Educational Methods & Psychometrics—Vol. 3, Article No. 14
January 2025
https://dx.doi.org/10.61186/emp.2025.1

A TUTORIAL ON ITEM RESPONSE MODELING WITH MULTIPLE GROUPS USING TAM

PURYA BAGHAEI

IEA HAMBURG, HAMBURG, GERMANY

ALEXANDER ROBITZSCH

IPN−LEIBNIZ INSTITUTE FOR SCIENCE AND MATHEMATICS EDUCATION, KIEL, GERMANY

CENTRE FOR INTERNATIONAL STUDENT COMPARISONS (ZIB), KIEL, GERMANY

Multiple-group item response theory (MGIRT) is the standard psychometric model for the analysis of large-scale assessments in education. In this tutorial, a short nontechnical introduction to MGIRT is first provided. Second, the R package TAM and the included relevant functions to estimate MGIRT are introduced and applied to a small data set from the science assessment of TIMSS 2019 for Grade 8. Specifically, the following techniques are illustrated: the estimation of population parameters, analysis of differential item functioning (DIF; also referred to as non-invariance), and linking under full non-invariance. The hands-on training in this walkthrough enables researchers to estimate MGIRT with TAM and interpret the outputs confidently.

Key words: Multiple-group item response theory (MGIRT), RMSD, partial invariance, Haberman linking method, Stocking-Lord linking method

1. Introduction

In large-scale educational surveys, the assessment aims to examine and compare groups such as schools, districts, states, or countries in a given subject domain. To fulfil their goal and measure groups' skills and abilities in their entirety, such assessments cover extremely wide content domains, in the magnitude of 100 to 400 items. Since the administration of so many items to each individual student is impossible, a test design referred to as multiple-matrix sampling is usually employed (Shoemaker, 1973). In multiple-matrix sampling, to avoid overwhelming the examinees, the items are distributed in several test booklets, and examinees answer only a fraction of the many items that are selected to represent a content domain. Multiple-matrix sampling gives a wide coverage of the content domain in the population since all the items that represent a domain are administrated. This is achieved by presenting a portion of the items to a portion of the examinees to prevent overwhelming the examinees. However, all the items representing a specific content domain are collectively covered. This procedure is very efficient and economical as the objective of large scale assessments is group assessments as opposed to individual assessments (Shoemaker & Shoemaker, 1981).

Correspondence should be made to Purya Baghaei, Research and Analysis Unit, IEA Hamburg, Überseering 27, 22297 Hamburg, Germany . Email: purya.baghaei@iea-hamburg.de

Multiple-matrix sampling is, however, suboptimal for individual assessment (see Frey et al., 2009), and as Gonzalez and Rutkowski (2010) wrote, "…in most large-scale assessments, individual measurement precision is sacrificed in the interest of increased content coverage" (p.126).

In international large-scale assessments (ILSA) in education, the goal is to evaluate the effectiveness of educational systems or programs. In such testing contexts, since the focus of the assessment is on the collective performance of examinees (i.e., groups' means and standard deviations), there is no need to estimate individual students' scores. Early studies on multiple-matrix sampling showed that it is a reasonable approach for this purpose (Johnson & Lord, 1958; Knapp, 1968). In these early studies, raw score means, and standard deviations were estimated.

Standard single-group IRT models have been extended to model different populations. Single-group IRT models operate under the assumption that each examinee provides sufficient responses to yield stable parameters of the latent ability. Furthermore, these models are not applicable when heterogeneous populations are involved. When using MML estimation, MGIRT is essential because the population of individuals is heterogeneous and cannot be assumed to follow a normal distribution. Given the significant differences in mean competence levels across countries, a mixed distribution is required to accurately model the population. MGIRT facilitates this process.

Multiple-group item response theory (MGIRT) models are defined at the level of groups rather than individuals and are optimal for multiple-matrix data using full information maximum likelihood estimation algorithm (Andersen & Madsen, 1977; Bock et al., 1982; Bock & Mislevy, 1981; Bock & Zimowski, 1997). While in single-group IRT models, the probability of a correct response to an item by a person is modeled as a function of the ability of the person and one or more item parameters, in MGIRT, the probability of a correct response to an item by a person selected at random from a group is modeled as a function of the ability of the group and one or more item parameters (Mislevy, 1983). MGIRT has most of the advantages of the single-group IRT models and has successfully been applied to large-scale assessments since the early 1980s (Mislevy, 1983).

Apart from providing group-level parameters, MGIRT can also be used to evaluate differential item functioning (DIF; Holland & Wainer, 1993) or measurement invariance (MI; Meredith, 1993; Millsap, 2011). MGIRT models allow researchers to constrain or free the parameters across groups to adjust for measurement non-invariance. In the following sections, using an educational data set, we will show how MGIRT can be used to obtain population parameters and how to examine and adjust for DIF (measurement non-invariance. The TAM package (Robitzsch et al., 2022) in R (R Core Team, 2023) will be employed to estimate the model parameters. A step-by-step procedure is explained, relevant TAM functions are provided, and the outputs are interpreted.

2. Analysis

2.1 Data Source

The science section of TIMSS 2019 (Grade 8) (Martin et al., 2020) for five countries, namely, Chile (n=4115), Finland (n=4872), Georgia (n=3315), Hong Kong (n=3265), and Italy (n=3619) was used for this demonstration. All 14 booklets, containing 336 items, were entered into the MGIRT analysis. Twenty-two items were polytomous items with three response categories (0, 1, and 2), and the rest were binary items. The polytomous items were dichotomized for easier analysis. The values 0 and 1 were rescored to 0, and 2 was rescored to 1. The combined sample size for the five countries was 19186 (9240 female, 9942 male, and 4 unspecified).

2.2 Estimation of the Multiple-Group IRT with TAM

Before estimating the MGIRT model, the items that were not scaled in the official TIMSS scaling were omitted from the analysis (Fishbein et al., 2020). In addition, seven items were removed because they had a maximum value of zero for all the countries. This left us with 204 dichotomous items. To estimate MGIRT, the following codes were run:

```
library(TAM)# load the package
library (haven)
path<- "F:\\IEA\\Tutorial-MGIRT\\"
# load the data
data<- haven::read_sav(file.path( path, "5-count-dicho.sav"))
# specify the grouping variable. That is, the column in the data
# which contains the country codes (i.e., column "IDCNTRY")
country<- as.factor(data$IDCNTRY)
# estimate the multiple-group 2PL model and include the house weights
mod1<- TAM::tam.mml.2pl (data, group=country, pweights = data$HOUWGT)
summary(mod1)
```

Alternatively, the multiple-group Rasch model can be estimated:

```
mod2<- TAM::tam.mml (data, group=country, pweights = data$HOUWGT)
```

3. Examination of the Output

3.1 Item Parameters and Group Statistics

Figure 1 shows the item statistics and item difficulty parameters truncated for the first 10 items. The table shows the item labels, the sum of weights (2710), the item means or classical item easiness values, and the 2PL item difficulty and discrimination parameters. Note that *beta* is the IRT item difficulty parameter while *xsi* is the item intercept. The columns that follow show the threshold parameters. Since the data was dichotomous, the item intercepts and the thresholds are the same. In the case of polytomous items, the thresholds of the graded response model (Samejima, 1969) will be shown. To obtain item parameters' standard errors, run the following code:

```
mod1$xsi
```

Figure 2 shows the first part of the output for the multiple group 2PL model produced by TAM. Information criteria such as the *deviance*, AIC, BIC, and EAP reliability are provided, along with the number of items, persons, and parameters. Perhaps the most important statistics are provided under "Covariances and Variances", "Correlations and Standard Deviations", and "Regression Coefficients". As mentioned earlier, five countries were entered into the analysis using their numeric ISO codes. The code 152 is for Chile, 246 is Finland, 268 is Georgia, 344 is Hong Kong, and 380 is Italy.

The values under *Covariances and Variances* show the variances for each country, while the values under *Correlations and Standard Deviations* show country standard deviations. The most important statistic is in the *Regression Coefficients* table which shows the means of the countries. The mean of the first group is always fixed to zero. These values are estimated from the population model without estimating the examinees' ability parameters. The table shows that Georgia has the lowest mean (−0.21), and Finland has the highest (1.18).

```
Item Parameters -A*Xsi
         item        N     M xsi.item AXsi_.Cat1 B.Cat1.Dim1
1     SE52006 2709.889 0.313    1.075      1.075       0.494
2     SE52069 2710.149 0.525    0.115      0.115       0.486
3     SE52012 2710.115 0.563    0.035      0.035       0.732
4     SE52021 2710.621 0.266    1.630      1.630       0.855
5    SE52095B 2710.149 0.732   -0.859     -0.859       0.510
6     SE52150 2710.621 0.307    0.903      0.903       0.171
7    SE52243A 2709.889 0.388    0.785      0.785       0.598
8    SE52243B 2709.889 0.314    1.105      1.105       0.538
9    SE52243C 2710.621 0.395    0.776      0.776       0.643
10    SE52206 2710.149 0.511    0.317      0.317       0.822


Item Parameters in IRT parameterization
         item alpha   beta
1     SE52006 0.494  2.176
2     SE52069 0.486  0.236
3     SE52012 0.732  0.048
4     SE52021 0.855  1.906
5    SE52095B 0.510 -1.685
6     SE52150 0.171  5.268
7    SE52243A 0.598  1.312
8    SE52243B 0.538  2.054
9    SE52243C 0.643  1.206
10    SE52206 0.822  0.385
```

FIGURE 1.
Multiple Group 2PL Model Item Parameters

```
Number of iterations = 209
Numeric integration with 21 integration points

Deviance = 615141.5
Log likelihood = -307570.8
Number of persons = 19186
Number of persons used = 19119
Number of items = 204
Number of estimated parameters = 416
    Item threshold parameters = 204
    Item slope parameters = 204
    Regression parameters = 4
    Variance/covariance parameters = 4

AIC = 615974  | penalty=832    | AIC=-2*LL + 2*p
AIC3 = 616390  | penalty=1248    | AIC3=-2*LL + 3*p
BIC = 619243  | penalty=4101.11    | BIC=-2*LL + log(n)*p
aBIC = 617920  | penalty=2779    | aBIC=-2*LL + log((n-2)/24)*p  (adjusted
BIC)
CAIC = 619659 | penalty=4517.11    | CAIC=-2*LL + [log(n)+1]*p  (consistent
AIC)
AICc = 615992  | penalty=850.55    | AICc=-2*LL + 2*p + 2*p*(p+1)/(n-p-1)
(bias corrected AIC)
GHP = 0.55293     | GHP=( -LL + p ) / (#Persons * #Items)  (Gilula-Haberman
log penalty)

------------------------------------------------------------
EAP Reliability
[1] 0.826
```

```
------------------------------------------------------------
Covariances and Variances
Group152 Group246 Group268 Group344 Group380
   1.000    1.678    1.008    1.744    1.012
------------------------------------------------------------
Correlations and Standard Deviations (in the diagonal)
Group152 Group246 Group268 Group344 Group380
   1.000    1.295    1.004    1.321    1.006
------------------------------------------------------------
Regression Coefficients
         [,1]
[1,]  0.00000
[2,]  1.18314
[3,] -0.21857
[4,]  0.60721
[5,]  0.48996
------------------------------------------------------------
```

FIGURE 2.
Multiple Group 2PL Model Population Parameters

MG Rasch model and MG 2PL IRT model can be compared with their deviances (i.e., defined as minus two times the log-likelihood value) using the following function:

```
anova(mod1, mod2)
```

which returns:

```
Model   loglike Deviance Npars      AIC      BIC    Chisq    df  p
1  mod2 -310941.4 621882.9   213 622308.9 623982.7 6741.386 203  0
2  mod1 -307570.8 615141.5   416 615973.5 619242.6       NA  NA NA
```

The output above shows that the 2PL MGIRT fits significantly better than then Rasch MG model, χ^2= 6741.386, *df*=203, *p*=.000.

3.2 Multiple-group Differential Item Functioning Analysis

Differential item functioning (DIF, Holland & Wainer, 1993) is usually examined across two groups. However, in large-scale international studies where more than two groups are involved, examining DIF is somewhat complicated. One strategy is to examine DIF across all pairs of groups. However, this becomes very difficult to interpret. In ILSAs and in cross-cultural investigations in psychology, where sometimes more than 50 countries take part, pairwise comparisons are not very helpful (Halamová et al., 2019; OECD, 2017). Recently, OECD (2017) used a strategy involving MGIRT to examine DIF or measurement invariance (MI) when more than two groups are involved.

This strategy involves estimating an MGIRT model where all the groups have the same item parameters (i.e., assuming invariant item parameters). However, countries are allowed to have different means and variances. In the next step, an empirical ICC for each item is estimated within each group. Then, this group-specific ICC is compared with the joint ICC estimated from the MGIRT model, in which all groups were involved. A few statistics are computed which quantify the distance between the group-specific ICC and the international or joint ICC. A lack of DIF or MI is established if the difference between the group-specific ICC and the overall ICC is minimal. When the two ICCs overlap or are very close, the group performance can be safely explained by the overall joint parameters (Khorramdel et al., 2019; OECD, 2017).

The statistics that show the distance between the group-specific ICC and the joint ICC are Root Mean Square Difference (RMSD) and Mean Deviation (MD) (OECD, 2017). Both RMSD and MD

quantify the distance between the empirical ICC and the model ICC. They are sample-independent and range between zero and 1. MD is, however, based on the weighted sum of these differences. An RMSD value of zero indicates a perfect fit for the item and the presence of MI. OECD (2017) recommended an RMSD value of greater than .12 as a cutoff criterion to identify DIF items. However, in PISA 2015, a cutoff value of .30 was used for non-cognitive scales. The same cutoff values between –.12 to .12 is also used for MD. Khorramdel et al. (2019) state that the RMSD is sensitive to the deviations of both item difficulty and item slope from the joint item ICC, but MD is most sensitive to the deviations of item difficulty parameters. To get RMSD and MD values run the following codes:

```
fmod1 <- IRT.itemfit(mod1)
summary (fmod1)

Root Mean Square Deviation (RMSD)

    Parm     M    SD   Min   Max
1 Group1 0.067 0.040 0.012 0.196
2 Group2 0.065 0.038 0.009 0.261
3 Group3 0.083 0.051 0.010 0.287
4 Group4 0.095 0.058 0.011 0.351
5 Group5 0.065 0.044 0.009 0.298
6  WRMSD 0.082 0.031 0.024 0.176

         item Group1 Group2 Group3 Group4 Group5 WRMSD
1     SE52006  0.075  0.034  0.102  0.049  0.060 0.067
2     SE52069  0.088  0.039  0.074  0.136  0.220 0.123
3     SE52012  0.196  0.064  0.213  0.206  0.064 0.159
4     SE52021  0.071  0.080  0.194  0.088  0.048 0.105
5    SE52095B  0.034  0.040  0.037  0.099  0.028 0.052
6     SE52150  0.099  0.148  0.287  0.084  0.077 0.156
7    SE52243A  0.047  0.070  0.034  0.097  0.035 0.061
8    SE52243B  0.147  0.180  0.081  0.095  0.096 0.131
9    SE52243C  0.074  0.033  0.067  0.074  0.023 0.057
10    SE52206  0.019  0.084  0.100  0.210  0.018 0.105
```

FIGURE 3.
RMSD Values for the First 10 Items

To plot the RMSDs for visual inspection and comparison run the following codes:

```
RMSD<-fmod1$RMSD# all RMSD values
RMSDs<-RMSD[, 2:ncol(RMSD)] # RMSDs without item labels (from column 2)
boxplot(RMSDs)
```

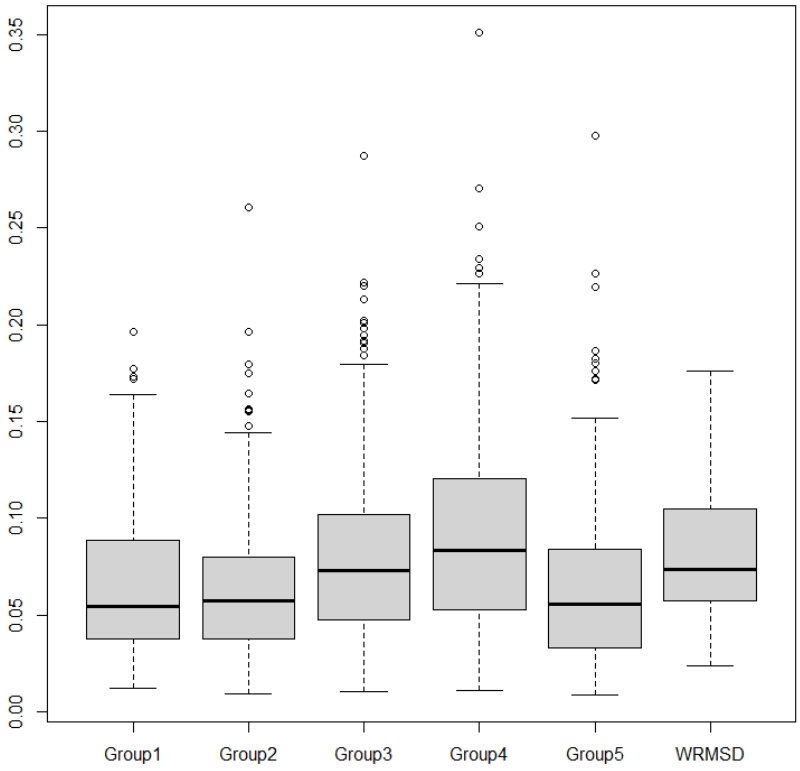

FIGURE 4.
Boxplot of RSMD Values in the Five Groups

The first table in Figure 3 shows the descriptive statistics for the RMSD values within each group (country). The mean of the RMSD values in Group 1 (Chile) is .067 (SD=.040). The lowest RMSD in this country is .012, and the highest is .196. The table shows that overall, the items had the best fit in Group 2 (Finland) and Group 5 (Italy), which have the lowest RMSD means, and have the worst fit in Group 4 (Hong Kong) with the highest mean. WRMSD stands for weighted RMSD which is the weighted average of RMSDs across countries and serves as an overall fit value for the item across all countries. Figure 4 is a boxplot of the distribution of RMSD values in each group.

The second part of the table shows the RMSD values for each item within each group. The table is truncated, and the first 10 items are only shown. Considering the cutoff value of RMSD>.12 as an indicator of DIF, (among the first 10 items), items 3 and 8 were identified as DIF items in Group 1 (Chile). In Group 2 (Finland), the DIF items were items 6 and 8. In Group 3 (Georgia), the DIF items were 3, 4, and 6. In Group 4 (Hong Kong), 2, 3, and 10 exhibits DIF, and in Group 5 (Italy), items 2, 3, 6, and 8 exhibit DIF. To identify the noninvariant or DIF items within each group, we can use the following codes:

```
rmsd<-fmod1$RMSD # extract RMSD values from object 'fmod1'
rmsd.1<-rmsd$Group1# extract RMSDs for Group 1 from object 'rmsd'
g1.flag<- ifelse(rmsd.1 > .12, "DIF", "OK") # within Group 1, flag items
with RMSD values greater than .12 with "DIF", otherwise with "OK"
```

which returns:

```
 [1] "Ok"  "Ok"  "DIF" "Ok" "Ok"  "Ok" "Ok"  "DIF" "Ok" "Ok" "DIF"
[12] "Ok"  "Ok"  "Ok"  "Ok" "Ok"  "Ok" "DIF" "Ok"  "Ok" "Ok" "DIF" [23]
"DIF" "DIF" "Ok"  "Ok" "DIF" "Ok" "Ok"  "Ok"  "Ok" "Ok" "DIF"
```

The next part of the output (Figure 5) shows the bias-corrected RMSD values. Robitzsch (2022) showed that the RMSD statistic is dependent on the test length and the number and proportion of misfitting items in the test. Using several simulation studies, he showed that the RMSD values of misfitting items become larger when the number of misfitting items is small. When the number of misfitting items increases, the RMSD values of misfitting items become smaller. The study also showed that the RMSD value of fitting items increases when the proportion of misfitting items increases. To remedy this problem, he suggested bias-corrected RMSD values, substantially reducing the bias for the fitting items.

```
Bias Corrected Root Mean Square Deviation (RMSD)

    Parm     M    SD    Min   Max
1 Group1 0.063 0.042 -0.016 0.195
2 Group2 0.062 0.040 -0.013 0.260
3 Group3 0.079 0.054 -0.017 0.286
4 Group4 0.092 0.061 -0.014 0.350
5 Group5 0.060 0.046 -0.013 0.297
6  WRMSD 0.079 0.032  0.019 0.175

         item Group1 Group2 Group3 Group4 Group5 WRMSD
1     SE52006  0.073  0.029  0.101  0.044  0.056 0.064
2     SE52069  0.086  0.035  0.070  0.134  0.219 0.121
3     SE52012  0.195  0.062  0.212  0.205  0.061 0.158
4     SE52021  0.069  0.078  0.194  0.086  0.044 0.104
5    SE52095B  0.028  0.037  0.030  0.097  0.021 0.048
6     SE52150  0.097  0.147  0.286  0.082  0.075 0.155
7    SE52243A  0.043  0.067  0.027  0.094  0.028 0.057
8    SE52243B  0.146  0.179  0.079  0.092  0.094 0.130
9    SE52243C  0.071  0.028  0.064  0.071  0.010 0.054
10    SE52206 -0.004  0.082  0.097  0.209 -0.010 0.103
```

FIGURE 5.
Bias-Corrected RMSD Values

The last part of the output (Figure 6) shows the MD values. These values have the same interpretation and cut-off criteria as the RMSD values. While RMSD values are always positive, MD values could be negative. The advantage of the MD values is that they show the direction of the DIF. A positive MD value means that the item is easier for the group and a negative MD indicates that it is harder.

```
Mean Deviation (MD)

    Parm      M    SD    Min   Max
1 Group1  0.002 0.067 -0.171 0.194
2 Group2 -0.005 0.065 -0.249 0.154
3 Group3  0.000 0.084 -0.206 0.269
4 Group4  0.002 0.097 -0.333 0.257
5 Group5  0.002 0.069 -0.287 0.181

         item Group1 Group2 Group3 Group4 Group5
1     SE52006 -0.069 -0.014  0.100 -0.046  0.047
2     SE52069  0.088  0.033 -0.064  0.133 -0.205
3     SE52012  0.194  0.060 -0.200 -0.196  0.055
4     SE52021 -0.064 -0.077  0.154  0.077 -0.038
5    SE52095B  0.027  0.004  0.010 -0.076  0.024
```

```
6     SE52150 -0.082 -0.130  0.269 -0.061  0.075
7    SE52243A  0.040 -0.067  0.026  0.050 -0.025
8    SE52243B  0.133 -0.177 -0.055  0.041  0.091
9    SE52243C  0.067  0.013 -0.060 -0.048  0.003
10    SE52206  0.002  0.074  0.091 -0.195 -0.012
```

FIGURE 6.
MD Values for the First 10 Items

3.3 MAD Outlier Detection Method

Another method to evaluate RMSD item fit statistic is the MAD (mean absolute difference) outlier detection method (von Davier & Bezirhan, 2023). To circumvent the problem of using a fixed RMSD threshold, MAD relies on removing items if they are identified as outliers with reference to the median of the item fit statistics. Thus, for example, if most of the items have RMSD values around .02 and a few items have RMSD values of say .08, these are identified as outliers (misfit) and removed. MAD is a robust measure of dispersion that flags an item as misfit if its deviation from the median of the absolute deviations of all other observations is greater than a cutoff value. Using a simulation study, (von Davier & Bezirhan, 2023) showed that the MAD can detect item misfit when the misfit is very small, while fixed RMSD cut-off can detect it when misfit is large. They concluded that MAD outlier detection performed better than a fixed RMSD cutoff value in detecting misfit. They also showed that in the context of ILSAs, a cutoff value of 2.5 for MAD resulted in very good to perfect detection rates. In PIRLS 2021, a MAD value of 4.5 was considered as a threshold for identifying misfitting items (Bristol et al., 2023). To compute the MAD outlier detection statistic, run the following codes:

```
RMSD<-fmod1$RMSD[, "Group1"]# Choose the RMSD values for Group 1.

med<- median(RMSD) # Compute the median of RMSDs.

absdif<-abs(RMSD -med) # Compute the absolute differences between the RMSDs
and their median.

MAD<-median(absdif)*2.4826# Compute the median of the absolute differences
(MAD) with the constant 2.4826 to align the transformation with the
standard normal distribution (Equation 5 in von Davier and Bezirhan
[2023]).

Robfit<-absdif/MAD # This is Equation 8 in von Davier and Bezirhan (2023).
Values smaller than 4.5 indicate invariance.
```

3.4 Modeling Measurement Noninvariance

Recently, Robitzsch and Lüdtke (2023) stated that measurement invariance, or in this case absence of item-by-country interactions or DIF, is not a prerequisite for group comparisons. They argue that measurement noninvariance is equivalent to the presence of interaction effects in ANOVA models which does not prevent researchers from computing means and comparing groups. In principle, there are different defensible approaches for modeling DIF, and it is up to the researchers to choose their strategy for modeling measurement noninvariance (Robitzsch, 2022). Robitzsch and Lüdtke (2023) propose inspecting all items for DIF and model it only when DIF can be explained by technical reasons or translation issues. Otherwise, non-invariance should be ignored. In this tutorial, we introduce two strategies for modeling item-by-country interactions, namely, partial invariance and linking methods.

One possible strategy when MI fails or when there is DIF, as was observed above, is partial invariance. In partial invariance, the items which exhibit DIF are freed to have their group-specific item parameters while the rest of the items are fixed to have the same parameters across all groups. Using this strategy, the non-invariance problem is solved, and group comparison on the same scale is made

possible. However, Robitzsch and Lüdtke (2022) state that the partial invariance method is not an optimal approach because items with group-specific parameters do not contribute to the linking process. This is a threat to validity as group comparisons will depend on different sets of items. Furthermore, partial invariance only works if DIF effects are sparsely distributed, i.e., only a small number of items exhibit DIF.

Another approach is linking under the assumption of full non-invariance. In this approach, no assumptions concerning MI need to be made. In this modeling approach, groups are calibrated separately in the first step, and in the second step, item parameters are brought onto a common scale, and group means are estimated. In other words, this strategy does not require items to have equal parameters across groups. In a simulation study, Robitzsch and Lüdtke (2020) demonstrated that robust linking approaches (robust Haberman and robust Haebara) outperformed partial invariance approaches in some conditions. Thus, they questioned the superiority of the partial invariance approaches for yielding stable estimates of group means to linking approaches based on separate calibrations in the context of ILSAs. The advantage of robust linking methods is that they circumvent the definition of the RMSD cutoff values. Since the robust linking approaches also rely on the partial invariance approaches, the question is whether one should believe in partial invariance modeling or not. In the following section, the functions and procedures for linking approaches are provided. As noted above, in the first step, each group should be analyzed separately. First, we estimate the 2PL model for each country:

```
library(dplyr)

chile.data <- dplyr::filter(data, IDCNTRY==152)
# takes only those participants with IDCNTRY=152 (i.e., Chilean students)

mod3<- TAM::tam.mml.2pl (chile.data, pweights = data$HOUWGT)# estimate the
2PL model for Chile
```

The procedure above is repeated for all the countries:

```
finland.data <- dplyr::filter(data, IDCNTRY==246)
mod4<- TAM::tam.mml.2pl (finland.data, pweights = data$HOUWGT)
georgia.data <- dplyr::filter(data, IDCNTRY==268)
mod5<- TAM::tam.mml.2pl (georgia.data, pweights = data$HOUWGT)
hongkong.data <- dplyr::filter(data, IDCNTRY==344)
mod6<- TAM::tam.mml.2pl (hongkong.data, pweights = data$HOUWGT)
italy.data <- dplyr::filter(data, IDCNTRY==380)
mod7<- TAM::tam.mml.2pl (italy.data, pweights = data$HOUWGT)
```

The data for the five groups/countries were analyzed with the 2PL model separately and stored under objects ‘mod3’, ‘mod4’, ‘mod5’, ‘mod6’, and ‘mod7’. In the next step, these analyses are linked:

```
models <- list (mod3, mod4, mod5, mod6, mod7)# a list of the separate
models is defined
lmod1 <- TAM::tam.linking(models, type="SL")# separate analyses are linked
with the Stocking-Lord linking method
summary(lmod1)
```

Alternatively, Haebara and robust Haebara linking methods can be used:

```
lmod2 <- TAM::tam.linking(models, type="Hae")# Haebara linking method
lmod3 <- TAM::tam.linking(models, type="RobHae")# robust Haebara linking
method
```

The Stocking-Lord linking method returns the output displayed in Figure 7. The first table shows the number of items used for linking. Transformation constants for item parameters and transformation constants for person parameters refer to the constants needed for adjusting the item parameters. The last

table in Figure 7 shows the group means and standard deviations on the same metric. As the table shows Finland (Study2) has the highest mean and Georgia (Study 3) has the lowest. This is in line with the MGIRT results before linking.

```
method = joint
type = SL  | Stocking Lord Linking Method

-------------------------------------------------------------
Number of Linking Items
       study1 study2 study3 study4 study5
study1      0    204    204    204    204
study2    204      0    204    204    204
study3    204    204      0    204    204
study4    204    204    204      0    204
study5    204    204    204    204      0
-------------------------------------------------------------
Transformation Constants for Item Parameters
           a      b
study1 1.000  0.000
study2 0.790 -0.802
study3 1.074  0.345
study4 0.826 -0.362
study5 1.029 -0.422
-------------------------------------------------------------
Transformation Constants for Person Parameters
           a      b
study1 1.000  0.000
study2 1.265  1.014
study3 0.931 -0.321
study4 1.211  0.438
study5 0.972  0.410
-------------------------------------------------------------
Means and Standard Deviations of Studies
            M    SD      d
study1  0.000 1.000  0.000
study2  1.014 1.265  0.943
study3 -0.321 0.931 -0.298
study4  0.438 1.211  0.407
study5  0.410 0.972  0.381
```

FIGURE 7.
Output from the Stocking-Lord Linking Method

3.5 Population Modeling and Plausible Values

In ILSAs, students' abilities are estimated on the basis of a small portion of the items and, thus, the ability parameters are not accurate. Consequently, population parameters computed based on the point estimates of proficiency will be biased. To circumvent this problem, students' test scores are estimated in the form of plausible values (PV). To obtain PVs, a probability distribution for students' θ

parameter is estimated. In other words, instead of estimating a single θ for each student, a range of possible thetas are estimated for each examinee. Plausible values are a random selection from this possible range of θ values which is referred to as the posterior distribution (Wu, 2004)[1].

Since examinees with the same raw scores might have different PVs, they are not appropriate to be used as individual ability scores for reporting back to the students. PVs are used to estimate population parameters such as mean, variance, and percentiles, and in secondary analyses to uncover the relationships between proficiency and some background variables (Wu, 2005).

Wu (2005) states that when the aim is to estimate statistics for subgroups of test takers, such as age, gender, etc., these groups must be included in the estimation of PVs. She further argues that secondary analyses on the relationship between PVs and some background variables will be reliable if the background variables of interest were included in the generation of PVs. Thus, to further improve the accuracy of PVs in ILSAs, latent regression modeling also known as population modeling is employed (Khoramdel et al., 2020). In the population model it is assumed that the posterior distribution of the ability depends both on the item responses and some background variables such as gender, education, socioeconomic status, etc. In this approach, the ability variable is regressed on the background variables, i.e., background variables are used as predictors to predict the IRT item parameters. After the regression parameters are estimated, PVs are drawn.

4. Conclusion

In large-scale assessments, to efficiently and cost-effectively estimate group parameters, a test design referred to as multiple matrix sampling is employed. In multiple-matrix sampling, each examinee answers a portion of the items sampled from the pool of items that represent the domain of interest. In this design, all the items are answered by each group while each examinee is exposed to a small number of items to avoid student fatigue. Since single-group IRT analysis is not suitable for heterogeneous populations, Bock and Mislevy (1981) and Bock and Zimowski (1997) proposed an IRT model suitable for group-level analysis. MGIRT has several advantages, including applicability to matrix sampled data, straightforward differential item functioning analysis, and test equating, amongst others.

In this tutorial, a nontechnical introduction to the multiple-group item response theory model and its applications in large-scale assessments was provided. Using the TAM package (Robitzsch et al., 2022), a small TIMSS 2019 grade 8 science assessment dataset was analyzed with the MGIRT, and the outputs were elaborated. In this walkthrough, group parameters, analysis of DIF or non-invariance using RMSD, and linking with separate calibrations under the assumption full non-invariance was demonstrated were demonstrated. This tutorial should enable researchers and practitioners to apply MGIRT in research and large-scale projects.

Funding

The author(s) received no specific funding for this work from any funding agencies.

Conflict of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1 Wu (2005), using a simulation study, showed that when a test is well-targeted (i.e., test difficulty and population ability match), MLE (Maximum Likelihood Estimate) and WLE (Weighted Maximum Likelihood Estimates (Warm,1985, 1989) are as good as PVs to recover population mean but not the population variance. When a correction factor was applied, MLE and WLE could also recover the population variance. However, they also showed that when a test is off-target, only PVs can recover population parameters, and the application of the correction factor does not fix the problem.

Data Availability Statements

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

How to Cite

Baghaei, P., & Robitzsch, A. (2025). A tutorial on item response modeling with multiple groups using TAM. *Educational Methods & Psychometrics, 3:*14. https://dx.doi.org/10.61186/emp.2025.1

References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42*, 357−374. https://doi.org/10.1007/BF02293656

Bristol, J., Mullis, I. V. S., Fishbein, B., & Foy, P. (2023). Reviewing the PIRLS 2021 achievement item statistics. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and procedures: PIRLS 2021 technical report* (pp. 9.1–9.45). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5892

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In van der Linden, W. J. & Hambleton, R. K. (Eds.) *Handbook of modern item response theory* (pp. 433–448). Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Bock, R. D., Mislevy, R. J., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher, 11*(3), 4−11. https://doi.org/10.3102/0013189X011003004

Bock, R. D., & Mislevy, R. J. (1981). An item response model for matrix sampling data: The California grade-three assessment. *New Directions for Testing and Measurement, 10,* 65–90.

Fishbein, B., Foy, P., & Tyack, L. (2020). Reviewing the TIMSS 2019 achievement item statistics. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 10.1-10.70). Retrieved from Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-10.html

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In von Davier, M. & Hastedt, D. (Eds.) *IERI monograph series: Issues and methodologies in large-scale assessments* (pp. 125–156). IEA-ETS Research Institute.

Halamová, J., Kanovský, M., Gilbert, P., Troop, N. A., Zuroff, D. C., Petrocchi, N., Hermanto, N., Krieger, T., Kirby, J. N., Asano, K., Matos, M., Yu, F., Sommers-Spijkerman, M., Shahar, B., Basran, J., & Kupeli, N. (2019). Multiple group IRT measurement invariance analysis of the Forms of Self-Criticising/Attacking and Self-Reassuring Scale in thirteen international samples. *Journal of Rational-Emotive & Cognitive-Behavior Therapy, 37*(4), 411–444. https://doi.org/10.1007/s10942-019-00319-1

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Johnson, M., & Lord, F. (1958). An empirical study of the stability of a group mean in relation to the distribution of test items among students. *Educational and Psychological Measurement, 18*(2), 325−329. https://doi.org/10.1177/001316445801800209

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software *mdltm* including parallel EM algorithm. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models. Methodology of educational measurement and assessment* (pp. 603–628). Springer. https://doi.org/10.1007/978-3-030-05584-4_30

Khorramdel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: Principles of item response theory and multiple imputations. In: Maehler, D., & Rammstedt, B. (Eds.), *Large-scale cognitive assessment. Methodology of educational measurement and assessment* (pp. 27–47). Springer, Cham. https://doi.org/10.1007/978-3-030-47515-4_3

Knapp, T. (1968). An application of balanced incomplete block design to the estimation of test norms. *Educational and Psychological Measurement, 28*, 265−272. https://doi.org/10.1177/001316446802800206

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational and Behavioral Statistics, 8*, 271–288. https://doi.org/10.3102/10769986008004271

OECD (2017). *PISA 2015 technical report.* Paris: OECD Publishing.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 30*(6), 859–870. https://doi.org/10.1080/10705511.2023.2191292

Robitzsch, A., & Lüdtke, O. (2022). Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *Journal of Educational and Behavioral Statistics*, *47*, 36–68. https://doi.org/10.3102/10769986211017479

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test Analysis Modules*. R package version 4.1-4. https://CRAN.R-project.org/package=TAM

Robitzsch, A. (2022). Estimation methods of the multiple-group one-dimensional factor model: Implied identification constraints in the violation of measurement invariance. *Axioms, 11*, 119. https://doi.org/10.3390/axioms11030119

Robitzsch, A. (2022). Statistical properties of estimators of the RMSD item fit statistic. *Foundations*, *2*(2), 488–503. https://doi.org/10.3390/foundations2020032

Robitzsch, A., & Lüdtke, O. (2020). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, *62*(2), 233–279.

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

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Ballinger Publishing Company.

Shoemaker, D. M., & Shoemaker, J. S. (1981). Applicability of multiple matrix sampling to estimating effectiveness of educational programs. *Evaluation and Program Planning*, *4*(2), 151–161. doi:10.1016/0149-7189(81)90005-7

von Davier, M., & Bezirhan, U. (2023). A Robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement, 83*(4), 740–765. https://doi.org/10.1177/00131644221105819

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions, 18*(2), 976–978.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Manuscript Received: 18 May 2024
Final Version Received: 15 Nov 2024
Published Online Date: 30 Jan 2024