4-2020

Examining the Performance of the Alignment Method in DIF Analyses

Paulius Satkus

Christine E. DeMars
demarsce@jmu.edu

Recommended Citation

Satkus, P., & DeMars, C. E. (2020, April). Examining the Performance of the Alignment Method in DIF Analyses. Paper presented at the annual meeting of the the National Council on Measurement in Education, San Francisco, CA. (virtual, conference cancelled)

Examining the Performance of the Alignment Method in DIF Analyses

Paulius Satkus

Christine E. DeMars

Department of Graduate Psychology and Center for Assessment and Research Studies,

James Madison University

Satkus, P., & DeMars, C. E. (2020, April). *Examining the Performance of the Alignment Method in DIF Analyses.* Paper presented at the annual meeting of the the National Council on Measurement in Education, San Francisco, CA. (virtual, conference cancelled)

Abstract

The alignment procedure is a new method for multiple group invariance models. An important advantage of alignment over the traditional methods is that alignment does not require full measurement invariance to estimate group means and variances (Muthén & Asparouhov, 2014). Simulation studies have supported that alignment performs adequately in situations when few items are noninvariant (or function differentially across groups – DIF). In most other studies, the tests were simulated to represent attitudinal surveys (e.g., fewer items, continuous data). In this study, we evaluated how alignment would perform with a typical educational cognitive test – 40 items scored dichotomously. Different patterns of DIF were examined. Results suggest that alignment is fairly robust; the only condition where the recovered parameters were non-trivially biased was when, in addition, to few large DIF items, several moderately DIF items were present. Our findings add to the growing body of evidence that alignment is an adequate procedure for multiple group analyses.

Examining the Performance of the Alignment Method in DIF Analyses

Background

A new method for multiple-group measurement models called alignment was proposed by Asparouhov and Muthén (2014). The alignment procedure offers an advantage over the traditional methods (e.g., nested multiple-group confirmatory factor analysis (CFA) or differential item functioning (DIF) analyses in IRT) as it does not require full measurement invariance (i.e., a well-fitting scalar measurement model in CFA or multiple purification steps in IRT) to estimate group means and variances. It does so by utilizing the total loss (or simplicity) function. If separate models are estimated for each group, the metric is identified separately for each group. Given that, the group means cannot be compared. The point of alignment is to choose linear constants for each group that will best align them. The constants for each group are chosen based on minimizing a total loss function that selects a model for which the majority of item loadings and intercepts are equivalent across multiple groups. Asparouhov and Muthén (2014) explain that this loss function should yield a few large differences between item parameter estimates for different groups and many minimal differences. The purpose of this study is to explore the accuracy of the alignment procedure for varying patterns and magnitudes of DIF, beyond contexts where a small proportion of items show large DIF and other items show little or no DIF.

Broadly speaking, the estimated model is only one of an infinite amount of models with equal fit. Initially each group's mean (α_g) is fixed to 0 and each group's variance (ψ_g) is fixed to 1 to identify the model within each group. During the alignment step, the group means and variances are then adjusted (no longer fixed to 0 and 1 for all groups), and correspondingly the item loadings and thresholds are transformed via a linear transformation:

$$\lambda_{pg,1} = \frac{\lambda_{pg,0}}{\sqrt{\psi_g}} \quad (1)$$

$$\nu_{pg,1} = \nu_{pg,0} - \alpha_g \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}, \quad (2)$$

where $\lambda_{pg,0}$ and $\nu_{pg,0}$ are the estimated loading and threshold,[1] respectively, for item *p* and group *g* in the initial model and $\lambda_{pg,1}$ and $\nu_{pg,1}$ are the aligned estimates. The α_g and ψ_g (group mean and variance) are chosen to minimize the weighted differences between the item parameters across all groups. The transformation does not change the model-data fit. The loss function is:

$$F = \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(\lambda_{pg_1,1-} \lambda_{pg_2}) + \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(\nu_{pg_1,1-} \nu_{pg_2}) \quad (3)$$

Inevitably, some estimated item parameters may remain noninvariant across groups. In other words, the alignment procedure estimates a model where many estimated item parameters are invariant across groups, but a few may remain noninvariant (i.e., displaying DIF). The estimated model has the same exact fit as a model where all items are freely estimated (i.e., the configural model) but using alignment procedure factor means and variances are on the same metric and therefore comparable across groups.

One of the advantages of using the alignment approach (over the traditional multiple group models) is that alignment allows comparison of many groups (whereas using the traditional methods, comparison of more than 5 groups becomes cumbersome; Muthén & Asparouhov, 2013). Researchers used alignment with as many as 92 groups to compare attitudes towards immigrants among youth across Europe (Munck, Barber, & Turney-Purta, 2018). In this study, and another study by Lomazzi (2018) examining attitudes towards gender roles, the

[1] λ is related to the slope in the 2PL IRT model: $\lambda = \frac{a}{\sqrt{a^2 + \pi^2/3}}$. Similarly, $\nu = \frac{d}{\sqrt{a^2 + \pi^2/3}}$

attitudes were measured using short (i.e., five item) 5-point Likert scales that were treated as continuous data. Other studies used alignment with categorical data: dichotomous items (Muthén & Asparouhov, 2014) and 4-point polytomous items (Flake & McCoach, 2018; Tay Jayasuriya, Jayasuriya, & Silove, 2017).

So far, the performance of alignment has fared well in simulation studies. Asparouhov and Muthen (2014) reported on three studies where they examined how well group means and variances, along with invariant item parameters (loadings and intercepts) were recovered under several conditions. In the first study using maximum likelihood (ML) estimation, three levels of noninvariance were examined (0%, 10%, or 20% of items parameters exhibiting noninvariance). With as many as 60 groups, they found that when $n = 1000$ per group, even with 20% noninvariant item parameters, the group means were recovered relatively well. Additionally, the item parameters were also recovered well. This simulation study lead the authors to suggest that with as much as one fifth noninvariant items, the alignment procedure works well.

The second study in Asparouhov and Muthén (2014) compared the item parameter standard errors from ML estimation (as in the first study) and Bayesian estimation. They used simulation conditions from the first study (e.g., 20% noninvariant items, 3 groups). They found that using Bayesian estimation, the standard errors were estimated better – average standard error was closer to the empirical standard deviation. In the third simulation study, Asparouhov and Muthén (2014) compared FIXED and FREE options of the alignment method. The FIXED option constrains the group mean and variance of the first group to 0 and 1 respectively. This is done for identifying the estimated model. With the FREE option, the group mean is estimated for each group, so there is an additional parameter. They discussed the advantages and

disadvantages for the FREE and FIXED options, and thus the third study was used to illustrate them. All applied studies discussed in this paper used the FIXED option.

Flake and McCoach (2018) studied alignment with polytomous items. In this study, they also varied the amount of noninvariance (DIF) in the item loadings and the item thresholds. They found that the absolute bias was greater than 0.06 only when the magnitude of DIF was large (i.e., .40 for loadings) for 3 of 7 items (43%). For thresholds, when DIF was medium (.5) or large (.8) for either 2 or 3 items the absolute bias was greater than .006. In other words, large (and medium) DIF for thresholds had greater impact for item recovery as alignment was not performing adequately when 29% or more of the items were noninvariant. This finding led the authors to recommend that alignment would perform well if no more than than 29% of items had DIF. Similar recommendations were made for adequate recovery for group means and variances.

The simulation studies described here primarily studied contexts where a few items had large magnitudes of DIF, which is what the loss function is designed for. The purpose of this study is to explore how accurately the alignment procedure adjusts the group means and variances, given a fixed total quantity of DIF but different patterns of DIF magnitude in individual items. Additionally, the prior research mostly focuses on polytomous or continuous data with short scales. Educational assessments typically use dichotomous items and longer tests. Thus, this study will focus on conditions s more typical of educational assessments.

Data

40 items were simulated to follow a 2PL IRT model: $P(\theta_j) = \frac{e^{a_i\theta_j - d_i}}{1 + e^{a_i\theta_j - d_i}}$, where P(θ) is the probability of correct response given ability θ, a_i is the item discrimination, and d_i is the item difficulty. Item parameters were chosen to represent two levels of discrimination and a broad range of difficulty. Table 1 shows the item parameters. 12 items contained a degree of DIF,

whereas the other 28 were DIF-free. We simulated 8 groups with the following true means: 0.2, 0 (reference group), -0.2, -0.4, -0.4, -0.6, -0.8, -1. The θs within each group were normally distributed with variance = 1. The simulated sample sizes for the 8 groups were: 500, 2000, 2000, 500, 200, 300, 2000, 200.

The focus of the present study was to examine how different patterns and magnitudes of DIF affected the alignment of the group means and variances. Thus, we varied whether DIF was balanced or unbalanced among the 12 items. Only the item difficulties showed DIF, sometimes called uniform DIF or *d*-DIF.

Table 2 shows the pattern of *d*-DIF for each condition. The first three conditions were balanced—the total magnitude of DIF was zero because the sum of the *d*-DIF for items favoring groups 1-4 was offset by the sum of the *d*-DIF for items favoring groups 5-8. The other three conditions favored groups 1-4 more than groups 5-8, thus the *d*-DIF was unbalanced.

All data were simulated using SAS 9.4. The alignment procedure was conducted using Mplus 8.2. Group 2 was fixed as the reference group. Each condition was replicated 500 times.

Results

Figure 1 shows the resulting bias in the estimated group means for all six conditions. Results for all but one condition followed the same pattern, but the magnitude of the bias varied somewhat. That is, it appears that whether *d*-DIF was balanced (conditions 1, 2, and 3) or unbalanced (conditions 5 and 6) did not change the pattern of bias. The group means were slightly positively biased for all other groups relative to the reference group, which was fixed. The bias increased from group 1 to group 6 as the true group mean decreased and then decreased for groups 7 and 8, the groups with the lowest means. The DIF for groups 5-8 was in the opposite direction of the DIF for groups 1-4, and bias in groups 5-8 was slightly larger than the

bias for groups 1-4. The difference between the groupings can be seen in the contrast between groups 4 and 5, which had the same true mean. In all five conditions, group 6 had the greatest bias, however the magnitude was not alarming (i.e., greatest bias in condition 3, group 6 bias was 0.086).

A different pattern of bias was observed for condition 4, where the *d*-DIF was unbalanced with groups 1-4 favored with six items by 0.8 and six items by 0.4. In this condition, the estimated group means for groups 1-4 followed the same increasing trajectory as in the other conditions, but groups 5-8 did not show the same pattern of bias. Overall, the bias in condition 4 was lower than in other conditions; it may be easier for the alignment procedure to detect and correct for DIF in conditions like this where the DIF is relatively large and in the same direction for all DIF items.

Discussion and Educational Importance

The alignment procedure offers a promising new approach to estimating group means in the presence of DIF items. However, few studies have been conducted evaluating the performance of alignment, especially with dichotomous items and moderate-length tests. In the current simulation study, we examined how different DIF patterns affect the estimated group means. Our results suggest that the alignment procedure adjusted the group means well. In condition 4, where the pattern of bias was noticeably different, the bias was greater for groups 1-4 than for groups 5-8, which is the opposite result as for the other five conditions. Condition 3, in which two large DIF items were balanced by ten smaller items with DIF in the opposite direction, showed slightly more bias for groups 5-8 than Conditions 1 and 2, in which six items were biased in each direction with the same amount of DIF. Similarly, Condition 6, which had the same total DIF as Condition 5 but several items with smaller DIF, showed slightly more bias

for groups 5-8. Alignment may not be as effective when items with moderate DIF occur in the same test as items with large DIF. Further research is warranted to continue studying the alignment procedure. If the group means can be accurately aligned, next steps include detecting DIF items.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495-508.

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 56-70.

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *12*, 77-103.

Munck, I., Barber, C., & Torney-Purta, J. (2017). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*, *47*, 687-728.

Muthén, B., & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups (*MPLUS Technical Report). http://www.statmodel.com

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology*, *5*, 1-7.

Tay, A. K., Jayasuriya, R., Jayasuriya, D., & Silove, D. (2017). Assessing the factorial structure and measurement invariance of PTSD by gender and ethnic groups in Sri Lanka: An analysis of the modified Harvard Trauma Questionnaire (HTQ). *Journal of anxiety disorders*, *47*, 45-53.

Table 1
Simulated item parameters

Item	*a*	*d*
1	0.6	-1.95
2	1.2	-1.95
3	0.6	-1.65
4	1.2	-1.65
5	0.6	-1.35
6	1.2	-1.35
7	0.6	-1.05
8	1.2	-1.05
9	0.6	-0.75
10	1.2	-0.75
11	0.6	-0.45
12	1.2	-0.45
13	0.6	-0.15
14	1.2	-0.15
15	0.6	0.15
16	1.2	0.15
17	0.6	0.45
18	1.2	0.45
19	0.6	0.75
20	1.2	0.75
21	0.6	1.05
22	1.2	1.05
23	0.6	1.35
24	1.2	1.35
25	0.6	1.65
26	1.2	1.65
27	0.6	1.95
28	1.2	1.95
29	0.6	-1±.5*d*-DIF*
30	1.2	-1±.5*d*-DIF
31	0.6	0±.5*d*-DIF
32	1.2	0±.5*d*-DIF
33	0.6	1±.5*d*-DIF
34	1.2	1±.5*d*-DIF
35	0.6	-1±.5*d*-DIF
36	1.2	-1±.5*d*-DIF
37	0.6	0±.5*d*-DIF
38	1.2	0±.5*d*-DIF
39	0.6	-1±.5*d*-DIF
40	1.2	-1±.5*d*-DIF

Note. * The item difficulty parameters (*d)* for items 29-40 varied by the DIF condition as detailed in Table 2.

Table 2
Summary of d-DIF in DIF items in six conditions

	Balanced DIF			Unbalanced DIF		
Item	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
29	0.4	0.8	0.8	0.8	1.1	1.4
30	0.4	0.8	0.8	0.8	1.1	1.4
31	0.4	0.8	-0.16	0.8	1.1	1.4
32	0.4	0.8	-0.16	0.8	1.1	1.4
33	0.4	0.8	-0.16	0.8	1.1	1.4
34	0.4	0.8	-0.16	0.8	1.1	1.4
35	-0.4	-0.8	-0.16	0.4	1.1	-0.2
36	-0.4	-0.8	-0.16	0.4	1.1	-0.2
37	-0.4	-0.8	-0.16	0.4	-0.4	-0.2
38	-0.4	-0.8	-0.16	0.4	-0.4	-0.2
39	-0.4	-0.8	-0.16	0.4	-0.4	-0.2
40	-0.4	-0.8	-0.16	0.4	-0.4	-0.2

Note: *d*-DIF is the difference between the *d*-parameters. Half of this value was subtracted from the *d* for groups 1-4 and half was added to the *d* for groups 5-8. Thus, positive *d*-DIF favors groups 1-4.

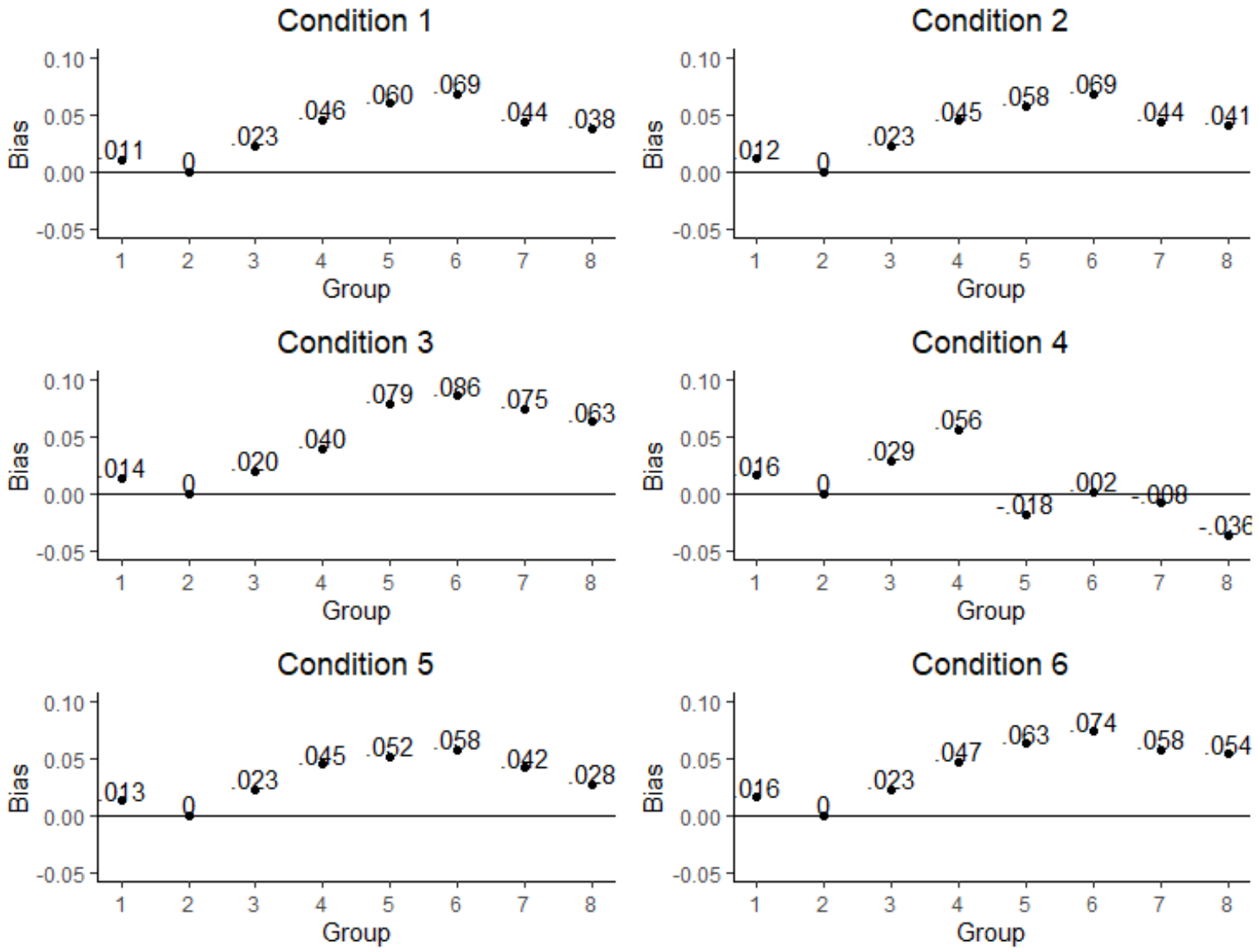

Figure 1. The bias in aligned group means. Note group two was fixed to 0 in each condition.