mathematics

MDPI

Article

Identifying Bias in Social and Health Research: Measurement Invariance and Latent Mean Differences Using the Alignment Approach

Ioannis Tsaousis [1,*] and Fathima M. Jaffari [2]

[1] Department of Psychology, National and Kapodistrian University of Athens, 15784 Athens, Greece
[2] Education & Training Evaluation Commission (ETEC), Riyadh 12395, Saudi Arabia; f.jaffari@etec.gov.sa
* Correspondence: ioantsaousis@psych.uoa.gr; Tel.: +30-2107277533

Abstract: When comparison among groups is of major importance, it is necessary to ensure that the measuring tool exhibits measurement invariance. This means that it measures the same construct in the same way for all groups. In the opposite case, the test results in measurement error and bias toward a particular group of respondents. In this study, a new approach to examine measurement invariance was applied, which was appropriately designed when a large number of group comparisons are involved: the alignment approach. We used this approach to examine whether the factor structure of a cognitive ability test exhibited measurement invariance across the 26 universities of the Kingdom of Saudi Arabia. The results indicated that the P-GAT subscales were invariant across the 26 universities. Moreover, the aligned factor mean values were estimated, and factor mean comparisons of every group's mean with all the other group means were conducted. The findings from this study showed that the alignment procedure is a valuable method to assess measurement invariance and latent mean differences when a large number of groups are involved. This technique provides an unbiased statistical estimation of group means, with significance tests between group pairs that adjust for sampling errors and missing data.

Keywords: measurement invariance; bias; alignment method; configural invariance; latent means comparisons

MSC: 97C40

Citation: Tsaousis, I.; Jaffari, F.M. Identifying Bias in Social and Health Research: Measurement Invariance and Latent Mean Differences Using the Alignment Approach. *Mathematics* **2023**, *11*, 4007. https://doi.org/10.3390/math11184007

Academic Editors: Carmen Patino-Alonso and Dan Vilenchik

Received: 10 July 2023
Revised: 27 August 2023
Accepted: 19 September 2023
Published: 21 September 2023

1. Introduction

The purpose of any psychometric scale, test, or inventory is to produce a valid score that reflects the degree to which a person possesses a given attribute [1]. For scores to be useful, they should reveal differences among test takers in their attribute if they indeed vary on the attribute. Because an item is the building block of any instrument, the quality of items needs to be checked. That is, items of any assessment should function similarly for different groups of individuals (i.e., ethnicity, gender, school type, universities, etc.), assuming that the grouping condition is irrelevant to the attribute being measured. In the event that this is not the case, the item results in measurement error and has a bias toward a particular group of respondents [2].

To investigate whether or not an item is biased in relation to a grouping condition, the individuals belonging to varying groups should be matched first on the attribute of interest. Thus, a biased item produces differences in the probability of a correct response despite equivalence in the attribute among the groups. The methods for investigating item bias are rooted in varying psychometric modeling theories, including classical test theory (CTT) and item response theory (IRT), in both of which the primary focus is on item-by-item analysis. One of the most prominent approaches to examining whether an item (or a scale) produces bias toward a group of participants is to test for measurement

invariance. Measurement invariance implies that all items of an assessment function are the same matter across the grouping of individuals. If an item violates measurement invariance, systematic inaccuracy in measurement is introduced. Thus, the item is labeled as biased in relation to the grouping of individuals [3].

1.1. Measurement Invariance: Literature Review

Measurement invariance is a key idea in many scientific fields, such as psychology, education, and sociology. When comparison among groups or populations is of major importance in a study, it is necessary to ensure that the measuring tool measures the same thing in the same way for all groups. If measurement invariance exists, this means that the construct under investigation is the same across the compared groups. On the other hand, a lack of measurement invariance means that these comparisons are biased since respondents in one group provide systematically different responses than respondents from another group, although they share the same level of the latent trait. In such a case, the obtained differences among the groups are not meaningful since they are the product of bias [4].

The most frequently used method to test measurement invariance across groups is the multi-group confirmatory factor analysis (MGCFA) method [5,6]. Within this approach, various (nested) models are tested by constraining different parameters (e.g., factor loadings, intercepts, residual variances, etc.), and the subsequent loss of fit is compared. The most frequently examined models are the configural model (i.e., whether the scale's structure is conceptualized similarly across groups), the metric or weak model (i.e., whether each item contributes to the latent construct in the same manner and the same degree across groups), and the scalar or strong invariance (i.e., whether all groups exhibit the same mean level in the latent construct) [7]. It should be noted that although additional constraints on certain parameters (e.g., item residuals, factor variances, and covariances) could be imposed, these additional forms of invariance are useful only when specific hypotheses regarding the relationship among the dimensions of the construct being measured may be of interest [8,9].

Finally, to be able to compare group means at the latent level across different groups, scalar invariance must be supported since only then could one be confident that any statistically significant differences in group means are not due to idiosyncratic scale characteristics (e.g., poor-quality items, low reliability, vague factor structure, etc.) but reflect true mean differences across groups [8–10]. On the other hand, if scalar invariance fails, multigroup equivalence cannot be assumed, and as a result, no comparisons at the mean level can be undertaken.

1.2. The Alignment Method

An important limitation of the MGCFA approach as a method of examining measurement invariance, especially with large-scale studies, is that it is extremely difficult to satisfy the assumption of scalar invariance when a large number of group comparisons are involved [11]. To overcome this problem and make group comparisons feasible, a new method for comparing latent variables across a large number of groups was introduced without requiring measurement invariance [12]. This is called the alignment method and finds great application, especially in cross-cultural research where large-scale and widely diverse cultural groups are examined, e.g., [13,14]. Interestingly, previous research has shown that measurement invariance with the alignment method can be possible even when the number of groups is as large as 92 [15].

Unlike traditional measurement invariance testing, where both factor loadings and intercepts are constrained, the alignment approach typically allows the intercepts to vary across groups. This recognizes that variations in response styles or item biases may exist between groups. In practice, the alignment method involves estimating a configural model that assumes the same overall factor structure for all groups and aligning this model to the specific factor structures of each group. To achieve this, the alignment technique uses an alignment optimization function to look for invariant item loadings and intercepts, which,

in turn, look for latent means and standard deviations. (e.g., a quadratic loss function). With this method, all groups may be compared at once, and latent means can be aligned and compared even if some loadings and intercepts are non-invariant. The function minimizes some non-invariances while leaving some of them large; its logic is similar to factor rotation. Once the groups have been aligned, the factor loadings can be compared directly to assess measurement invariance. Moreover, the resulting aligned model can be used to compare the factor means and variances across groups.

Mathematically, the goal of the alignment method is to align the factor loading matrices (e.g., λ_1 and λ_2) so that they are comparable across groups. To achieve this, the method attempts to align the groups on a common factor space using orthogonal Procrustes rotation. This involves finding a matrix that minimizes the difference between the factor loadings of the items in the different groups while preserving the overall structure of the factor space. This can be expressed as:

$$\left|\left|\lambda_1 - \lambda_2' \mathrm{R}\right|\right|^2, \tag{1}$$

subject to the constraint that R′ R = I, where I is the identity matrix.

The degree of non-invariance in pattern coefficients between each pair of groups is estimated using a *loss function*, and Bayesian estimation is used to re-weight the estimates in the configural invariance model to minimize the non-invariance in the aligned model. Equation (1) is part of the loss function and is used to measure the degree of non-invariance between two groups. Specifically, Equation (1) measures the squared difference between the factor loading matrix for group 1 (λ_1) and the factor loading matrix for group 2 (λ_2) multiplied by the rotation matrix (R) that aligns the factor structures. The alignment method iteratively adjusts the rotation matrix to minimize the loss function and align the factor structures across all groups. Once we have found the Procrustes rotation matrix, R, we can apply it to the factor loading matrix, λ_2, to obtain the aligned factor loading matrix, $\lambda_2{}'$:

$$\lambda_2{}' = \lambda_2 \, \mathrm{R} \tag{2}$$

We can then compare the factor loadings of the items between the two groups by testing whether the aligned factor loading matrix, $\lambda_2{}'$, is equal to the factor loading matrix, λ_1, or whether it differs by a constant factor. If the factor loading matrices are equal up to a constant factor, then the measurement instrument exhibits configural invariance. If the factor loadings are equal in magnitude up to a constant factor, then the measurement instrument exhibits metric invariance. If the factor loadings are equal in magnitude and intercept up to a constant factor, then the measurement instrument exhibits scalar invariance.

In its simplest application (i.e., a one-factor model), the alignment method can be mathematically illustrated as follows [16]:

$$y_{ipg} = v_{pg} + \lambda_{pg}\eta_{ig} + \varepsilon_{ipg} \tag{3}$$

where y_{ipg} is the pth observed variable for participant i in group g, v_{pg} represents the intercept, λ_{pg} is the factor loading for the pth observed variable in group g, $\varepsilon_{\mathrm{ipg}}$~N $(0, \theta_{pg})$ represents the error term for individual i in group g, and η_{ig} is the factor for individual i in group g. The alignment method estimates all the parameters, including v_{pg}, λ_{pg}, a_g, ψ_g, and θ_{pg} as group-specific parameters. This means the method estimates the factor mean and variance separately for each group without assuming measurement invariance. In other words, the alignment method allows each group to have its own unique factor structure rather than assuming all groups.

The initial stage of the alignment method involves estimating the configural model, which assumes that all groups have the same overall structure. The configural model sets certain parameters in each group, g, to specific values ($\alpha_g = 1, \psi_g = 1$) and estimates group-specific parameters for all other parameters. The configural factor model is represented as follows:

$$\eta_{ig} \quad = \alpha_g \quad + \sqrt{\psi_g}\eta_{ig,\ configural} \tag{4}$$

Since the aligned model has the same fit as the configural model, certain relationships must be held between these parameters.

$$v(y_{ipg}) = \lambda_{pg}^2\ \psi_g + \theta_{pg} = \lambda_{pg,\ configural}^2 \ + \theta_{pg,\ configural} \tag{5}$$

$$\mathrm{E}(y_{ipg}) = V_{pg}\ + \lambda_{pg}\alpha_g = V_{pg,\ configural} \tag{6}$$

where $v(y_{ipg})$ and $\mathrm{E}(y_{ipg})$ are the y_{ipg} model estimated variance and mean.

By imposing equality constraints, $\theta_{pg} = \theta_{pg,\ configural}$, Equation (5) will be:

$$\lambda_{pg}^2\ \psi_g = \lambda_{pg,\ configural}^2$$

$$\lambda_{pg}^2\ = \frac{\lambda_{pg,\ configural}^2}{\psi_g} \tag{7}$$

$$\lambda_{pg}\ = \frac{\lambda_{pg,\ configural}}{\sqrt{\psi_g}} \tag{8}$$

Putting the value of λ_{pg} obtained from Equation (7) into Equation (6),

$$V_{pg}\ = V_{pg,\ configural}\ - \alpha_g \frac{\lambda_{pg,\ configural}}{\sqrt{\psi_g}} \tag{9}$$

To make this more precise, the alignment function, *f*, will be minimized in terms of α_g and ψ_g. This function takes into account all sources of non-invariance in the measurements.

$$F = \sum_p \sum_{g1<g2} W_{g1,g2} f\left(\lambda_{pg1} - \lambda_{pg2}\right) + \sum_p \sum_{g1<g2} W_{g1,g2} f\left(v_{pg1} - v_{pg2}\right) \tag{10}$$

$$\mathrm{W_{g1,g2}} = \sqrt{(N_1\ N_2)} \tag{11}$$

where W is the factor weight, N is the sample size of the group, and f is a component loss function:

$$f = \sqrt[4]{(x^2 + \in)} \tag{12}$$

where $\in$ represents a small value, typically around 0.0001.

The alignment function, *f*, is designed to be approximately equal to the absolute value of x. To ensure that the function has a continuous first derivative, which makes optimization easier and more stable, we use a positive value for $\in$.

Previous studies have introduced the alignment method as a powerful tool for analyzing measurement invariance across multiple groups, especially when the number of compared groups is large, e.g., [11,12,15]. It seems a sophisticated method for examining measurement invariance in latent variable models because it provides researchers with a flexible way to align specific components of the measurement model while allowing for intercept variability, ensuring meaningful and valid comparisons across different groups or time points. The alignment approach allows for a nuanced examination of measurement invariance by selectively aligning parameters.

From this perspective, this study aimed to demonstrate the empirical usefulness of this method when a large number of group comparisons are necessary. For that, we examined the measurement invariance of an instrument test, measuring general cognitive ability across 36 universities in the Kingdom of Saudi Arabia. The findings from this study will help researchers evaluate the psychometric robustness of the method in examining measurement invariance across multiple groups using real-life data. Additionally, it will help them determine whether this strategy is useful in actual practice when meaningful comparisons between groups (e.g., universities) are important. For instance, the results of this study

may be a strong warrant for a better understanding of students' academic performance and score differences among universities, helping policy educators and national governmental agencies explore the possible factors that might have caused these gaps.

2. Methodology

2.1. Participants and Procedure

The sample consisted of 9849 graduate students from all universities across the Kingdom. Of them, 4682 (47.5%) were females, and 5167 (52.5%) were males. The mean age of the participants was 25.78 (S.D. = 5.58). With regards to the region of residence, 169 (1.7%) came from the Albahah region, 163 (1.7%) from the Aljawf region, 812 (8.2%) from the Almadinah region, 491 (5.0%) from the Alqasim region, 2515 (25.5%) from the Alriyadh region, 1000 (10.2%) from the Asir region, 532 (5.4%) from the Eastern Province, 208 (2.1%) from the Hail region, 286 (2.9%) from the Jizan region, 3097 (31.5%) from the Makkah region, 92 (0.9%) from the Najran region, 122 (1.2%) from the Northern Borders region, and 359 (3.6%) from the Tabuk region. Three participants (0.01%) did not report their region of residence. Participants came from all 35 universities in the Kingdom. However, only universities with more than 59 participants were retained in this study to ensure adequate statistical power [11]. Therefore, nine universities with less than sixty participants (ranging from three to twenty-seven) were removed from the analysis. Table 1 presents the university code and its corresponding sample size.

Table 1. Reference code and sample size by university.

Code	University	N
1	Albaha University	223
3	Arab Open University	150
4	Dammam University	90
8	Hail University	182
9	Imam Mohammed Bin Saud Islamic University	1088
10	Islamic University	191
11	Jazan University	263
12	Jeddah University	202
13	Jouf University	180
14	King Abdulaziz University	1142
16	King Faisal University	237
17	King Khalid University	771
19	King Saud University	523
20	Majmaah University	173
21	Najran University	77
22	Northern Border University	102
24	Prince Sattam Bin Abdulaziz University	198
26	Prince Nourah Bin Abdulrahman University	295
27	Qassim University	496
29	Shaqra University	262
30	Tabouk University	358
31	Taiba University	574
32	Taif University	666
33	Umm Al-Qura University	1155
34	University of Bisha	192
35	University of Hafr Batin	59
Total		9849

2.2. Measure

The Post-Graduate General Aptitude Test (PGAT; Education & Training Evaluation Commission): The P-GAT is a psychometric tool measuring graduate students' analytical and deductive skills. It consists of one hundred and four dichotomous items covering ten different content areas organized into two broader cognitive domains: (a) *verbal* (*linguistic*)

and (b) quantitative (*numerical*). The verbal domain comprises four sub-scales (i.e., *analogy*, *sentence completion*, *context errors*, *and reading comprehension*). The quantitative domain comprises six sub-scales (i.e., *arithmetic*, *analysis*, *comparison*, *critical thinking*, *spatial*, and *logic*). All PGAT items are in a multiple-choice format and scored as either correct (1) or wrong (0). The test has a 2.5 h duration and is presented in Arabic.

2.3. Data Analysis Strategy

First, the fit of the conceptual model for P-GAT (i.e., a higher-order two-factor model) was examined via conventional CFA. To assess the fit, the following indices were used: the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). In general, a CFI and TLI value above 0.90 indicates an acceptable fit (with values >0.95 being ideal). Further, RMSEA and SRMR values up to 0.06 indicate a reasonable fit to the data, while values up to 0.05 indicate an excellent fit [17]. Next, the measurement invariance across the 26 universities was examined following the conventional MGCFA criteria of configural, metric, and scalar invariance [3,5,6]. For evaluating invariance among different consecutive models, loglikelihood and chi-square statistics were used [6].

Next, the alignment approach was applied to examine the measurement invariance. From a technical perspective, as previously stated, the alignment method allows for a pattern of *approximate measurement invariance*, in contrast to MGCFA, where full or partial measurement invariance (particularly at the scalar level) is a required criterion when group means are attempted to be compared. Particularly, the alignment method focuses only on the configural model and then automates the closeness of the factor loading estimates in establishing the most optimal measurement invariance pattern [12].

First, the factor loadings and intercepts of a configural invariance CFA model were estimated, and the alignment process used these values as inputs. Then, the factor means and variances were freely estimated across the different groups, with the objective of choosing corresponding parameters that minimized the total amount of measurement non-invariance. During this process, the alignment estimation model identified for each measurement parameter (e.g., factor loading, intercept, etc.) the largest invariant set of groups for which the specific parameter was not statistically significant from the mean value for that parameter across all groups. At the final stage, this minimization process ended up with many approximately non-invariant parameters and very few large non-invariant measurement parameters. The alignment approach to these analyses began with the configural model and, consistent with the CFA method, was based on robust ML estimation (MLR).

The R^2 (as an indicator of effect size) and the average correlation of aligned item characteristics among groups was also estimated to evaluate the degree of the approximate invariance. The R^2 values represent the variation in these parameters across groups that can be accounted for by variation in the factor means and variances across groups. If the R^2 for factor loadings is close to 1 and the average correlation of the aligned factor loadings is high, then all the aligned item factor loadings are approximately invariant (metric invariance). If the R^2 for the intercepts is close to 1 and the average correlation of the aligned intercepts is high, then all the aligned item intercepts are approximately invariant (scalar invariance).

Additionally, the fit information contribution (FIC), an index that represents the contribution that each parameter makes to the final simplicity function, and the total contribution function (TCF), an index that represents the total contribution of each variable to the fitting model (taking into account together the factor loadings and intercepts) was estimated. In both indices, the higher the value, the higher the contribution [11].

In terms of latent mean comparisons, the alignment method, via the optimization process, simplifies the invariance examination by taking the non-invariance of all factor loadings and intercepts parameters into account in the process of means estimation, thereby yielding mean values that are more trustworthy than those calculated without this strategy. This optimization process enables the estimation of trustworthy means despite the presence

of some measurement non-invariance [12]. To test for possible differences across the universities at the latent mean level, factor mean comparisons of every group's mean with all other group means were examined.

It should also be noted that the fixed alignment optimization method was applied, in which the factor means and variances in the reference group were fixed to 0 and 1, respectively. We preferred this approach instead of the alternative of the free alignment optimization method (for more information, see [13]) following the suggestion from the Mplus software after a warning message that the free alignment method (that we first attempted) was poorly identified. All analyses were conducted with the Mplus (8.5) software [18]. In the Appendix A, the corresponding code is provided.

3. Results

3.1. Descriptive Statistics and Normality

Descriptive statistics, normality indices, and inter-correlations among the study variables are presented in Table 2. No violation of the univariate normality of all the variables was found (values < 2.0).

Table 2. Descriptive statistics and inter-correlations among the variables of the study (N = 9849).

	Subscales	1	2	3	4	5	6	7	8	9	10
1	Analogy	-									
2	Sentence Completion	0.48	-								
3	Context Errors	0.54	0.48	-							
4	Reading Comprehension	0.55	0.49	0.55	-						
5	Arithmetic	0.51	0.36	0.42	0.46	-					
6	Analysis	0.41	0.30	0.35	0.38	0.57	-				
7	Comparison	0.42	0.30	0.35	0.38	0.53	0.44	-			
8	Critical Thinking	0.42	0.39	0.39	0.41	0.35	0.31	0.30	-		
9	Spatial	0.44	0.28	0.36	0.38	0.54	0.44	0.44	0.29	-	
10	Logic	0.46	0.33	0.42	0.45	0.56	0.45	0.42	0.36	0.49	-
Mean		90.77	30.99	60.14	110.04	60.75	30.37	20.95	50.84	40.86	40.37
SD		30.15	10.47	20.03	30.24	20.66	10.53	10.53	20.04	200.3	20.18
Skewness		−0.64	−0.51	−0.44	−0.29	−0.06	−0.09	0.18	−0.04	0.001	0.23
Kurtosis		−0.18	−0.41	−0.58	−0.38	−0.74	−0.73	−0.66	−0.30	−0.65	−0.59

Note: All correlation coefficients were significant at $p < 0.001$ level.

In terms of multivariate normality, the obtained results are presented in Table 3. As can be seen, both tests revealed that the data were not multivariate normal. For that, the maximum likelihood with robust standard errors (MLR) estimation method was used since it is robust to non-normality and non-independence of observations. Finally, we used the Mahalanobis distance statistic (D^2) to identify possible irrelevant response patterns [19]. The results revealed no outliers. No missing data were observed.

Table 3. Test for normality using Mardia's test.

Test	Skewness	Kurtosis
Test Statistic	1.51	118.11
p-value	0.001	0.001

3.2. P-GAT Measurement Model

First, the P-GAT measurement model was tested via CFA. The CFA model of the P-GAT structure is shown schematically in Figure 1.

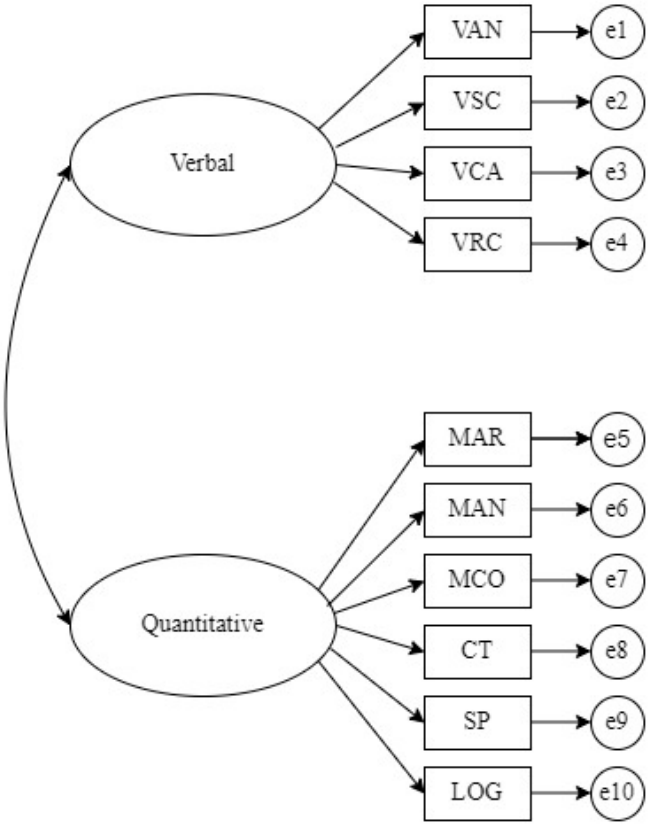

Figure 1. The conceptual model of the P-GAT method. *Note.* VAN = analogy, VSC = sentence completion, VCA = context errors, VRC = reading comprehension, MAR = arithmetic, MAN = analysis, MCO = comparisons, CT = critical thinking, SP=spatial thinking, LOG = logic.

As can be seen, the P-GAT is an instrument measuring ten analytical and deductive skills organized into two broader cognitive domains: (a) verbal (linguistic) and (b) quantitative (numerical). This is a higher-order conceptualization, where the ten cognitive variables (subscale scores) are treated as observed variables. It was hypothesized that this model would be a robust conceptualization for every university group. Previous findings have shown that this theoretical conceptualization provides an acceptable fit [20,21]. The results from the analysis revealed that the P-GAT conceptual model exhibited an excellent fit (χ^2 = 1149.73 (34); CFI = 0.969, TLI = 0.959, RMSEA (90% C.I.s) = 0.058 (0.055–0.061), SRMR = 0.035). Therefore, we examined the test's measurement invariance across the universities.

3.3. P-GAT Measurement Invariance Results

Before applying the alignment method, MGCFA, the most commonly used approach to measuring invariance [6], was applied to examine whether configural, metric, and scalar invariance was supported. Table 4 shows the results of the analysis. As shown, it is clear that both the metric and the scalar invariance model were rejected, a finding that is not surprising due to the large number of contrasted groups (i.e., 26) and the large sample size (i.e., overpower).

Then, the alignment method was applied. The major advantage of this method is that metric and scalar invariance are not required. Only the configural model is necessary to be supported to compare group means meaningfully. A 26-group alignment analysis of the P-GAT subscales was performed across the 26 universities in Saudi Arabia. First, the results of the approximate measurement invariance (non-invariance) analysis for the intercept in each P-GAT subscale are shown in Table 5. The numbers in the parentheses represent the different universities and designate which item parameters (i.e., item threshold) were

non-invariant in which groups. Universities that had a measurement parameter that was considered to be significantly non-invariant are shown in boldface within parentheses.

Table 4. Measurement invariance results (N = 9849).

Model	**No Par**	**Loglikelihood**	
Configural	806	−194,124.095	
Metric	606	−194,247.290	
Scalar	406	−194,476.274	
Models Compared	χ^2	*df*	*p*
Metric vs. Configural	262.051	200	0.002
Scalar vs. Configural	725.901	400	0.001
Scalar vs. Metric	457.782	200	0.001

Table 5. Invariance results for aligned intercept parameters for P-GAT subscales (VAN–LOG).

Scales	**University Identification Number**												
VAN	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VSC	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	**(17)**	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VCA	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VRC	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MAR	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MAN	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MCO	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
CT	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	**(14)**	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
LOG	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)

Note. VAN = analogy, VSC = sentence completion, VCA = context errors, VRC = reading comprehension, MAR = arithmetic, MAN = analysis, MCO = comparisons, CT = critical thinking, SP = spatial thinking, LOG = logic.

The results indicate that only in two groups (universities) the item intercepts were non-invariant. Specifically, the sentence completion subscale was significantly non-invariant at the King Khalid University (i.e., 17), and the critical thinking subscale was significantly non-invariant at the King Abdulaziz University (i.e., 14). Next, the results of the approximate measurement invariance (non-invariance) analysis for the factor loadings in each P-GAT subscale are shown in Table 6. There were no non-invariant factor loadings at any P-GAT subscale across the universities.

3.4. P-GAT Latent Mean Difference Results

Given that the configural invariance assumption was satisfied, the next step was to test for possible differences across the universities at the latent mean level. The aligned factor mean values and factor mean comparisons of every group's mean with all the other group means for the verbal and quantitative domains are shown in Tables 7 and 8. For ease of presentation, the factor means are arranged from high to low, and groups with significantly different factor means at the 5% level are presented in the last columns of the tables.

Table 6. Invariance results for aligned factor loading parameters for P-GAT subscales (VAN–LOG).

Scales	University Identification Number												
VAN	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VSC	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VCA	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
VRC	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MAR	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MAN	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
MCO	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
CT	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)
LOG	(1)	(3)	(4)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(16)	(17)	(19)
	(20)	(21)	(22)	(24)	(26)	(27)	(29)	(30)	(31)	(32)	(33)	(34)	(35)

Note. VAN = analogy, VSC = sentence completion, VCA = context errors, VRC = reading comprehension, MAR = arithmetic, MAN = analysis, MCO = comparisons, CT = critical thinking, SP = spatial thinking, LOG = logic.

Table 7. Factor mean comparisons among the 26 universities at the 5% significance level for the verbal domain.

Ranking	University Code	Factor Mean	Universities with Significantly Smaller Factor Mean
1	19	0.454	(26) (27) (14) (17) (31) (1) (21) (16) (35) (12) (9) (8) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
2	4	0.333	(17) (31) (1) (21) (16) (35) (12) (9) (8) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
3	26	0.279	(17) (31) (1) (21) (16) (35) (12) (9) (8) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
4	27	0.214	(17) (31) (1) (16) (35) (12) (9) (8) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
5	14	0.174	(17) (31) (1) (16) (12) (9) (8) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
6	17	0.064	(16) (12) (9) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
7	31	0.026	(9) (32) (30) (3) (33) (20) (24) (34) (10) (11) (22) (13) (29)
8	1	0.000	(33) (20) (24) (34) (10) (11) (22) (13) (29)
9	21	−0.078	(13) (29)
10	16	−0.112	(22) (13) (29)
11	35	−0.114	(29)
12	12	−0.117	(13) (29)
13	9	−0.119	(33) (11) (22) (13) (29)
14	8	−0.120	(13) (29)
15	32	−0.161	(13) (29)
16	30	−0.173	(13) (29)
17	3	−0.191	(29)
18	33	−0.233	(13) (29)
19	20	−0.238	(29)
20	24	−0.272	(29)
21	34	−0.282	(29)
22	10	−0.306	
23	11	−0.308	
24	22	−0.401	
25	13	−0.449	
26	29	−0.521	

Note. The numbers in parentheses represent each university's identification number (See Table 1).

Table 8. Factor mean comparisons among the 26 universities at the 5% significance level for the quantitative domain.

Ranking	University Code	Factor Mean	Universities with Significantly Smaller Factor Mean
1	19	0.578	(14) (27) (26) (35) (17) (31) (3) (12) (21) (8) (1) (24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
2	14	0.340	(17) (31) (3) (12) (21) (8) (1) (24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
3	4	0.324	(1) (24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
4	27	0.244	(31) (12) (8) (1) (24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
5	26	0.197	(1) (24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
6	35	0.179	(29) (13) (22)
7	17	0.146	(24) (32) (16) (9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
8	31	0.071	(9) (20) (30) (33) (11) (10) (34) (29) (13) (22)
9	3	0.070	(29) (13) (22)
10	12	0.049	(33) (29) (13) (22)
11	21	0.028	(29) (13) (22)
12	8	0.012	(29) (13) (22)
13	1	0.000	(29) (13) (22)
14	24	−0.026	(29) (13) (22)
15	32	−0.043	(29) (13) (22)
16	16	−0.058	(29) (13) (22)
17	9	−0.075	(29) (13) (22)
18	20	−0.093	(29)
19	30	−0.102	(29)
20	33	−0.123	(29)
21	11	−0.152	
22	10	−0.157	
23	34	−0.168	
24	29	−0.306	
25	13	−0.312	
26	22	−0.342	

Note. The numbers in parentheses represent each university's identification number (See Table 1).

Regarding the verbal domain, the results showed that King Saud University (code no. 19) had the highest mean score among all the universities (0.454). More interestingly, almost all the other universities (except Dammam University—code no. 4) had P-GAT factor means that were significantly lower than King Saud's factor mean. On the other hand, the Islamic University, the Jazan University, the Jouf University, the Northern Border University, and the Shaqra University had the lowest P-GAT mean scores among all the universities. Moreover, their factor mean scores were not significantly higher than any other university's factor means.

In terms of the quantitative domain, the results showed again that King Saud University (code no. 19) had the highest mean score among all the universities (0.578). Moreover, its factor mean was significantly higher than almost all the other universities (except Dammam University—code no. 4). On the other hand, the Jazan University, the Islamic University, the University of Bisha, the Shaqra University, the Jouf University, and the Northern Border University exhibited the lowest mean scores. Moreover, their factor mean scores were not significantly higher than any other university's factor means.

Next, we examined the fit of the solution provided by the alignment analysis. Particularly, the alignment method provides some fitting statistics of both the factor loading and intercept for each observed variable to evaluate the robustness of the fitting function. The results are shown in Table 9. First, the *fit information contribution* (FIC) provides information values separately for the factor loading and the intercept of each observed variable, representing each parameter's contribution to the final solution. As can be seen, the variable VCA (contextual errors) for the factor loading parameters and the variable VRC (reading comprehension) for the intercept parameters contributed the least to the fitting function (−114.00 and −112.20, respectively). Similarly, the *total contribution function* (TC)

represents the total contribution of each variable to the fitting model (taking into account together the factor loadings and intercepts). Again, the results showed that the variable VRC (reading comprehension) contributed the least to the fitting function (−229.43). The above results can be interpreted as an indication that these variables exhibited the least amount of non-invariance.

Table 9. Alignment fit statistics for the P-GAT across universities.

	Factor Loadings				**Intercepts**		**Loadings + Intercepts**
	Verbal		**Quantitative**				
	FIC	R^2	**FIC**	R^2	**FIC**	R^2	**TC**
VAN	−120.45	0.45			−121.13	0.91	−241.58
VSC	−114.07	0.57			−130.10	0.84	−244.08
VCA	−114.00	0.52			−119.02	0.85	−233.02
VRC	−117.24	0.45			−112.20	0.93	−229.43
MAR			−113.08	0.42	−119.17	0.72	−232.25
MAN			−125.33	0.36	−131.26	0.73	−256.59
MCO			−121.73	0.50	−127.04	0.88	−248.77
CT			−134.24	0.00	−149.18	0.75	−283.43
SP			−122.11	0.27	−118.40	0.88	−240.51
LOG			−112.62	0.29	−122.71	0.81	−235.33

Note: VAN = analogy, VSC = sentence completion, VCA = context errors, VRC = reading comprehension, MAR = arithmetic, MAN = analysis, MCO = comparisons, CT = critical thinking, SP = spatial thinking, LOG = logic, FIC = fit information contribution, TC = total contribution.

Finally, the R^2 *value* indicates the degree of invariance of a given parameter. A value close to 1 designates a high degree of invariance, while a value close to 0 designates a low degree of invariance [12]. As seen in Table 4, all variables exhibited a high degree of invariance in terms of the intercept parameter. This finding suggests that the latent means could be meaningfully compared across the universities. In terms of the factor loading parameter, however, this statistic showed that several variables exhibited a low degree of invariance (i.e., CT (critical thinking), SP (spatial thinking), and LOG (logic)). For the scope of this study, however, the intercept invariance was considered more important since it was the prerequisite assumption before we conducted the latent mean comparisons across universities. Therefore, from this perspective, the results from this analysis are considered acceptable.

4. Discussion

The scope of this study was twofold: The first was to introduce a relatively new method in examining measurement invariance, especially with large-scale studies where a large number of group comparisons are involved. In those situations, when traditional approaches for examining measurement invariance are applied (e.g., MGCFA), scalar invariance is rarely satisfied, and as a result, latent mean differences across groups cannot be examined. To overcome this problem and make group comparisons feasible, the alignment approach was introduced [12], where only configural invariance was necessary to be satisfied. The second was to evaluate the psychometric robustness of this approach using real-life data. Particularly, we applied this approach to examine whether the factor structure of a cognitive ability test (PGAT) exhibited measurement invariance across the 26 universities of the Kingdom of Saudi Arabia.

The main advantage of this method is that metric and, most importantly, scalar invariance are not prerequisites for comparing group means meaningfully. Only the configural model must be established. The obtained results indicated a robust configural model. Particularly, all the factor loadings of the P-GAT subscales were invariant across the 26 universities. Most importantly, almost all the P-GAT intercepts were invariant across the 26 universities. Only two universities (i.e., King Khalid University and King Abdulaziz University) showed non-invariant intercepts for the sentence completion and

critical thinking subscales, respectively. This indicates that out of a total of two hundred and sixty parameters (ten items multiplied by twenty-six universities), just two (0.8 percent) were found to be non-invariant. These findings fell considerably below the 25% cutoff rule of thumb that is provided as a general guideline for the minimum required non-invariant parameters in order to move forward with comparisons at the latent mean level [22].

Next, given that the configural invariance assumption was satisfied, the P-GAT latent mean differences across the universities were examined. The results showed that the five universities with the highest mean values in terms of verbal P-GAT scores were (1) the King Saud University (0.454), (2) the Dammam University (0.333), (3) the Prince Nourah Bin Abdulrahman University (0.279), (4) the Qassim University (0.214), and (5) the King Abdulaziz University (0.174). On the other hand, the five universities with the lowest verbal P-GAT scores were (1) the Islamic University (−0.306), (2) the Jazan University (−0.308), (3) the Jouf University (−0.401), (4) the Northern Border University (−0.449), and (5) the Shaqra University (−0.521).

Regarding the quantitative P-GAT domain, at the top of the list appeared the same five universities, although some of them were at different positions: (1) the King Saud University (0.578), (2) the King Abdulaziz University (0.340), (3) the Dammam University (0.324), (4) the Qassim University (0.244), and (5) the Prince Nourah Bin Abdulrahman University (0.197). Finally, the six universities with the lowest mean scores in the quantitative domain were: (1) the Jazan University (−0.152), (2) the Islamic University (−0.157), (3) the University of Bisha (−0.168), (4) the Shaqra University (−0.306), (5) the Jouf University (−0.312), and (6) the Northern Border University (−0.342). Interestingly, King Saud University's P-GAT factor means were significantly higher than any other university apart from Dammam University in both domains (i.e., verbal and quantitative).

The main potential contributions of this study are twofold. First, it was demonstrated that the alignment approach is a valuable method to assess measurement invariance and latent mean differences when a large number of groups are involved. The main advantage of this method over the already established methods for testing measurement invariance (e.g., MGCFA) is that metric and, most importantly, scalar invariance are not prerequisites for comparing group means meaningfully. Only a configural model must be established. This technique provides an unbiased statistical estimation of group means, with significance tests between group pairs that adjust for sampling errors and missing data. This is particularly important since it ensures that scores across different groups (e.g., different cultures, different schools, etc.) are comparable and that valid inferences regarding the differences and similarities between these groups can be made from their comparisons.

Second, it provides valuable empirical information for policymakers and educators to examine the performance of each of the Kingdom's universities in terms of their P-GAT verbal and quantitative mean scores [15]. The results of this study may be a strong warrant for a better understanding of students' academic performances and score differences among universities, helping policy educators and national governmental agencies in exploring the possible factors that might have caused these gaps. For example, it seems that almost all the universities that exhibited low mean scores are located at the borders of the country (north and south), compared to the universities that are located in big cities (e.g., King Saud or King Abdulaziz universities, which are located in the capital of the Kingdom, Riyadh or Dammam University, which is the biggest city of the Eastern Province and its port is one of the biggest ports in the Arabian Gulf). These findings may help experts to understand the possible educational and socio-economic factors affecting individuals' performances and design appropriate actions. For example, previous studies have shown that the economic prosperity of a region can significantly impact academic performance through improved funding for education, enhanced learning materials, advanced technology, smaller class sizes, qualified teaching staff, enrichment opportunities, support services, cultural exposure, and a supportive learning environment [23,24].

5. Conclusions

In conclusion, the alignment method is pivotal in testing measurement invariance, offering practical implications that extend to various domains such as cross-cultural research and policy-making studies. This method enables researchers to make accurate and meaningful comparisons while maintaining the integrity of the measured constructs across large and diverse groups and contexts. Moreover, ensuring measurement equivalence across diverse groups helps to maintain fairness and equity in assessment practices.

Author Contributions: Conceptualization, I.T.; methodology, F.M.J.; formal analysis, I.T.; data curation, F.M.J.; writing—original draft preparation, I.T.; writing—review and editing, I.T. and F.M.J.; supervision, I.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Education & Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the Education & Training Evaluation Commission (ETEC).

Data Availability Statement: The data that support the findings of this study are available from the Education & Training Evaluation Commission (ETEC). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors upon reasonable request and with the permission of the ETEC.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Mplus syntax (input) file for testing measurement in variance of the P-GAT two-factor model across 26 universities using the Alignment (fixed) method.

```
title: P-GAT_Measurement Invariance across Universities: The Alignment Method
data: file is PGAT_alignment file.dat;
variable: names are gender school age uni van vsc vca vrc mar man mco ct sp log;
USEVARIABLES ARE van vsc vca vrc mar man mco ct sp log;
classes = c (26);
knownclass = c(uni = 1 3 4 8 9 10 11 12 13 14 16 17 19 20 21 22 24 26 27 29 30 31 32 33 34 35);
analysis: TYPE = MIXTURE;
ESTIMATOR is MLR;
ALIGNMENT = FIXED;
model:
%overall%
verbal by van vsc vca vrc;
quant by mar man mco ct sp log;
output: align TECH1 TECH8;
```

References

1. Nunnally, J.C.; Bernstein, I.H. *Psychometric Theory*, 3rd ed.; McGraw-Hill, Inc.: New York, NY, USA, 1994.
2. Bandalos, D.L. *Measurement Theory and Applications for the Social Sciences*; The Guilford Press: New York, NY, USA, 2018.
3. Millsap, R.E. *Statistical Approaches to Measurement Invariance*; Routledge/Taylor & Francis Group: Abingdon, UK, 2011.
4. Van De Schoot, R.; Schmidt, P.; De Beuckelaer, A.; Lek, K.; Zondervan-Zwijnenburg, M. Measurement invariance. *Front. Psychol.* **2015**, *6*, 1064. [CrossRef] [PubMed]
5. Vandenberg, R.J.; Lance, C.E. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* **2000**, *3*, 4–70. [CrossRef]
6. Milfont, T.L.; Fischer, R. Testing measurement invariance across groups: Applications in cross-cultural research. *Int. J. Psychol. Res.* **2010**, *3*, 2011–2084. [CrossRef]
7. Putnick, D.L.; Bornstein, M.H. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Dev. Rev.* **2016**, *41*, 71–90. [CrossRef] [PubMed]

8. Byrne, B.M. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*; Routledge/Taylor & Francis Group: Abingdon, UK, 2012.
9. Cheung, G.W.; Rensvold, R.B. Evaluating goodness of fit indexes for testing measurement invariance. *Struct. Equ. Model. Multidiscip. J.* **2002**, *9*, 233–255. [CrossRef]
10. Chen, F.F. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* **2007**, *14*, 464–504. [CrossRef]
11. Byrne, B.M.; van de Vijver, F.J.R. Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *Int. J. Test.* **2010**, *10*, 107–132. [CrossRef]
12. Asparouhov, T.; Muthén, B. Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* **2014**, *21*, 495–508. [CrossRef]
13. Kim, E.S.; Cao, C.; Wang, Y.; Nguyen, D.T. Alignment optimization in multiple-group analysis of structural equation models. *Struct. Equ. Model. Multidiscip. J.* **2017**, *24*, 183–197.
14. Sirganci, G.; Uyumaz, G.; Yandi, A. Measurement invariance testing with alignment method: Many groups comparison. *Int. J. Assess. Tools Educ.* **2020**, *7*, 657–673. [CrossRef]
15. Munck, I.; Barber, C.; Torney-Purta, J. Measurement Invariance in Comparing Attitudes Toward Immigrants Among Youth Across Europe in 1999 and 2009: The Alignment Method Applied to IEA CIVED and ICCS. *Sociol. Methods Res.* **2018**, *47*, 687–728. [CrossRef]
16. Asparouhov, T.; Muthén, B. Multiple group alignment for exploratory and structural equation models. *Struct. Equ. Model. Multidiscip. J.* **2023**, *30*, 169–191. [CrossRef]
17. Hu, L.T.; Bentler, P.M. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Struct. Equ. Model.* **1999**, *6*, 1–55. [CrossRef]
18. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 8th ed.; Muthén & Muthén: Los Angeles, CA, USA, 2017.
19. Mahalanobis, P.C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci.* **1936**, *2*, 49–55.
20. Tsaousis, I. *Using Confirmatory Factor Analysis in Testing for the Reliability and Validity of General Aptitude Test (GAT) Scores for Postgraduate Students*; Publication No. TR050-2014; National Center for Assessment in Higher Education: Riyadh, Saudi Arabia, 2014.
21. Tsaousis, I. *Using a Multiple Indicators Multiple Causes (MIMIC) Model to Examine Item and Scale Performance across Different Response Time Groups*; Publication No. TR164-2016; National Center for Assessment in Higher Education: Riyadh, Saudi Arabia, 2016.
22. Muthén, B.; Asparouhov, T. Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociol. Methods Res.* **2018**, *47*, 637–664. [CrossRef]
23. Hanushek, E.A.; Woessmann, L. The economics of international differences in educational achievement. In *Handbook of the Economics of Education*; Hanushek, E.A., Machin, S., Woessmann, L., Eds.; Elsevier: Amsterdam, The Netherlands, 2011; Volume 3, pp. 89–200.
24. Glewwe, P.; Kremer, M. Schools, teachers, and education outcomes in developing countries. In *Handbook of the Economics of Education*; Hanushek, E.A., Machin, S., Woessmann, L., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Volume 2, pp. 945–1017.