydeu-Olivares, 2015](#ref-maydeu-olivaresEvaluatingFitIRT2015)), which may well be the case since we have more than 9,000 cases in our dataset.

Overall, our model is neither a poorly fit nor a good fit. RMSEA shows a poor fit, while CFI and TLI indicate a good fit. However, the most important goodness-of-fit statistics we should pay attention to is SRMSR, since it provides the average effect size of model misfit ([Maydeu-Olivares, 2015](#ref-maydeu-olivaresEvaluatingFitIRT2015)). SRMSR is only slightly above the 0.05 cutoff, suggesting that some areas of our model may be misfitting, but the misfit is not so severe. With this in mind, it is noteworthy to identify the source of the misfit, and local dependency between items is the main suspect.

Next, we calculate item-level fit statistics, i.e., the signed 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 {apatb-table3}, which shows that the signed 2 for item fit ([Orlando & Thissen, 2000](#X1e1b7a71c2ddf66bdc23924de21bc94a49230b1)) for each item. All items have very good *RMSEA* values ( 0.05), indicating that each individual item fits the underlying IRT model. Normally, we also want the *p* values of the signed 2 to be insignificant (*p* > .05), suggesting that there is no discrepancy between the observed response pattern and what the model predicts. In {apatb-table3}, however, only three items here are not significant (#2, #4, and #12), which is expected when we have a very large sample ([Kline, 2023](#X841dc8cf9fe2c9aa03f8adf3ba844f5d6f66257)). Nonetheless, it may also be the case that some of these items are not locally independent. We will explore this issue further in the next part.

| Item | Sχ2 | df | p | RMSEA |
| --- | --- | --- | --- | --- |
| 001 | 1,187.23 | 957 | 0.0000 | 0.005 |
| 002 | 917.58 | 915 | 0.4698 | 0.001 |
| 003 | 983.47 | 847 | 0.0008 | 0.004 |
| 004 | 978.46 | 930 | 0.1313 | 0.002 |
| 005 | 1,072.31 | 957 | 0.0054 | 0.004 |
| 006 | 1,310.80 | 1,041 | 0.0000 | 0.005 |
| 007 | 972.91 | 780 | 0.0000 | 0.005 |
| 008 | 1,191.41 | 1,041 | 0.0008 | 0.004 |
| 009 | 1,254.66 | 981 | 0.0000 | 0.005 |
| 010 | 981.24 | 770 | 0.0000 | 0.005 |
| 011 | 1,206.09 | 988 | 0.0000 | 0.005 |
| 012 | 842.01 | 811 | 0.2187 | 0.002 |
| 013 | 1,088.69 | 904 | 0.0000 | 0.005 |
| 014 | 1,031.66 | 903 | 0.0018 | 0.004 |
| 015 | 1,152.22 | 932 | 0.0000 | 0.005 |
| 016 | 997.76 | 891 | 0.0071 | 0.004 |
| 017 | 1,142.47 | 906 | 0.0000 | 0.005 |
| 018 | 1,182.62 | 936 | 0.0000 | 0.005 |
| 019 | 912.71 | 793 | 0.0020 | 0.004 |
| 020 | 1,262.10 | 1,046 | 0.0000 | 0.005 |
| 021 | 1,031.48 | 878 | 0.0002 | 0.004 |
| 022 | 1,122.72 | 832 | 0.0000 | 0.006 |

*Note*. Signed χ2 Statistics. RMSEA = Root Mean Square Error of Approximation, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, SRMR = Standardized Root Mean Square Residual.

## Step 4: Model Residuals

To test whether the item pool is locally independent, *mirt* provides several alternatives for examining the behavior of the residuals. The first option is to run the Local Dependency (LD) 2 statistic ([Chen & Thissen, 1997](#ref-chenLocalDependenceIndexes1997)), which looks at the covariance between pairs of items after accounting for . In this tutorial, we demonstrate the use of Yen’s *Q*3 statistic ([Yen, 1984](#ref-yenEffectsLocalItem1984)), which is deemed more powerful for detecting underlying local dependency in unidimensional models, especially when dealing with polytomous data, than the LD 2 statistic ([Chen & Thissen, 1997](#ref-chenLocalDependenceIndexes1997)). For comparison, we also demonstrate the use of LD 2 statistics in the annotated R Markdown file. To run Yen’s *Q*3, readers can type and run this command:

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

As a practical guideline, Chen and Thissen ([1997](#ref-chenLocalDependenceIndexes1997)) suggests that a critical value for the *Q*3 statistic is around |0.2|, so we are interested in item pairwise correlations above |0.2|. To help us flag problematic, locally dependent pairs, we can use the findCorrelation() function from the *caret* package 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 the residual correlations between items #3, #4, #5, #7, #11, #13, #14, #18, #19, and #21 are above |0.2|. As suspected, all of these items, except #4, also have significant *S*- $2 statistics (see {apatb-table3}), which tells us that we need to look at these items closely and then decide whether these items have common characteristics beyond what is explained by the model. While it is worth looking at all of these item pairs, we recommend that readers prioritize the investigation of severe local dependencies, as indicated by item pair correlations above |0.3| (i.e., #14 and #3, #11 and #20, #5 and #1, #18 and #2), before addressing the rest of problematic pairs.

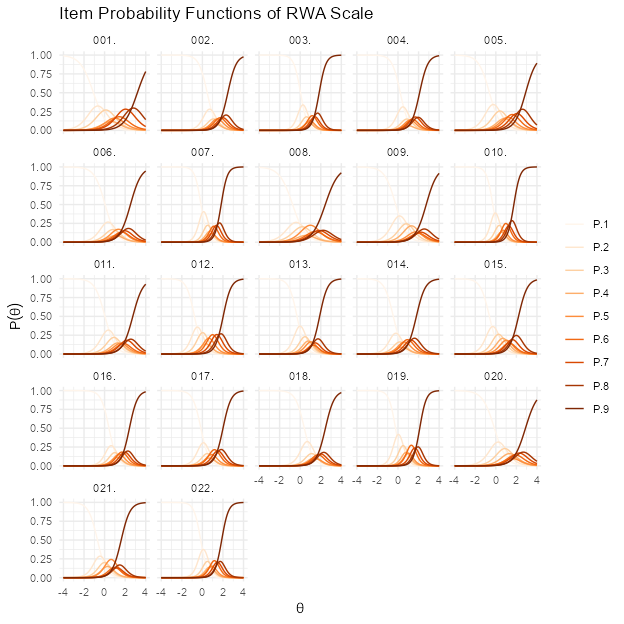
## Step 5: IRT Plots

*mirt* offers features to visualize model parameters but the options for customizing these plots are rather limited. *ggmirt* package 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 output of this function is {apafg-fig3}, which we can see that as increases, participants 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, which is consistent with the gradual increase in threshold parameters we saw in the previous step. 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.

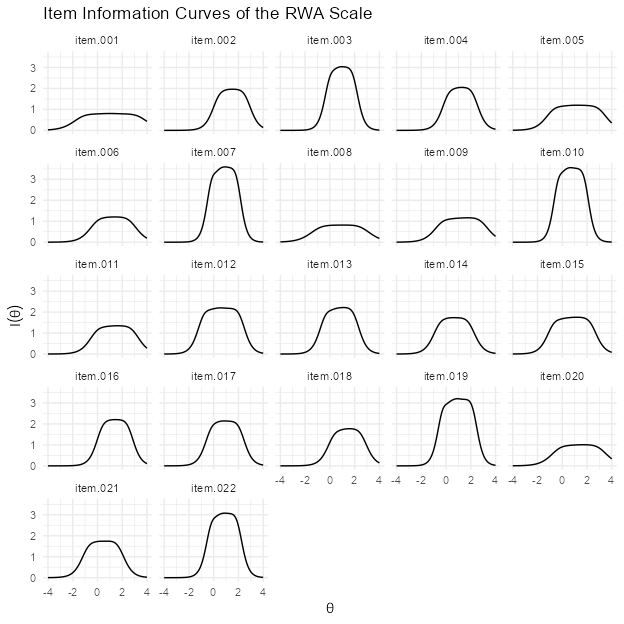


*Note*. {}

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 output for this function is {apafg-fig4} 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.

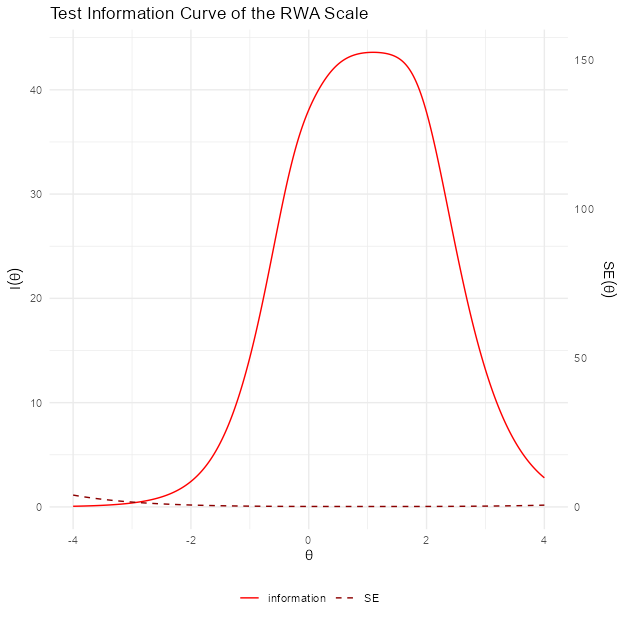


*Note*. {}

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 {apafg-fig5}, which we can see that 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* (low) to +4*SD* (very high). Moreover, the RWA scale is not optimal for measuring individuals with RWA levels below or beyond this range, as we can see in {apafg-fig5} that standard error increases below -3*SD*.



*Note*. {}

## Step 6: Computing Reliability

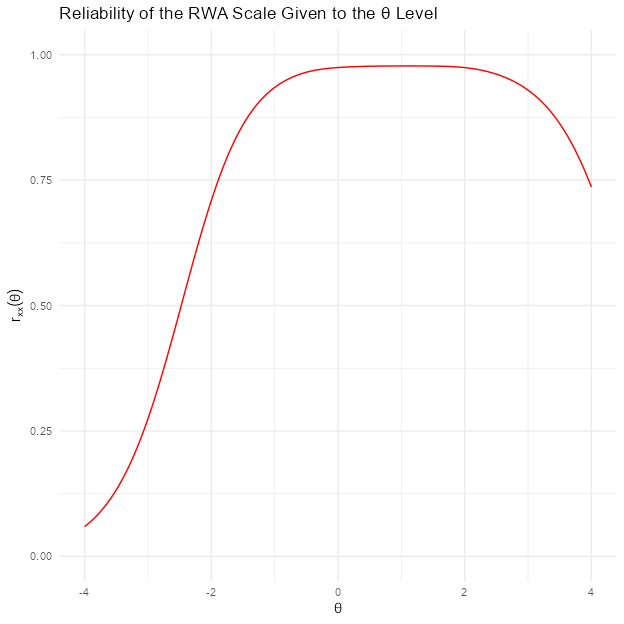
The heart of an IRT analysis is to estimate the precision of the measure, which can vary given different levels of . *mirt* provides two strategies to compute reliability that differ in their assumption regarding the distribution of . First, readers may calculate marginal reliability by running this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

marginal\_rxx(fit)

The output of the above code is 0.948, which indicates that the overall RWA scale is highly reliable, assuming that the underlying distribution follows the Gaussian or normal distribution. Since marginal reliability can vary across different levels of , readers can also visualize the marginal 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 {apafg-fig6}, which shows that the RWA scale can measure individuals with levels between approximately -1.75*SD* and +2.75*SD* with optimal reliability, i.e., *r*xx 0.80. This means that the RWA scale is most reliable for participants with low to high levels of RWA, but it becomes less reliable when measuring participants with extremely low ( < -1.75*SD*) or extremely high levels of RWA ( > 2.75*SD*).



*Note*. {}

*mirt* can also compute the overall reliability of the RWA scale, assuming the model-predicted distribution (i.e., factor scores) and its associated standard errors. If item discrimination (*a*) and thresholds (*b*) we showed earlier reflect *item parameters*, then the estimated or factor scores we compute here are the *person parameters*, since they represent each participant’s “position” on the continuum. To compute empirical reliability, one needs to calculate the estimated value, and its associated standard errors, of each participant first, and then, calculate empirical 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.

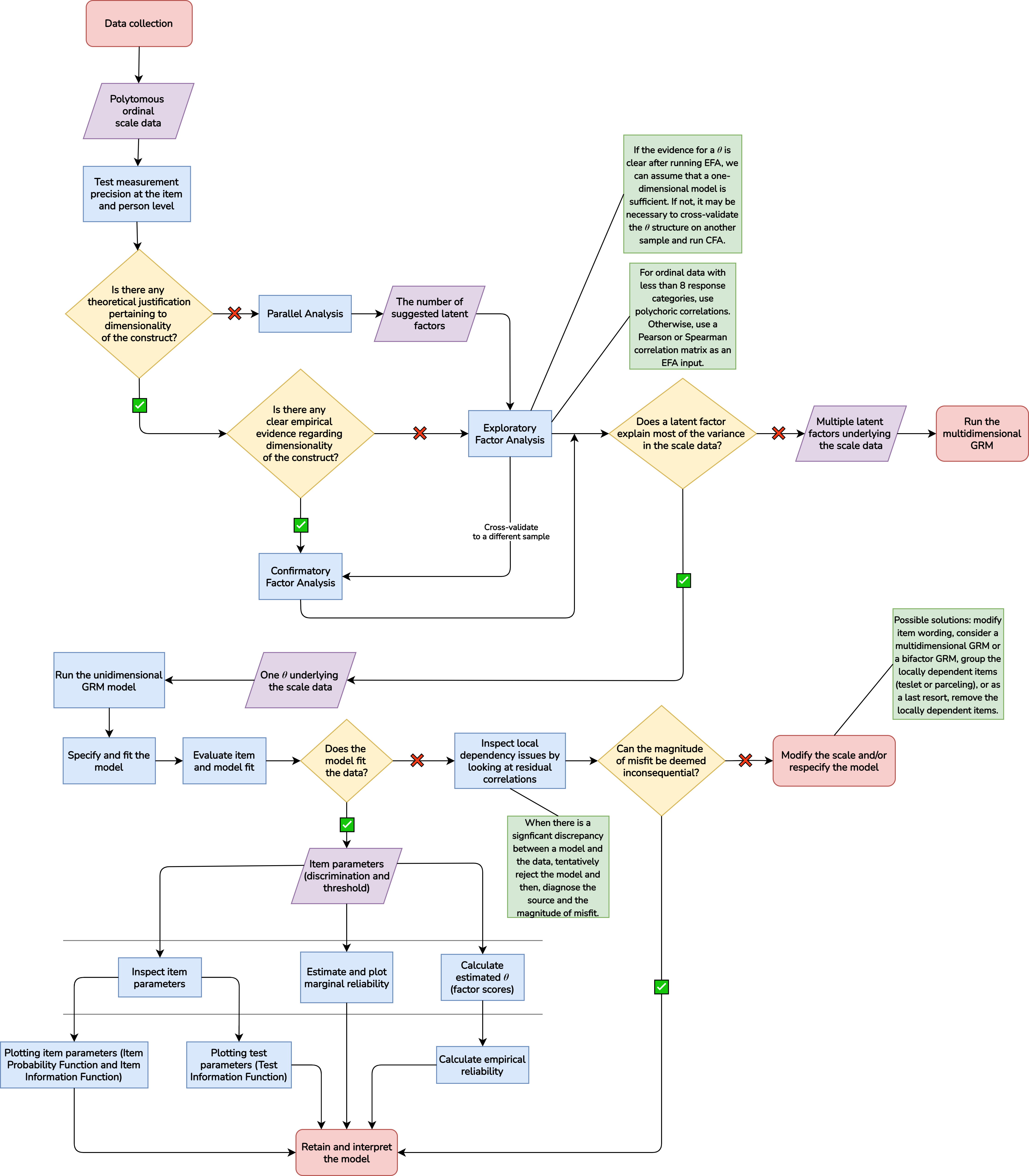
The output of these codes is 0.953, which is rather close to marginal reliability we estimated before. This implies that RWA scale is overall highly reliable in measuring RWA trait across all participants in the sample. Generally, reliability analysis aligns with the IPFs and IIFs we saw in the previous step, i.e., the RWA scale is highly reliable overall, but works best to measure the RWA construct across participants with RWA levels between ~2*SD* below the mean and ~3*SD* above the mean.

# Discussion

In this tutorial, we have demonstrated 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 6-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, an open-source statistical software, and make all materials publicly available.

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, including an explanation of how IRT differs from CTT in its assumptions. In general, GRM offers several benefits and is useful as a complement to CTT analysis, especially when researchers focus on closely examining the precision of their measures at the item and person levels. However, we emphasize that our goal here is not to argue for the superiority of IRT over CTT, as the choice of analytic tool depends largely on the specific research question at hand. It is important to emphasize that researchers should be aware of the merits and limitations of their chosen methods. Therefore, researchers should always justify why they choose a particular method over many available alternatives.

We have described the process of conducting a GRM analysis, starting with testing dimensionality, fitting the model, calculating and plotting item parameters, and calculating scale reliability. To help readers intuitively understand the process, we have summarized the steps we demonstrate in this tutorial in an annotated flowchart ({apafg-fig7}).



*Note*. {}

Note that the model we estimate does not fit the data perfectly due to local dependency, as we showed in steps #3 and #4. If researchers encounter this problem when evaluating their scale data, we recommend tentatively rejecting the model and then locating the source of the model misfit before interpreting measurement precision or drawing substantive conclusions based on these estimates ([Kline, 2023](#X841dc8cf9fe2c9aa03f8adf3ba844f5d6f66257)). In doing so, researchers can take a closer look at residual correlations to identify the source and magnitude of misfit ([Kline, 2023](#X841dc8cf9fe2c9aa03f8adf3ba844f5d6f66257); [Maydeu-Olivares, 2015](#ref-maydeu-olivaresEvaluatingFitIRT2015)).

Once the sources of misfit are identified, one solution to address them is to rephrase or combine the content of the problematic, locally dependent items and then cross-validate the modified scale on another sample. Alternatively, researchers may need to reconsider the structure. If the data are unidimensional, but some of the items are locally dependent, this may indicate the existence of multiple within the data structure. In this sense, modeling the test data as a correlated multidimensional model or a bifactor model (i.e., multidimensional models with a *g* factor) may be a viable solution. Another solution to consider is to group locally dependent items by modeling them together as a “testlet” ([Cook et al., 1999](#ref-cookComparisonThreePolytomous1999)), or to combine locally dependent items into a single composite score (i.e., item parceling), especially when the goal of the analysis is to understand the construct being measured rather than to identify the relationship between items and the ([Little et al., 2013](#ref-littleWhyItemsParcels2013)). Removing locally dependent items should be a last resort, as it may improve model fit, but can jeopardize measurement precision.

Finally, while it is important to acknowledge that while a poorly fitted model may result in biased parameters, in reality, no IRT model can be expected to fit the data perfectly ([Maydeu-Olivares, 2015](#ref-maydeu-olivaresEvaluatingFitIRT2015)). That said, some models may be useful even if they are wrong to some degree ([Box, 1976](#ref-boxScienceStatistics1976)). Therefore, researchers are encouraged to not only evaluate the overall fit of their model, but also perform a piece-wise evaluation in some parts of their model ([Maydeu-Olivares, 2015](#ref-maydeu-olivaresEvaluatingFitIRT2015)) to get a comprehensive picture of the performance of their measures.

# References

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

Arias, V. B., Nuñez, D. E., Martínez-Molina, A., Ponce, F. P., & Arias, B. (2016). Hierarchy and Psychometric Properties of ADHD Symptoms in Spanish Children: An Application of the Graded Response Model. *PLoS ONE*, *11*(10), e0164474. <https://doi.org/10.1371/journal.pone.0164474>

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>

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

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>

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). [A comparison of three polytomous item response theory models in the context of testlet scoring](https://www.ncbi.nlm.nih.gov/pubmed/10063769). *J Outcome Meas*, *3*(1), 1–20.

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational Judgement Tests for Selection. In *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention* (pp. 226–246). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118972472.ch11>

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>

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

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., 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.

Kline, R. B. (2023). *Principles and Practice of Structural Equation Modeling* (5th ed.). The Guilford Press.

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>

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the Items versus Parcels Controversy Needn’t Be One. *Psychol Methods*, *18*(3), 285–300. <https://doi.org/10.1037/a0033266>

Liu, O. L., Rijmen, F., & Kong, N. (2007). An Initial Investigation of a Modified Procedure for Parallel Analysis. *ETS Research Report Series*, *2007*(2), i–19. <https://doi.org/10.1002/j.2333-8504.2007.tb02083.x>

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the Source of Misfit in Item Response Theory Models. *Multivariate Behavioral Research*, *49*(4), 354–371. <https://doi.org/10.1080/00273171.2014.910744>

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

Maydeu-Olivares, A. (2015). Evaluating the Fit of IRT Models. In *Handbook of Item Response Theory Modeling: Application to Typical Performance Assessment* (pp. 111–127). Routledge.

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>

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/.

Normand, S.-L. T., Belanger, A. J., & Eisen, S. V. (2006). Graded response model-based item selection for behavior and symptom identification. *Health Services and Outcomes Research Methodology*, *6*(1), 1–19. <https://doi.org/10.1007/s10742-006-0005-0>

Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. <https://doi.org/10.1177/01466216000241003>

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

Rauthmann, J. F. (2013). Investigating the MACH–IV With Item Response Theory and Proposing the Trimmed MACH\*. *Journal of Personality Assessment*, *95*(4), 388–397. <https://doi.org/10.1080/00223891.2012.742905>

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*.

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.

Schneider, W. J. (2024). *Wjschne/apaquarto*.

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>

Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2024). *Skimr: Compact and Flexible Summaries of Data*. rOpenSci.

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>