International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 167–180

https://doi.org/10.21449/ijate.690865

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

The Problem of Measurement Equivalence or Invariance in Instruments

Tulin Otbicer-Acar [1,*]

[1]Parantez Education Research Consultancy&Publishing, Ankara, Turkey

ARTICLE HISTORY

Received: Feb. 18, 2020
Revised: Nov. 15, 2020
Accepted: Jan. 27, 2021

Keywords:

Measurement equivalence,
Measurement invariance,
Cross-validation.

Abstract: The purpose of this study is to discuss the validity of equivalence in the sample groups of young and adult; females and males in the scale of assessing the attitudes towards foreign language skills and to offer the researchers that will use this scale certain evidence based on data. No measurement equivalence/invariance was found in adult and young groups. Consequently, measurement equivalence / invariance based on gender variable was not present, either. The absence of measurement equivalence/invariance is in fact a fundamental proof that the measurement instrument is specific to the group that it is intended for. For this reason, researchers should evaluate cross-validity or multi-group analyses on the basis of the traits that are measured using the measurement instrument. It is not always negative not to have measurement equivalence/invariance during the process of gathering validity evidences.

1. INTRODUCTION

Numerous instruments of measurement have been developed by researchers to measure the psychological structures of individuals, such as interest, attitude, success, anxiety, and motivation. A measurement instrument is sometimes considered within the scope of adaptation studies. Developing or adapting an instrument is a time consuming and rigorous process in which whether the measurement instrument is capable of measuring the same conceptual structure in different groups and cultures signifies the validity of the instrument. In validity studies, it is desirable for the structure that is being measured under measurement by the instrument to be invariant and unbiased. When the measurements vary among the subgroups of the populations that are measured or among different populations, there is a certain amount of *bias*. The potential for bias in test items is the most significant element. They arise from systematic errors. Also, other sources should be taken into consideration for the validity of the instruments of measurement. The sources of bias are studied under the categories of construct bias, method bias (namely sample bias, administration bias, and instrument bias) and, item bias (Vijver &Tanzer, 2004).

Item bias is typically referred to as Differential Item Functioning (DIF). However, educational experts, test developers make a difference between the concept of item bias and DIF. The concept of item bias has a negative meaning in everyday life and it is associated with a negative idea. The difference between technical use of item bias and the everyday use of it is uncertain.

CONTACT: Tulin Otbicer Acar ✉ totbicer@gmail.com 🖃 Parantez Education Research Consultancy Publishing, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

The conceptual difference between DIF and item bias is as follows (Hambleton et al., 1991, p.109):

> Investigation of bias involve gathering empirical evidence concerning the relative performances on the test item of members of minority group of interest and members of the group that represents the majoriy. Empirical evidence of differential performance is necesssary, but not sufficient, to draw the conclusion that bias is present; this conclusion involves an inference that goes beyond the data. To distinguish the empirical evidence from the conclusion, the term differential item functioning (DIF) rather than bias is used commonly to described the empirical evidence obtained in investigations of bias.

It is understood that item response theory (IRT) and structural equation modeling (SEM) are used in the studies that are intended to determine the systematic errors that interfere in the results of measurement. The traits measured in these two methods are defined as latent traits. According to the IRT, item bias is determined by DIF. DIF is a function that is used to determine whether the probability of responding an item differs among subgroups in each skill level of the psychological structure that is to be measured by an item (Lord, 1980; Embretson & Reise, 2000). Likelihood Ratio according to IRT (Thissen et al., 1988), Lord's chi-square test (Lord, 1980), and Raju's area measures (Raju, 1988) are among the techniques that are used in the literature to determine DIF. In addition, there are techniques of DIF determination such as Mantel-Haenszel, Logistic Regression, and SIBTEST in the classical test theory based on observed scores in metrology (Gierl et al., 1999).

A different approach, according to DIF techniques in IRT, is measurement equivalence/invariance. In the literature, the term 'measurement invariance' is usually used as the synonym of measurement equivalence (Davidov et al., 2014). Wadenberg and Lance (2000, p.5) stated that "measurement equivalence-ME (or alternately, measurement invariance-MI) across populations". In addition, measurement equivalence is also called structural equivalence (Kankaraš & Moors, 2010). Measurement equivalence denotes similarity of observed variables and latent structures among groups (Koh & Zumbo, 2008). A method based on covariance structures is used in measurement equivalence research (French & Finch, 2008). This covariance-based method is resolved by SEM analyses. Studying multiple group equivalence by SEM method corresponds to the concept of measurement method. According to the definition made by Byrne (2008), measurement equivalence or invariance (ME/I) implies that the measurement instrument has the same psychological meaning and theoretical structure in the groups of interest. It is an approach that is based on restriction of structural parameters (factor loadings, observed variable error variances, error covariances) produced by multiple group invariance – an extension of Confirmatory Factor Analysis – CFA). This approach is associated with measurement equivalence/invariance. Two types of techniques are used for measurement equivalence/invariance in SEM. The first one is Multi-Group Confirmatory Factor Analysis (MGCFA, see Jöreskog, 1971; Cheung & Rensvold, 2002) where the equivalence of covariance structures is tested. The second one is Mean and Covariance Structure (MACS, see Byrne et al., 1989; Little, 2010; Yuan & Bentler, 1997) where the equivalence of mean and covariance structures is tested. Both MGCFA and MACS are cross-validation techniques. These analyzes are resolved by SEM. MACS analysis is used to assess differences between group in terms of the constructs' mean, variances and covariances. MGCFA tests the invariance of estimated parameters across groups.

There are numerous studies focusing on whether several instruments of measurement that measure different psychological characteristics ensure measurement equivalence/invariance in different subgroups (see Akyıldız, 2009; Asil & Gelbal, 2012; Baumgartner & Steenkamp, 1998; Lomax, 1983; Mullen, 1995; Önen, 2007; Uyar & Doğan, 2014; Yoo, 2002). It is observed that these studies offer a comparison of the models that are made up of restricted parameters. The steps of the measurement equivalence/invariance that shows a series of progressivity depending on the number of restricted parameters are as follows (Byrne & Stewart, 2006):

Model 1 – Configural invariance: The first stage. Factor loads, regression constants and error variances are released among groups. However, the number of factors, and the factor loading pattern are defined similarly among groups. Therefore, structural invariance is ensured among groups. Measurement of the same structure is measured among groups.

Model 2 – Weak factorial invariance or Metric invariance: Factor loads are restricted in addition to the first stage. When metric invariance is not ensured, the items in the groups are not considered to be interpreted at the same level. Factor loads correspond to the Discrimination Parameter, and non-uniform DIF of factor load is present among groups (Steinmetz et al., 2009).

Model 3 – Strong factorial invariance or scalar invariance: This is the stage where the regression constant equates between groups. On the other hand, for straightforward interpretation of latent means and correlations across groups, both the factor loadings and intercepts should be the same across groups (Van de Schoot, Lugtig & Hox, 2012, p.490). Variance of regression constant among groups signifies presence of uniform DIF on the items and means that scalar invariance is not present (Kankaraš & Moors, 2010).

Model 4 – Strict factorial invariance: At this stage, critical error variances have been restricted as well. In this model, the error variances of the second group stabilize on the error variances of the first group.

Cheung and Rensvold (2002, p.236) stated that “the statistic for testing the hypothesis is the difference between the fit of the constrained model and that of a less constrained model. Many fit indices are obtained for each of the four models mentioned above. The most frequently used criterion is that the difference between the values of RMSEA and CFI – fit indices – in comparison of models is smaller than 0.01 (Byrne & Stewart, 2006; Hirschfeld & Brachel, 2014). Since RMSEA, CFI, and SRMR are not affected by the sample size of fit indices, these indices are suggested to be taken into consideration in comparing these models (Hu & Bentler, 1999). Similarly, the chi-square difference between the two models, the insignificance of the chi-square difference test, and the difference between the degrees of freedom are considered as an indication of the invariance between the models. Byrne and Stewart (2006, p.305) noted that “$\Delta\chi^2$ value is as sensitive to sample size and nonnormality as the chi-square statistic itself, thereby rendering it an impractical and unrealistic criterion on which to base evidence of invariance.”

An examination of the literature reveals that multi-group analyzes are also called cross-validation (Fiala et al., 2002; Gandek et al., 1998). It is obvious that these techniques provide extra data in gathering data for validity. However, it should be noted that reducing the sample size makes a major disadvantage in cross-validation or multi-group studies. Varoquaux (2018, p.68) stated that “the shortcoming of small samples are more stringent and inherent as they are not offset with large effect sizes”.

1.1. Aim of the Study

Sample characteristics of subjects must be taken into consideration for the future usage of the same scale. Otherwise, measurements errors are likely to appear. So, the scope of present study is to discuss the validity of equivalence in the sample groups of young and adult, and females and males in the scale of assessing the attitudes towards foreign language skills and to offer the researchers that will use this scale certain evidence based on data. It should be noted that this study was carried out for evidence of measurement equivalence/invariance for the scale developed. The empirical outcomes of this study will make important contributions to both psychological test developers and psychometrists.

2. METHOD

2.1. Researh Design

In this research, measurement equivalence/invariance was investigated for gender and two groups. Thus, present research is a descriptive research. Descriptive research is "current events and that the research questions and problems raised are based on the appreciation of the present phenomena, events, or state of affairs" (Ariola, 2006, p.47). The scale developed for 15-16 year old people cannot be applied to the scale developed for 18-60 year old people without being tested and applied. Like age variable, gender variable plays a significant role in measurement equivalence/invariance due to the fact that gender difference embraces both biologic and cultural implications.

2.2. The Characteristics of Participants

The researcher collected data on from 563 participants to test the equivalence in the group of adults aged 18 to 60 in the scale which was developed for the student groups of 15-16 years of age for determining attitudes towards foreign language skills. 15-16-year-old students were high school students who continued secondary education. Therefore, this group of students was named *young* in this study.

The scale was administered to the participants in Turkey. Some of the participants are employed and some are out of employment. They belong to various occupational groups such as academicians, entrepreneurs and business people. The scale was administered online and in printed form. Missing values were excluded from the data set. Therefore, the data set includes 481 participants – 275 young students (57.2%) and 206 adults (42.8%). The frequency of gender distributions by groups and the result of the chi-square test is given in Table 1.

Table 1. *Gender distribution by groups.*

			Group		Total
			Young	Adult	
Gender	Female	Count	136	109	245
		%	55.5%	44.5%	100.0%
	Male	Count	139	97	236
		%	58.9%	41.1%	100,0%
Total		Count	275	206	481
		%	57.2%	42.8%	100.0%
χ^2=0.564	Sig.=0.453				

55.5% of the female participants are young and 44.5% are adults. 58.9% of male participants are young and 41.1% are adults. 49.5% of the young are females. 52.9% of the adults are females. Statistically, no significant difference was found between the gender distributions of the individuals according to the groups (p>0.05). For both young and adult groups, according to the gender of the participants, the difference between age averages was tested by t test for independent samples. The results are given in Table 2.

Table 2. *Results of the difference between age averages according to the gender.*

Group		N	Mean	Std. Deviation	*t* value	df	*p*
Young	Female	136	15.54	0.50	.073	273	.942
	Male	139	15.53	0.50			
	Total	275	15.53	0.50			
Adult	Female	106	27.34	7.14	-.279	197	.780
	Male	93	27.63	7.74			
	Total	199	27.48	7.41			

The average age of the young group is 16-years old. The average age of the adult's group is 28-years old. Statistically, there was no significant difference between the average age of male and female young (p>0.05). Also, there was no significant difference between the mean age of male and female adults (p>0.05).

2.3. Instrument

The developed scale comprises 29 items that are structured on a 5-point scale ranging from 1 to 5. The original purpose of the scale was to identify the attitudes of 15- and 16-year-old students towards Foreign Language Skills. For 15-16-years old students, the reliability-validity analyses of the development process of the scale are available in the reference Acar (2016). The implementation scale is given in the Table A1. The scale has 4 sub-factors: reading, writing, speaking, and listening. In this study (for 481 participants), the internal consistency of the scale's Cronbach Alpha reliability is 0.923 for the adult group; 0.900 for the young group; 0.922 for the females' group and 0.899 for the males' group. Sub-scales reliabilities were showed in Table 3. It is observed that the internal consistency of the subscales is at appropriate values.

Table 3. *Cronbach's Alpha coefficients.*

	Sub-Groups			
	Young	Adult	Female	Male
Reading	0.768	0.729	0.782	0.724
Writting	0.758	0.756	0.783	0.744
Speaking	0.758	0.623	0.722	0.692
Listening	0.804	0.786	0.789	0.793

Item-total correlations are shown in Table A2. The variation between 0.140 and 0.655 in the subscales of item total correlations was measured. No item was removed in this study, although the number of items in the subscales was relatively low. Due to the fact that the purpose of the research is the measurement invariance of the instruments. In addition, the reliability and validity of the scale were tested in another sample, too.

2.4. Data Analysis

For measurement equivalence/invariance, all procedures were based on the analysis of MACS within the framework of CFA modeling. The LISREL (Jöroskog&Sörbom, 2003) program was used for all analyses. First of all, the dataset was completely cleared of missing values. It was observed that item scores ranged from 1 to 5, and there were no extreme values. Through confirmatory factor analysis, four sub-dimensional scale was tested for the all data before multi-group CFA was carried out. The multivariate assumption of normality was not met. Because, Mardia's measure of multivariate skewness and kurtosis was not found significant (χ^2=2664.719 p<0.000). Thus, the observed scores of scale items were converted into normal scores in LISREL. Estimations of parameters were carried out through maximum likelihood. Asymptotic covariance matrix was used for parameter estimations. Fit indices was presented Table 4.

Root mean square approximation error was calculated as (RMSEA) = 0.074. Van de Schoot, Lugtig and Hox (2012, p.488) stated that "the RMSEA is insensitive to sample size, but sensitive to model complexity". Bialosiewicz, Murphy, and Berry (2013) pointed out that an RMSEA around 0.10 is acceptable. Standardized root mean square residual was calculated as (S-RMR) = 0.068; comparative fit index was calculated as (CFI) = 0.93; normed fit index was calculated as (NFI) = 0.91 and relative fit index was calculated as (RFI) = 0.90. Chi-square statistics of the similarity rate was calculated as χ^2 (371) = 1339.65 p<0.01 and χ^2 / df is 3.61.

Table 4. *Goodness of fit indices.*

Goodness of fit indices	Cut off value [*]	Values
χ^2/df	<5 Moderate <3 Perfect fit	1339.65/371= 3.61
GFI	>0.90	0.79
CFI	>0.90	0.93
NFI	>0.90	0.91
RFI	>0.85	0.90
S-RMR	< 0.08	0.068
RMSEA	< 0.08	0.074

[*] Resources: Kline, 2011; Bentler, 1980

Goodness-of-fit index was calculated as (GFI)= 0.79 and only this index was found below 0.90. GFI involve terms that adjust for degrees of freedom. Thus, GFI is highly dependent on sample size. In addition, Cheung and Rensvold (2002) showed that number of items per factor and number of factors in the model affect GFI values. Bollen and Long (1983) pointed out, "the test statistics and fit indices are very beneficial, they are no replacement for sound judgment and substantive expertise". It was observed that 4-factor structure attitude scale concerning the English language skills was acceptable according to the standard criteria. Baumgartner and Homburg (1996, p.153) suggest that general rules of thumb (e.g., that GFI be greater than 0.9) may be misleading because they ignore such contingencies. χ^2 /df and RMSEA seem to be effective in controlling for model complexity by assessing fit per degree of freedom. t values indicating the significance of the relationship between the items and the latent variable are presented in the Figure A1.

3. FINDINGS

In the invariance studies, the RMSEA value is not interpreted alone. According to, literature for comparison of the four models, difference values or difference tests (for example $\Delta\chi^2$, ΔGFI, ΔCFI, ΔTLI, ΔBBI or ΔRMSEA) are used. Rijkeboer and van den Bergh (2006) suggested the use of Chi-Square difference test which is the most efficient one with respect to both goodness-of-fit and parsimony. The choice of difference tests remains at the expertise the researcher. The dataset was divided in two groups – namely, females and males, then the measurement equivalence/invariance of the scale for determining the attitudes towards foreign language skills was tested on the basis of the gender variable, and the results of the fit indices were specified in Table 5.

Table 5. *Measurement equivalence/invariance based on gender variable.*

Models	χ^2	df	RMSEA	CFI	*ΔCFI*	*Δ* RMSEA	$\Delta\chi^2$	*Δdf*	*p*
1: Configural invariance	2527.59	781	0.097	0.92	-	-	-	-	-
2: Metric invariance	2579.36	806	0.096	0.92	0.00	-0.001	51.77	25	0.001
3: Scalar invariance	2731.79	834	0.097	0.91	-0.01	0.001	152.43	28	0.000
4: Strict factorial invariance	2759.81	835	0.098	0.91	0.00	0.001	28.02	1	0.000

When comparing *Model 2* versus *Model 1*, Cheung and Rensvold (2002, p.251) pointed out "a value of smaller than or equal to –0.01 indicates that the null hypothesis of invariance should not be rejected". A comparison of *Model 1* - *Model 2*, *Model 2* - *Model 3*, and *Model 3* - *Model 4* reveal that Δ RMSEA and ΔCFI values were in appropriate ranges. However, p value of the chi-square difference test was found to be significant. It is seen that $\Delta\chi^2$, ΔCFI, and ΔRMSEA values provide different interpretations. In this study, final comments are made according to $\Delta\chi^2$ values. It was observed that metric, scalar, and strict factorial invariances could not be ensured in the multi-group analysis based on the gender variable. At this stage, it was suggested to test whether there are any items that contain uniform and non-uniform DIF.

According to the system of progressivity, it is not significant to skip to the next stage when a stage is not appropriate. It is observed that in certain studies, partial invariance models are attempted to be used where invariance cannot be ensured (Murayama et al., 2009; Milfont & Fischer, 2010). However, partial invariance models were not used in this study. Measurement equivalence/invariance of the scale of determining the attitudes towards foreign language skills in young and adult groups and the results of the fit indices are given in Table 6.

Table 6. *Measurement equivalence/invariance based on the group variable.*

Models	χ^2	df	RMSEA	CFI	*ΔCFI*	*Δ* RMSEA	*Δ*χ^2	*Δdf*	*p*
1: Configural invariance	2502.04	781	0.096	0.92	-	-	-	-	-
2: Metric invariance	2565.62	806	0.095	0.92	0.00	-0.001	63.58	25	0.000
3: Scalar invariance	2459.67	835	0.103	0.91	-0.01	0.008	105.95	29	0.000
4: Strict factorial invariance	2459.67	835	0.103	0.91	0.00	0.000			

A comparison of Model 1 and Model 2 reveals that ΔRMSEA and ΔCFI values didn't not exceed the 0.01 threshold. However according to chi-square difference test (*Δ*χ^2), metric invariance was not ensured for factor number, factor loading pattern, and factor loads among young and adult groups for the scale for determining the attitudes towards foreign language skills. It is seen that Δχ^2, ΔCFI, and ΔRMSEA values provide different interpretations. In this study, final comments are made according to Δχ^2 values. Therefore, it was found that certain items in young and adult groups may be biased. This result offers a clue in identifying the items that contain uniform DIF. Since metric invariance was not ensured it was understood that factors do not mean the same in different groups.

When *Model 2* was compared with *Model 3*, it eas revealed that ΔRMSEA and ΔCFI values didn't not exceed the threshold. However, according to chi-square difference test, scalar invariance was not ensured for factor number, factor loading pattern, factor loads, and regression constants among young and adult groups for the scale for determining the attitudes towards foreign language skills. Therefore, it was found again that certain items in young and adult groups may be biased. These results offer clues on identifying the items that contain non-uniform DIF. In other words, the mean values of latent structures vary among the groups. It is not appropriate to make a comparison between the means of youngs and adults.

Accoording to chi-square difference test, scale item equivalence could not be ensured on the basis of groups. The results of the discriminant analysis were used to decide which group the developed scale is appropriate for. The correct classification ratio, equality of covariance matrices, and log determinant tables were evaluated according to the discriminant analysis results. According to the discriminant analysis, the correct classification ratio of original and predicted group memberships was 81.1% for youngs and 66% for adults. Also, 74.6% of original grouped cases correctly classified. The results indicate a higher classification consistence for the young group. An examination of Box'M results in the equation of covariance matrices leads to rejection of the equation of covariance matrices in young and adult groups ($F_{(2;\ 592459.\ 45)}$ =833.362 sig=0.000). Log Determinant values are given in Table 7.

Table 7. *Log determinants.*

Group	Rank	Log Determinant
Young	29	-1.781
Adult	29	-7.182
Pooled within-groups	29	-2.353

In the multi-group model, log determinant values provide an indication of which groups' covariance matrices differ most. For each group, its log determinant is the product of the eigenvalues of its within group covariance matrix. In this research, the log determinant value for adult group is very small relative to that of the young group. Therefore, it is fair to say that scale items are suitable for the young group that was developed initially.

4. CONCLUSION and RECOMMENDATIONS

In the study, analysis of MACS was used to test for measurement invariance of the scale items across group and gender variables. $\Delta\chi^2$, ΔCFI, and ΔRMSEA values provided different interpretations. In this study, final comments were made according to chi-square difference test values. No measurement equivalence/invariance was found in adult and young groups. Consequently, measurement equivalence/invariance based on gender variable was not presented, either. Female and male datasets include adult and young as well. For this reason, it is a predictable result that measurement equivalence/invariance is absent for groups, and that the measurement equivalence/invariance based on the gender of the same individuals is not in compliance. The finding bears similarity to the finding of Feingold (1992) who emphased that cognitive abilities arise from gender differences.

Little (2010, p.53) said that "the nature of sociocultural differences and similarities can be confidently and meaningfully tested among the constructs' moments in each sociocultural sample". But in this study, the measurement equivalence/invariance of the scale in different cultures was not tested. Since the scale was intended to measure the attitudes of 15-16-year-old Turkish students towards foreign language skills, it is restricted with the psychological traits of Turkish students. It is considered that the reasons for the absence of measurement equivalence/invariance in the scale include different interests, motivations, and attitudes towards foreign language skills among adult and young groups. Since the young group is made up of individuals who receive formal education, it is quite likely that they have different perceptions of foreign language skills compared to adults. Students’ respective success in English courses is considered to have an impact on their attitude to foreign language skills. On the other hand, adults’ perspective of foreign language skills is generally influenced by their occupational development, because they are not in formal education anymore due to their age.

Metric and scalar invariance was not present based on groups of adults and young, and on genders. There is evidence of the presence of uniform and non-uniform DIF items. However, a detailed study on DIF was not conducted due to the purpose of this study. The measurement instrument may be redesigned later. Certain items may be added, removed, or modified depending on the psychological traits of the implementation group. Equivalence trait of the measurement instrument may be abandoned in different groups. In this respect, the scale may be used for the target group for which it was originally intended.

It is not always negative not to have measurement equivalence/invariance during the process of gathering validity evidence. The absence of measurement equivalence/invariance is, in fact, a fundamental proof that the measurement instrument is specific to the group that it is intended for. For this reason, researchers should evaluate cross-validity or multi-group analyses on the basis of the traits that are measured using the measurement instrument. The validity of the instruments is the evidence gathering process. An ad-hoc measurement instrument should not be developed or used. It is recommended that any kind of information be used in gathering evidence and data for examination of the instruments of measurement.

Scale developing is a process. The most important stage of this process is ensuring the validity of the measurement instrument. Validity analysis should be examined through different techniques. In this process, items can be regulated. The qualifications of the application group may vary. Even the application area of the scales may expand.

Acknowledgments

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests and Ethics

The author(s) declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Tulin Otbicer Acar: All chapters are written by the author.

ORCID

Tülin Otbiçer Acar https://orcid.org/0000-0001-7976-5521

5. REFERENCES

Acar, T. (2016). Measurement of attitudes regarding foreign language skills and its relation with success. *International Journal of Evaluation and Research in Education, 5*(4), 310-322. https://doi.org/10.11591/ijere.v5i4.5959

Akyıldız, D. (2009). PIRLS 2001 testinin yapı geçerliliğinin ülkelerarası karşılaştırılması [The comparison of construct validities of the PIRLS 2001 test between countries]. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6*(1), 18-47. https://dergipark.org.tr/tr/pub/yyuefd/issue/13711/165993

Ariola, M. M. (2006). *Principles and methods of research*. Rex Book Store.

Asil, M., & Gelbal, S. (2012). Crosscultural equivalence of the PISA student questionnaire. *Education and Science, 37*, 236-249.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139-161.

Baumgartner, H., & Steenkamp, J.-B. E. M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters, 9*, 21-35. https://doi.org/10.1023/A:1007911903032

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31*(1), 419-456.

Bialosiewicz, S., Murphy, K., & Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants*. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Bollen, K. A., & Long, J. S. (Eds.). (1983). *Testing structural equation models*. Sage.

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,* 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Byrne, B. M. (2008) Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20,* 872-882.

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: the_macs approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287-321. https://doi.org/10.1207/s15328007sem1302_7

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

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1) 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin, 111*(2), 304–341. https://doi.org/10.1037/0033-2909.111.2.304

Fiala, W. E., Bjorck, J. P., & Gorsuch, R. (2002). The religious support scale: construction, validation, and cross-validation. *American Journal of community Psychology, 30*, 761-786. https://doi.org/10.1023/A:1020264718397

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. Structural Equation Modeling: *A Multidisciplinary Journal, 15*(1), 96-113. https://doi.org/10.1080/10705510701758349

Gandek, B., Ware J. E., Aaronson N. K., Apolone, G. B., Brazier J. E., et al. (1998). Cross validation of item selection and scoring for the SF-12 health survey in nine countries: results from the iqola project, international quality of life assessment. *Journal of Clinical Epidemiology, 51*(11), 1171-1180. https://doi.org/10.1016/S0895-4356(98)00109-7

Gierl, M., Khaliq, S. N., & Boughton, K. (1999). *Gender differential ıtem functioning in mathematics and science: prevalence and policy ımplications.* Paper Presented at the Symposium entitled "Improving Large-Scale Assessment in Education" at the Annual Meeting of the Canadian Society for the Study of Education, Canada

Hambleton, R. K., Swaminathan, H., & J. H. Rogers. (1991). *Fundamentals of Item Response Theory*. Sage Publications.

Hirschfeld, G., & Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R-A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation, 19*(7), 1-12.

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

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426. https://doi.org/10.1007/BF02291366

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross cultural studies. *Psihologija, 43*(2), 121-136. https://doi.org/10.2298/PSI1002121K

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. The Guilford Press.

Koh, K., & Zumbo, B. D. (2008). Multi-group confirmatory factor analysis for testing measurement invariance in mixed item format data. *Journal of Modern Applied Statistical Methods, 7*(2), 471-477. https://doi.org/10.22237/jmasm/1225512660

Little, T.D. (2010). Mean and covariance structures (MACS) analyses of crosscultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53-76. https://doi.org/10.1207/s15327906mbr3201_3

Lomax, R. G. (1983). A guide to multiple-sample structural equation modeling. *Behavior Research Methods & Instrumentation, 15,* 580-584. https://doi.org/10.3758/BF03203726

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge. https://doi.org/10.4324/9780203056615

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-130. https://doi.org/10.21500/20112084.857

Mullen, M. R. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies, 26,* 573-596. https://doi.org/10.1057/palgrave.jibs.8490187

Murayama, K., Zhou, M., & Nesbit, J. C. (2009) A cross-cultural examination of the psychometric properties of responses to the Achievement Goal Questionnaire. *Educational and Psychological Measurement, 69*(2), 266-286. https://doi.org/10.1177/0013164408322017

Önen, E. (2007). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin incelenmesi: Epistemolojik inançlar envanteri üzerine bir çalışma [Examination of measurement invariance at groups' comparisions: A study on epistemological beliefs inventory]. *Ege Eğitim Dergisi*, *8*(2), 87-109. https://dergipark.org.tr/tr/pub/egeefd/issue/4913/67270

Raju, N. S. (1988). The area between two item response functions. *Psychometrika, 53*, 495-502. https://doi.org/10.1007/BF02294403

Rijkeboer, M. M., & van den Bergh, H. (2006). Multiple group confirmatory factor analysis of the young schema-questionnaire in a Dutch clinical versus non-clinical population. *Cogn. Ther. Res., 30*, 263–278. https://doi.org/10.1007/s10608-006-9051-8

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality and Quantity, 42,* 599-616. https://doi.org/10.1007/s11135-007-9143-x

Thissen, D., Steinberg, L., & Wainer, H (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 147-169). Lawrence Erlbaum Associates, Inc.

Uyar, Ş., & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi [An investigation of measurement invariance of learning strategies model across different groups in PISA turkey sample]. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2*, 30-43.

Varoquaux, G. (2018). Cross-validation failure: small sample sizes lead to large error bars. *Neuroimage,180*, 68–77.

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

Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée, 54*(2), 119-135. https://doi.org/10.1016/j.erap.2003.12.004

Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology & Marketing, 19*(4), 357-368.https://doi.org/10.1002/mar.10014

Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*(438*)*, 767-774. https://doi.org/10.1080/01621459.1997.10474029

6. APPENDIX

Table A1. *Form of the Attitude Scale Regarding English Language Skills.*

		Strongly Disagree	Disagree	Neither	Agree	Strongly Agree
1	I can answer the questions asked, after listening to a dialogue.					
2	I listen to a tourist if I encounter one.					
3	I look up the words in the dictionary, whose English meanings I don’t know.					
4	I make an effort to watch an English movie or listen to English news or music.					
5*	I’m anxious about writing a letter, petition or resume in English.					
6	When I listen to a text or music in English, I make an effort to understand its meaning.					
7*	Writing in English in English exams, makes me anxious.					
8*	I close the English pages I encounter while making a search in the search engines.					
9*	I get bored with English listening activities.					
10	Speaking English, increases my self-confidence.					
11*	Speaking English, makes me anxious.					
12	I like reading English story books.					
13	I read a lot, in order to learn English words.					
14*	It is boring for me to listen to someone speaking English.					
15	I care about summarizing what I’ve heard in English, and writing them correctly.					
16*	I immediately walk away when I see someone speaking English.					
17*	I don’t prefer having foreign friends to speak English with.					
18	I enjoy speaking English.					
19	I’d like to be a listener in a conference where English is spoken.					
20*	Reading and perceiving what is written in English, does not take an important place in my daily life.					
21*	I can’t express my opinions easily while writing an English text.					
22*	Writing in English, is not important in daily life.					
23	I’d like the English reading activities to be more.					
24	I do not hesitate from answering the questions asked in English.					
25	I pay attention to the grammar rules while writing in English.					
26*	It is not important to speak English fluently.					
27*	I don’t like reading equipment manuals that are written in English.					
28	I can write an English text about myself.					
29	I try to speak in accordance with the grammar rules.					

*Reverse coded items.

Attitude items related to the reading skill=3,8,12,13,20,23,27
Attitude items related to the writing skill =5,7,15,21,22,25,28
Attitude items related to the speaking skill =10,11,17,18,24,26,29
Attitude items related to the listening skill =1,2,4,6,9,14,16,19

Figure A1. *The Path Diagram which is Factor Load per Item for All Dataset.*

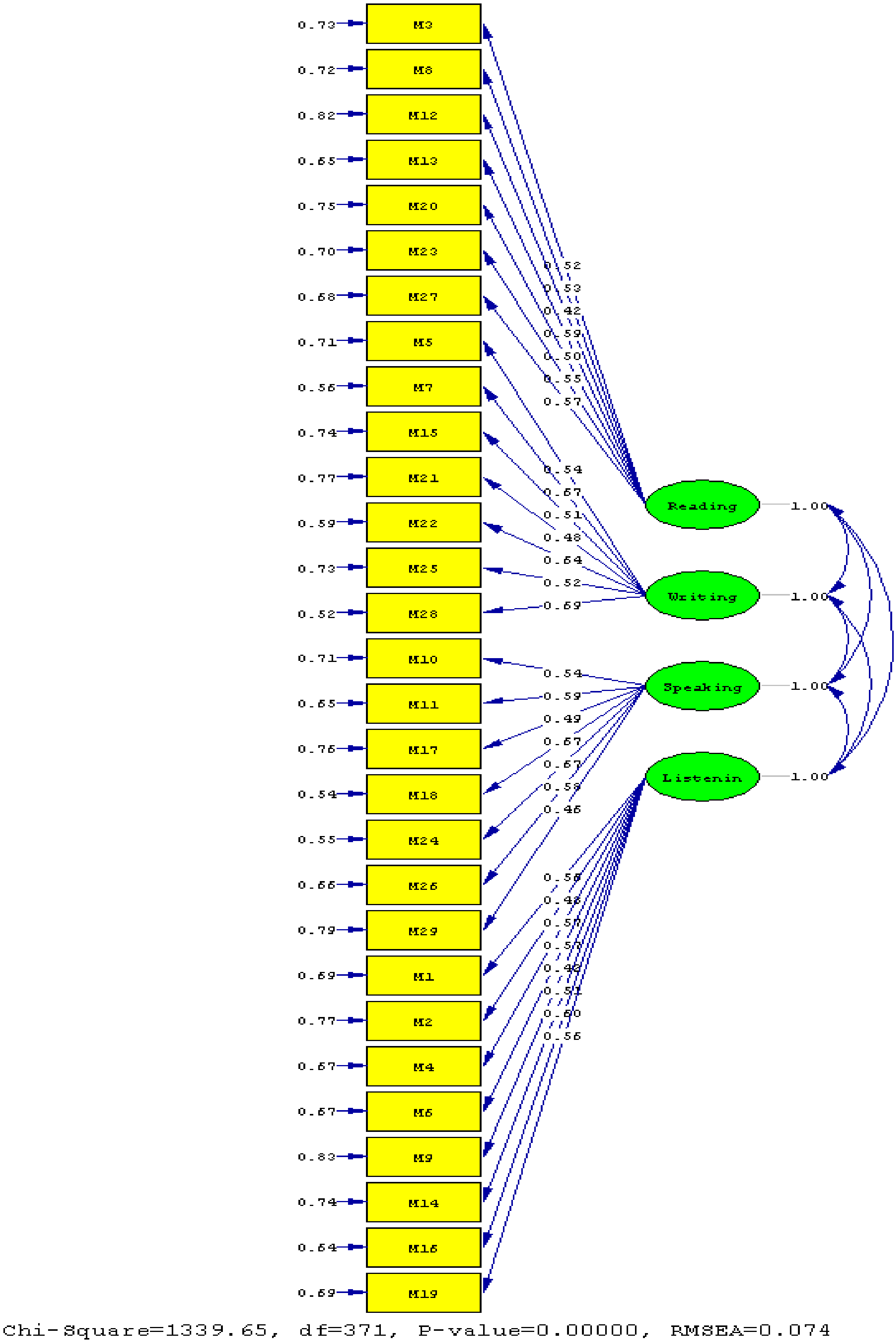

Table A2. *Item-Total Correlations.*

Sub-scales	Item No	Young Group	Adult Group	Female Group	Male Group
Reading	m3	.395	.348	.427	.341
	m8	.470	.428	.486	.383
	m12	.613	.633	.655	.562
	m13	.585	.531	.567	.580
	m20	.386	.398	.409	.367
	m23	.530	.330	.428	.508
	m27	.436	.425	.573	.307
Writing	m5	.406	.640	.504	.548
	m7	.498	.647	.606	.568
	m15	.476	.290	.475	.348
	m21	.457	.601	.573	.503
	m22	.451	.313	.413	.361
	m25	.537	.381	.420	.476
	m28	.511	.456	.562	.405
Speaking	m10	.535	.334	.493	.408
	m11	.455	.339	.422	.413
	m17	.437	.317	.345	.383
	m18	.577	.611	.592	.572
	m24	.466	.447	.543	.414
	m26	.431	.140	.296	.311
	m29	.434	.214	.361	.318
Listening	m1	.390	.375	.457	.322
	m2	.503	.409	.457	.489
	m4	.551	.538	.566	.511
	m6	.541	.426	.469	.499
	m9	.489	.604	.515	.520
	m14	.585	.601	.559	.588
	m16	.523	.417	.444	.504
	m19	.549	.552	.505	.568