Educational Methods & Practice

https://emp-open.de
Volume 2 (2024)

Research Article

Assessing Measurement Invariance in a University Entrance Exam: A Comparison of Multigroup Confirmatory Factor Analysis Alignment Method vs. Multigroup Item Response Theory

Hamdollah Ravand*
Vali-e-Asr University of Rafsanjan, Iran

Abstract

This study aimed to assess measurement invariance in a university entrance examination while comparing the effectiveness of multigroup item response theory (MGIRT) and multigroup confirmatory factor analysis (MGCFA) alignment methods in identifying non-invariant items. Data from four groups, representing different university types, were analyzed using the TAM package in R and Mplus. The findings indicated that both MGIRT and MGCFA alignment methods converged on one non-invariant item, with MGCFA alignment detecting an additional three cases of non-invariance. Sensitivity analyses revealed that the presence of these non-invariant items did not significantly affect the equity of the tests, which are norm-referenced. Methodological and substantive implications of the study findings are discussed.

Keywords

Alignment, measurement invariance, multigroup confirmatory factor analysis, multigroup item response theory

1 | Introduction

Measurement invariance (MI) is a crucial consideration in the field of psychometrics, particularly when comparing latent traits across different groups. It ensures that the construct being measured is interpreted consistently across diverse populations. Without establishing MI, comparisons of scores across groups may be misleading, as differences in scores could be attributed to the measurement instrument rather than the constructs themselves (Meredith, 1993; Vandenberg & Lance, 2000). This paper examines measurement invariance in a language test using two advanced statistical methods, namely, Multigroup Confirmatory Factor Analysis (MGCFA) alignment and Multigroup Item Response Theory (MGIRT).

Corresponding author:
*Hamdollah Ravand, English Department, Faculty of Foreign Languages, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
Email: ravand@vru.ac.ir
Hamdollah Ravand: https://orcid.org/0000-0002-8757-3850

Received 5 March 2024; Received in revised form 2 June 2024; Accepted 18 June 2024
Available online 23 June 2024

Ensuring MI is critical for the validity of cross-group comparisons. Without MI, differences in scores may be attributed to the measurement instrument rather than actual differences in the constructs being measured. This misattribution can lead to erroneous conclusions and potentially flawed policy decisions, especially in fields like education where such measurements inform educational practices and interventions (Vandenberg & Lance, 2000). Ignoring MI can result in biased parameter estimates, distorting the interpretation of research findings. For example, in educational assessments, failing to ensure MI could mean that test scores reflect not only students' abilities but also irrelevant group characteristics such as cultural background or language proficiency, thereby undermining the fairness and equity of the assessment process (Steenkamp & Baumgartner, 1998).

2 | Methodological Approaches: MGCFA and MGIRT

MGCFA is a traditional and widely used method for testing measurement invariance. It involves comparing nested models with increasingly restrictive equality constraints on parameters across groups (Meredith, 1993). The process typically starts with assessing configural invariance, which checks if the same factor structure holds across groups without imposing equality constraints on factor loadings or intercepts/thresholds. If configural invariance is supported, the next step is to test for metric invariance by constraining factor loadings to be equal across groups. Finally, scalar invariance is tested by additionally constraining item intercepts/thresholds/difficulties to be equal across groups. These steps ensure that the construct is measured equivalently across groups, allowing meaningful comparison of latent means and relationships between constructs across these groups (Davidov et al., 2014). Detection of partial invariance is one of the challenges that arise when testing for measurement invariance using MGCFA. Partial invariance occurs when some, but not all, parameters are invariant across groups. Traditional methods such as MGCFA often assume full invariance, which can lead to inaccurate conclusions if this assumption is violated (Millsap, 2011).

MGCFA alignment, a recent advancement in the field, addresses this limitation of the MGCFA approach by allowing for approximate invariance. This method involves testing configural invariance as the initial step. However, instead of imposing strict equality constraints, the alignment method allows for small differences in item parameters across groups, provided these differences do not substantially impact the construct being measured. This approach is particularly useful when dealing with large and diverse groups where strict invariance is often unrealistic (Asparouhov & Muthén, 2014). The alignment approach, first presented by Asparouhov and Muthén (2014), initially focused on the multiple group factor analysis model utilizing continuous variables and employing both maximum-likelihood (ML) and Bayesian estimators. Subsequently, Muthén and Asparouhov (2014) extended this method to encompass the multiple group factor analysis model incorporating ordered categorical variables. Alignment is designed to facilitate the comparison of latent variables across groups without necessitating measurement invariance.

The alignment optimization procedure simplifies comparing factor means across different groups by minimizing differences in factor loadings and item intercepts. This

involves two models: the original model (M0), based on a baseline configural model assuming similar item configurations across all groups, and the optimized model (M1), derived by minimizing differences in factor loadings and item intercepts among groups. The optimized factor means and variances in M1 are then used for group comparisons. Unlike traditional MGCFA, which requires invariant factor configuration, factor loadings, and item intercepts for scalar invariance, the alignment procedure allows for an adequate configural model with minimal differences in these parameters, sufficient for valid factor mean comparisons. Similar to rotation algorithms in exploratory factor analysis, which aim for a simple structure by maximizing large loadings and minimizing small ones, alignment optimization seeks to minimize differences in loadings and intercepts across groups, ensuring most item parameters are approximately equal with only a few substantial differences. The alignment method proves advantageous in discerning meaningful differences in factor means, even amidst non-invariance, contrasting with traditional MGCFA that necessitates scalar invariance for accurate factor mean comparisons. This flexibility enables meaningful comparisons of factor means even when scalar invariance is not achieved, which is significant in real-world data analysis where perfect invariance is often elusive.

MGIRT (Bock & Zimowski, 1997) offers a robust framework for examining item-level measurement invariance, commonly referred to as Differential Item Functioning (DIF). This method accounts for the possibility that item characteristics, such as difficulty and discrimination, might vary across groups, potentially affecting the validity of cross-group comparisons. In MGIRT, item parameters are estimated regardless of the groupings. Statistics such as Root Mean Square Deviation (RMSD) and Mean Deviation (MD) are employed to detect non-invariance or DIF, providing detailed information on which items function differently across groups (OECD, 2017). This strategy involves estimating a MGIRT model where item parameters are consistent across groups, but group means and variances differ. Next, empirical Item Characteristic Curves (ICCs) for each item are estimated within each group and compared with the joint ICC from the MGIRT model. Statistics quantify the distance between group-specific and joint ICCs. Measurement invariance (MI) is confirmed if the difference between the group-specific ICC and the overall ICC is minimal, indicating that group performance is adequately explained by the joint parameters (Yamamoto et al., 2013).

3 | Comparison of MGIRT and MGCFA

Several studies have highlighted discrepancies between MGCFA and MGIRT in detecting measurement invariance with ordered-categorical variables (Meade & Lautenschlager, 2004; Oishi, 2006; Raju et al., 2002; Reise et al., 1993). Kim and Yoon (2011) compared the two methods using CFA adapted for ordered categorical variables. They found that IRT demonstrated superior performance in terms of both true positive (TP) and false positive (FP) rates. While MCCFA showed nearly identical TP rates to IRT, it had significantly higher FP rates, especially as sample size and DIF increased. This inflation of FP rates in MCCFA was attributed to its use of an incorrect baseline model, which magnified chi-square

differences, particularly in DIF cases. In contrast, IRT showed less severe FP rate inflation under similar conditions, consistent with previous findings.

Buchholz and Hartig (2020), in research using both simulation and real data, compared the performances of MGCFA and MGIRT. They found that both methods demonstrated high agreement in identifying group differences in the threshold parameter, which represents the difficulty of responding in a high category of an item. However, the agreement between the two methods in identifying group differences in the slope parameter, which represents the strength of the relationship between an indicator and the latent construct, was substantially lower. Despite these insights, no study has yet compared the MGCFA alignment with MGIRT.

4 | Measurement Invariance Studies on the UEE Tests

The University Entrance Examination (UEE), locally known as Konkoor, refers to the suite of matriculation examinations developed and administered by the Assessment Organization (AO) in Iran. The UEE is designed to screen candidates aspiring to pursue higher education at Iranian universities at the Bachelor's, Master's, or Ph.D. levels, both in public and private institutions. For applicants to English programs, the UEE typically includes a General English (GE) section and a content knowledge part, with all items presented in multiple-choice format. Notably, the GE section does not assess writing or speaking skills.

Most measurement invariance (MI) studies on the UEE have employed differential item functioning (DIF) analysis to examine the comparability of the constructs measured across different observable groups, such as gender and major (e.g., Ahmadi et al., 2015; Barati & Ahmadi, 2010; Ravand, 2015; Ravand & Firoozi, 2016; Ravand et al., 2019). For instance, Barati and Ahmadi (2010) investigated gender DIF in the UEE for applicants to English programs at the Bachelor's level and found that approximately 44% of the 70 items exhibited DIF. Similarly, Ravand et al. (2019) examined major and gender DIF in the UEE for applicants to English programs at Master's level and identified DIF in 12 of the 60 items. Ahmadi et al. (2015) analyzed both the GE and content knowledge sections (totaling 100 items) of the UEE for Ph.D. applicants to English programs, finding that 12 items exhibited DIF.

To the best of the author's knowledge, only one study has explored MI in the UEE using MGCFA alignment and MGIRT. Ravand et al. (2018) investigated the factor structure invariance of the GE section of the UEE for Ph.D. applicants to English programs across two proficiency levels. Using the traditional MGCFA, they analyzed responses from a random sample of participants (N=1009) who took the test in 2014. The results demonstrated that configural, metric, and scalar invariance held across both proficiency levels, indicating that both the factor loadings (metric invariance) and the thresholds/difficulties (scalar invariance) were invariant across the groups.

The present study holds both practical and methodological significance. Practically, it aims to determine whether the University Entrance Examination for Master's applicants into English programs at Iranian universities (MAUEE) measures the same construct across different groups of test takers, thereby informing test developers and users about potential

item bias, which can influence the test development process. Methodologically, it sheds more light on how detection of non-invariant items is consistent across the two methods of MFCFA alignment and MGIRT. The significance of measurement invariance in high stakes testing, particularly for selection purposes, is underscored by Meredith and Teresi (2006) and Borsboom (2006). They highlight the critical role of measurement invariance in preventing measurement bias, which can distort comparisons across different groups. The absence of measurement invariance jeopardizes both fairness and equity across groups, as noted by Millsap and Kwok (2004), who observed that violations could lead to unjust selection outcomes based on group affiliation. Thus, confirming measurement invariance is paramount in ensuring the validity and fairness of high stakes testing processes.

Additionally, it investigates whether the two popular methods of MI—namely, MGCFA alignment and MGIRT—identify the same items as functioning differentially across groups. The following sections will outline the procedures involved in each method. For MGCFA, using Mplus, I will detail the steps of testing measurement invariance, through the MGCFA alignment method. For MGIRT, using the *TAM* package in R, I will describe the process of detecting DIF through RMSD and MD. By comparing these methods, it is intended to provide insights into their relative strengths and limitations, contributing to a broader understanding of measurement invariance in psychometric evaluations.

5 | Data

For the present study, data from MAUEE were employed. Specifically, the data came from the general English (GE) section of the MAUEE, which consisted of 60 multiple-choice items: 10 grammar, 20 vocabulary, 10 cloze, and 20 reading comprehension questions. Test takers were required to complete the GE section within 60 minutes. The Assessment Organization, responsible for developing and administering all high-stakes tests, including university entrance examinations in Iran, oversees the MAUEE.

The study included data from four groups of applicants, totaling 20,395 individuals, comprising 5,810 males and 15,831 females, who attended different types of universities for their Bachelor's studies.

1. State universities (*n*=4,615)
2. Azad universities (*n*=9,238)
3. Payame Noor Universities (*n*=3,887)
4. Non-profit Non-government Universities (*n*=2,655)

State universities (referred to as Group 1 hereafter) in Iran are typically funded and managed by the government. They often have relatively strict admission criteria and highly competitive entrance exams, attracting top-performing students. Azad universities (Group 2), also known as Islamic Azad University, are private institutions in Iran. They offer a wide range of programs and are known for their relatively flexible admission policies compared to state universities. Payame Noor University (Group 3) operates with a unique model where attendance at classes is not mandatory, and the majority of the course load is learned through self-study. This model accommodates individuals seeking a university degree while

balancing other professional commitments in government and private sectors. Additionally, there are Non-government non-profit universities (Group 4) in Iran, which are also private institutions. Admission to these private universities is less competitive than to state universities.

Given the absence of official reports on the validity of the MAUEE, it was necessary to assess the quality of the items before proceeding with MGCFA and MGIRT. Researchers should only consider measurement invariance testing for scales that have an established factor structure in at least one group or sample, preferably with existing confirmatory evidence (Luong & Flake, 2023). While it is ideal to use scales with robust validity evidence, this is not always feasible. Therefore, it is recommended that researchers confirm the factor structure within their own sample by performing a CFA on the entire sample. This step is crucial as a known factor structure is essential for testing configural invariance (Luong & Flake, 2023). To ensure the validity of the MAUEE items, exploratory factor analysis (EFA) was conducted first, followed by single-group CFA, and finally, MGCFA using the alignment method.

6| Results

6.1 | Exploratory Factor Analysis

EFA analysis was carried out using Mplus 8 (Muthén & Muthén, 2017). Since the data were binary (0 and 1), the WLSMV estimator and the Geomin rotation method, which takes into account the correlations between factors, were used. To determine the number of factors to retain in the EFA, a set of fit indices were checked, the eigenvalues (Kaiser, 1960) and scree plot were consulted, and the final decision was made based on the considerations of the simple structure (Thurstone, 1947).

As Table 1 shows, the chi-square values for all three models are significant, which is expected given the large sample size in the present study. However, the RMSEA (Root Mean Square Error of Approximation), TLI (Tucker-Lewis Index), CFI (Comparative Fit Index), and SRMR (Standardized Root Mean Square Residual) indices for all models indicate well-fitting models, with the indices for the 3- and 4-factor models being better than those for the 2-factor model. Specifically, RMSEA values less than 0.06 indicate a good fit (Hu & Bentler, 1999), TLI values above 0.95 are considered excellent (Hu & Bentler, 1999), CFI values above 0.95 denote a good fit (Hu & Bentler, 1999), and SRMR values less than 0.08 are deemed acceptable (Hu & Bentler, 1999). Specifically, RMSEA values less than 0.06 indicate a good fit (Hu & Bentler, 1999), TLI values above 0.95 are considered excellent (Hu & Bentler, 1999), CFI values above 0.95 denote a good fit (Hu & Bentler, 1999), and SRMR values less than 0.08 are deemed acceptable (Hu & Bentler, 1999).

There were 12 eigenvalues greater than 1. Kaiser (1960) suggested that all factors with an eigenvalue greater than 1.0 or more can be retained. This method has been criticized since it leads to over-estimation of the number of factors (Fabrigar et al., 1999; Zwick & Velicer, 1982, 1986). However, since the extraction of this many factors was not compatible with the current understanding of language proficiency, Scree Plots of the eigenvalues were

consulted. There was no sharp inflexion point on the plot. One could say there is a sharp bent at the third or fourth nodes, indicating two, or three factors to be retained, respectively.

According to the current understanding of language proficiency, the results from two-, three-, and four-factor solutions were compared. Tabachnik and Fidell (2021) suggested 0.32 as a rule of thumb for the factor loadings, However, Cornrey and Lee (I992) suggest that loadings in excess of 0.71 (50% overlapping variance) are considered excellent, 0.63 (40% overlapping variance) very good, 0.55 (30% overlapping variance) good, 0.45 (20% overlapping variance) fair, and 0.32 (10% overlapping variance) poor.

In the four-factor solution, there were only two items loading onto the second and fourth factor each with loadings of at least 0.32. According to Tabachnick and Fidell (2021), the reliability of a factor, when only two variables load on it, hinges on the correlation pattern between these variables and with other variables in the correlation matrix (R). If these two variables exhibit a high inter-correlation (e.g., $r > 0.70$) and show minimal correlation with other variables, the factor may be deemed reliable. Nonetheless, interpreting factors that are defined by just one or two variables remains problematic, even in the context of exploratory factor analysis. Thus, since the correlations between the two items loading onto each factor were 0.2 and 0.4, the four-factor solution was not considered anymore.

In the three-factor solution, from among the items with factor loadings of at least 0.32, six grammar items, three cloze test items, two vocabulary items, and two reading comprehension items loaded on the first factor. Upon further inspection of the content of the items, it was found that the two vocabulary items measured knowledge of idiomatic expressions and the two reading comprehension items measured understanding the main idea of the text and the ability to infer information, respectively, which require mainly knowledge of reading comprehension than grammar. Thus, it seems that the first factor could be interpreted as grammar knowledge. In the case of the second factor, among items with factor loadings of at least 0.32, there were seven vocabulary, one grammar, and two cloze items. Inspecting the content of the grammar item revealed that it mainly measures knowledge of grammar rather than vocabulary. Finally, as to the third factor, from among the factors with loadings at least 0.32, seven items measured reading comprehension and one item measured vocabulary knowledge. It should be noted that the other items either had loadings of below 0.3 on the corresponding factors or had cross loadings as big as 0.3 on two factors.

Therefore, the two-factor model was deemed the best solution since the factor loadings were relatively strong, there were relatively smaller number of cross-loadings, and the interpretation of the factors was simpler. As Table 2 shows, the first factor was labelled lexico-grammatical knowledge (grammar and vocabulary) and the second factor was interpreted as reading comprehension.

Table 1

Fit Indices of the EFA Models

Model	Parameters	Chi-Square	df	*p*-Value	RMSEA	TLI	CFI	SRMSR
2-factor	111	4302.495	1429	0	0.010	0.979	0.978	0.026
3-factor	165	2729.335	1375	0	0.007	0.990	0.989	0.020
4-factor	218	2298.195	1322	0	0.006	0.993	0.992	0.019

Note. df = Degrees of Freedom; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; CFI = Comparative Fit Index; SRMSR = Standardized Root Mean Square Residual.

Table 2

Factor Structure and Factor Loadings

Item	Lexico-Grammatical Knowledge	Reading Comprehension
3	0.47	
5	0.54	
6	0.54	
7	0.44	
10	0.56	
11	0.47	
12	0.49	
15	0.56	
18	0.69	
21	0.44	
25	0.54	
29	0.36	
30	0.38	
33	0.34	
34	0.50	
37	0.34	
38	0.33	
41		0.36
44		0.37
45		0.34
46		0.50
49		0.44
50		0.35
54		0.56
60		0.65

It should be noted that according to the results of EFA from among the initial 60 items only 25 items had loadings of at least 0.32 on either of the two factors. The majority of the items had either low loadings or cross loadings on the factors and the loadings of few items on the factors did not make sense as evidenced by further inspection of the content of the items. Therefore, only 25 items were kept and the rest of the items were discarded.

6.2 | MGCFA Alignment

As recommended by Asparouhov and Muthén (2023), a configural model was initially examined before executing the alignment procedure. If the proposed configural model does not achieve an acceptable fit, it can be adjusted. Based on the outcomes of EFA, the configural model was delineated as a two-factor model comprising lexico-grammatical knowledge and reading comprehension as the factors. The pattern depicting the relationships between these factors and the items is illustrated in Table 2. Configural invariance assumes uniformity in the number and arrangement of parameters across groups, albeit allowing for variations in parameter values within specified identification constraints. Despite its significant chi-square test result ($\chi^2 = 2035.639$, df = 1096, $p < 0.001$), likely influenced by the large sample sizes in the current study, the configural measurement invariance model exhibited a satisfactory fit (CFI = 0.97, RMSEA = 0.015). I then ran the metric invariance model ($\chi^2 = 11609.421$, df = 1219, $p < 0.001$; CFI = 0.726, RMSEA = 0.041). Since the chi-square values in both models are likely to have been influenced by the large sample size, differences in RMSEA and CFI were considered for model comparison. Rutkowski and Svetina (2017) recommend adjusted criteria for evaluating MI in settings with a large number of groups, particularly when dealing with ordered categorical data. They advise that for metric invariance, changes in CFI greater than or equal to −0.004 and RMSEA less than or equal to 0.050 should be considered, while for scalar invariance, changes in CFI greater than or equal to −0.010 and RMSEA less than or equal to 0.010 are appropriate. These thresholds help determine if differences in fit indices are significant enough to reject the hypothesis of MI across groups. Although the difference in RMSEA between the configural and metric models does not meet the cutoff suggested by Rutkowski and Svetina (2017), the CFI is well below the 0.95 cutoff suggested by Hu and Bentler (1999) for a well-fitting model, and the difference between the CFI of the two models is significantly greater than the 0.004 threshold. Therefore, imposing equality constraints on the factor loadings across the groups resulted in a significant degradation of model fit indices.

In the next step an alignment model with "free alignment" specification was run. There are two possible approaches as to the alignment specification. According to Asparouhov and Muthén (2023) in fixed alignment, the factor means and variances are set to 0 and 1, respectively, in the reference group. Conversely, in free alignment, the factor means in the reference group are estimated while only the factor variances are fixed to 1. Alternatively, under free alignment, both the factor means and variances can be estimated across all groups, but the product of factor variances across groups is constrained to 1, eliminating the need for a reference group. This parameterization is referred to as the product metric. Free alignment estimation is generally feasible when approximate metric invariance is not upheld, indicating significant loading non-invariance across groups. However, if metric invariance holds approximately, the standard errors in free alignment estimation may substantially increase compared to the configural model, potentially compromising inference. Mplus will issue a warning in such instances, prompting the replacement of free alignment with fixed alignment. I received the warning "Standard error comparison indicates that the free alignment model may be poorly identified. Using the fixed alignment option may resolve this problem." Thus, I reran the model with the fixed specification. In this case the group

with the smallest means (as indicated by the results of the free alignment specifications) should be specified as the reference group. In the present study according to the results of the free alignment output, Group 2 was specified as the reference group. According to Asparouhov and Muthén (2023) in fixed alignment, the factor means are set to 0, as indicated.

Table 3 presents the (non)invariance results for the measurement intercepts and factor loadings. The numbers in the table represent groups, and bold and parenthesized numbers in front of each item indicate that the parameter(s) of the corresponding item are non-invariant in the given group. As indicated in Table 3, most items demonstrate measurement invariance for both intercepts and loadings, with only a few exceptions. Regarding the threshold/difficulty measures, Items 11 and 29 are noninvariant in Group 1, and Item 25 is noninvariant in Group 3. However, for the factor loadings, Items 25, 29, and 60 are noninvariant in Group 1. A notable pattern in the MGCFA results is that from among the six non-invariant cases five pertained to Group 1 and only one belonged to Group 3.

Table 3

Approximate Measurement (Non)Invariance for Intercepts and Loadings Across Groups

Item	**Intercept/Threshold/Difficulty**					**Loadings**			
3	1	2	3	4	Lexico-	1	2	3	4
5	1	2	3	4	Grammatical	1	2	3	4
6	1	2	3	4	Knowledge	1	2	3	4
7	1	2	3	4		1	2	3	4
10	1	2	3	4		1	2	3	4
11	**(1)**	2	3	4		1	2	3	4
12	1	2	3	4		1	2	3	4
15	1	2	3	4		1	2	3	4
18	1	2	3	4		1	2	3	4
21	1	2	3	4		1	2	3	4
25	1	2	**(3)**	4		**(1)**	2	3	4
29	**(1)**	2	3	4		**(1)**	2	3	4
30	1	2	3	4		1	2	3	4
33	1	2	3	4		1	2	3	4
34	1	2	3	4		1	2	3	4
37	1	2	3	4		1	2	3	4
38	1	2	3	4		1	2	3	4
41	1	2	3	4	Reading	1	2	3	4
44	1	2	3	4	Comprehension	1	2	3	4
45	1	2	3	4		1	2	3	4
46	1	2	3	4		1	2	3	4
49	1	2	3	4		1	2	3	4
50	1	2	3	4		1	2	3	4
54	1	2	3	4		**(1)**	2	3	4

60	1	2	3	4		1	2	3	4

Table 4 displays the factor means estimated using the MGCFA alignment and MGIRT method. The "Factor mean" columns under "Lexico-grammatical Knowledge" and "Reading Comprehension" headings represent means obtained from the MGCFA alignment and the MGIRT column represents the means obtained from the MGIRT. As one can see for the purpose of MGCFA alignment the mean of the second group has been fixed to zero for modeling considerations. Conventionally, the group with the lowest mean based on the results of the free alignment specification is set as the reference group, hence its mean is set to zero. For convenience, the factor means are arranged from highest to lowest, and groups with factor means that significantly differ at the 5 percent level are indicated.

The results indicate that, according to the MGCFA, for both factors, the means follow the same order, with Group 1 having the highest and Group 2 the lowest means. Specifically, applicants who attended a state university during their Bachelor's program outperformed students from other university types on average, while those who attended Azad University trailed behind students from all other university types. As shown in the table, all means are significantly different from each other in both lexico-grammatical knowledge and reading comprehension.

Interestingly, as the MGIRT column in Table 4 shows, the means generated by the MGIRT mirror the order of the means generated by MGCFA alignment. For modeling purposes, in MGIRT, the first group is set as the reference group, and the mean of the other groups is calculated as the difference between the mean of the respective group and the reference group. Note that in MGIRT, a unidimensional model was estimated and thus there is only one mean per group.

Table 4

Factor Mean Comparison across the Groups

	Lexico-Grammatical Knowledge		**Reading Comprehension**		**MGIRT**
Group	Factor mean	Groups with Significantly Smaller Factor Mean	Factor mean	Groups with Significantly Smaller Factor Mean	
1	1.326	3 4 2	0.925	3 4 2	0.000
3	0.356	4 2	0.299	4 2	-0.758
4	0.137	2	0.144	2	-0.92
2	0.000		0.000		-1.043

To examine to what extent the no-invariant items would impact the performance of the subjects from the different groups, a sensitivity analysis was conducted. The mean scores of the individual with and without the non-invariant items were calculated for the subjects. Then withing each group a paired samples *t*-test and a correlation was run to see the impact. Table 5 shows the results of the paired samples t-tests and Pearson correlations across the

groups. As Table 5 shows, mean performances of the persons within each group before and after removing the non-invariant items are statistically significant whereas the correlation coefficients are at least 0.97. As the effect sizes show, according to Cohen's (1988) guidelines, there is a large impact of removing the non-invariant items on the subjects performances.

Table 5

Sensitivity Analysis

Group	Mean Difference	SD Difference	*t*	df	*p*	Eta squared	Pearson Correlation
1	2.09	1.10	40.08	446	< 0.001	0.859	0.99
2	1.41	0.99	44.34	970	< 0.001	0.836	0.97
3	1.62	0.96	39.01	530	< 0.001	0.878	0.97
4	1.40	0.93	22.11	215	< 0.001	0.867	0.98

Note. SD = Standard Deviation; df = Degrees of Freedom.

6.3 | Multigroup IRT

Prior to running multigroup analysis, single group 2PL IRT was conducted and four items which had negative item discriminations were deleted. All the items had infit and outfit values within the range of 0.8 and 1.2. MGIRT can also be used to evaluate MI which is referred to as differential item functioning (DIF) in IRT. The *TAM* package (Robitzsch et al, 2022) in R (R Core Team, 2023) was employed to estimate the model parameters.

The metrics used to measure the distance between the group-specific ICC and the joint ICC include RMSD and MD (Baghaei & Robitzsch, in press). These metrics evaluate the discrepancy between any given ICC and the joint ICC, are independent of the sample, and range from zero to one. Notably, MD is calculated based on a weighted sum of these differences. An RMSD value of zero denotes a perfect fit for the item and indicates MI. suggest using an RMSD value greater than 0.10 to identify items with DIF. However, in some international large-scale assessments, such as PISA 2015, more relaxed cutoff values for RMSD and MD were set at 0.12 for cognitive scales and 0.30 for non-cognitive scales. Expressed in terms of MGCFA, RMSD is sensitive to scalar invariance (the deviations of both item difficulty and item slope from the joint item ICC), but MD is most sensitive to metric invariance (the deviations of item difficulty parameters).

Table 6 displays the RMSDs and MDs for select items from the test. According to the table, only the RMSD for Item 25 for Group 3 exceeds the cutoff of 0.12, indicating differential functioning for this item in Group 3. The same item in the same group has a MD value of 0.1, which is just below the 0.12 cutoff. Therefore, it can be concluded that Item 25 shows metric invariance, at the level of factor loading, but there is an issue with its scalar invariance. In Table 6, a positive MD value indicates that the item is more difficult for the comparison group compared to the reference group, while a negative MD value indicates that the item is less difficult for the comparison group compared to the reference group. For example, for Item 1, Group 2 finds this item slightly easier (-0.02) compared to the reference group (i.e., Group 1), while Group 3 finds it slightly more difficult (0.03). Group 4 has the same difficulty level as Group 1 (0.00). It should be noted that, in the interest of space, the

table has been truncated to include only five items. However, none of the other items were non-invariant across any of the groups.

Table 6

RMSD and MD Statistics

	RMSD					MD			
Items	Group 1	Group 2	Group 3	Group 4	WRMSD	Group 1	Group 2	Group 3	Group 4
1	0.018	0.025	0.035	0.020	0.025	0.00	-0.02	0.03	0.00
2	0.069	0.043	0.040	0.061	0.052	0.02	-0.01	-0.02	0.02
25	0.060	0.039	0.123	0.050	0.071	0.02	0.01	-0.10	0.03
26	0.068	0.038	0.033	0.046	0.046	0.02	0.00	-0.01	-0.02
27	0.029	0.021	0.026	0.035	0.026	0.00	-0.02	0.02	0.01

6 | Discussion

The present study aimed to examine MI in a general English test that is part of a university entrance exam. The test comprised items assessing vocabulary, grammar, cloze, and reading comprehension. Both the alignment method, an extension of MGCFA, and MGIRT were utilized. Prior to the main analyses, preliminary analyses were conducted to ascertain the quality of the items. Due to the absence of an official report on the validity of the UEE and the quality of its items, an EFA was carried out using Mplus. The results indicated a two-factor solution. Of the 60 items, only 25 loaded on either of the factors, while the remaining items had cross-loadings or very low loadings and were thus removed from further analysis. Following the recommendation by Asparouhov and Muthén (2023), a configural model was established, and the alignment method was subsequently applied.

For the MGIRT analysis, a two-parameter logistic (2-PL) model was initially run, and items showing negative discrimination were removed. The results revealed that the only item exhibiting DIF according to the MGIRT (i.e., Item 25) was also flagged for DIF in the MGCFA results using the alignment method. Items flagged for DIF by RMSD in MGIRT are equivalent to being scalar non-invariant in MGCFA. This was corroborated by the results in Table 3, which showed that Item 25 is non-invariant in terms of both threshold and loading.

However, the MGCFA alignment results also identified additional items showing non-invariance. Specifically, Item 29 demonstrated non-invariance in both loading and threshold parameters, Item 11 showed threshold non-invariance, and Item 54 showed non-invariance in loading. This highlights potential discrepancies between the two methods in identifying DIF/non-invariance. It should be noted that according to both methods, Group 3 was consistently identified as the source of non-invariance, while Group 1 was identified as the source of non-invariance for the additional items found by the MGCFA alignment. Based on the results of both methods, Groups 1 and 3 appear to be the primary sources of non-invariance.

The higher levels of non-invariance observed in Group 1 (state universities) can be attributed to several factors inherent to their unique position within the Iranian higher education system. State universities, being government-funded and highly competitive, have stringent admission criteria and attract top-performing students. Moreover, state universities uniformly enjoy better educational quality, superior instruction, and a much higher student-to-faculty ratio compared to other university types.

In contrast, Groups 2, 3, and 4 comprise universities with lower admission standards, limited faculty support due to low student-to-faculty ratios, and relatively lower quality of instruction. Many students in these universities prioritize commitments other than studying. The combination of these factors, including variance in educational preparedness and resource availability, leads to diverse educational experiences and outcomes, thereby explaining the observed non-invariance in state universities compared to other groups.

Overall, 16% of the items (4 out of 25) exhibited non-invariance, with discrepancies found in both factor loadings (items 25, 29, and 54) and intercepts (items 11, 25, and 29). Notably, items 25 and 29 showed non-invariance in both loading and threshold parameters. According to some psychometric standards, having up to 20-25% of items displaying some form of non-invariance might still be considered acceptable, particularly in large-scale assessments where perfect invariance is rarely achieved (Reise et al., 1993; Raju et al., 2002).

Items with non-invariant intercepts indicate that different groups might have different baseline levels on the latent trait, which can affect the comparison of group means (Cheung & Rensvold, 2002). On the other hand, items with non-invariant loadings suggest that the items might measure the construct differently across groups, affecting the validity of the construct being measured (Byrne et al., 1989).

The sensitivity analysis revealed that, although the paired samples t-test results indicated statistically significant differences in the mean performances of individuals, the correlation coefficients between the scores with and without the non-invariant items were 0.97 and above. This suggests that the ranking of individuals remained consistent. In the context of the present high-stakes, norm-referenced test, where the purpose is to screen applicants for university admission, the ranking of applicants is the crucial factor. Therefore, the differences in mean scores may not be a cause for concern in this case.

The results indicate that statistically significant t-tests between the sum scores of all items and the sum scores after removing non-invariant items suggest that the presence of non-invariant items affects the overall test scores to a detectable degree, potentially impacting criterion-referenced decisions and leading to incorrect classifications of individuals. However, for the UEE in Iran, which is used to rank order test takers and typically does not have a cutoff score, the significant mean differences do not have negative implications for decisions based on the test results. The high Pearson correlations (above 0.97) between the two sets of scores indicate a strong linear relationship, meaning the rank order of individuals remains largely unchanged despite the statistically significant differences. This implies that while significant mean differences raise concerns for criterion-referenced decisions due to potential bias, the test remains effective for norm-referenced

interpretations, as the relative positions and rank ordering of individuals are consistent and unaffected by the presence of non-invariant items.

One possible reason for the discrepancy in the number of items flagged for DIF or non-invariance between MGIRT and MGCFA alignment could be the differing sensitivity of the methods. MGIRT is highly sensitive to item-level discrepancies and focuses on detecting DIF at the granular level of item parameters (Magis & Facon, 2012). In contrast, MGCFA alignment method, may identify broader patterns of non-invariance across groups, which can include both loading and threshold differences (Asparouhov & Muthén, 2014). It should be noted that the data input into the MGCFA were also item-level data.

Additionally, MGIRT's reliance on statistical measures such as RMSD and MD to detect DIF may provide a more stringent criterion, resulting in fewer items being flagged compared to MGCFA alignment (Stark et al., 2006). The MGCFA alignment is designed to optimize the detection of non-invariance without requiring full scalar invariance, which might lead to a higher number of items being identified as non-invariant (Asparouhov & Muthén, 2014).

The inconsistencies between the methods might be attributed to the different numbers of items analyzed. With MGIRT, one item was removed due to its negative discrimination, resulting in 59 items being analyzed. In contrast, with MGCFA alignment, 25 items remained after the EFA analysis and were analyzed. Although no study has directly compared MGCFA alignment with MGIRT, the results of the present study align with those comparing traditional MGCFA and MGIRT. Most studies found some discrepancies in the performance of the two models. Kim and Yoon (2011) found that MGCFA had significantly higher false positive rates, especially as sample size and DIF increased. The larger number of items flagged for non-invariance in the MGCFA alignment might be an artifact of the statistical procedure and the relatively large sample sizes in the groups.

Comparison of factor means reveals that all the means are significantly different from each other in both lexico-grammatical knowledge and reading comprehension with state university students scoring the highest and Azad University students scoring the lowest. The top performance of state university students is anticipated, as admission to these universities is far more competitive than to the other three types of universities. State university students are primarily focused on attending classes and studying their courses, with higher standards and strict requirements to meet these standards. In contrast, students at other types of universities may not prioritize attending classes and studying, as they often have other commitments and must balance work and study.

The results displayed in Table 4 provide evidence of the alignment method's advantage over the traditional MGCFA. The alignment method excels at discerning meaningful differences in factor means, even amidst instances of non-invariance. This is in stark contrast to traditional MGCFA, which requires scalar invariance for an accurate comparison of factor means. In the present study, metric invariance did not hold, preventing the traditional MGCFA from estimating means.

A significant challenge with the traditional MGCFA, especially when dealing with a large number of groups, is its inability to pinpoint non-invariant parameters. The presence

of many large modification indices implies that numerous model modifications are needed to achieve an acceptable fit, which can lead to the identification of incorrect models. Consequently, traditional MGCFA makes it very difficult to properly identify the sources of non-invariance due to the extensive model modifications required. This is a typical outcome when applying a scalar invariance model to multiple groups, making it impossible to compare factor means across these groups.

7 | Limitations and Suggestions for Further Research

The present study has several limitations that should be addressed in future research. It should be noted that MI was examined across only four groups in this study. Future research could benefit from exploring the application of both the MGCFA alignment method and MGIRT in studies involving a larger number of groups. Additionally, while the present study compared MGCFA alignment and MGIRT using real data, future studies could enhance understanding by employing simulated data. Simulated data allow for comparisons under controlled conditions, which can illuminate how these methods perform across varying scenarios.

Moreover, there is a need for further investigation into the discrepancies observed in the number of items flagged for non-invariance by different methods. Exploring factors such as sample size, test length, and the nature of DIF could provide valuable insights into these discrepancies. Understanding these factors can inform improvements in MI detection methods.

Furthermore, it is crucial to explore the practical implications of non-invariance in high-stakes testing environments. This includes assessing how non-invariance impacts the fairness and validity of decisions based on test results. Addressing these issues will contribute to the refinement and application of MI detection methods in educational and psychological assessment contexts.

Funding

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

Conflict of Interest

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

Data Availability Statements

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

How to Cite:

Ravand, H. (2024). Assessing measurement invariance in a university entrance exam: A comparison of multigroup confirmatory factor analysis alignment method vs. multigroup item response theory. *Educational Methods & Practice, 2*:11, 1–15.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547–573. https://doi.org/10.1177/0013164411432166

Ahmadi, A., Darabi Bazvand, A., Sahragard, R., & Razmjoo, A. (2015). Investigating the validity of PhD. entrance exam of ELT in Iran in light of argument-based validity and theory of action. *Journal of Teaching Language Skills, 34*(2), 1–37. https://doi.org/10.22099/jtls.2015.3581

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

Asparouhov, T., & Muthén, B. (2023). Advances in measurement invariance testing: Alignment and beyond. Structural Equation Modeling: A Multidisciplinary Journal, 30(1), 1–20. https://doi.org/10.1080/10705511.2023.2158759

Baghaei, P., & Robitzsch, A. (in press). A tutorial on item response modeling with multiple groups using TAM. *Educational Methods and Practice*.

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the Subject Area: A study of the Iranian National University Entrance Exam. *Journal of Teaching Language Skills, 29*(3), 1–26. https://doi.org/10.22099/jtls.2012.413

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

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd Ed.). Guilford Press.

Buchholz, J., & Hartig, J. (2020). Measurement invariance testing in questionnaires: A comparison of three multigroup CFA and IRT-based approaches. *Psychological Test and Assessment Modeling*, 62(1), 29–53. URL: https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/03_Buchholz.pdf

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Lawrence Erlbaum Associates.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd Ed.). Lawrence Erlbaum Associates, Inc.

Davidov, E., Schmidt, P., Billiet, J., & Meuleman, B. (2014). *Cross-cultural analysis: Methods and applications*. Routledge.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kim, S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

Luong, G., & Flake, J. K. (2023). The role of measurement invariance in psychological research: Applications, advances, and challenges. *Psychological Methods, 28*(2), 157–176. https://doi.org/10.1037/met0000442

Magis, D., & Facon, B. (2012). Angoff's delta plot and the Likelihood Ratio Test for differential item functioning detection: Two concurrent approaches. *Educational and Psychological Measurement, 72*(6), 930–950. https://doi.org/10.1111/j.2044-8317.2011.02025.x

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361–388. https://doi.org/10.1177/1094428104268027

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11 Suppl 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93–115. https://doi.org/10.1037/1082-989X.9.1.93

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th Ed.). Muthén & Muthén.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality, 40*(4), 411–423. https://doi.org/10.1016/j.jrp.2005.02.002

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *Sage Open*, *5*(2). https://doi.org/10.1177/2158244015585607

Ravand, H., Baghaei, P., & Doebler, P. (2020). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology*, *10*: 2930. https://doi.org/10.3389/fpsyg.2019.02930

Ravand, H., Faryabi, F., Rohani, G. (2018). On the factor structure invariance of the general English section of university entrance examination for Ph.D. applicants into the English programs in Iran. *Journal of Teaching Language Skills 36(4)*, 141-170. doi*:* 10.22099/jtls.2018.27029.2372

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the Master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testing*, *6*(1), 1–23. URL: https://www.ijlt.ir/article_114414.html

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules*. R package version 4.0-0. https://CRAN.R-project.org/package=TAM

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fitmeasure performance. *Applied Measurement in Education*, *30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Tabachnick, B. G., & Fidell, L. S. (2021). *Using multivariate statistics* (7th Ed.). Pearson.

Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. https://doi.org/10.1177/109442810031002

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432