The resulting MA-PMI distribution is a mixture of posterior predictive distributions. Although such mixtures are complex (e.g., can be multimodal), their first two moments can easily be described. The point estimate of our parameter of interest $p$ can be readily derived from the mean of the MA-PMI distribution. For any finite distribution it is:

$$p = \sum_{k=1}^{K} w_k p_k \tag{C.5}$$

where $p_k$ is the parameter of interest for model $k$ and $w$ its weight. The variance of the BMA distribution is given as follows (Frühwirth-Schnatter, 2006, p.11):

$$\mathrm{Var}(p) = \sum_{k=1}^{K} w_k \sigma_{pk}^2 + \sum_{k=1}^{K} w_k p_k^2 - \left( \sum_{k=1}^{K} w_k p_k \right)^2 \tag{C.6}$$

where $\sigma_{pk}^2$ is the variance of the parameter in model $k$.

### C.3.2 Illustration of the approach

In order to illustrate the proposed approach, consider estimating the latent mean difference $\delta$ between two groups. The latent variables are assessed by six items that conform to the Rasch model. For simplicity, assume that the first three items and the last three items function homogeneously within their group, but not between the item groups. Thus, there are two sets of possible anchor items and we want to consider both. Applying MA-PMI to two models, the point estimate of $\delta$ is a weighted average of the $\delta_i$s of the two models (Eq. C.5). As to the variance of the distribution resulting from MA-PMI, in the case of two models Eq. C.6 simplifies to (see Proof 1 in the Appendix):

$$\mathrm{Var}(\delta) = w_1 \sigma_{\delta_1}^2 + w_2 \sigma_{\delta_2}^2 + w_1 w_2 (\delta_1 - \delta_2)^2. \tag{C.7}$$

These equations can be used to illustrate in which way the different terms impact point and interval estimation in the MA-PMI.

*Impact of differences in the parameter estimate across models*. Assume a situation, in which the variance of the posterior predictive distribution of the mean difference parameter $\sigma_{\delta_i}^2$ is unity in both models, but their estimated mean differences $\delta$ are zero and one, respectively. If the two models are weighted equally, $\mathrm{E}(\delta) = 0.5$ (see Eq. C.5) and $\mathrm{Var}(\delta) = 1.25$ result (see Eq. C.6). This increase in variance (from 1 in each single solution to 1.25 in the averaged solution) is a direct result of model selection uncertainty - a source of variance in the parameter which is neglected when only a single anchor item set is chosen. Naturally, the MA-PMI variance increases monotonically in the form of a quadratic function with increasing parameter differences in the models (see Figure C.2).

*Impact of parameter variances*. While the mean of the MA-PMI distribution is not affected by the parameter's variances nor differences of those variances across different models, $\mathrm{Var}(p)$ depends on them, as shown in Eq. C.6. Imagine a setup like the initial example from the last paragraph. Given everything is kept constant but $\sigma_{\delta_2}^2$, the lowest possible value of $\mathrm{Var}(\delta)$ is 0.75 for $\sigma_{\delta_2}^2$ approaching zero. When $\sigma_{\delta_2}^2$ increases, $\mathrm{Var}(\delta)$ increases linearly.

*Impact of weighting*. Weights impact both the point estimate as well as the variance of this estimate in MA-PMI. To illustrate the impact of weights, we use the example from above with a

difference of $\delta_2 - \delta_1 = 1$ and $\sigma^2_{\delta_1} = \sigma^2_{\delta_2} = 1$, but vary the weights. As can be seen in Eq. 4, the mean of the BMI distribution moves linearly by weight toward the higher weighted model parameter value. As can be seen in Fig. C.2, given both models have the same variance of the parameter, the parameter's variance is highest for equal weights. It decreases as either model is weighted higher. If one of the models receives a weight of 1, which corresponds to choosing one of the anchor items sets, the variance of the respective model (here 1) results. The greater variance of $\delta$ for equal weights reflects the fact that this expresses a state of maximum uncertainty. On the contrary, more unequal weights express increased certainty.

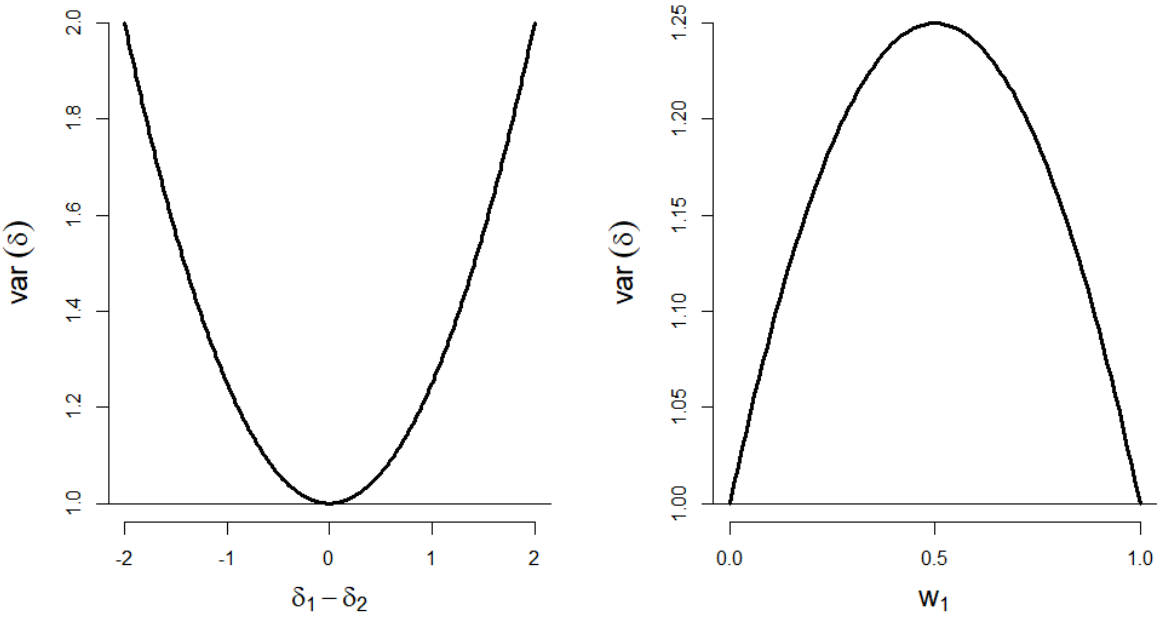


Figure C.2: Dependence of the variance of a parameter in the MA-PMI distribution on the difference of parameter estimates between models (left) and on the chosen weights (right). This is shown for a two-model case, in which both models provide distributions with a variance of 1.

## C.4 Empirical Example

In the following, we will illustrate our approach with data of the assessment of obsessive-compulsive disorder (OCD). We used data on $N = 1036$ outpatients at the start of therapy, who received treatment at a university outpatient clinic specialised in OCD (for more info see Schulze et al., 2018). In the course of initial diagnostics, the German Yale-Brown Obsessive Compulsive Inventory (Y-BOCS, Goodman, Price, Rasmussen, Mazure, Delgado, et al., 1989; Goodman, Price, Rasmussen, Mazure, Fleischmann, et al., 1989; Hand & Büttner-Westphal, 1991) was applied. The Y-BOCS interview is a gold standard in the diagnostic process of OCD (Fatori et al., 2020) and consists of a symptom catalogue among others. The items are dichotomously scored with a score of 1 indicating the presence of the symptom and a score of 0 indicating its absence. Schulze et al. (2018) showed that OCD symptoms can be reduced to a total of 10 distinct dimensions. Here, we focus on the Cleanliness subscale, which consists of 12 items subsuming contaminant obsessessions and

cleaning compulsions (see Table C.1).

Table C.1: Items of the Y-BOCS Cleanliness scale and item cluster analysis results with gender as MI covariate.

| # | Item | Cluster | rel. DIF$_\lambda$ | rel. DIF$_\tau$ | $\hat{\delta}$ (SE) |
|---|---|---|---|---|---|
| 1 | Concerns with bodily waste or secretions | A | -0.18 | 0.46 | 0.10 (0.08) |
| 6 | Bothered by sticky substances or residues | | -0.05 | 0.42 | |
| 9 | Excessive or ritualized hand washing | | -0.12 | 0.30 | |
| 10 | Excessive or ritualized showering, bathing, tooth brushing, or grooming | | -0.19 | 0.16 | |
| 12 | Other measures to prevent or remove contact with contaminants | | -0.19 | 0.10 | |
| 2 | Concerns with germs or dirt | B | -0.02 | -0.46 | 0.28 (0.08) |
| 7 | Concerned will get ill because of contaminant | | -0.08 | -0.14 | |
| 8 | Concerned will get others ill by spreading contaminant | | -0.08 | -0.10 | |
| 11 | Involves cleaning of household items and inanimate objects | | -0.05 | -0.16 | |
| 3 | Excessive concerns with environmental contaminants | C | 0.25 | 0.22 | 0.24 (0.10) |
| 4 | Excessive concern with household items | | 0.09 | -0.18 | |
| 5 | Excessive concern with animals | D | 0.62 | -0.61 | 0.43 (0.11) |

*Note.* Partial MI analysis results in four clusters found by the item cluster approach. rel. DIF = relative DIF effect as defined in Schulze and Pohl (2020) as the foundation of the item cluster approach. Effects were centered on zero. $\hat{\delta}$ = estimated standardized mean difference with female as focus group (standard error).

For the illustration of our approach, we compared Cleanliness symptoms between the two groups of self-identified females and males ($n_f = 573, n_m = 463$). The applied research question at hand is thus to detect a possible difference in Cleanliness symptoms between female and male patients.

### C.4.1 MI Analysis

In the first step, we applied a model as given in Eq. C.3 to scale the responses to the items of the Y-BOCS. We tested for violations of MI in both item difficulties $\tau$ and item loadings $\lambda$. We used the weighted least squares estimator as implemented in the package lavaan (Rosseel, 2012) in R (R Core Team, 2020). Syntax can be found in the Appendix. A global test of MI of the Cleanliness subscale compared the configural model with a strong MI model in which loadings and thresholds were constrained to equality. It yielded a significant difference indicating substantial violation of MI ($\chi^2(10) = 25.19, p = 0.005$). This finding was supported by large ranges in DIF effects in both loadings and difficulties (DIF range $\log(\lambda)$ = .81, DIF range $\tau$ = 1.07, see Table C.1).

In order to identify sets of items that are functioning homogeneously in both item parameters within each set, we applied the item cluster approach for two groups (Pohl & Schulze, 2020). This approach has to main steps: 1) Calculate DIF effects for each item parameter that are relative to an arbitrarily chosen item. Such relative DIFs have the advantage to be invariant to changing model constraints, i.e. to circumvent the scale indeterminacy issue. On the downside, such DIF effects cannot be interpreted on an absolute scale, i.e. a value of zero does not convey zero

true DIF. 2) These relative DIF effects provide the source for clustering the items: Items with approximately equal relative DIF are closely measurement invariant to each other. The cluster process is steered by setting maximum ranges of relative DIF values within each item cluster (corresponds to the small amount of residual DIF still allowed). We set these ranges to be 0.2 for item loadings and of 0.5 for item intercepts, which can be considered moderate (for rating of DIF size see González-Betanzos & Abad, 2012).

The item cluster approach yielded four distinct item clusters in this example (see Table C.1), which upon closer inspection depicted different contents of the overarching Cleanliness symptom domain. Cluster A involved obsessions and compulsions directed at touch and having contaminants on the body's surface while cluster B subsumed mostly items concerning general illness. Cluster C consisted of two items on hazards in the environment, and cluster D covered a single item on animals.

### C.4.2 Bayesian Analysis and Model Averaging

Using Bayesian estimation, we implemented four partial MI models in STAN (Carpenter et al., 2017). In each of the models one of the item clusters was used for anchoring.[19] This resulted in four group mean differences $\delta$ in Cleanliness symptoms between female and male patients with all four models indicating a higher symptom prevalence for females. Still, the models differed substantially in the magnitude of standardized mean differences (see Table C.1). The standard errors of the mean difference ranged from 0.08 to 0.11 across the models, yielding a significant $\delta$ for all clusters but the model of cluster A. Figure C.3A depicts the posterior distributions of the four partial MI models.

When naively assuming MI for all items (full MI, Figure C.3B), a significant standardized mean difference in Cleanliness scores of $\delta = 0.19$ (SE=0.07) in favor of women was estimated.

Figure C.3C depicts the MA-PMI distribution with equal and thus uninformative weights $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Here, an estimated standardized mean difference of $\delta = 0.26$ (SE=0.15) resulted. Note that the mean difference was close to the one when assuming MI for all items[20], but the standard error was substantially larger, leaving the credibility interval to include zero.

We also incorporated substantive and expert knowledge about the plausibility of anchor item sets by using informed weights (Figure C.3D). When looking at the item content differences between the item clusters, we found that cluster A and B resembled the general concept of contamination obsessions and cleaning compulsions more closely than cluster C and D which contained rather specific niches of this symptom domain. On the one hand this was due to higher item counts in cluster A and B, but these clusters furthermore reflected generalized core aspects of touching, cleaning, and illness. The results of the item clusters approach thus unraveled a factor structure within the Cleanliness factor. Such unaccounted latent structures are likely sources of MI violations. We chose the weights 0.45, 0.45, 0.10, 0.00 and thus decided to put emphasis on cluster A and

[19] We ran multi-group 2PL models in STAN for 100000 iterations (warm-up = 50000). There were no signs of non-convergence (maxRhat = 1.001, medianRhat = 1.00001, minEffN = 836, medianEffN = 72052). Both the STAN and R syntax can be found in the Appendix.

[20] The small difference occurs due to the fact that when assuming full MI all items equally contribute to the estimate, while in the cluster approach each cluster equally contributes to the aggregated estimate; the number of items, however, differs between clusters.

B without differentiating in their importance. Cluster C adds a little to the content domain of Cleanliness symptoms but was not discarded completely as anchoring candidate. Cluster D, which contained only a single item on a niche phenomenon (animals), was excluded as anchor candidate by setting its weight to zero. The resulting MA-PMI distribution was shifted towards the mean difference found by cluster A and B. Thus, a somewhat smaller effect of $\delta = 0.19$ compared to equal weighting resulted. It was statistically insignificant.

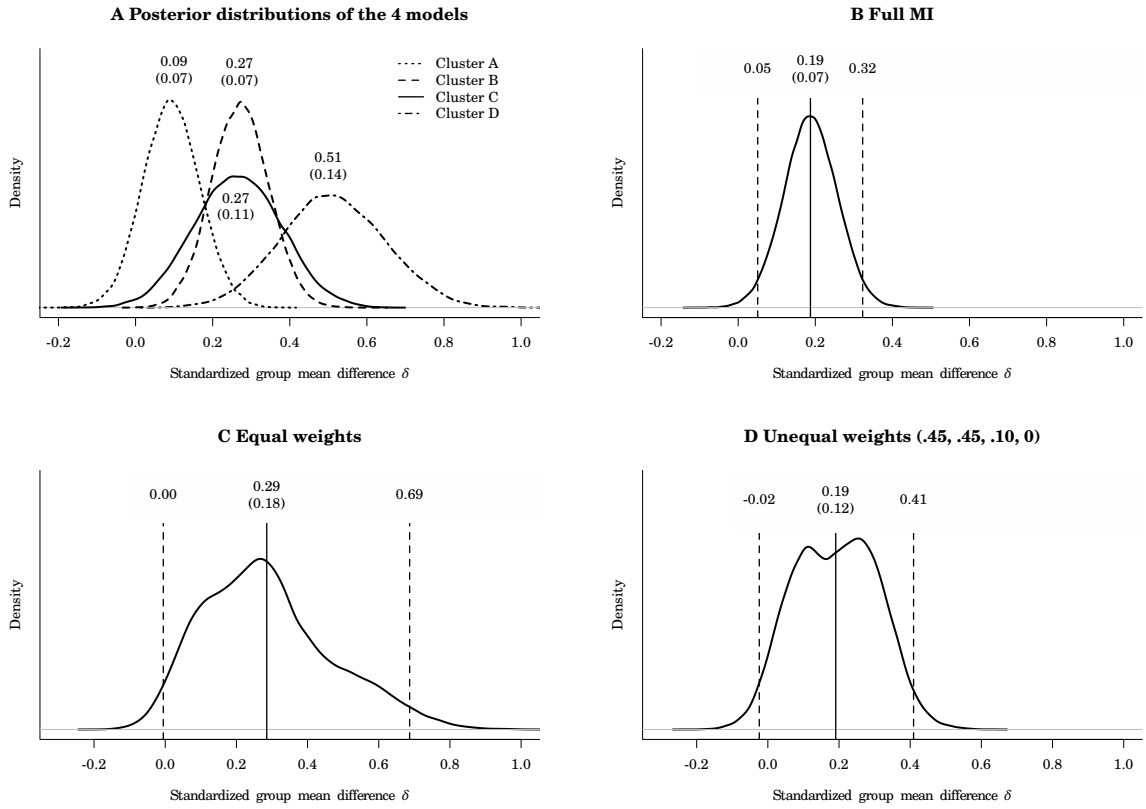


Figure C.3: Distributions of the standardized mean difference for female and male patients on the Cleanliness scale. C and D are MA-PMI distributions. Values in parentheses are parameter variances in terms of standard deviations. Dashed lines indicate the bounds of 95% credibility intervals.

## C.5 Discussion

Here we introduced a model averaging approach for estimating latent differences while incorporating uncertainty in anchor item choice. As common partial MI analysis techniques demand making a choice for a single anchor item set, these approaches fall prone to making unstated, unchecked, or unfounded assumption on the nature of partial MI. We rely on the idea that different assumptions may be plausible and that the variety of choices of plausible assumptions should be depicted in the results. From this perspective, unavailable strong expert knowledge results in all items or item clusters being plausible anchors to some extent. Our approach provides a tool to make flexible choices about the degree of belief in different sets of anchor items. By making different assumptions and their impact on analysis results transparent fosters much more informed results that may be used for replication studies or for further studies on the topic.

There exists an extensive theoretical literature on model combination techniques, primarily discussing the automated delineation of model weights from model fit in the sense of machine learning approaches. This strand of research has produced various elaborate mathematical approaches, like BMA (Madigan et al., 1996), model stacking (Wolpert, 1992; Yao et al., 2018), bagging (Breiman, 1996), and Bayesian model combination (Monteith et al., 2011). These approaches use model fit or prediction accuracy as determinants of model weight and are not suited for partial MI modeling: Due to scale indeterminacy, model fit does not display bias of latent group differences resulting from anchor item choice. As such, it is not a suitable measure for plausibility of assumptions. Instead, we propose using substantive considerations about prior beliefs as weights. This notion refers to core aspects of Bayesian statistics in the sense of probability as a mathematical expression of a person's degree of belief.

We have illustrated the properties of the proposed approach - which are straightforward in the sense that they are grounded in basic mixture distribution calculation. The averaged parameter mean depends on the parameters estimated in the different models and the weights assigned to the models. Parameter variance increases with a) an increase in variance of the parameter estimates in the different models, b) an increase in differences in parameter estimates across the candidate models and c) decreasing information (i.e., equality of weights) on the plausibility of the candidate models.

In our applied example, we found four inherently measurement invariant item clusters which reflected multidimensionality of the scale. These four item sets provided viable candidates for anchors. With respect to the substantive research question at hand, the the latent difference in Cleanliness symptoms between females and males, different anchor choices led to varying effect sizes and, more importantly, to different decisions on the statistical significance of this effect. When choosing one of the clusters B, C, or D for anchoring, a significant effect resulted. The same conclusion would have been drawn when neglecting the violation of MI all together. The latent difference was insignificant when anchoring with cluster A (which is the largest one and would thus pose a natural candidate aside from content-wise considerations) or when applying model averaging of any kind. As the latter takes the uncertainty due to anchor item choice into account, it seems to be a sound approach on the conservative side of statistical testing.

MA-PMI uses a priori weights stemming from expert knowledge. If one has no grounds to decide between two or more given anchor item sets, their weights will be equal in order to express this maximum uncertainty. Setting a weight to zero on the other hand expresses maximum certainty about excluding it as an anchor candidate. We illustrated such reasoning in the applied example via two different weighting schemes. While these two weight sets were for the purpose of illustration, an applied researcher should only use a single set of weights for which they can state substantive arguments. These arguments are then open for scientific debate. In order to avoid the impression of p-hacking and likewise unscientific endeavors, the rationale for finding weights and the extent of expert knowledge could be laid down as part of a study's preregistration.

Although we made use of models for dichotomous items in our illustrations and the example, the described principles of MA-PMI can be applied to any other latent variable model type as well. In this respect, we supply code for models with dichotomous as well as continuous items in the supplements. Further models of interest are the case with more than two groups as well as

continuous MI covariates ((like age, Schulze & Pohl, 2020). Furthermore, MA-PMI is a general approach and can be applied to any parameter of interest and is not restricted to the latent mean difference.

We illustrated the MA-PMI approach using the anchor item sets identified by the item cluster approach. It is of course also possible to identify candidates for anchor item sets by other ways (e.g., substantive considerations or other MI approaches) and apply MA-PMI to these different options. For example, researcher may be aware of the different underlying sub-dimensions and may deliberately choose some or all of these as possible anchor item sets. Furthermore, MA-PMI could also be applied without using anchor *sets* but single items instead. Single-item anchors come at the cost of increased standard errors for the latent parameter under scrutiny (e.g., $\delta$) when compared to longer anchors.

Similarly to setting prior distributions for parameters in Bayesian analysis, future research may address how to systematically translate expert knowledge and knowledge from previous studies into weights for different anchor item sets (i.e. models). This may require applications that serve as blueprints but also reviews of the amount of measurement invariance, typical differences in estimated parameters as well as their standard errors in different partial MI models. These may serve for simulation studies investigating the effect on approaches used and weights chosen and can result in more detailed guidelines for how to deal with violations of MI in applications.

Future applications of this approach will show in which way making assumptions transparent and depicting uncertainty helps understanding the issue at hand. Furthermore, such applications could also help in performing meta-analyses on reasons for measurement invariance. This could in turn help test developers to construct tests for which measurement invariance either does not occur or is deliberately introduced (e.g., in multidimensional scales). We also believe that the transparency inherent in our approach will facilitate designing replication studies and to estimate the effect of changing design factors on replicability of the results (Steiner et al., 2019).

## References


Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Bernardo, J., & Smith, A. (1994). *Bayesian theory*. John Wiley; Sons: New York.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement*, *43*(4), 303–321. https://doi.org/10.1177/0146621618795727

Fatori, D., Costa, D. L., Asbahr, F. R., Ferrão, Y. A., Rosário, M. C., Miguel, E. C., Shavitt, R. G., & Batistuzzo, M. C. (2020). Is it time to change the gold standard of obsessive-compulsive disorder severity assessment? Factor structure of the Yale-Brown Obsessive-Compulsive Scale. *Australian & New Zealand Journal of Psychiatry*, *54*(7), 732–742. https://doi.org/10.1177/0004867420924113

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology*, *8*(4), 134–145. https://doi.org/10.1027/1614-2241/a000046

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown obsessive compulsive scale: II. Validity. *Archives of General Psychiatry*, *46*(11), 1012–1016. https://doi.org/10.1001/archpsyc.1989.01810110054008

Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown Obsessive Compulsive Scale: I. Development, use, and reliability. *Archives of General Psychiatry*, *46*(11), 1006–1011. https://doi.org/10.1001/archpsyc.1989.01810110048007

Hand, I., & Büttner-Westphal, H. (1991). Die Yale-Brown Obsessive Compulsive Scale (Y-BOCS): Ein halbstrukturiertes Interview zur Beurteilung des Schweregrades von Denk-und Handlungszwängen. *Verhaltenstherapie*, *1*(3), 223–225.

Hidalgo, M. D., & Gómez-Benito, J. (2010). Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd, pp. 36–44). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00242-6

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Ithaya Rani, P., & Muneeswaran, K. (2016). Facial emotion recognition based on eye and mouth regions. *International Journal of Pattern Recognition and Artificial Intelligence*, *30*(07), 1655020. https://doi.org/10.1142/S021800141655020X

Jin, S., & Ankargren, S. (2019). Frequentist model averaging in structural equation modelling. *Psychometrika*, *84*(1), 84–104. https://doi.org/10.1007/s11336-018-9624-y

Kaplan, D., & Lee, C. (2016). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 343–353. https://doi.org/10.1080/10705511.2015.1092088

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103. https://doi.org/10.1177/0146621614544195

Madigan, D., Raftery, A. E., Volinsky, C., & Hoeting, J. (1996). Bayesian model averaging. *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, 77–83.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

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

Monteith, K., Carroll, J. L., Seppi, K., & Martinez, T. (2011). Turning Bayesian model averaging into Bayesian model combination. *The 2011 International Joint Conference on Neural Networks*, 2657–2663.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Pohl, S., & Schulze, D. (2020). Assessing group comparisons or change over time under measurement non-invariance: The cluster approach for nonuniform DIF. *Psychological Test*

*Assessment and Modelling*, *2*(62), 281–303. https://www.psychologie-aktuell.com/fileadmin/%20Redaktion/Journale/ptam-2020-2/04%5C_Pohl.pdf

Pohl, S., Schulze, D., & Stets, E. (in press). Partial measurement invariance: Extending and evaluating the cluster approach for identifying anchor items. *Applied Psychological Assessment*.

Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2016). Testing students with special educational needs in large-scale assessments–Psychometric properties of test scores and associations with test taking behavior. *Frontiers in Psychology*, *7*, 154. https://doi.org/10.3389/fpsyg.2016.00154

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724–744.

Pokropek, A., Lüdtke, O., & Robitzsch, A. (2020). An extension of the invariance alignment method for scale linking. *Psychological Test and Assessment Modelling*, *62*, 303–334. https://www.psychologie-aktuell.com/fileadmin/Redaktion/%20Journale/ptam-2020-2/05%5C_Pokropek.pdf

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rasch. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Institute of Educational Research.

Rights, J. D., Sterba, S. K., Cho, S.-J., & Preacher, K. J. (2018). Addressing model uncertainty in item response theory person scores through model averaging. *Behaviormetrika*, *45*(2), 495–503. https://doi.org/10.1007/s41237-018-0052-1

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Schnittjer, I., Gerken, A.-L., & Petersen, L. A. (2020). *NEPS technical report for mathematics–Scaling results of starting cohort 2 in fourth grade*. (NEPS Survey Paper No. 69). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Schultze, M., & Eid, M. (2018). Identifying measurement invariant item sets in cross-cultural settings using an automated item selection procedure. *Methodology*, *18*, 177–188.

Schulze, D., Kathmann, N., & Reuter, B. (2018). Getting it just right: A reevaluation of OCD symptom dimensions integrating traditional and Bayesian approaches. *Journal of Anxiety Disorders*, *56*, 63–73. https://doi.org/10.1016/j.janxdis.2018.04.003

Schulze, D., & Pohl, S. (2020). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 219–228. https://doi.org/10.1080/10705511.2020.1771186

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, *227*, 280–292. https://doi.org/10.1027/2151-2604/a000385

Strobl, C., Kopf, J., Kohler, L., von Oertzen, T., & Zeileis, A. (in press). Anchor point selection: An approach for anchoring without anchor items. *Applied Psychological Measurement*. https://www2.uibk.ac.at/downloads/c4041030/wpaper/2018-03.pdf

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

Verhagen, A., & Fox, J. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401. https://doi.org/10.1111/j.2044-8317.2012.02059.x

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

Yeung, K. Y., Bumgarner, R. E., & Raftery, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, *21*(10), 2394–2402. https://doi.org/10.1093/bioinformatics/bti319

## C.6 Appendix

### C.6.1 Proof

We show that the variance of the by MA-PMI averaged parameter of interest $p = \delta$ over two models is

$$\begin{aligned}\text{Var}(\delta) &\sum_{k=1}^{K} w_k \sigma_{\delta k}^2 + \sum_{k=1}^{K} w_k \delta_k^2 - \left(\sum_{k=1}^{K} w_k \delta_k\right)^2 \\ &= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1 w_2(\delta_1 - \delta_2)^2\end{aligned}$$

Proof:

$$\begin{aligned}\text{Var}(\delta) &= w_1\sigma_1^2 + w_2\sigma_2^2 + w_1\delta_1^2 + w_2\delta_2^2 - (w_1\delta_1 + w_2\delta_2)^2 \\ &= w_1\sigma_1^2 + w_2\sigma_2^2 + x \\ x &= w_1\delta_1^2 + w_2\delta_2^2 - \left(w_1^2\delta_1^2 + w_2^2\delta_2^2 - 2w_1\delta_1 w_2\delta_2\right)\end{aligned}$$

As due to the definition $\sum_{k=1}^{K} w_k = 1$ for $K = 2 \Rightarrow w_2 = 1 - w_1$

$$\begin{aligned}x &= w_1\delta_1^2 + (1-w_1)\delta_2^2 - \left[w_1^2\delta_1^2 + (1-w_1)^2\delta_2^2 + (1-w_1)(2w_1\delta_1\delta_2)\right] \\ &= w_1\delta_1^2 + \delta_2^2 - w_1\delta_2^2 - w_1^2\delta_1^2 - \left(1 + w_1^2 - 2w_1\right)\delta_2^2 - 2w_1\delta_1\delta_2 + 2w_1^2\delta_1\delta_2 \\ &= w_1\left(\delta_1^2 - \delta_2^2 + 2\delta_2^2 - 2\delta_1\delta_2\right) - w_1^2\left(\delta_1^2 + \delta_2^2 - 2\delta_1\delta_2\right) \\ &= w_1(\delta_1 - \delta_2)^2 - w_1^2(\delta_1 - \delta_2)^2 \\ &= (1-w_1)w_1(\delta_1 - \delta_2)^2 \\ &= w_1 w_2(\delta_1 - \delta_2)^2\end{aligned}$$

### C.6.2 Bayesian Model Averaging for Measurement Invariance Models - cont. items

```
# Bayesian partial measurement invariance models for two groups and classic CFA
    models
# and Bayesian model averaging for such models
library(blavaan)

model <- "Factor␣=~␣item1␣+␣item2␣+␣item3␣+␣item4␣+␣item5" # set up measurement
    model in lavaan syntax
items <- c("item1", "item2", "item3", "item4", "item5")
clustering <- c(1, 1, 1, 2, 2) # cluster coding
weights <- c(0.5, 0.5)
iter <- 5000
muAver <- NULL #averaged mean difference

for (currCluster in unique(clustering)) {
    currNonAnchor <- which(clustering != currCluster)
    partialItems <- append(partialItems,
                           paste0(lv, "=~", items[currNonAnchor]),
                           paste0(items[currNonAnchor], "~␣1"))

    res_blavaan <- bcfa(model,
                        data = Data,
                        group = "group",
                        std.lv = TRUE,
                        group.equal = c("loadings", "intercepts"),
                        group.partial = partialItems,
                        burnin = iter/2,
                        sample = iter/2,
                        n.chains = 2,
                        bcontrol = list(cores = 2))

    fit <- res_blavaan @external$mcmcout

    # Bayesian model averaging by weighting

    post <- extract(fit)
    nSims <- weights[currCluster]*(iter/2)
    draws <- sample((iter/2):iter, nSims)

    muAver <- append(muAver,
                     rnorm(nSims,
                           post$Alpha_free[draws],
                           post$Psi_var[draws]/sqrt(nrow(Data))))
}
```

### C.6.3 Bayesian Model Averaging for Measurement Invariance Models - binary items

```
# Bayesian partial measurement invariance models for two groups and 2PL models
# and Bayesian model averaging for such models
library(reshape2)
library(rstan)
options(mc.cores = 2)
rstan_options(auto_write = TRUE)

# prepare data for STAN:
# data frame should contain only the items in a dichotomous 0/1 coding
# Transformation as model syntax expects a long data format with subject and item
    indicators. Group indicator is seperate.
Data$sbjct <- 1:nrow(Data)
DataInput <- melt(Data, value.name="y", id.vars="sbjct")
DatInput$item <- rep(1:ncol(Data), each=nrow(Data))

group <- as.numeric(group) # group indicator has to be numeric with 1 and 2
table(group)

# Estimation:
# setting up input for STAN
Dat2plMG <- list(N =nrow(Data),
                 K =ncol(Data),
                 Ntot=nrow(Data)*ncol(Data),
                 jj=DatInput$item,
                 ii=DatInput$pbn,
```

```
                y =DatInput$y,
                g =group,
                gg=rep(group, ncol(Data)),
                G =2,
                NconItem = 2,       # 2 anchor items
                conItem  = c(1, 6), # ... which are item 1 and 6
                freeItem = 2:5)     # ... while all other items are freely
                    estimated
# Estimate (takes a while)
Sys.time()
fit <- stan(file = '2PL multigroup partial MI.stan',
            data = Dat2plMG,
            iter=20000, chains=2)
Sys.time()

# Having a look at the aggregated results for the core parameters
res <- summary(fit)
# mean difference
round(res$summary[grep("mu", attr(res$summary, "dimnames")[[1]]), ], 3)
# variance ratio
round(res$summary[grep("sigma", attr(res$summary, "dimnames")[[1]]), ], 3)
# item difficulties
round(res$summary[grep("^b\\[", attr(res$summary, "dimnames")[[1]]), ], 3)
# item discriminations
round(res$summary[grep("^v", attr(res$summary, "dimnames")[[1]]), ], 3)
# look for convergence issues via Rhat and Effective sample size
summary((res$summary[, 9:10]))

# Bayesian model averaging for partial MI models:
# Repeat the above analysis multiple times and obtain multiple models, e.g., 2.
post1 <- extract(fit1)
post2 <- extract(fit2)
iter <- 20000

# Obtain a weighted averaged posterior for the mean difference parameter
weights <- c(0.5, 0.5)
muAver <- NULL
for (currCluster in unique(clustering)) {
  nSims <- weights[currCluster]*(iter/2)
  draws <- sample((iter/2):iter, nSims)

  muAver <- append(muAver,
                   rnorm(nSims,
                         get(paste0("post", currCluster))$mu2[draws],
                         get(paste0("post", currCluster))$sigma2[draws]/sqrt(nrow(
                             DataInput))))
}


############## STAN file, saved as "2PL multigroup partial MI.stan" ######
data{
  int<lower = 1> N; // number of examinees
  int<lower = 1> K; // number of items
  int<lower = 1> Ntot; // number of data points
  int<lower = 1> jj[Ntot]; // item id
  int<lower = 1> ii[Ntot]; // person id
  int<lower = 0> y[Ntot]; // responses
  int<lower = 1> g[N]; // group
  int<lower = 1> gg[Ntot]; // group
  int<lower = 1> G; // number of groups
  int<lower=1> NconItem; // number of constrained items
  int<lower=1> conItem[NconItem]; // item constraint
  int<lower=1> freeItem[K - NconItem];
}
parameters{
  vector[N] PersPar; // person parameters
  real mu2; // freely estimated
  real<lower=0> sigma2; // freely estimated
  vector[NconItem] equalB;
  matrix[K - NconItem, 2] unequalB;
  vector<lower=0>[NconItem] equalV;
  matrix<lower=0>[K - NconItem, 2] unequalV;
}
transformed parameters{
  matrix<lower=0>[K, G] v;
  matrix[K, G] b;
b[conItem, 1] = equalB;
v[conItem, 1] = equalV;
b[conItem, 2] = equalB;
v[conItem, 2] = equalV;
b[freeItem, ] = unequalB;
v[freeItem, ] = unequalV;
}
```

```
model{
  // prior person parameter
  for(i in 1:N){
    PersPar[i] ~ normal((g[i]-1)*mu2, (g[i]-1)*sigma2+((g[i]-2)*(-1)));
   }
    mu2 ~ normal(0, 5);
    sigma2 ~ cauchy(0, 5);
  // prior item parameter
  to_vector(unequalB) ~ normal(0, 5); //  discrimination
  equalB ~ normal(0, 5); //  discrimination
  to_vector(unequalV) ~ normal(0, 5); // difficulties
  equalV ~ normal(0, 5);
  for(n in 1:Ntot){
    target += bernoulli_logit_lpmf(y[n] | (v[jj[n], gg[n]]*(PersPar[ii[n]]) - b[jj[n
        ], gg[n]]));
  }
}

########
```

### C.6.4 Item Cluster Analysis for 2PL Models and Two Groups

```
# The IRT estimators used come from mirt.
# No data is provided in these examples.
# requested packages:
library(mirt)
library(doParallel)
library(plyr)
library(rlist)
library(Ckmeans.1d.dp)
library(shape)
# Main function: twoStepThreshold for the 2 groups case to apply the cluster
    algorithm with various thresholds at once.
######################### Two Groups ##########################
# DIF analysis for 2 groups. Returns a list with the results.
# Prints a duration estimate. Has its own summary class.
res <- twoStepThreshold(
data,                       # observed data (items answers only)
group,                      # grouping variable
aThreshold = 0.3,           # alpha threshold (can be a scalar or vector)
bThreshold = c(0.5, 0.6), # beta threshold (can be a scalar or vector)
nCPUs = 4)                  # (maximum) number of processors to be used in parallel in
     the second step (the first step is single core only)
# An overview over the final cluster results for varying threshold combinations:
summary(res)
# A detailed look at the returned list. It is generally structured in the way: list(
    alphaThresholds = list(betaThresholds))
str(res, 2)

############ Estimate 2PL Model for Specific Cluster Solution ############
cluster <- res$`aThresh=0.3`$`bThresh=0.5`$cluster2ndStep
anchor <- which(cluster == 1)  # Get positions of the anchor items. Here, cluster 1
    is used as an example.

mod <- mirt:::multipleGroup(data,
group = group,
invariance = c("free_means", "free_var", names(data)[anchor]), # the invariant items
model = 1,
itemtype = "2PL")
coef(mod, simplify=T)  # show item parameters and latent moments

################## Functions ##############
# 1)  2step algorithm for 2PL model for groups.
# Yields a list with the resulting cluster memberships and parameter estimates per
    step.
twoStepThreshold <- function(
Dat,
groups,
aThresholds,    # can be a single value or a vector of values (i.e., a sequence)
bThresholds,    # can be a single value or a vector of values (i.e., a sequence)
nCPUs = 1
) {    # (maximum) number of processors to be used in 2nd step

threshL <- list()   #  intialize level 2 list
if (length(unique(groups)) > 2) stop("Group␣number␣is␣wrong.␣Exactly␣2␣groups␣are␣
    needed.")

# initial model
cat("Estimating␣initial␣model...")
t0 <- proc.time()
mod1 <- multipleGroup(Dat,
model = 1,
itemtype = "2PL",
#method = "MHRM",
group = groups,
invariance = character(),
SE = T, verbose=F)
t1 <- proc.time()
tDiff <- as.numeric(t1 - t0)[3]

params1 <- getItemParams(mod1, SE=F)

if (min(params1[c(2, 4)]) < 0) {
"Negative␣loadings.␣Cannot␣continue."
stop()
}
##  2step approach
#  1st step: alphas

for (aThresh in aThresholds) {  #  main loop for various a-thresholds
dR_alpha <- drids2PL(mod1)$dR_alpha
cluster1stStep <- kMeansThresh(dR_alpha[, 1], aThresh)
```

```
cat("\r", "Current alpha threshold =", aThresh, "with approx. duration =",
ceiling(ceiling(max(cluster1stStep) / nCPUs)*tDiff/60), "minutes")

#  2nd step: betas
cluster2ndStepL <- list(cluster1stStep = cluster1stStep, model1stStep = list(params
    = params1,
output = capture.output(mod1)))
for (bThresh in bThresholds)  cluster2ndStepL[[paste0("bThresh=", bThresh)]][["
    cluster2ndStep"]] <- vector(length = length(cluster1stStep))    #  initialize
    lower storage list
cluster1stStep_1 <- unname(which(table(cluster1stStep) == 1))           #  check for
     1-item alpha clusters...
cluster1stStep_for2ndStep <- unname(which(table(cluster1stStep) > 1))   #  ...and
    here all clusters with more than 1 item

# parallized 2nd step estimation (only for alpha clusters with at least 2 items)
cl <- makeCluster(nCPUs)   # initialize parallization
registerDoParallel(cl)

mod2L <- foreach(curr1step = cluster1stStep_for2ndStep, .inorder=T, .combine=append,
.packages=c("mirt","Ckmeans.1d.dp"), .errorhandling='stop') %dopar% {

currCluster <- which(cluster1stStep == curr1step)   # get current items
anchorMod <- paste0("F1 =  1-", length(cluster1stStep),"
CONSTRAINB =  (", paste(curr1step, collapse=",  "), ",  a1),  (", sample(curr1step, 1),
    ",  d)") #  constrain item of current alpha cluster and a random intercept from
    the current cluster
mod2 <- multipleGroup(Dat,
model = anchorMod,
itemtype = "2PL",
#method = "MHRM",
group = groups,
invariance = c("free_var"),  #"free_means",
SE = T, verbose=F)
}
stopCluster(cl) # cleanly exit parallel threads

if (length(cluster1stStep_for2ndStep) == 1) mod2L <- list(mod2L) # achieve list
    structure even when there is only 1 entry

for (i in cluster1stStep_for2ndStep) {
curr1step <- which(cluster1stStep == i)   # get current items

params2 <- getItemParams(mod2L[[which(cluster1stStep_for2ndStep == i)]], SE=F)
dR_beta <- drids2PL(mod2L[[which(cluster1stStep_for2ndStep == i)]])$dR_beta
cluster2ndStepL[["model2ndStep"]][[paste0("1st cluster=", i)]] <- list(params=
    params2,
output=capture.output(mod2L[[which(cluster1stStep_for2ndStep == i)]]))

for (bThresh in bThresholds) {
curr2step <- kMeansThresh(dR_beta[curr1step, 1], bThresh)
cluster2ndStepL[[paste0("bThresh=", bThresh)]][["cluster2ndStep"]][curr1step] <-
    curr2step + i*100   #  i*100 produces unique cluster labels
historySteps <- rep(NA, length(cluster1stStep))  #  save history
historySteps[curr1step] <- curr2step
cluster2ndStepL[[paste0("bThresh=", bThresh)]][[paste0("history: 1st cluster=", i)]]
     <- historySteps
}
}

for (i in cluster1stStep_1) {
for (bThresh in bThresholds) {
cluster2ndStepL[[paste0("bThresh=", bThresh)]][["cluster2ndStep"]][cluster1stStep==i
    ] <- rep(i, 1)  #  for 1-item clusters: give cluster codes that will not occur
    otherwise (below 100)

}
}
# unify cluster labels
for (bThresh in bThresholds) cluster2ndStepL[[paste0("bThresh=", bThresh)]][["
    cluster2ndStep"]] <- as.integer(factor(cluster2ndStepL[[paste0("bThresh=",
    bThresh)]][["cluster2ndStep"]]))
##  end 2step approach

threshL[[paste0("aThresh=", aThresh)]] <- cluster2ndStepL
}

class(threshL) <- "twoStepClass"   # give the result list a class attribute in order
     to be able to apply a customized summary-function
return(threshL)
}
```

```
summary.twoStepClass <- function(threshL,
inRows = F
) {
stopifnot(inherits(threshL, "twoStepClass"))
Res <- data.frame()
i <- 0
nItems <- length(threshL[[c(1, 3, 1)]])

if (inRows == F) {
for (aThresh in 1:length(threshL)) {
for (bThresh in 1:(length(threshL[[1]]) - 3)) {
i <- i + 1
Res[1, i] <- sub(".......=", "", names(threshL[aThresh]))                    # save a-
    threshold
Res[2, i] <- sub(".......=", "", names(threshL[[aThresh]][bThresh + 2]))    # save b-
    threshold
Res[3, i] <- max(t(t(threshL[[c(aThresh, (bThresh + 2), 1)]])))             # save
    number of clusters
Res[4, i] <- ""
Res[5:(nItems + 4), i] <- t(t(threshL[[c(aThresh, (bThresh + 2), 1)]]))     # save
    cluster code
}
}
rownames(Res) <- c("aThresh", "bThresh", "nClusters", "", paste0("i", 1:nItems))
colnames(Res) <- NULL
cat("####␣Final␣clusters␣found␣in␣2step␣algorithm␣####", "\n")
} else {
for (aThresh in 1:length(threshL)) {
for (bThresh in 1:(length(threshL[[1]]) - 3)) {
i <- i + 1
Res <- rbind(Res, as.numeric(c(sub(".......=", "", names(threshL[aThresh])),
                  # save a-threshold
sub(".......=", "", names(threshL[[aThresh]][bThresh + 2])), # save b-threshold
max(t(t(threshL[[c(aThresh, (bThresh + 2), 1)]]))),           # save number of
    clusters
threshL[[c(aThresh, (bThresh + 2), 1)]])))                    # save cluster code
}
}
colnames(Res) <- c("aThresh", "bThresh", "nClusters", paste0("i", 1:nItems))
cat("####␣Final␣clusters␣found␣in␣2step␣algorithm␣####", "\n", "\n")
}
return(Res)
}

################# Auxilary functions ##########################
## retrieve parameters from mirt in a digestable way
getItemParams <- function(mod,                          # mirt-model
params = c("d", "a1"), # character[i]: names of the item parameters to be extracted
    (in general: d/a1/g/u)
SE = TRUE,             # logical: should parameter standard errors be reported as
    well?
addVar = NULL,         # numeric[k]: add an additional/ multiple variables to the
    output (for ecample to compare simulated values), use cbind() to give them
    useful names
roundTo = 5            # numeric[1]: decimals, the values should be rounded to
#-> Q4
){

containsSE <- !is.na(vcov(mod)[1,1]) # does the model include SEs? (estimated with
    SE = TRUE?)
if (!containsSE && SE) {
warning("Standard␣errors␣were␣requested␣but␣not␣estimated.␣They␣will␣not␣be␣reported
    ...")
SE <- FALSE
}
itNames <- extract.mirt(mod, "itemnames")
grpNames <- extract.mirt(mod, "groupNames")
out <- data.frame(matrix(NA, nrow = length(itNames), ncol = 1))
cfg <- coef(mod, printSE = TRUE)
res <- data.frame(t(data.frame(cfg)))
for (g in grpNames){
if (length(grpNames) > 1) {
grpIdx <- grep(g, rownames(res))
} else {
grpIdx <- 1:nrow(res)
}
s <- strsplit(rownames(res),".", fixed = TRUE)
parVec <- sapply(s,function(x) x[length(x)])
for (p in params){
parIdx <- which(parVec == p)
curr <- res[intersect(parIdx,grpIdx),]
if (!is.numeric(curr)) curr <- curr[,"par"]
```

```
if (length(grpNames) == 1) {
colName <- p
} else {
colName <- paste(g, p, sep = ".")
}
out[colName] <- curr
if (SE && containsSE){
curr <- res[intersect(parIdx,grpIdx),]
out[paste(colName, "SE",sep = ".")] <- curr[, "SE"]
}
}
}
out <- out[-1]
rownames(out) <- itNames
if (!is.null(addVar)) out <- cbind(out, addVar)
return(as.data.frame(round(out, roundTo)))
}

## apply k-means clustering with a threshold
kMeansThresh <- function(drids,          # relative DIF-values for the items (taken
    from delta-R matrix)
thresh = NULL  # threshhold as a stopping rule. Threshold is maximum cluster width
    on logit scale
){
k <- 1
currRange <- max(drids) - min(drids)
while(max(currRange) > thresh & k < length(drids)) {
k <- k + 1
clusCode <- Ckmeans.1d.dp(drids, k=k)$cluster
currRange <- NULL
for(i in 1:max(clusCode)) {
clusCode2 <- clusCode     # helper copy
clusCode2[clusCode2 != i] <- NA         # make picking vector
clusCode2[clusCode2 == i] <- 1
currRange[i] <- max(clusCode2*drids, na.rm=T) - min(clusCode2*drids, na.rm=T)   #
    get current range
}
}
res <- Ckmeans.1d.dp(drids, k=k)$cluster
res
}
```

# Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorgelegte Dissertation selbständig und nur unter Verwendung der angegebenen Hilfsmittel verfasst habe. Ich besitze noch keinen Doktorgrad im Fach Psychologie und habe mich auch nicht darum beworben. Die Arbeit oder Teile von ihr sind in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden. Mir ist die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 13.02.2006 bekannt. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsbearbeiterinnen/ Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

---

Daniel Schulze