Guidelines for measurement invariance and aligment methods using library(rd3c3)

Sandoval-Hernández, Carrasco & Eryılmaz

2024-12-14

Table of contents

Preface

`library(rd3c3)` is A ***wrapper function library*** designed to fit model-based invariance models and alignment methods to assess cross cultural comparability of responses between countries and groups, on international large-scale assessments.

It helps researchers to write Mplus code, via MplusAutomation to implement invariance analysis.

It uses graded response models with a probit link (i.e., confirmatory factor analysis with ordinal indicators).

How to install

```
# -----------------------------------------------
# install rd3c3
# -----------------------------------------------

devtools::install_github(
  'dacarras/rd3c3',
  force = TRUE
)
```

How to cite the current document

Sandoval-Hernández, A.; Carrasco, D; & Eryilmaz, N. (2024). Guidelines for measurement invariance and aligment methods using `library(rd3c3)`. figshare. Software. https://doi.org/10.6084/m9.figshare.28028564

Acknowledgement

This work was supported by the IEA Research and Development Fund.

1 Introduction

Invariance is a desired property of total scores generated with polytomous items, to allows comparisons among collections of observations on those total scores (Tse, Lai, & Chang, 2024). These collections of observations include meaningful factors such as sex, age, the language of the test or survey, and the participating country, among different groups of interest. Ideally, response models are used to provide evidence (or the lack of it) that comparisons among these groups can be made on the same scale when total scores are used.

There are different approaches to assess if the assume comparability is tenable. The most popular approaches include differential item functioning (DIF), and factorial invariance (Thissen, 2024). The most simple version of differential item functioning (uniform DIF), consist of the study of item location parameters conditional to group membership of the observations. DIF is a procedure to produce evidence that the expected response onto items, is similar among groups if the level of the attribute is similar between groups, or not. Thus, if two (or more) groups of the same attribute level do differ on the expected responses to an item, this scenario is taken as evidence of DIF (e.g., Wu et al, 2016) . Factorial invariance or measurement invariance, is a model based strategy, where different response models are fitted onto the groups of interest, to assess the equivalence of the response model besides groups latent mean differences. As such, comparisons of interest among groups includes not only location item paratemers, but also factor loadings, and residuals or item uniqueness (Kline, 2023). Traditionally, a descriptive model is specified (i.e. configural model), and a sequence of more constrained models are fitted, where different model parameters are held equal among groups till the most equivalent model specification is fitted where factor loadings (metric invariance), item location (scalar invariance) and item uniques parameters are held equal, and only latent mean differences are allow to vary (strict invariance) (e.g., Dimitrov, 2010).

Within this later approach there are current development that deviate from the common practice of confirmatory factor alysis (CFA) for continous indicators, specially recommended for confirmatory factor analysis fitted onto ordinal indicators (e.g., Wu & Estabrook, 2016; Svetina, Rutkowski & Rutkowski, 2020; Tse, Lai & Chang, 2024). Two points of contention are of special relevance for the present guideline. These are the order in which the different model specifications should fitted; and if the different invariance model specifications describe for CFA with continous indicators are identified for response models with ordinal indicators.

The present guidelines are focus on how to produce different response model specifications, using the `library(rd3c3)` in R. This is an R library, with a collection of different wrapper functions that helps to speed up the process of fitting different model specifications (i.e., strict, scalar, configural) onto large scale assessment studies. `library(rd3c3)` follows the work of Wu & Estabrook (2016) on model identification, and starts first with item threshold invariance as the first step to build the model sequence of measurent invariance. An follows the Svetina et al (2020) and Tse et al (2024) on model specification to produce scalar, and strict invariance model specifications.

Additionally, `library(rd3c3)` provides wrapper functions to fit alignment methods. These are invariance model specification that relaxes the models parameter constrains among groups in search of the least discrepant model estimates across the compared groups (Muthén & Asparouhov, 2014).

The following guideline is structure as follows: we first review measurement invariance within the propose response model, the graded response model (section 1); we then proceed to discuss the limitations of partially invariant models (section 2); we described the library in general terms (section 3); we provide a full example of invariance analysis (section 4); and finally we close the present guidelines with a section of intended uses (section 5).

1.1 References

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. Measurement and Evaluation in Counseling and Development, 43, 121–149. https://doi.org/10.1177/0748175610373459

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

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. Supplemental Material. Frontiers in Psychology, 5, 23–31. https://doi.org/10.3389/fpsyg.2014.00978

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling: A Multidisciplinary Journal, 27(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Thissen, D. (2024). A Review of Some of the History of Factorial Invariance and Differential Item Functioning. Multivariate Behavioral Research, 0(0), 1–25. https://doi.org/10.1080/00273171.2024.239614

Tse, W. W. Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. Behavior Research Methods, 56(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

Wang, J., & Wang, X. (2020). Confirmatory Factor Analysis. In Structural Equation Modeling: Applications Using Mplus (pp. 33–117). John Wiley & Sons, Inc. https://doi.org/10.4324/9781315832746-25

Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. Psychometrika, 81(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

2 Response model

The present guideline is focused on measurement invariance models for confirmatory factor analysis for ordinal indicators. In particular, we are focusing on the graded response model with probit link (Bovaird & Koziol, 2012).

2.1 Graded Response Model

The graded response model can be found in the literature also under the name of confirmatory factor analysis for ordinal indicators, the two parameter normal ogive form of the graded response model (Wang & Wang, 2020), and as a graded response model with probit link (Bovaird & Koziol, 2012). The graded response model is a response model propose for ordinal indicators, proposed by Samejima (1968). Historically, it appears before the partial credit model (Masters, 1982), which is the most popular response model to generate scores across several large scale assessment studies (Carrasco, Irribarra & González, 2022). This model includes different model variants. These variants include the homogenous and the heterogeneous case (Samejima, 2016). The homogenous case, is a model where factor loadings (or item slopes) are fixed to be common across items; while the heteorogenous case leaves the item slope parameters to vary freely. Moreover, this model can be specified with different link functions, the logit function and the probit function (Bovaird & Koziol, 2012).

We will review the formal presentation of these two variants, so is easier to make a bridge between polytomous item response theory models, and confirmatory factor models. Formally, these two models can be expressed with the following equations:

$$Pr(y_{ip} > k) = \frac{exp[a_i(\theta_p - b_{ik})]}{exp[1 + a_i(\theta_p - b_{ik})]} \tag{2.1}$$

The GRM model with a logit link expressed the probability of responses y to item i from person p as a ratio. In the exponentiated numerator we have the propensity to choose the ordered categories in a direction (θ_p), minus the boundary category parameter b for the ith item category k or higher, multiplied by the slope of the items a_i. The parameter a_i is often interpreted as a discrimination parameter, because the higher is its value, the higher is the separation between low and high attribute persons in their expected response probability. In the denominator of the previous formula, we repeat the previous term, the exponentiation of the propensity minus the boundary category parameter times the slope, plus a unity. The present formula can be expressed in a more concise manner, by calling the logit link function.

$$logit[Pr(y_{ip} \leq k)] = a_i(\theta_p - b_{ik}) \tag{2.2}$$

Graded response models (GRM) with logit link are very similar to partial credit model (PCM), under the homogeneous case. In the homogenous case, the a_i can be constrained to one, and then only the person locations (θ_p) and item locations (b_{ik}) are included in the model. The main difference between these two models is their logit link function. While the PCM includes the adjacent logit link; the GRM relies on the cumulative category link (Mellenbergh, 1994). Thus, for items with three ordered response categories, the item locations are the natural logarithms of the odds of answering 1 vs 2, and 2 vs 3 for the adjacent logit link; while for the cumulative link function consists of natural logarithms contrasting the odds of answering 1 vs 2, 3; and 1, 2 vs 3 (Carrasco et al., 2022).

An alternative formulation for the present model and the focus of the present guideline is teh GRM with the probit link. Following Bovaird & Koziol (2012), we express this model with the next equation:

$$Pr(y_{ip} > k) = \phi(-\tau_{ik} + \lambda_i\theta_p) \tag{2.3}$$

Similary to the previous equation, we can express the previous formula in a more concise manner by using the probit link in the equation:

$$probit[Pr(y_{ip} > k)] = \tau_{ik} - \lambda_i\theta_p \tag{2.4}$$

This second formulation is more akin to the confirmatory factor analysis tradition, where is common to include item intercepts (i.e., item location parameters), item slopes (i.e., factor loadings) and a term for the theoretical factor.

2.2 Invariance model specifications with the GRM model

A response model can be considered invariant if all response model parameters can be held equal across groups, besides the group latent means. This general idea applies to polytomous item response models such as confirmatory factor analysis, graded response models (e.g., Wu & Estabrook, 2016; Tse et al., 2024) and to mixture variants of response models (e.g., Masyn, 2017; Torres Irribarra & Carrasco, 2021); that is response models with latent factors that are discrete instead of normally distributed (Torres Irribarra, 2021). This is the most demanding for equivalence of responses models between groups, usually described as strict invariance. A more relaxed version of invariance model specification is *scalar invariance*, in which all response model parameters are held equal across groups, beside latent means and item uniqueness or scale factors (e.g., Grimm et al. 2016; Tse et al., 2024). Moreover, model specification with more parameters allow to vary freely are not able to provide latent mean comparisons. These includes models with common thresholds but free factor loadings, and purely descriptive models were all response models parameters are allowed to vary freely.

Grouping variables can include sociodemographic variables such as age, students sex, and parents education. Yet, in large scale assessment, a popular grouping variable of interest is countries. Thus, if response model parameters can be considered invariant, one can assure that countries can be compared in a common scale.

A common practice in measurement invariance with CFA for continuous indicators is to start with the model with less constrains (e.g., Dimitrov, 2010), and continue further till the most constrained model (i.e., strict invariance). In essence, this is a model building sequence (Kline, 2023). In this sequence, difference model specifications are included. The first, is the **configural** model specification where only the model structure is common, yet all response model parameters vary freely between groups. Then is followed by the **metric** model specification where only factor loadings are held equal. Yet, in this model specificification there are no parameters in the multi-group model to compare latent means between groups (e.g. Wu & Estabrook, 2016). Hence, this model specification is similar to a latent centered means. In a third stage, the **scalar** model specification is included. In the scalar model, factor loadings, indicators intercepts, are held common among groups, while latent means are constrained (i.e., one group has a latent of mean zero and is used as a point of reference). This model allows for latent mean comparison among groups. Finally, the most constrained model specification, the **strict** model, held as common parameters all response model parameters with the exemption of latent means. In the model building sequence, the new parameter that is held common among groups are indicators uniqueness. This last model specification also allowes to compare groups on latent means, while assuming residual error of the response model is common among groups.

Model specification sequence for assessing invariance on CFA with ordinal indicators, is different from CFA with continuous indicators. Wu & Estabrook (2016) asserts that invariance within the CFA for ordinal indicators common thresholds are needed before common factor loadings can be introduced in the model building sequence (Wu & Estabrook, 2016; Svetina, et al. 2020; Tse, et al, 2024). In practice, common factor loadings between group cannot be tested alone (Wu & Estabrook, 2016, p1023). Complementary, Tse et al. (2024) recommends to assess if strict invariance holds among groups, before relying on total scores (e.g., observed means) for group comparisons. Then, if strict invariance fails, then proceed to search for partially invariant solutions such as, partially strict invariance, and scalar invariance if latent means can be used instead of observed mean scores. Following Tse et al. (2024) one can alter the model sequence for a model trimming sequence instead (Kline, 2023). That is, instead of starting with the model with the most freely estimated parameters, one can start with the model with the most held equal parameters among groups (the most constrained). As such, the model sequence for GRM would be: strict, scalar, configural (with common thresholds), and a base model (with freely estimated response model parameters).

In the following figure, we summarize the parameters of the response models that can be held equal between groups in each of the model specification for CFA with continuous and for CFA with ordinal indicators.

CFA continous indicators			CFA categorical indicators		
Specification	Parameters	Constriction	Specification	Parameters	Constriction
configural	loadings	free	base	loadings	free
	intercepts	free		thresholds	free
	uniqueness	free		uniqueness	free
	latent means	centered		latent means	centered
metric	loadings	equal	metric	loadings	equal
	intercepts	free		thresholds	free
	uniqueness	free		residuals	free
	latent means	centered		latent means	free
			threshold	loadings	free
				thresholds	equal
				uniqueness	free
				latent means	centered
scalar	loadings	equal	scalar	loadings	equal
	intercepts	equal		thresholds	equal
	uniqueness	free		uniqueness	free
	latent means	constrained		latent means	constrained
strict	loadings	equal	strict	loadings	equal
	intercepts	equal		thresholds	equal
	uniqueness	equal		uniqueness	equal
	latent means	constrained		latent means	constrained

Figure 2.1: Figure 1: response model parameters being held equal in each model specification.

The present table is a rough summary of the different response model parameters that are held equal among groups to specify each model solution. For example, in Wu & Estabrook (2016) the configural model specification for the CFA with ordinal indicators consists of a model where thresholds are held equal among groups. This is the baseline model from which model comparisons can be made in contrast to scalar and strict solutions of the GRM model. However, the configural model has more constraints than solely common thresholds, this includes factor means constrained to zero (i.e., centered) and factor variances fix to 1 on all groups (see Svetina et al., 2020). Additionally, Tse et al. (2024) discusses alternative model specification for the configural solution, using the theta parametrization in which factor loadings are held common between groups, and thresholds are held common for marker indicators. In the present guidelines we will review these model specifications in more detail in section 4, following Svetina et al. (2020) and Wu & Estabrook (2016).

It should be clear that model specifications propose for CFA with continuous indicators are not equivalent for other response models. The weak invariance (e.g., Dimitrov, 2010) or metric invariance model specification (Wu & Estabrook, 2016), where common factor loadings are held equal across groups, do not reach a model specification that holds the same interpretation of traditional CFA, for CFA with ordinal indicators (Wu & Estabrook, 2016; Svetina, et al. 2020; Tse, et al, 2024) if thresholds are allow to vary freely. A similar observation can be done for the assumed interpretation of the metric model specification of latent class models (e.g., Hooghe & Oser, 2015; Hooghe et al. 2016), which is an special case of a non-invariant solution (Masyn, 2017), and doesn't hold the same interpretation of the random term across groups, the configuration of the latent classes (Torres Irribarra, et al., 2021). If invariance holds, the purpose is to assert that group differences are on the random term of the response model in a common scale (i.e., factors, latent means, latent classes); in contrast, if the model specification doesn't provide group differences estimates on the same scale, then substantive conclusions are not tenable, because these do not have a common meaning between groups. In summary, the interpretation one can hold over response models fitted between groups with varying equality constrains are not equivalent between response models.

In the following section (section 2) we will describe what are partially invariant solutions.

2.3 References

Carrasco, D., Irribarra, D. T., & González, J. (2022). Continuation Ratio Model for Polytomous Items Under Complex Sampling Design. In Quantitative Psychology (pp. 95–110). https://doi.org/10.1007/978-3-031-04572-1_8

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. Measurement and Evaluation in Counseling and Development, 43, 121–149. https://doi.org/10.1177/0748175610373459

Bovaird, J. A., & Koziol, N. A. (2012). Measurement Models for Ordered-Categorical Indicators. In R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling (pp. 495–511). Guilford Press.

Grimm, K. J., & Liu, Y. (2016). Residual Structures in Growth Models With Ordinal Outcomes. Structural Equation Modeling, 23(3), 466–475. https://doi.org/10.1080/10705511.2015.1103192

Hooghe, M., & Oser, J. (2015). The rise of engaged citizenship: The evolution of citizenship norms among adolescents in 21 countries between 1999 and 2009. International Journal of Comparative Sociology, 56(1), 29–52. https://doi.org/10.1177/0020715215578488.

Hooghe, M., Oser, J., & Marien, S. (2016). A comparative analysis of 'good citizenship': A latent class analysis of adolescents' citizenship norms in 38 countries. International Political Science Review, 37(1), 115–129. https://doi.org/10.1177/0192512114541562.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. Psychometrika, 47(2), 149–174. https://doi.org/10.1007/BF02296272

Masyn, K. E. (2017). Measurement Invariance and Differential Item Functioning in Latent Class Analysis With Stepwise Multiple Indicator Multiple Cause Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 24(2), 180–197. https://doi.org/10.1080/10705511.2016.1254049

Samejima, F. (1968). Estimation of latent Ability using a response pattern of graded scores. ETS Research Bulletin Series, 1968(1), i–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Samejima, F. (2016). Graded Response Models. In W. J. van der Linden (Ed.), Handbook of Item Response Theory. Volume One. Models (pp. 95–107). CRC Press. https://doi.org/10.1201/9781315374512-16

Torres Irribarra, D. (2021). A Pragmatic Perspective of Measurement. Springer International Publishing. https://doi.org/10.1007/978-3-030-74025-2

Torres Irribarra, D., & Carrasco, D. (2021). Profiles of Good Citizenship. In E. Treviño, D. Carrasco, E. Claes, & K. J. Kennedy (Eds.), Good Citizenship for the Next Generation. A Global Perspective Using IEA ICCS 2016 Data (pp. 33–50). Springer International Publishing. https://doi.org/10.1007/978-3-030-75746-5_3

Tse, W. W. Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. Behavior Research Methods, 56(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

Wang, J., & Wang, X. (2020). Confirmatory Factor Analysis. In Structural Equation Modeling: Applications Using Mplus (pp. 33–117). John Wiley & Sons, Inc. https://doi.org/10.4324/9781315832746-25

Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. Psychometrika, 81(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

3 Partial invariance

Partially invariant models are model specifications that allows for some item indicators parameters to vary freely, while still having enough common parameters in the response model among groups. These models offer a way to generalize findings based on the invariant parameters while excluding those that exhibit non-invariance in the results generalization (Meredith, 1993; Millsap, 2011). Hence, the name partially invariant.

These models allow researchers to identify group differences or treatment effects reliably for responses to items that remain consistent across groups, while keeping non invariant indicators in the model. For instance, if a treatment effect holds equally for 10 out of 12 items within a scale, assertions between treated and non treated can be made safely for these 10 items, while not making such claims onto the two non-invariant indicators (e.g., Gilbert, 2024).

However, if the number of non-invariant parameters in the response models is too large, then is less credible the ability of making claims that are applicable to all compared groups across all items. As a point reference, Muthén, & Asparouhov (2014) suggest that if .75 of the parameters between groups are held common (while .25 of the response parameters are non-invariant), latent means comparisons between groups are of good enough quality. Yet, such a threshold can be put to the test with Monte Carlo studies, for the speficities of a response model, and the amount of groups being compared (Muhen & Asparouhov, 2014, p3), and the research purpose.

There are few caveats to considered regarding partially invariant models when group comparisons are of interest. If non-invariant items are excluded, this selective exclusion can alter the meaning of the total score. Exclusion of non-invariant items can narrow the scope of the attribute of interest. For example, if one has three groups, and twelve items, is possible to have a scenario in which the response model is invariant for two of the three groups. And simultaneously, the measurement equivalence could hold with only four out of the twelve items for the three groups. As such, researchers can have the dilemma of narrowing the amount of indicators at the cost of reliability; and compare the three groups. Or, to do comparison across all the items for only two of the three groups. Yet, with partially invariant model while is allowed to keep all indicators in the response model, the meaning of the group differences would partially generalize to the responses where the response models are held equal. Restricting the comparison to only the common items among the three groups can restrict the scope of the intended interpretations. This a central problem for cross-cultural studies, where including more diverse groups can augment the chances of non-comparability (Van De Vijver & Matsumoto, 2011). In essence, partially invariant models can pose interpretive challenges. The treatment effects or group differences identified in these models may not fully represent the intended construct's complexity. Researchers must exercise caution in claiming generalizability across indicators, ensuring transparency in reporting the extent of invariance and acknowledging the limitations of their results (Fischer et al., 2019; Van de Schoot et al., 2012).

In practical terms, the process of implementing partially invariant models is time-consuming and often requires significant manual intervention (Svetina et al., 2020; Robitzsch & Lüdtke, 2023). Arranging and adjusting the model to exclude non-invariant items, or to freely estimate response model indicators between partially comparable groups, demands meticulous attention to detail and and iterative testing process. As a whole, is a procedure which can hinder efficiency. This labor-intensive aspect of the method underscores the need for more streamlined analytical tools or automated procedures to facilitate its application in large-scale studies.

Alignment methods (Muthén & Asparouhov, 2014; Asparouhov & Muthén, 2014; Asparouhov & Muthén, 2022) are a collection of procedures which are helpful in finding the least discrepant solution among groups for a given response model. This method searches for an optimal solution where the amount of discrepant parameters (i.e., non-invariant) is minimized. Although, is a procedures which helps to searchs partially invariant solutions, the resulting partially invariant solutions are conditional to the selection algorithm (Pokropek, Lüdtke, & Robitzsch, 2020). Thus, is not a method which would yield non debatable partially invariant solutions, but plausible partially invariant solutions. As such, is the researcher who would need to make judgment call regarding if the reached solution is a useful model specification for their purposes, taking into account its limitations.

In conclusion, while partially invariant models offer a practical approach to addressing measurement invariance challenges, their limitations highlight the importance of careful interpretation and methodological rigor. Aligment methods offer an interisting tool to search for partially invariant model specification in cases where strict and scalar invariance is not held. Apart from aligment methods, there are more alternatives that are aim at addressing the challenges of comparing many groups such as bayesian aproximate invariance, measurement invariance via multilevel models, mixture multigroup factor analysis among others (see Leitgöb et al, 2023). These other alternatives, besides aignment methods are out of the scope of the present guidelines

In the following section (section 3), we describe what is the `library(rd3c3)`, and how it can help to fit model based measurement invariance onto graded response models, and how it helps to fit alignment method optimization onto the same response models.

3.1 References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. Structural Equation Modeling: A Multidisciplinary Journal, 21(4), 495–508. https://doi.org/10.1080/10705511.2014.91921(

Asparouhov, T., & Muthén, B. (2022). Multiple group alignment for exploratory and structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 1–23. https://doi.org/10.1080/10705511.2022.2127100

Fischer, J., Praetorius, A. K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. Educational Assessment, Evaluation and Accountability, 31, 201–220.

Gilbert, J. B. (2024). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. Behavior Research Methods, 56(5), 5055–5067. https://doi.org/10.3758/s13428-023-02245-8

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. Social Science Research, 110(October 2022). https://doi.org/10.1016/j.ssresearch.2022.102805

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. Psychometrika, 58(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). Statistical Approaches to Measurement Invariance. New York, NY: Routledge.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. Supplemental Material. Frontiers in Psychology, 5, 23–31. https://doi.org/10.3389/fpsyg.2014.00978

Pokropek, A., Lüdtke, O., & Robitzsch, A. (2020). An extension of the invariance alignment method for scale linking. Psychological Test and Assessment Modeling, 62(2), 305–334.

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, 30(2), 190–204. https://doi.org/10.1080/10705511.2023.2191292

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. Structural Equation Modeling: A Multidisciplinary Journal, 27(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. European Journal of Developmental Psychology, 9(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Van De Vijver, F.J.R., Matsumoto, D. (2011) Introduction to the methodological issues associated with cross-cultural research. In: Matsumoto, D., van de Vijver, F.J.R. (Eds.), Cross-Cultural Research Methods in Psychology, 1st ed..,. Cambridge University Press, pp. 1–14.

4 Library overview

4.1 Summary

- `library(rd3c3)` is a collection of wrapper functions
- these wrapper functions are included within template to produce item analysis reports
- the item analysis reports provide different statistical and psychometric results of interest to make judgments regarding the quality of scale scores

4.2 `rd3c3` as a collection of wrappers

The `library(rd3c3)` is a collection of wrapper functions (Stanton, 2017) that helps to streamline the task of genereting code to fit different response model specifications. Wrapper functions are a way to call more complex functions and code into a simpler user interface. The main aim of the library is to ease the coding time needed when assessing measurement invariance of polytomous batteries of items from background questionnaires of large scale assessment studies. The different model specifications follow Wu & Estabrook (2014), Svetina, Rutkowski & Rutkowski (2020) and Tse, Lai & Chang (2024) recommendations to fit different multigroup models of the graded response model with probit link, to fit strict, scalar, common threshold and base (i.e., descriptive) multigroup models. Moreover, the library also contain functions to apply alignment methods onto the same response model among groups. It relies on `MplusAutomation` (Hallquist & Wiley, 2018), to fit response models using Mplus (Muthén & Muthén, 2017), so all fitted models can take into account sample design features of large scale assessment studies such as stratification variables, clustering variables, and survey weights (e.g.., Stapleton, 2013).

The handler name of the library is `rd3c3`. This handler stems from the research development call number 3 (rd3), focused on cross cultural comparison (c3).

4.3 Code applicable to a set of items

Most of the wrapper functions included in `library(rd3c3)` are not intended to be use onto sole objects, such as vectors or data frames. These are design to be fitted onto a set of elements, define in a table. Once the table, which we call generally `scale_info`, is filled-in and is called into the R session, the wrapper functions can resolve which items are subject to an analysis, within a define data object a particular data frame. We illustrate the general logic of the wrapper functions with the following diagram.

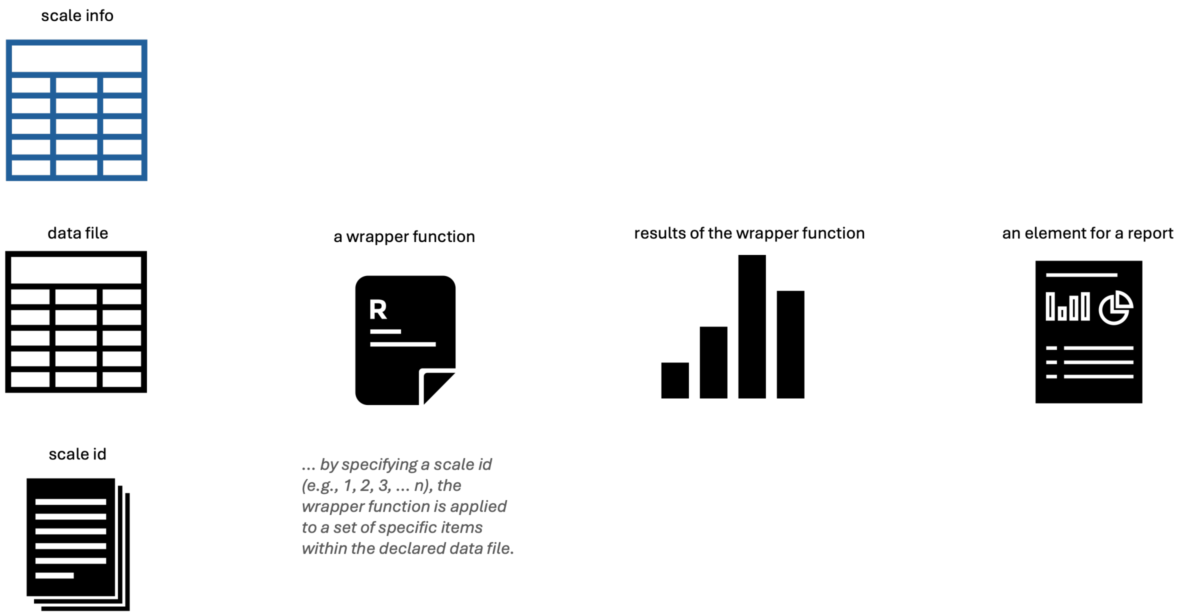

Figure 4.1: Figure 2: wrapper function logic

For the wrapper functions to work, one needs to provide:

- a `scale_info` table
- to specify within the table the `data_file` name with the responses of interest
- choose a particular scale, via a `scale_id`

As such, the library contains different wrapper functions that have the same arguments. These arguments usually present the following form:

```
rd3c3::name_of_the_function(
  data = data_responses,
  scale_num = scale_id,
  scale_info = scales_data
)
```

This is the case for different functions within `rd3c3`. For example, to name a few:

- `rd3c3::get_descriptives()`

  - is a wrapper function to produce nominal descriptives of items

- `rd3c3::missing_summary()`
  - it generates a summary of missing data across items, distinguishing complete, partial and missing response patterns by observations

- `rd3c3::fit_grm2()`
  - is a wrapper function that fits a graded response model with probit link onto a set of items declared in the `scale_info` table

- `rd3c3::ctt_table()`
  - it produces classical test theory statistics such as biserial correlations and alpha if deleted for set of items declared in the `scale_info` table

For example, the `rd3c3::get_descriptives()` function generates item descriptives including means, standard deviations, and histograms; and instead of being applicable to a matrix of responses is applied onto a whole data object. Thanks to the additional arguments included in the wrapper function, the code is able to select the items that are indexed with a `scale_id` number within the `scale_info` table. Thus, just by specifing the desired `scale_id` the user can get the results for these collection of items contained in a particular data file.

```
#--------------------------------------------------------------------------------
# descriptives
#--------------------------------------------------------------------------------

# -----------------------------------------------
# scale id
# -----------------------------------------------

rd3c3::silent(library(dplyr))

# -----------------------------------------------
# scale id
# -----------------------------------------------

scale_id <- 1

# -----------------------------------------------
# scales info
# -----------------------------------------------

scales_data <- readxl::read_xlsx(
               'guideline_scale_info_example.xlsx',
```

```r
               sheet = 'scales_data'
               )

# -----------------------------------------------
# data file
# -----------------------------------------------

data_file <- scales_data %>%
dplyr::filter(scale_num == scale_id) %>%
dplyr::select(data_file) %>%
unique() %>%
dplyr::pull()

# -----------------------------------------------
# response matrix
# -----------------------------------------------

data_responses <- readRDS(data_file)

# -----------------------------------------------
# descriptives
# -----------------------------------------------

rd3c3::get_descriptives(
  data = data_responses,
  scale_num = scale_id,
  scale_info = scales_data) %>%
dplyr::select(var, missing, complete, n, mean, sd, skew, kurt, hist) %>%
knitr::kable(.,
  digits = 2,
  caption = 'Table 1: descriptives example')
```

Table 4.1: Table 1: descriptives example

var	missing	complete	n	mean	sd	skew	kurt	hist
BSBG13A	0.04	0.96	7480	2.13	0.84	0.61	2.97	
BSBG13B	0.04	0.96	7480	1.82	0.77	0.78	3.34	
BSBG13C	0.05	0.95	7480	1.84	0.83	0.83	3.16	
BSBG13E	0.04	0.96	7480	1.88	0.85	0.81	3.12	

The generated results are the descriptives of items BSBG13A, BSBG13B, BSBG13C, BSBG13E from the “Sense of School Belonging” scale, present in TIMSS 2019, for Chile and England.

4.4 *Template based workflow*

The `library(rd3c3)` is intended to produce **item analys reports** of polytomous scales, in the spirit of dynamic reports (Xie, 2017). These are reproducible statistical analysis, that fills-in a define template. For any `scale_id`, a collection of items, one can generate an **item analys report** that include different sets of results.

This is in stark contrast to a *manually coded workflow*, where the user needs to code every function to a set of specific items within a data frame, many times to build a single item analysis report to every scale. In comparison, a *template based workflow* already contains an opinionated set of analysis (Parker, 2017) selected with a purpose. In this case, to make judgments of the quality of scale scores in terms of unidimensionality, reliability, comparability and inference limitations. The following diagram depicts the contrast between these two manners to reach the set of intended item analysis reports.

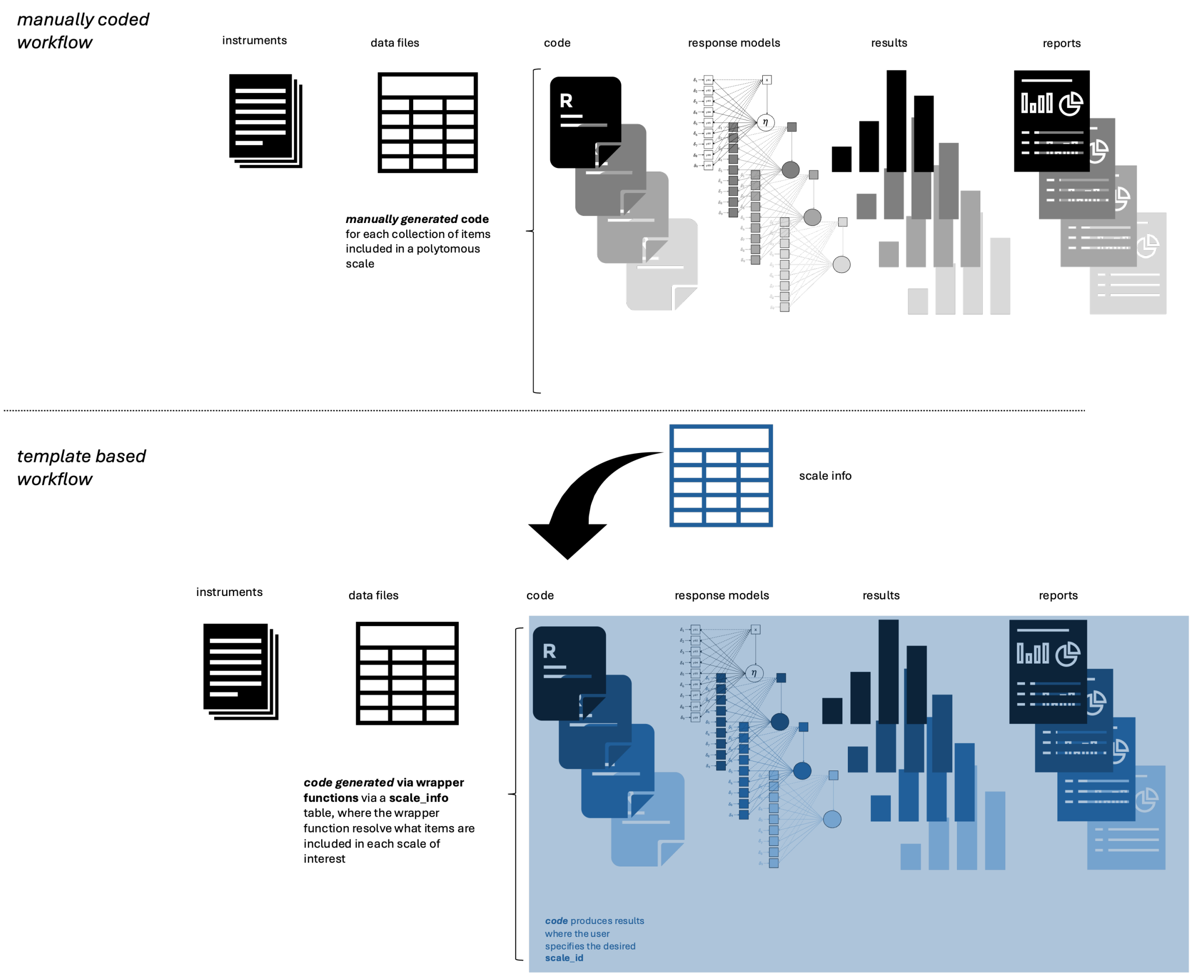

Figure 4.2: Figure 3: wrapper function logic within a dynamic report

In the *manually coded workflow* the user needs to code each set of analysis for each set of items. While in the second workflow, the *template based workflow*, the user just needs to choose the respective `scale_id` and can get the intended **item analys reports** that should be already planned.

The logic of a template is that allows a user to get a series of results regarding the responses to a collection of items that are intended to measure a known attribute (i.e., a construct). This templates includes a selection of the different results a researcher can use to make judgements regarding the quality of a total score from polytomous items scale.

A template for assesing the quality of scale scores should follow different design data analysis principles such as reproducibility and exhaustivity (McGowan et al., 2023). This template should be reproducible in the sense that other user, with the same data file, library, and needed software (i.e., Mplus), should reach the same results. Thus, such a template should be able to get the same results present in a generated report. Ideally a template for item analysis reports should follows the design principle of exhaustivity, in the sense of including different results that helps to make a judgment regarding the quality of the total score one can produce with the responses to a set of indicators. Using `library(rd3c3)` the users can include results from a graded response models, under different model specifications (i.e., strict, scalar, configural, base), and alignment methods for graded response models.

In essence, to build a dynamic **item analysis report**, the user needs to define the `scale_info`, define the `data_responses` of interest, and by using `library(rd3c3)` within a define **template** the user can include all the statistical analysis relevant for its purpose. As a whole, the user can generate dynamic results reports per scale. The following diagram summarize the minimal elements to produce these dynamic reports.

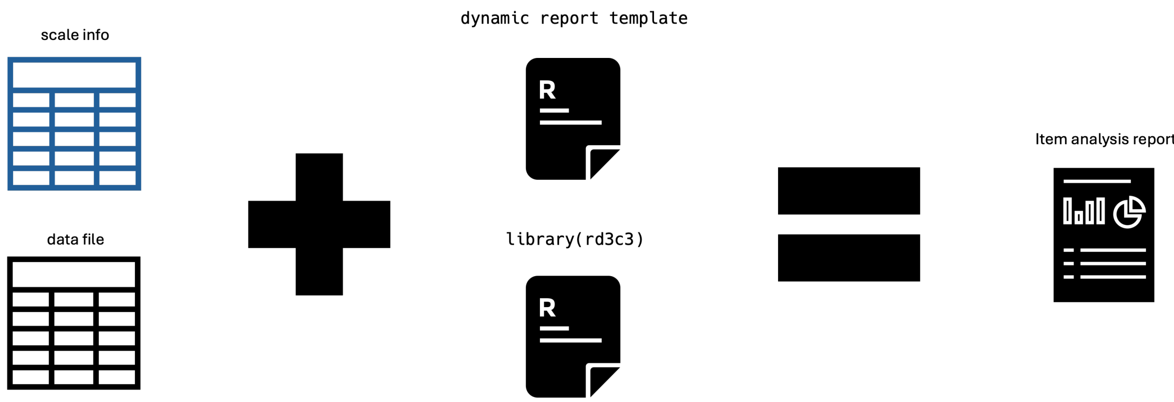

Figure 4.3: Figure 4: dynamic reports

4.5 Main pieces of evidence to judge the quality of scales scores

The main pieces of evidence that `librar(rd3c3)` can produce for item analysis reports are:

- Unidimensionality
- Reliability
- Comparability
- Inference limitations

Unidimensionality is judge based on parallel analysis for ordinal indicators (Lubbe, 2019).

Reliability of scale scores is judge by inspecting the distribution of errors of the latent factor of the GRM model, its summary via person separation reliability (Verhavert et al., 2018) and via the Cronbach's alpha index (Cronbach, 1951).

Comparability among participating countries is assess via the results of the model based measurement invariance results (Wu & Estabrook, 2016; Svetina, Rutkowski & Rutkowski, 2020; Tse, Lai & Chang, 2024), and the complementary results of the alignment analysis results (Muthén & Asparouhov, 2014).

Inference limitations can be made based on an holistic judgment of previous results, regarding to which locations of the latent continuum the scale score is more informative, via inspection of the person item map, and the distribution of standard errors of the theta realizations of the response model. For example if the distribution of item location is concentraded in one of the tails a researcher can identify possible ceiling or floor effects of the scale (e.g., Carrasco, Rutkowski, and Rutkowski, 2023). Moreover, the item analysis reports provides results of measurement invariance and alignment methods providing information regarding the tenability of assuming a common response model among the compared groups. Thus, the researcher can spot scales scores where the comparison among groups may not be guarantee and further research is needed. Further research could isolate points of support for comparisons, by iterating the decisions on item and group selections and exclusions.

In the following section (section 4), we illustrate the application of the `library(rd3c3)` to produce invariance analysis, and build an **item analysis report** with the described characteristics.

4.6 References

Carrasco, D., Rutkowski, D., & Rutkowski, L. (2023). The advantages of regional large-scale assessments: Evidence from the ERCE learning survey. International Journal of Educational Development, 102(May), 102867. https://doi.org/10.1016/j.ijedudev.2023.102867

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. Structural Equation Modeling: A Multidisciplinary Journal, 25(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Lubbe, D. (2019). Parallel analysis with categorical variables: Impact of category probability proportions on dimensionality assessment accuracy. Psychological Methods, 24(3), 339–351. https://doi.org/10.1037/met0000171

McGowan, L. D. A., Peng, R. D., & Hicks, S. C. (2023). Design principles for data analysis. Journal of Computational and Graphical Statistics, 32(2), 754-761.

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide (8th ed.). Muthén & Muthén.

Parker, H. (2017). Opinionated analysis development (pp. 1–13). https://doi.org/10.7287/peerj.preprints.3210

Stanton, J. M. (2017). Reasoning with Data. An Introduction to Traditional and Bayesian Statistics Using R. Guilford Press.

Stapleton, L. M. (2013). Incorporating Sampling Weights into Single- and Multilevel Analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of International Large scale Assessment: background, technical issues, and methods of data analysis (pp. 363–388). Chapman and Hall/CRC.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling: A Multidisciplinary Journal, 27(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Tse, W. W. Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. Behavior Research Methods, 56(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment?. Applied Psychological Measurement, 42(6), 428-445.

Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. Psychometrika, 81(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Xie, Y. (2017). Dynamic Documents with R and knitr. CRC Press.

5 Applied Examples

5.1 Summary

- We use data from TIMSS 2019, from the student background questionnaire.
  - In particular we are using student responses to the items of the scale "Sense of School Belonging", from Chile and England.
  - The data file includes 4115 students from Chile, and 3365 students from England, from 8th grade
  - We are using a tailor-made data file where we include specific clustering and survey design variables:
    * id_i = unique id number for students
    * id_j = unique id number for schools
    * id_s = unique id number for stratification factors (i.e., JKZONES)
    * id_k = unique id number for country samples
    * ws = scale survey weights to a constant of 1000
    * `data_ex1.rds`
- To install the library, a user can use the following code, to download the development version of the library:

```
# ---------------------------------------------
# install rd3c3
# ---------------------------------------------

devtools::install_github(
  'dacarras/rd3c3',
  force = TRUE
)
```

- We include three applied examples:
  - Invariance Analysis
    * Model based invariance analysis
    * Alignment method analsyis
  - Item report analysis

* The item report analysis includes a larger set of analysis, besides Model based invariance analysis and Alignment method analsyis, such as item descriptives, missing data analysis, parallel analysis and reliability analysis among others.

5.2 Invariance analysis

5.2.1 Model based invariance analysis

Model based invariance analysis includes five model specifications:

- **pooled** is a graded response model with probit link, including survey design variables (i.e., stratification factors, primary sampling unit and student survey weights) (e.g., Stapleton, 2013). *Pooled* a single response model, fitted onto pooled sample of cases were sampling weights have been scale to a common constant so including countries contributes equally to estimates (Gonzalez, 2012).
- **strict** is a multigroup graded response model, where all response model parameters are held equal between compared groups (Tse et al., 2024), including the scale factors.
- **scalar** is a multigroup graded response model, where all response model parameters are held equal among groups, with the exemption of scale factors. It follows model specifications described in Svetina et al. (2020, proposition 7).
- **configural** is a multigroup graded response model, where thresholds are held common among the compared countries (i.e., threshold invariance). Is the baseline model for a model building sequence for assesing model based measurement invariance (Wu & Estabrook, 2016). It follows model specifications described in Svetina et al. (2020, proposition 4).
- **base** is a a multigroup graded response model, where all response model parameters are freely estimated among the compare groups.

Simulations studies from Rutkowski & Svetina (2017) with the graded response model with probit link and a larger amount of compare groups (e.g., 10, 20) suggest that RMSEA < .055 serves as a rule of thumb to select well fitting response models with invariant parameters among compare groups.

To fit the following models we use as inputs:

- data_responses `data_ex1.rds`
- scale_info = `guideline_scale_info_example.xlsx`
- scale_id = 1
- and is using the functions:
  - `rd3c3::fit_grm2` for the pooled model

- rd3c3::fit_grm2_m01_strict for the strict model
- rd3c3::fit_grm2_m02_scalar for the scalar model
- rd3c3::fit_grm2_m03_configural for the configural model
- rd3c3::fit_grm2_m04_base for the base model (i.e., descriptive model)

In the following section we include code (*folded*) to produce the invariance model fit indexes table (see Table 1).

```
#-------------------------------------------------------------------------------
# define objects
#-------------------------------------------------------------------------------

# -----------------------------------------------
# scale id
# -----------------------------------------------

rd3c3::silent(library(dplyr))

# -----------------------------------------------
# scale id
# -----------------------------------------------

scale_id <- 1

# -----------------------------------------------
# scales info
# -----------------------------------------------

scales_data <- readxl::read_xlsx(
               'guideline_scale_info_example.xlsx',
               sheet = 'scales_data'
               )

# -----------------------------------------------
# data file
# -----------------------------------------------

data_file <- scales_data %>%
dplyr::filter(scale_num == scale_id) %>%
dplyr::select(data_file) %>%
unique() %>%
dplyr::pull()
```

```
# ------------------------------------------------
# response matrix
# ------------------------------------------------

data_responses <- readRDS(data_file) %>%
mutate(grp = paste0(COUNTRY)) %>%
mutate(grp = as.numeric(as.factor(COUNTRY))) %>%
mutate(grp_name = paste0(COUNTRY))

#-------------------------------------------------------------------------------
# response models
#-------------------------------------------------------------------------------

# ------------------------------------------------
# most centered
# ------------------------------------------------

grp_centered <- 'CHL'

# ------------------------------------------------
# pooled
# ------------------------------------------------

inv_0 <- rd3c3::silent(
             rd3c3::fit_grm2(
           data = data_responses,
           scale_num  = scale_id,
             scale_info = scales_data
              )
             )

# ------------------------------------------------
# strict
# ------------------------------------------------

inv_1 <- rd3c3::silent(
           rd3c3::fit_grm2_m01_strict(
           data = data_responses,
           scale_num = scale_id,
           scale_info = scales_data,
           grp_var = 'id_k',
           grp_txt = 'grp_name',
```

```
           grp_ref = grp_centered
           )
           )

# -----------------------------------------------
# scalar
# -----------------------------------------------

inv_2 <- rd3c3::silent(
           rd3c3::fit_grm2_m02_scalar(
           data = data_responses,
           scale_num = scale_id,
           scale_info = scales_data,
           grp_var = 'id_k',
           grp_txt = 'grp_name',
           grp_ref = grp_centered
           )
           )

# -----------------------------------------------
# configural
# -----------------------------------------------

inv_3 <- rd3c3::silent(
           rd3c3::fit_grm2_m03_config(
           data = data_responses,
           scale_num = scale_id,
           scale_info = scales_data,
           grp_var = 'id_k',
           grp_txt = 'grp_name',
           grp_ref = grp_centered
           )
           )

# -----------------------------------------------
# base
# -----------------------------------------------

inv_4 <- rd3c3::silent(
           rd3c3::fit_grm2_m04_base(
           data = data_responses,
           scale_num = scale_id,
```

```
           scale_info = scales_data,
           grp_var = 'id_k',
           grp_txt = 'grp_name',
           grp_ref = grp_centered
           )
           )

# ----------------------------------------------
# retrieve fit indexes per model
# ----------------------------------------------

fit_0 <- rd3c3::get_inv_fit(inv_0, model_name = 'pooled')
fit_1 <- rd3c3::get_inv_fit(inv_1, model_name = 'strict')
fit_2 <- rd3c3::get_inv_fit(inv_2, model_name = 'scalar')
fit_3 <- rd3c3::get_inv_fit(inv_3, model_name = 'config')
fit_4 <- rd3c3::get_inv_fit(inv_4, model_name = 'base')

# ----------------------------------------------
# general table
# ----------------------------------------------

fit_table <- dplyr::bind_rows(
             dplyr::select(fit_0, model, RMSEA, CFI, TLI, SRMR, x2, df, p_val),
             dplyr::select(fit_1, model, RMSEA, CFI, TLI, SRMR, x2, df, p_val),
             dplyr::select(fit_2, model, RMSEA, CFI, TLI, SRMR, x2, df, p_val),
             dplyr::select(fit_3, model, RMSEA, CFI, TLI, SRMR, x2, df, p_val),
             dplyr::select(fit_4, model, RMSEA, CFI, TLI, SRMR, x2, df, p_val)
             )


# ----------------------------------------------
# model fit
# ----------------------------------------------

fit_table %>%
knitr::kable(.,
  digits = c(0,3,2,2,2,2,0,2),
    caption = 'Table 1: invariance model fit indexes between compared groups'
  )
```

Table 5.1: Table 1: invariance model fit indexes between compared groups

model	RMSEA	CFI	TLI	SRMR	x2	df	p_val
pooled	0.034	1.00	1.00	0.01	18.61	2	0
strict	0.042	0.99	1.00	0.02	112.97	15	0
scalar	0.027	1.00	1.00	0.01	40.66	11	0
config	0.032	1.00	1.00	0.01	37.18	8	0
base	0.044	1.00	0.99	0.01	31.78	4	0

Note: pooled = is the response model fitted onto the pooled sample. strict = is a multigroup response model with common thresholds, common loadings, and a common scale. This response model suffice mean score comparisons (Tse et al., 2024). scalar = is a multigroup esponse model with common thresholds, common loadings, and free scales for each item. This models supports latent mean comparisons (Tse et al., 2024). config = is a multigroup response model with common thresholds. base = is a multigroup descriptive model where all response model parameter are free to vary. Metric model specification is not identified under graded response models (Wu & Estabrook, 2016), thus metric solution is not included. A RMSEA of .055 or less has been found to be good threshold for fit for graded response models with many groups of 10 or 20 compared groups (see Rutkowski & Svetina, 2017).

5.2.2 Alignment

The aligment method is optimizing for the least discrepance response model parameters among the compared groups. Is fitting a graded response model with probit link, and using the most optimal group as a reference. We are using the statement `ALIGNMENT = FIXED(*);` within Mplus for these purposes. We rely on the *Measurement invariance explorer* (https://github.com/MaksimRudnev/MIE.package) to retrieve aligment results.

The following code (*folded*) is using as inputs:

- data_responses `data_ex1.rds`
- scale_info = `guideline_scale_info_example.xlsx`
- scale_id = 1
- and is using the function `rd3c3::fit_grm2_align_wlsmv()` to run an alignment method analysis

```
#------------------------------------------------------------------------------
# define objects
#------------------------------------------------------------------------------
```

```
# ----------------------------------------------
# scale id
# ----------------------------------------------

rd3c3::silent(library(dplyr))

# ----------------------------------------------
# scale id
# ----------------------------------------------

scale_id <- 1

# ----------------------------------------------
# scales info
# ----------------------------------------------

scales_data <- readxl::read_xlsx(
               'guideline_scale_info_example.xlsx',
               sheet = 'scales_data'
               )

# ----------------------------------------------
# data file
# ----------------------------------------------

data_file <- scales_data %>%
dplyr::filter(scale_num == scale_id) %>%
dplyr::select(data_file) %>%
unique() %>%
dplyr::pull()

# ----------------------------------------------
# response matrix
# ----------------------------------------------

data_responses <- readRDS(data_file) %>%
mutate(grp = paste0(COUNTRY)) %>%
mutate(grp = as.numeric(as.factor(COUNTRY))) %>%
mutate(grp_name = paste0(COUNTRY))

#-----------------------------------------------------------------------------
# alignment
```

```r
#-------------------------------------------------------------------------------

# -----------------------------------------------
# aligned
# -----------------------------------------------


fitted_align <- rd3c3::silent(
rd3c3::fit_grm2_align_wlsmv(
data = data_responses,
scale_num = scale_id,
scale_info = scales_data)
)

# -----------------------------------------------
# retrieve output
# -----------------------------------------------

scale_file <- scales_data %>%
dplyr::filter(scale_num == scale_id) %>%
dplyr::select(mplus_file) %>%
unique() %>%
dplyr::pull()


alignment_out <- MIE::extractAlignment(paste0(scale_file,'_align.out'), silent = TRUE)

# -----------------------------------------------
# display
# -----------------------------------------------

alignment_table <- alignment_out$summary %>%
                   tibble::rownames_to_column("terms") %>%
                   tibble::as_tibble() %>%
                   rename(
                   term        = 1,
                   a_par       = 2,
                   R2          = 3,
                   n_inv       = 4,
                   n_dis       = 5,
                   inv_grp     = 6,
                   dis_grp     = 7
```

```
                    ) %>%
                    mutate(type = case_when(
                    stringr::str_detect(term, 'Threshold') ~ 'tau',
                    stringr::str_detect(term, 'Loadings')  ~ 'lambda'
                    )) %>%
                    mutate(term = stringr::str_replace(term, '\\$', '_')) %>%
                    mutate(term = stringr::str_replace(term, 'Threshold', '')) %>%
                    mutate(term = stringr::str_replace(term, 'Loadings', '')) %>%
                    mutate(term = stringr::str_replace(term, 'ETA by ', '')) %>%
                    dplyr::select(
                    type, term, a_par, R2, n_inv, n_dis, inv_grp, dis_grp)

# ------------------------------------------------
# display
# ------------------------------------------------

alignment_table %>%
knitr::kable(.,
  digits = 2,
    caption = 'Table 2: alignment comparisons'
)
```

Table 5.2: Table 2: alignment comparisons

type	term	a_par	R2	n_inv	n_dis	inv_grp	dis_grp
tau	I01_1	NA	NA	0	2		18 11
tau	I01_2	NA	NA	0	2		18 11
tau	I01_3	0.73	1.00	2	0	11 18	
tau	I02_1	-1.89	0.94	2	0	11 18	
tau	I02_2	-1.09	0.00	2	0	11 18	
tau	I02_3	NA	NA	0	2		18 11
tau	I03_1	-1.73	0.83	2	0	11 18	
tau	I03_2	-1.00	0.91	2	0	11 18	
tau	I03_3	0.20	0.85	2	0	11 18	
tau	I05_1	-1.64	0.88	2	0	11 18	
tau	I05_2	-0.97	0.01	2	0	11 18	
tau	I05_3	0.23	0.93	2	0	11 18	
lambda	I01	0.72	0.00	2	0	11 18	
lambda	I02	0.77	0.20	2	0	11 18	
lambda	I03	0.84	0.26	2	0	11 18	
lambda	I05	0.80	0.00	2	0	11 18	

5.3 Item report analysis

In the following section we include a **template** example, to produce **item analysis reports**. This template includes:

- **Scale description**
  - a presentation of the name of the collection of items (i.e., the scale name)
  - a presentation of items as these where presented to the participants
  - a table with the item text, with the public data file names, and the shortened variable names
- **Analysis of responses**
  - descriptives
  - missing data descriptives
- **Response model**
  - dimensionality analysis via parallel analysis for ordinal indicators (Lubbe, 2019)
  - response model parameters for a graded response model
  - reliability analysis
  - item person maps
- **Item analysis**
  - item test correlation
  - item fit based on partial credit model
- **Comparability**
  - model based measurement invariance
  - alignment analysis of GRM among groups

The following figure depicts an overview of the generated report.

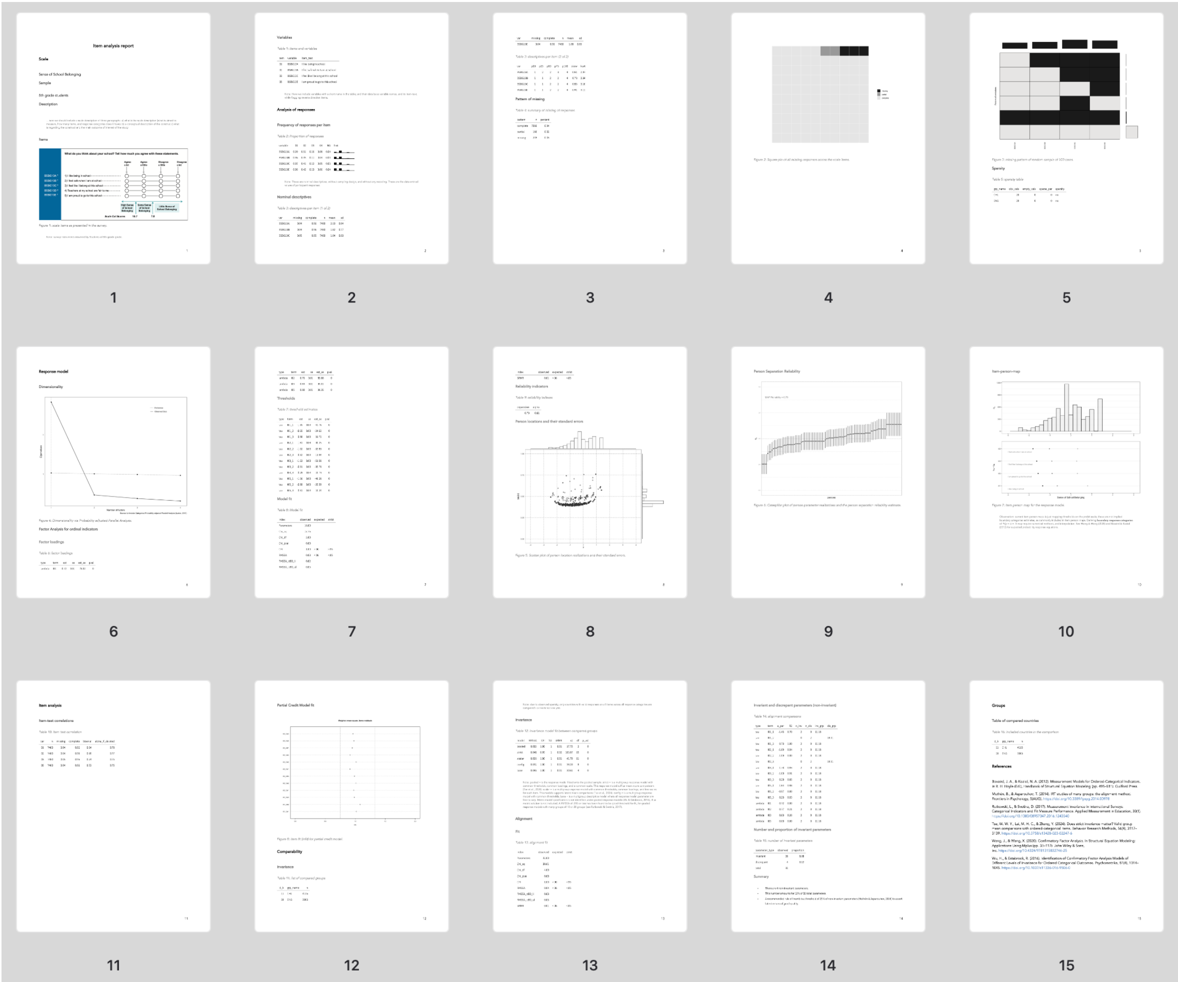

Figure 5.1: Figure 5: overview of a dynamic item report

To produce this examplary report the user needs as inputs:

- data_responses data_ex1.rds
- scale_info = guideline_scale_info_example.xlsx
- scale_id = 1
- the template guideline_item_report_example.rmd
  - and the word template report_template.docx

The end product of this procedure can be inspected in the following file guideline_item_report_example.docx

5.4 Additional examples

The current examples are just toy examples, to illustrate the basic capabilities of the library. We include two other with more realistic examples. One example with three countries, were is possible to see that three countries do not obtained strict invariance (example with "Sense of School Belonging" for three countries). And a third example of a template based workflow, where we procede with a full fledge item analysis report including all participating countries (example with "Bullying Scale" for all participating countries).

- Example with "Sense of School Belonging" for three countries
  - data: `data_example.rds`
  - code template: `template_example_2.rmd`
    - word template: `report_template.docx`
  - resulting report: `template_example_2.docx`
- Example with "Bullying Scale" for all participating countries
  - data`survey_1_g8.rds`
  - code template: `template_example_3.rmd`
    - word template: `report_template.docx`
  - resulting report `template_example_3.docx`.

All example files can be downloaded from the following link.

5.5 References

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. IERI Monograph Series Issues and Methodologies in Large-Scale Assessments, 5, 115–134.

Tse, W. W. Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. Behavior Research Methods, 56(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

Stapleton, L. M. (2013). Incorporating Sampling Weights into Single- and Multilevel Analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of International Large scale Assessment: background, technical issues, and methods of data analysis (pp. 363–388). Chapman and Hall/CRC.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the lavaan/semTools Packages. Structural Equation Modeling: A Multidisciplinary Journal, 27(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

Rutkowski, L., & Svetina, D. (2017). Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance. Applied Measurement in Education, 30(1). https://doi.org/10.1080/08957347.2016.1243540

6 Intended use

In this final section of the guidelines, we outlined the intended use of the library. We emphasize the responsibilities of the researcher in employing this tool. The primary goal of the library is to streamline and accelerate the process of generating results for multiple groups comparisons on item scales. By automating repetitive analytical tasks, the library facilitates the production of key outputs necessary for assessing the fit or misfit of measurement invariance models across various groups.

However, it is crucial to underscore that the library is not a substitute for sound judgment and expertise of the researcher and user. While it efficiently produces the primary results needed for evaluation, it does not draw definitive conclusions regarding whether models meet specific thresholds or satisfy measurement invariance criteria. This critical interpretive step remains the responsibility of the researcher and the library user, who must apply their expertise to analyze and contextualize the findings appropriately.

The library is specifically designed to assist users in generating foundational results for evaluating measurement invariance. It is not intended for purposes outside this scope. It should not be used to automate decision-making regarding the acceptability or applicability of response models. Users must approach the outputs with caution, ensuring that the analyses are tailored to the specific goals of their research and the nuances of their studies.

In conclusion, this library is a powerful tool to enhance efficiency in measurement invariance analyses, but it is not a substitute for thorough methodological understanding and critical interpretation. Researchers are encouraged to use this resource judiciously and within its intended purpose, recognizing its limitations and their own role in ensuring the validity and reliability of their conclusions.