Education and Science

Vol 45 (2020) No 203 287-308

Investigating Measurement Invariance of Ankara University Foreign Students Selection Test According to Latent Class and Rasch Model *

Özge Altıntaş [1], Ömer Kutlu [2]

Abstract

This study aimed to investigate the factor structure of the Basic Learning Skills Test included in the Ankara University Examination for Foreign Students and determine whether this test has measurement invariance and whether the items of the test shows differential item functioning according to the country and gender. Of the descriptive research models, screening was used in the design of the study. The population of the study consisted of 2134 individuals. Since the research involved intercultural comparisons, culture was taken into consideration in the selection of the sample. Accordingly, 1110 individuals from Germany, Azerbaijan and Iran constituted the sample. In the research, the first 60 items of the Basic Learning Skills Test concerning letters, numbers and figures were utilized. Latent Class and Latent Class Factor Analyses were employed to determine the factor structure of the test, and a Simultaneous Latent Class Analysis was undertaken to investigate the measurement invariance of the results according to country and gender. Besides, Recursive Partitioning Analysis based on Rasch Model was used to identify whether the items included in test show differential item functioning by country and gender. Before determining the factor structure of the Basic Learning Skills Test, an examination was conducted to ascertain whether the test as a whole fitted the scale structure, and which classes were included in the response patterns of the individuals taking the test. According to the results, the test had a latent model of four classes and a multiple-factor structure, comprising three factors of establishing part-to-whole relations, making inferences, and analytical thinking. Concerning the findings of analysis regarding the measurement invariance according to country and gender, it was found that the test had a homogeneous structure, and as a result, the structure measured by the test was equivalent across the groups compared. Studying whether the items in the test show differential item functioning according to country and gender, it was discovered that 14 items involve differential item functioning by country. Differential item functioning in regard to gender was not observed.

Keywords

Test development
Latent class factor analysis
Measurement invariance
Simultaneous latent class analysis
Differential item functioning
Recursive partitioning analysis based on Rasch model
International student selection tests

Article Info

Received: 04.19.2019
Accepted: 11.20.2019
Online Published: 05.04.2020

DOI: 10.15390/EB.2020.8685

* This article is derived from Özge Altıntaş's PhD dissertation entitled "Investigating the Measurement Invariance of Ankara University Foreign Students Selection Test by Latent Class and Rasch Model", conducted under the supervision of Ömer Kutlu.

[1] Ankara University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Turkey, oaltintas@ankara.edu.tr

[2] Ankara University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Turkey, omerkutlu@ankara.edu.tr

Introduction

In a society, individuals that develop and ensure the continuity of political, economic and cultural system are raised by educational institutions. This function of education, which provides social development, has increased the importance of educational institutions in all societies. While societies provide basic life skills that individuals need to acquire, especially in the preschool, primary school and middle school periods, they consider high school as a process of development of basic mental skills and orientation toward a profession. Higher education programs facilitate the selection of a profession and competence in that profession, while, at the same time, enriching the level of intellectual, actual and scientific knowledge.

In recent years, many graduates of high schools do not confine themselves to their native country to acquire a higher-quality university education, and they want to attend various higher education programs of universities abroad. In Turkey, the selection and placement procedures of foreign students who want to enroll in higher education programs can be chronologically examined in three groups as: 1980 and before, from 1981 to 2011, and 2011 and after. In 1980 and before, universities conducted foreign student selection and placement procedures according to the criteria they defined. Between 1981 and 2011, Examination for Foreign Students (YOS) tests were used for foreign student selection and placement procedures. However, in 2010, the Council of Higher Education (YOK) decided that the selection and placement of foreign students would be carried out by universities. According to this decision, universities designed their programs and determined the quotas after receiving approval from YOK (2013). In 2011, Ankara University took the decision to implement the Ankara University Examination for Foreign Students (AYOS) through the Measurement and Evaluation Application and Research Center (ANKUDEM) of the university (ANKUDEM, 2011).

The mental structures in the tests were developed to measure psychological characteristics and the dimensions of these structures concern construct validity, which is very important for the test development process. Developers design the test primarily in accordance with the logical structure they define, and they try to statistically confirm the existence of this structure. This process, which aims to determine the construct validity of the test, is also known as the determination of factor structures in test development (Bollen, 1989). Studies conducted on tests that measure psychological constructs are carried out at the level of the subdimensions and items, rather than the whole test in order to provide reliable and valid results from the scores obtained from the test. Therefore, it is essential to examine the psychometric properties of the tests included in examinations, such as AYOS taken by international students. Considering the psychological constructs measured by AYOS and the characteristics of the participating students, it is necessary to investigate whether the test structure as a whole varies according to student characteristics.

Tests that measure psychological constructs attempt to elucidate the implicit traits underlying these structures. It is important that these tests are able to measure the intended properties without being affected by the characteristics of the group, and this is addressed by the concept of measurement invariance in the literature. The items taking different roles from their core purpose by being affected by factors other than traits which the items aim to measure are analyzed through the aspects of differential item functioning (DIF) in the literature. (Mellenbergh, 1989; Meredith, 1993; Millsap, 2011; Shealy & Stout, 1993; Vandenberg & Lance, 2000). With the widespread use of tests measuring psychological traits at the international level, measurement invariance and differential item functioning have gained an important place in bias studies in the subjects of psychometrics and measurement science.

In 1966, the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement Education (NCME) established standards for educational and psychological testing to measure the quality of tests and test applications, and create test ethics. In the report on the *Standards for Educational and Psychological Testing* revised in 1974, 1985, 1999 and lastly in 2014, test fairness/fairness as a lack of bias is used as a technical term to refer to *individuals at the same level in terms of the measured construct obtaining the same score from a test, independent of their group* (AERA, APA, & NCME, 2014).

Taking many of the standards included in the above-mentioned guidelines as criteria, AYOS developed at Ankara University aims to attain an international level in performing fair testing for individuals from different subgroups with different language and cultural characteristics because it is essential for the international validity of such tests that the scores obtained are independent of individuals' demographic and cultural characteristics, and that they are fair and without bias. This is also a pre-requirement for comparing the scores from different individuals. Therefore, for a test developed to measure any psychological construct, determining whether the test as a whole is affected by the characteristics of individuals necessitates the evaluation of measurement invariance and study of differential item functioning. The findings of such studies are important for not only the test development process but also to ensure the elimination of conditions that threaten the validity of the test results. Determining whether the test used in AYOS and the items included this test taken by individuals from different countries of the world, differs in terms of demographic and cultural characteristics requires measurement invariance and differential item functioning studies.

Accordingly, the overall objective of the current study was to determine the factor structure of the Basic Learning Skills Test (TOBT) included in the AYOS and investigate whether this test has measurement invariance and whether it shows differential item functioning of the items of the test by country and gender.

Referring to this aim, answers to the questions below were sought:

1. What is factor structure of TOBT according to Germany, Azerbaijan and Iran?
2. Does TOBT have measurement invariance in respect to,
   a) countries (Germany, Azerbaijan and Iran) and
   b) gender?
3. Do the items in TOBT show differential item functioning in reference to,
   a) countries (Germany, Azerbaijan and Iran) and
   b) gender?

Psychological measurement tools, whether verbal-based or verbal-free, are influenced by demographic characteristics, such as age and gender, as well as cultural traits; e.g., ethnicity and country of the individual (Messick, 1989). This raises doubts concerning the validity of decisions based on scores obtained from these tools. It is therefore necessary to ensure that a measurement tool measures the intended psychological construct independently of these characteristics. For this purpose, the structural equivalence and differences that may arise from the groups should be examined for each group of the test responders. The current study is important for the examination of the structural equivalence of TOBT used in AYOS in terms of the groups of individuals taking the test and the inter-group differences.

Another important aspect of the study is that it is the first research that investigated one of the tests used for foreign student selection and placement in Turkey, an application that started in the 1980s. Turkey's efforts in increasing the standards of higher education (Bologna process, the European Union exchange programs, accreditation work, etc.) have increasingly attracted the attention of international students. The reliability and validity of the decisions based on the test results used for this purpose will encourage overseas students with advanced levels academic achievements to participate in higher

education programs in Turkey. Thus, this study is also significant because it is the first to examine the equivalence feature of a test and the items found in the test included in an examination for foreign student selection and will constitute an example for future studies to be undertaken in this area. Besides, the findings obtained from the study are significant to eliminate factors threatening both test development processes and measuring results.

Method

Research Model

This research aimed to describe the psychometric features of a test as presented, and therefore was of the screening type of descriptive research model. The description process undertaken in this research intended to draw attention to and strengthen the meaning of the information related to the items in the test, and refer to this information without changing its nature (Wolcott, 1994). Thus, the psychometric features of the test, which is the subject of this research, were described as they were in their own conditions.

Population and Sample

The population of the study comprised 2134 individuals (944 female, 1190 male), who participated in 2013 AYOS, and the research was conducted with individuals selected from this population. Depending on the subject of the study, in order to conduct in-depth research and obtain rich information relevant for the purpose of the research, a purposive sampling method of criterion sampling was employed. Since this research also involved intercultural comparisons, culture (country) was taken into account as a criterion in sample selection. Germany, Azerbaijan and Iran were selected as sample countries considering that they represent different cultures and a high number of students participated in 2013 AYOS from these countries. Table 1 presents the distribution of the sample group by country.

Table 1. Distribution of the Sample Group by Country and Gender

Country	Female		Male		Total	
	Frequency (f)	Percentage (%)	Frequency (f)	Percentage (%)	Frequency (f)	Percentage (%)
Germany	129	66.15	66	33.85	195	17.57
Azerbaijan	157	28.49	394	71.51	551	49.64
Iran	172	47.25	192	52.75	364	32.79
Total	458	41.26	652	58.74	1110	100.00

Of the 1110 individuals selected for the sample group, 458 (~41%) were female and 652 (~59%) were male. The number of individuals from each country was 551 for Azerbaijan (~50%), 364 for Iran (~33%), and 195 for Germany (~17%). Guilford (1954) stated that in the analysis of latent variable models, the minimum number of samples should be 200. Accordingly, the number of samples in the current study was considered sufficient for such analysis.

Data Collection

The data of the study consisted of the individuals' responses to the AYOS TOBT test administered simultaneously in a single session in three different examination centers located in Turkey (Ankara), Germany (Cologne), and Azerbaijan (Baku). The responses, obtained from two different test booklets presenting the same items in a different order, were merged according to the order in Booklet A and converted to the 1-0 scoring matrix to be made ready for analysis.

Data Collection Tool

The data collection tool used within the scope of the research was an academic skills test, TOBT, consisting of two subsections and a total of 100 items. The first 60-item part of the test utilizes letters,

numbers and figures to measure mental characteristics, such as analytical thinking, reasoning, and abstract and spatial thinking. The remaining 40 items aim to determine numerical thinking skills, which require the use of mathematics and geometry knowledge (ANKUDEM, 2012). In this study, only the first 60 items of TOBT were evaluated considering the similarity of the traits measured and the need to limit the number of items.

In the preparation of the items, the characteristics to be measured by the test were taken into account, and these items were grouped by the experts developing the test according to these characteristics. A table of specifications was prepared to show the distribution of the items according to the measured characteristics. Accordingly, table of specifications constitutes some traits such as finding the part fitting/unfitting for the whole, making inferences from figures, tables and suchlike information given in a whole, finding the rule or figure based on relations considering the connections and tables. It also aims to measure the traits mentioned above. The test developed should contain at least three times the number of items to be included in the final version. The final version of the test was achieved by selecting the items considered to best measure the corresponding psychological trait in the table of specifications. This also facilitates the management of both content and construct validity of the test through a rational and logical process.

Table 2 presents the descriptive statistics of the population and sample group including the reliability values in relation to the scores obtained from TOBT.

Table 2. Descriptive Statistics of the Population and Sample Group ($k = 60$)

Group	N	Arithmetic Mean	Standard Deviation	KR-20 Reliability
Population	2134	45.05	9.47	0.91
Sample	1110	45.48	9.01	0.90

The arithmetic mean and standard deviation values belonging to the population taking the test and the sample group selected for analysis in this research were similar, indicating that the sample represented the target population. In addition, the KR-20 reliability values were high and did not differ between the population and sample group. The reason to prefer this reliability value is that difficulty index of the items included in the test is different from each other (Kuder & Richardson, 1937).

Data Analysis

In this research, first, the factor structure of TOBT was determined by Latent Class Factor Analysis (LCFA) in terms of the sample consisting of individuals from Germany, Azerbaijan, and Iran. Then, whether or not TOBT had measurement invariance according to country and gender was investigated using Simultaneous Latent Class Analysis (SLCA). These methods were used because the structure of the test, from which the data was obtained, was at the rank-order level, and the test items had binary responses. For the analysis, the Basic and Advanced/Syntax versions of the Latent GOLD 5.1 program were used (Vermunt & Magidson, 2013a, 2013b, 2013c).

In order to determine whether the items in the test show differential item functioning according to country and gender, Rasch model developed referring to Recursive Partitioning Analysis was used. Data were obtained by making use of 0.12-1 version of package psychotree available in R programming which is a statistical open source software programme (Zeileis, Strobl, Wickelmaier, & Kopf, 2011).

Brief information about analysis used in the research was provided below.

Latent Class Factor Analysis: LCFA defines factors with two or more categories formed by the combination of variables with the same source of co-variability. In the literature, the models that incorporate traditional Latent Class Analysis (LCA) and Factor Analysis (FA) are referred to as LCFA models (Magidson & Vermunt, 2001, 2003a, 2003b, 2004; Vermunt & Magidson, 2000, 2005a).

Bartholomew, Steele, Moustaki, and Galbraith (2008) described the following similarities between LCA and LCFA in terms of their purpose of exploring and explaining the relationships between observed variables (1), determining whether these relationships can be explained by fewer latent variables (2), and obtaining each individual's score for each latent variable (3).

LCFA performs better with fewer variables compared to FA, which requires at least three continuous levels of variables that can only define a single factor. In LCFA, three dichotomous variables can similarly generate a factor; however, these models are not limited to dichotomous variables and are able to identify additional factors by including covariates in the model (Magidson & Vermunt, 2003a).

Apart from the purpose-related similarities and the differences in the structure of data matrices between the two methods, the application of FA has four important limitations compared to LCFA: (1) all variables being continuous, (2) assumptions of multivariate normality and linearity, (3) latent variables (factors) being at the level of a uniform or proportional scale, and (4) the researcher being required to choose one of the possible rotation options in the interpretation of analysis; i.e., producing a single result (Magidson & Vermunt, 2003b). With the combined approach in LCFA, these problems can be resolved within the limits of the method.

As in other latent class models, parameter estimation in LCFA is performed using the Maximum Likelihood (MLH) method employing the Expectation Maximization (EM) or Newton-Raphson (NR) algorithms or the combination of these two. Therefore, the only disadvantage of LCFA occurs during the interpretation of the estimated parameters because this method does not have easily interpretable parameters, such as factor loading values, factor-item correlations, factor correlations, or covariances that are included in FA. To overcome this problem, Vermunt and Magidson (2005a) recommended adopting a linear approach to the use of MLH estimates obtained from the LCFA model. This approach allows obtaining outputs similar to those in FA; thus, latent factor structures underlying a construct to be measured can be determined using a more reliable but non-linear factor analytical model.

Simultaneous Latent Class Analysis: In order to compare the latent structures between groups, Clogg and Goodman (1984, 1985) proposed the Multi-Group Latent Class Analysis (MGLCA) approach, which allows analyzing the latent structures based on observable variables of two or more groups. Using this method, not only can the latent traits of multiple groups be compared but also the latent traits of a single group can be comparatively analyzed at different times. This also makes it possible to perform an evaluation on the measurement invariance between the groups.

The models included in this method, often referred to as SLCA or MGLCA models in the literature, are developed using three different parameterization parameters: probabilistic, log-linear, or logistic (Kankaras & Vermunt, 2014). Only probabilistic parameterization can be used when the indicator and latent variable(s) are at the classification level; therefore, in the current study, the SLCA model in this study was explained using the probability parameters. Accordingly, the measurement invariance between the groups was examined by equalizing the probability of item responses in a given class (Clogg & Goodman, 1985).

In case of differences between the groups, SLCA provides the researcher with information regarding where these differences occur and how to interpret them (Collins & Lanza, 2010). Therefore, when performing inter-group comparisons, the researcher is required to determine whether the number of latent classes is equal, the conditional probabilities are equal and the latent class structures are similar. When differences are determined in one or more of these cases, it is necessary to identify and explain the reason for this situation.

The analysis of measurement invariance by SLCA is performed at different levels of homogeneity because the comparison of latent structures between the groups results in a number of possible outcomes, with the groups being completely discrete (heterogeneous model), partly similar (partially homogeneous), or completely homogeneous (Kankaras, Moors, & Vermunt, 2011). In the selection of a model, as in other latent class analyses, chi-square test-square statistics and information criteria are used. Similar to other methods for measurement invariance analysis, the aim of SLCA is to select the model producing the highest invariance.

The analysis of measurement invariance by SLCA is based on the comparison of models with equivalence at various levels, and therefore parameter estimation for model comparisons is undertaken using the MLH method employing the EM and NR algorithms, as in traditional LCA. Similarly, in the comparison of models, the degree of agreement between the models is determined using information criteria, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), as well as chi-square statistics.

Kankaras, Moors, and Vermunt (2011) emphasized that LCA had several advantages and offered a very valuable approach for testing measurement invariance in intercultural studies. The researchers provided three important reasons for this situation: First is that the items included in the measurement tools used by researchers mostly have discrete (rank-order or classification) response categories, and LCA is able to define the latent structures based on the relations between such discrete variables. Second, unlike the frequently used Multiple-Group Confirmatory Factor Analysis (MGCFA), a latent class model approaches latent variables as if they were at the classification level; e.g., allowing typological classification based on a group of categorical variables. Similarly, in order to investigate the scalability of categorical variables in a data set, LCA can also approach these variables as if they were at the rank-order level. The third important reason is that LCA-based methods offer a more flexible alternative to MGCFA and multiple-group IRT approaches, which are more commonly used but require stronger distribution assumptions.

One of the advantages of the use of SLCA in cross-cultural comparative studies is that the analysis is more flexible than in traditional approaches. For example, in order to make comparisons between groups in MGCFA, there should be at least two items with measurement invariance under each factor, whereas SLCA allows such comparisons even if the number of items that do not have measurement invariance is high. In addition, in order to interpret the findings obtained from MGCFA in terms of measurement invariance, at least a partial invariance level should be provided, in contrast to SLCA that does not require the satisfaction of structural invariance (Kankaras & Moors, 2009). However, at this point, it is worth mentioning that researchers are mostly interested in complete measurement invariance, and that they want to be able to achieve the full comparability of groups. For this purpose, structural equivalence must also be ensured in SLCA (Hagenaars & McCutcheon, 2002).

Recursive Partitioning Analysis based on Rasch Model: Zeileis, Hothorn, and Hornik (2008), identifying DIF within the frame of Rasch model, proposed a new statistical approach named Model Based Recursive Partitioning (MBRP) including tests related to both predetermined and all potential groups without making evaluation difficult. Parameter instabilities in the established model can be detected with the help of the approach. Similar to latent class analysis or mixture models, the main idea underlying this approach depending on identifying the groups which parameters in the model differ, is to test all groups consecutively, searching for all potential sources causing DIF. In recursive partitioning, groups aren't identified by a latent factor as in latent class model but defined by common

observed combinations of variables, using intuitive approach[1]. Therefore, model based partitioning, as an alternative to latent class or mixture model, is intuitive but offers options easy to evaluate.

In the comparative studies among groups, the approach, which is fairly new to determine DIF, is based on model based recursive partitioning method used structural alteration test adapted from econometrics. Model based recursive partitioning is strongly related to classification which variable field has recursively partitioning and the method of regression tree in order to identify the group whose response variable, at a perpetual or categorical level, has different values. In this method having a semi parametric approach, instead of values associated with single response variable, parameters of a parametric model which varies among groups plays a role. Such parameters can be not only item parameters of Rasch model varying among groups but also slope and intercept parameters of linear regression modelling (Strobl, Kopf, & Zeileis, 2015).

The purpose of analysis making use of model based recursive partitioning is to classify data matrix into subgroups (classes) showing a homogeneous structure in itself. Each subgroup is called node. Firstly, these subgroups are identified according to covariates (such as age and gender). Then, these nodes begin dividing as it occurs in classification and regression tree until they have an identical structure. This tree is defined as Rasch tree and the tree carries critical values (leaves) linked to covariates on its branches. The process continues until the variance takes the minimum value for each node and the variance among nodes takes the maximum value (Kopf, Augustin, & Strobl, 2010).

In this context, it is aimed to provide model compatibility by composing a Rasch tree whose nodes and leaves are correlated with a proper model (such as likelihood model or linear regression model). The main idea at this point is to relate each node to a single model. The sequential steps (algorithm) used to compose a Rasch tree are given below (Strobl, Wickelmaier, & Zeileis, 2011; Zeileis et al., 2008)

1. For all individuals taking part in the current subsample, item parameters in common are estimated.
2. According to each available covariate, the stability of item parameters is evaluated.
3. If the item parameters do not convey a stable structure on the significance level (if there is significant instability), the sample is split at its cutpoint according to covariates until model compatibility improves.
4. Steps from 1 to 3 are repeated recursively for the obtained subsamples until the stability is achieved on the significance level (or when the subsample is too small). In other words, partitioning continues till reaching to stopping criteria.

[1] The approach in which an optimal solution is tried to be obtained by making a complex problem simpler in an exploratory and experiential way that does not require an experimental proof.

Results

Before moving on to the factor structure analysis of TOBT using LCFA, as an a priori hypothesis to provide clues about this structure, it was investigated whether the classes formed based on the responses of individuals to the test showed a tendency to have a scale structure. It was expected that according to the classes formed, the value of any TOBT item in any category would consistently increase while the value of that item in another category would consistently decrease. This is interpreted as an indication that the test may have a scale structure (Vermunt & Magidson, 2005b). For this purpose, four[2] different latent class models, each with a different number of classes, were estimated. Table 3 presents the values for the estimated model.

Table 3. Latent Class Model Estimation for TOBT

Model	LL	BIC (LL)	AIC (LL)	Num. of parm.	L^2	df	*p-value*	Class. Error
1-class	-31077.55	62575.82	62275.09	60	46662.88	1050	0.00	0.00
2-class	-28251.32	57351.10	56744.64	121	41010.43	989	0.00	0.02
3-class	-26876.07	55028.35	54116.15	182	38259.94	928	0.00	0.01
4-class	-26141.19	53986.33	52768.39	243	36790.18	867	0.00	0.02

Table 3 presents models with a number of classes from 1 to 4, their fit measures, number of parameters, and degrees of freedom values. Prior to the model selection process, it should be examined whether the degrees of freedom for all the models established has a positive value since the degrees of freedom with a negative value leads to an identification problem in the estimation process of latent class models (McCutcheon, 1987). In the current study, these values were positive, and therefore the number of parameters in the model was considered equal to the column rank. The model evaluation process continues with the interpretation of fit measures, of which the L^2 statistic indicates the amount of the relationship between the unexplained variables after the model estimation. The lower the L^2 value, the better the model would fit into the data. The L^2 value should be considered when determining the number of latent classes since the p-value in Table 3 shows the significance level of each model with the assumption that the L^2 statistic follows the chi-square distribution (Vermunt & Magidson, 2005b). In model selection processes, where the L^2 statistic is used, model comparisons begin by establishing a hypothesis about model acceptance using the two most complex models. If there is a significant difference in the hypothesis test, then the fitness of one model is better than that of the other. This process continues until all models are tested or there is a significant difference between them (Lin, 2006). Hypotheses related to model acceptance are established as follows:

H_0: The model is appropriate.

H_1: The model is not appropriate.

In the current study, the significance level of all created models was below the pre-defined level of significance *($p<0.05$)*; thus, hypothesis H_0 was rejected. This shows that four-class model which was created is inappropriate. However, since the L^2 statistic is affected by the sample size, it will not produce significant results when the sample does not follow the chi-square distribution. Therefore, using only the L^2 value to evaluate the model fit is not very feasible due to the sampling-dependent structure and chi-square distribution requirement of this statistic (McCutcheon, 2002). Therefore, it would be better to conduct the fitness of model assessments with other measures; e.g., information criteria.

Similar to the case in the L^2 statistic, for the information criteria used in the model selection, a lower value indicates a better model fit for the data. The BIC and AIC values calculated according to the log-likelihood value (Table 3) were examined, and it was observed that as the number of classes increased, the BIC and AIC values decreased. However, it is important to note that this decreasing tendency does not mean that models with a higher number of classes should be selected. This means that if the model estimation is continued with models containing more classes, lower information

[2] In LCA, as a general starting rule, classes are estimated from 1 to 4 (Vermunt & Magidson, 2005b).

criteria values will be obtained, but it will be difficult to interpret the model due to the parallel increase in the number of parameters involved. Therefore, a model with too many parameters will lose its conservative nature. Accordingly, Lin (2006) suggested that a model with fewer parameters should be preferred for easy interpretation of the results. Thus, in the current study, the number of parameters was also taken into account in the model fit assessments for TOBT. The results revealed that the four-class model was appropriate in terms of both the BIC and AIC values and the number of parameters.

After the model selection, it is possible to use bootstrapping to re-test the significance value of the L^2 statistic that has diverged from the real significance due to the failure of the sample to comply with the chi-square distribution. This allows for a multi-dimensional decision-making approach in the model selection process. In order for the L^2 statistic to be recovered by the bootstrap method, an appropriate model must first be selected to compare the significance value of the selected model and that obtained from the sample randomly generated from the current sample. This can be easily achieved using the *Bootstrap L²* command in the Latent GOLD program. The recovered significance value must exceed the level of significance determined differently from the previous value; i.e., it should not be greater than *0.05*. In this study, after undertaking the bootstrapping approach, the L^2 statistic value was found to be *0.99*[3]. Based on the *p value* being *>0.05*, the four-class model was considered to be simpler, and therefore more appropriate. Table 4 shows the classification statistics for the four-class model obtained for TOBT.

Table 4. Classification Statistics for the Four-Class Model

Classification Error	0.02
Reduction of Errors (λ)	0.97
Entropy R^2 Value	0.97
Standard R^2 Value	0.96

The classification error related to the four-class model was very low (2%). In this case, it can be stated that this model performed classification at an accuracy of 98%. The reduction of errors, entropy R^2 and standard R^2 values, showing the estimates of how accurately the observed variables were assigned to latent classes, were calculated as 0.97, 0.97 and 0.96, respectively, indicating that the model estimations were very successful. These values being very close to 1.00 means that the estimation was very successful.

In explaining the four-class model obtained for TOBT, two important parameters of the latent class model, namely latent class probabilities and conditional probabilities, were utilized. These values are also important because they offer clues about the factor structure of TOBT. Thus, the latent class probabilities, providing information about the number of latent classes and the relative size of these classes, were calculated. The distribution of all the individuals in the current sample to the four classes obtained from TOBT were as follows: 0.31 for Class 1 (π_1^X), 0.30 for Class 2 (π_2^X), 0.21 for Class 3 (π_3^X), and 0.18 for Class 4 (π_4^X), with their sum being equal to 1.00. When the relative magnitudes of latent class probabilities differ, the distribution of the classes also varies in terms of the measured characteristics. In this study, the balanced distribution of individuals among the four classes suggests that there was no sharp differentiation concerning the measured characteristics.

The probability of an individual in any of the classes responding correctly or incorrectly to an item in TOBT is handled by conditional probabilities, which show the differences between the response patterns that differentiate classes, thus providing information about the nature of the latent variable. The sum of these probability values equals 1.00, depending on the response levels of each item. This also applies to the conditional probability values of individuals in other classes. The four classes

[3] Since the method randomly generates a sample every time, this value will change in the next attempt; however, it will continue to be very similar for each estimation (Vermunt & Magidson, 2005b).

demonstrating the profile of individuals in terms of their correct and incorrect responses to the TOBT items and the conditional probability values for the responses of the individuals in these classes to some of the TOBT items are given below.

Class 1 represents individuals with a high probability of correctly responding to the items in TOBT. The expected situation for this class is lower conditional probability values for level 0 (incorrect response) and higher conditional probability values for level 1 (correct response). For example, the probability of an individual in Class 1 to correctly respond to item 14 ($\pi_{2|1}^{14|X}$) is calculated as 1.00 and incorrectly as 0.00 ($\pi_{1|1}^{14|X}$), while the probability of the same individual responding to items 20, 32, and 35 correctly ($\pi_{2|1}^{20|X} = \pi_{2|1}^{32|X} = \pi_{2|1}^{35|X}$) is 0.99 with an incorrect response to these items having the probability of 0.01 ($\pi_{1|1}^{20|X} = \pi_{1|1}^{32|X} = \pi_{1|1}^{35|X}$).

The individuals in *Class 2* are similar to those in Class 1 in terms of their patterns of response to the TOBT items. However, they differ from Class 1 in that they had a lower probability of providing a correct response, especially to items 46 to 50. For example, an individual in Class 2 had high conditional probability values of 0.99 for item 12 ($\pi_{2|2}^{12|X}$), 0.90 for item 28 ($\pi_{2|2}^{28|X}$), and 0.93 for item 41 ($\pi_{2|2}^{41|X}$), but lower conditional probability values ranging from 0.03 to 0.16 for items 46 to 50.

Class 3 differed from Classes 1 and 2 concerning the probability of correct responses to the TOBT items. This class had a significantly lower probability of correctly responding to items 17 to 21 (range of probability: 0.01 to 0.03), compared to the response patterns of the individuals included in the remaining classes. Although the probability of the individuals in Class 3 to provide a correct response was lower for all items, it was found that their response tendency was similar to that obtained from the other classes.

The distribution of the correct responses of the individuals in *Class 4* to the items in TOBT was similar to the other classes; however, Class 4 was observed to have a much lower probability of responding to items than the other classes. For example, the conditional probability values were calculated as 0.35 for item 5, 0.31 for item 25, 0.15 for item 30, and 0.12 for item 46.

When the conditional probability values are examined as a whole, the individuals in Classes 1 to 4 generally have a similar response pattern in terms of accuracy, whereas concerning the responses to items 17-21 and 46-50 in particular; the conditional probabilities significantly differ between the individuals in different classes. For the remaining items, the correct responses of individuals were generally similar between Classes 1, 2, and 3, but Class 4 provided different results compared to the other three classes.

Another step to be taken in LCA is the calculation of the percentages of individuals for each class, who responded correctly or incorrectly to the TOBT items. Determination of classification percentages also provides clues about the factor structure of the test prior to the factor structure analysis using LCFA. Contrary to the conditional probabilities in which column percentages are examined, classification percentages are related to the percentages in the rows. Therefore, the sum of the rows in each item level equals 1.00. The expected situation in the distribution of percentages among classes is a consistent increase in the percentage value of an item at any level and a consistent decrease in the value of the same item at another level (Vermunt & Magidson, 2005b).

Accordingly, in terms of the classes included in this study, the expectation was that the percentage values would gradually decrease from Classes 1 to 4 for level 1 (correct response) in any item and gradually increase from Classes 1 to 4 for level 0 (incorrect response) in the same item. This is the ideal condition for a one-dimensional test to measure psychological constructs. According to the response levels of the items, the percentage distributions of the classes were in line with the expected finding, but there were also certain differences. This a priori assessment on TOBT before performing LCA for factor structure analysis revealed that the individuals who completed the test had different response patterns, indicating the multidimensionality of the test.

Factor Structure of TOBT: To determine the factor structure of TOBT, models with a different number of factors were established using LCFA. As a preliminary model, the findings of the four-class model obtained by LCA were used, taking into consideration that the factor structure of the test might have more than one dimension, although this was not a pre-requirement. The model selection process was initiated by creating a two-factor preliminary model. The values estimated for TOBT using latent class factor models are given in Table 5.

Table 5. Latent Class Factor Model Estimation for TOBT

Model	LL	BIC (LL)	AIC (LL)	Num. of parm.	L^2	df	*p-value*	Class. Error
2-factor	-26381.65	54039.51	53127.30	182	37271.09	928	0.00	0.00
3-factor	-25488.28	52680.50	51462.55	243	35484.34	867	0.00	0.00
4-factor	-25950.83	53033.34	51509.66	304	36790.18	806	0.00	0.00

Table 5 shows the models with a number of factors ranging from 2 to 4, and the corresponding values for the fit measures, number of parameters and the degrees of freedom for these models. It was observed that the degrees of freedom for all models had positive values, suggesting that the number of parameters in a model was equal to the corresponding column rank; thus, all models were identifiable.

Concerning the fit measures for model estimations, both the information criteria and the L^2 value decreased in the three-factor model established after the preliminary two-factor model. Since the significance value of the L^2 statistic for all models was below the predefined significance threshold ($p<0.05$), hypothesis H_0 was rejected for this statistic. This shows that the established model is inappropriate. However, as mentioned before, the L^2 statistic is affected by sample size and cannot produce significant results when the sample studied does not follow the chi-square distribution. Therefore, it would be better to conduct model fit assessments using other measures, such as information criteria. Bearing this in mind, the BIC and AIC values in Table 5 were examined, and it was observed that as the number of factors increased, the information criteria values gradually decreased, but they slightly increased in the four-factor model. Hence, it can be stated that the three-factor model obtained for TOBT was appropriate in terms of both BIC and AIC values. Table 6 shows the classification statistics for the three-factor model of TOBT.

Table 6. Classification Statistics for the Three-Factor Model

	Factor 1	Factor 2	Factor 3
Classification Error	0.00	0.04	0.03
Reduction of Errors (λ)	0.99	0.92	0.87
Entropy R^2 Value	0.99	0.86	0.86
Standard R^2 Value	0.99	0.89	0.88

The classification error of the three-factor model was very low for all factors (0%, 4%, and 3% for Factors 1 to 3, respectively). In other words, the three-factor model had classification accuracy of 100% for Factor 1, 96% for Factor 2, and 97% for Factor 3. The reduction of error values was 0.99, 0.92 and 0.87 for Factors 1 to 3, respectively. The entropy R^2 value was 0.99 for Factor 1 and 0.86 for Factors 2 and 3. The standard R^2 value was 0.99 for Factor 1, 0.89 for Factor 2, and 0.88 for Factor 3. These values being very close to 1.00 means that the estimation was very successful.

For the three-factor structure of TOBT, the factor loadings of the items ranged from 0.15 to 0.94 for Factor 1, 0.08 to 0.72 for Factor 2, and 0.14 to 0.53 for Factor 3. The factor loadings of items are reported as correlation coefficients indicating the square root value corresponding to the R^2 or common variance of the related item. These numbers can be interpreted as either linear regression coefficients or factor loadings similar to FA.

The R^2 values of the items ranged from 0.01 to 0.89. The R^2 value, explained by the model, shows how much each item contributes to the whole test. This value refers to standard R^2 measurements for sequential, continuous and countable variables. As was the case in this study, the interpretation of R^2 differs between variables at the classification level. Accordingly, each R^2, treated as a dichotomous response variable for each category, is expressed as the Goodman-Kruskal τ-b coefficient, representing the weighted average of the measurement. R^2 is similar to the variance explained in the variance analysis and item covariance in the factor analysis (Vermunt & Magidson, 2005b). Thus, the values obtained were interpreted using the general analysis logic of latent class models in terms of the factor in which each item was included and their overall contribution to the test, i.e., the psychological construct that was measured. In order to name the factors obtained, opinions were elicited from experts in generating items based on numbers, figures, and letters and academicians in the measurement and evaluation field. Table 7 presents the distribution of the psychological traits measured by items fitting the three named factors.

Table 7. Distribution of the Psychological Traits Measured by Items According to the Factors

Factor name	Psychological Trait	Item number	Total
Establishing part-whole relationships	Finding the part that is/is not suitable for the whole	1, 2, 17, 18, 19, 20, 21, 22, 23, 24, 25, 31	12
Making inferences	Reaching conclusions based on figures, numbers, and similar information presented as a whole	4, 5, 6, 9, 11, 26, 29, 32, 33, 46, 47, 48, 49, 50	14
Analytical thinking	Identifying the rule or number, on which the relationships are based, by utilizing the figures	3, 7, 8, 10, 12, 13, 14, 15, 16, 27, 28, 30, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60	34
Total			60

According to the results given in Table 7, TOBT aimed to measure the dimensions of establishing part-to-whole relationships, making inferences, and analytical thinking. The distribution of items into factors reveals that Factor 1 (12 items) and Factor 2 (14 items) had a similar number of items, whereas Factor 3 (34 items) had a higher number of items compared to the other two factors. If a test is not one-dimensional, each emerging subdimension is expected to measure the trait defined in it. Therefore, the correlation values between the subdimensions/factors are expected to be negative or low. In the current study, the following correlation values were obtained: -0.23 between Factor 1 and Factor 2, 0.25 between Factor 1 and Factor 3, and -0.30 between Factor 2 and Factor 3. Therefore, as expected, all correlation values were low and two were negative, indicating that each factor was homogeneous in itself. Thus, it can be stated that each factor contributed to the actual psychological construct evaluated by TOBT despite the differences between the factors in terms of the traits measured.

Examination of the Measurement Invariance of TOBT: SLCA was used to determine whether TOBT had measurement invariance in terms of the country and gender variables. Analysis was performed at different homogeneity levels of the latent class models. The first of these levels was the heterogeneous model, in which all the parameters were unconstrained in terms of the groups compared. The second was a partially homogeneous model with certain constraints. The last was a homogeneous model, in which the group variable had no effects (direct or common). The models used in country and gender comparisons were explained in a hierarchical order from the heterogeneous to the homogeneous model. Table 8 presents the results of analysis conducted to determine whether TOBT had measurement invariance according to the selected countries.

Table 8. Estimation of the Measurement Invariance of TOBT by Country

Model	LL	BIC (LL)	AIC (LL)	Num. of parm.	L^2	df	*p-value*	Class. Error
Heterogeneous	-26490.33	56809.28	54072.67	546	39718.17	564	0.00	0.01
Partially Hom.	-26643.30	55432.31	53898.60	306	40024.11	804	0.00	0.01
Homogeneous	-26876.07	55028.35	54116.15	182	40489.66	928	0.00	0.02

Table 8 shows the models with different levels of homogeneity, the fit measures for these models, number of parameters, degrees of freedom, and the significance value of the L^2 statistic. Before continuing to the model selection process, the degrees of freedom was examined and found to be positive for all models. This means that the number of parameters in each model was equal to the corresponding column rank, and there was no problem of identification in any of the models. Based on the L^2 statistic of all the established models being below the pre-defined significance threshold ($p<0.05$), it can be interpreted that the established model is not appropriate. However, as mentioned before, the information criteria were used in the model evaluation process due to the chi-square distribution requirement and sample-dependent nature of the L^2 statistic. When the information criteria values were examined, it was observed that as the homogeneity level of the model increased, the BIC values gradually decreased; however, the AIC value decreased from the heterogeneous model to the partially homogeneous model, but increased in transition to the homogeneous model. This suggests that the partially homogeneous model was appropriate. In other words, the test partly differed between the three countries, providing an indication that the items in the test were partially affected by culture. However, previous researchers reported that for a sample size of about 1000, BIC produces more consistent results than AIC (Dias, 2006; Kankaras, Vermunt, & Moors, 2011; Moors & Wennekers, 2003; Morren, Gelissen, & Vermunt, 2011). For this reason, the BIC values adjusted for sample size were taken into consideration in the model selection, and the homogeneous model was selected as the most appropriate.

In light of these determinations, it can be stated that TOBT had measurement invariance for Germany, Azerbaijan, and Iran; i.e., TOBT had the same factor structure and measured the same psychological traits across all three countries. Table 9 shows the classification statistics for the homogeneous model.

Table 9. Classification Statistics Obtained According to Countries for the Homogeneous Model

Classification Error	0.02
Reduction of Errors (λ)	0.97
Entropy R^2 Value	0.96
Standard R^2 Value	0.96

The classification error related to the homogeneous model was very low (2%); i.e., the classification accuracy of the model was 98%. The reduction of error, entropy R^2 and standard R^2 values were 0.97, 0.96, and 0.96, respectively, confirming that the model estimation was successful.

Table 10 shows the results of model estimation undertaken to determine whether TOBT had measurement invariance with respect to gender.

Table 10. Estimation of Measurement Invariance of TOBT by Gender

Model	LL	BIC (LL)	AIC (LL)	Num. of parm.	L^2	df	*p-value*	Class. Error
Heterogeneous	-26752.05	56056.51	54232.10	364	39493.91	746	0.00	0.01
Partially Hom.	-26824.13	55359.21	54136.25	244	39638.06	866	0.00	0.01
Homogeneous	-26876.07	55028.35	54116.15	182	39741.96	928	0.00	0.02

The degrees of freedom values of all three models being positive means that there was no problem concerning identifiability. The decrease in the BIC and AIC values of the models suggests that the most suitable model for the data obtained from TOBT was a homogeneous model. Based on these findings, it is possible to state that TOBT had measurement invariance in terms of gender; i.e., the factor structure of TOBT was the same for both female and male students and the test measured the same psychological traits in both groups. Table 11 gives the classification statistics for the homogeneous model.

Table 11. Classification Statistics Obtained According to Gender for the Homogeneous Model

Classification Error	0.02
Reduction of Errors (λ)	0.97
Entropy R^2 Value	0.96
Standard R^2 Value	0.96

The classification error of the homogeneous model was very low (2%). In other words, the model classified the data at 98% accuracy. The reduction of errors, entropy R^2 and standard R^2 values were 0.97, 0.96, and 0.96, respectively, indicating that the model estimation was very successful.

Examination of Differential Item Functioning of TOBT Items: It was determined through composing Rasch tree that whether each item taking part in TOBT shows DIF according to country and gender. However, Recursive Partitioning Analysis necessitates a starting variable to compose the tree. For the studies which DIF is evaluated in terms of two or more variables, the starting point is a significance value belonging to each variable. The smaller significance value which is calculated for each variable, R programme starts partitioning with this variable to create Rasch tree (Zeileis et al., 2008). In this research, DIF is evaluated separately in terms of country and gender variables.

To identify whether the items in the test shows DIF in terms of countries, in Figure 1 Rasch tree which is composed based on variables matrix including responses given to items and whose purpose is to divide this matrix into homogeneous sub groups is presented.

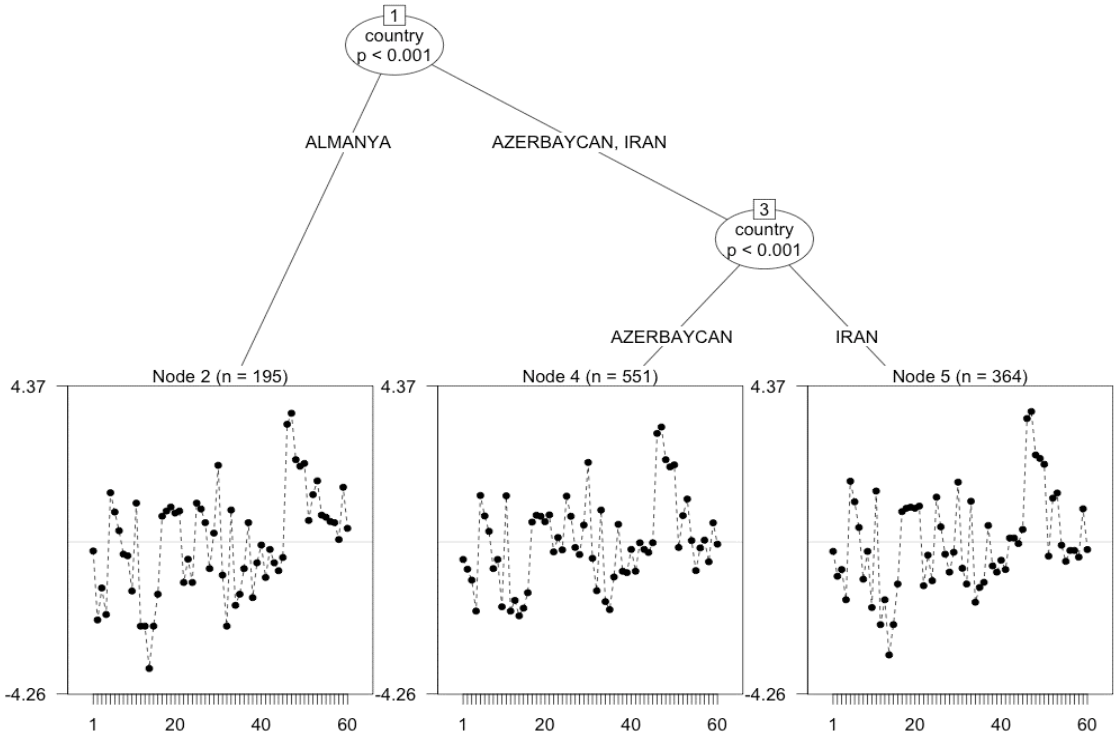

Figure 1. Rasch Tree According to Countries

The last nodes in the Rasch tree presented in Figure 1 show item parameter estimation obtained according to countries related to 60 items taking place in TOBT. Accordingly, estimation values associated with item difficulty parameters is between 4.37 and -4.26. High values indicate that items are difficult, low values mean that items are easy (Strobl et al., 2015). Also, the figure comprises significance values related to each partitioning.

In the light of given information, a differentiation in terms of countries which partitioning occurs on the significance level ($p<0.001$) is observed. Rasch tree which was composed also presents information in relation to the countries which differentiation occurs. As it is observed, on the first partitioning Germany differentiates according to two other countries (Azerbaijan and Iran), on the second partitioning Azerbaijan differentiates according to Iran. When Rasch tree is studied thoroughly, some items found in TOBT include DIF in terms of countries. Therefore, the items involving DIF are presented comparatively.

Item 23; It takes place below Factor 1 requiring to establish part-whole relationship. This item seems easy for German and Iranian individuals but relatively difficult for Azerbaijani individuals. Although the item is seen biased in favor of Azerbaijani individuals, considering the average item difficulty level, difficulty level of this item is about zero for three countries.

Item 27; It takes place below Factor 3 requiring analytical thinking. This item seems easy for Azerbaijani and Iranian individuals but difficult for German ones. Although the item is seen biased in favor of Azerbaijani and Iranian individuals, considering the average item difficulty level, difficulty level of this item is about zero for three countries.

Item 29; It takes place below Factor 2 requiring deduction (making inference). This item seems difficult for German and Azerbaijani individuals but easy for Iranian individuals. Although the item is seen biased in favor of Iranian individuals, the items are considerably close to average difficulty level with regards to individuals of three countries.

Items between 42 and 45; they take place below Factor 3 requiring analytical thinking. These items seem easy for German and Azerbaijani individuals but difficult for Iranian ones. The items cause bias in favor of Azerbaijani individuals to some extent and particularly of German individuals but the items are considerably close to average difficulty level with regards to Azerbaijani and Iranian individuals.

Item 51; it takes place below Factor 3 requiring analytical thinking. This item seems easy to Azerbaijani and Iranian individuals but difficult for German individuals. Although the item is seen biased in favor of Azerbaijani and Iranian individuals, the item is on the average difficulty level.

Items between 54 and 58; they take place below Factor 3 requiring analytical thinking. These items seem easy for Azerbaijani and Iranian individuals but difficult for German individuals. Although the items found in this group are seen biased in favor of Azerbaijani and Iranian individuals, it is obvious that the items are close to the average difficulty level.

Item 60; it takes place below Factor 3 requiring analytical thinking. This item seems easy for Azerbaijani and Iranian individuals to some extent but difficult for German individuals to a certain extent. Although the item is seen biased in favor of Azerbaijani and Iranian individuals, it is observed that the item is on the average difficulty level.

As 14 items consisting of DIF are observed as a whole, it is eye catching that difficulty and easiness levels of the items according to countries are close to, average value, zero, which shows that there is no distinctive bias regarding to the items according to countries. 46 items apart from these are observed to have approximate distribution and values.

To determine whether the items in the test shows DIF in terms of gender, in Figure 2 Rasch tree which is composed based on variables matrix including responses given to items and whose purpose is to divide the matrix into homogeneous sub groups is given.

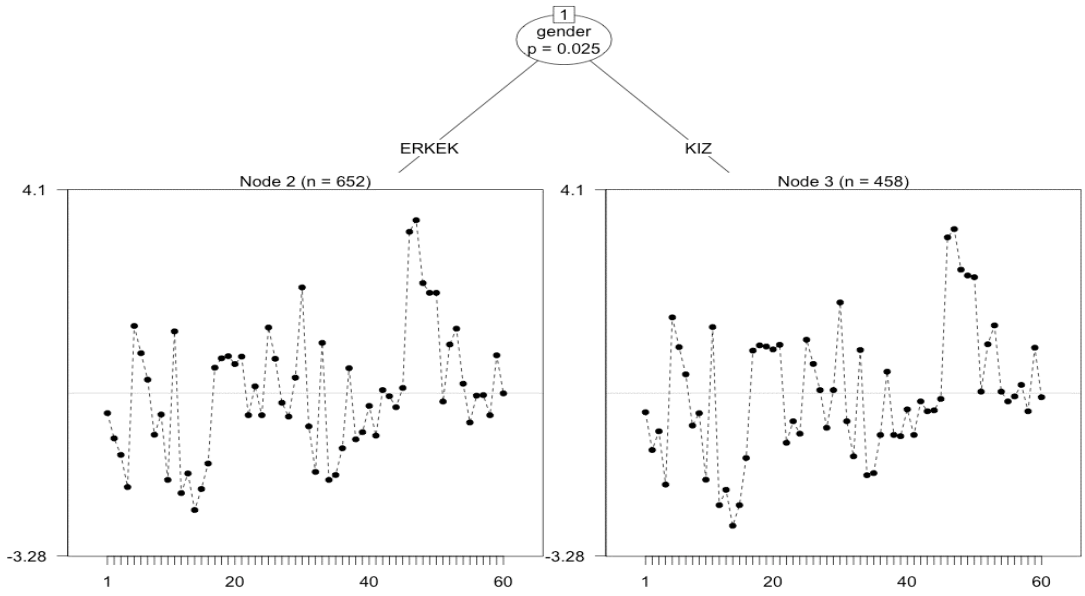

Figure 2. Rasch Tree According to Gender

The nodes found on Rasch tree in Figure 2 shows item parameter estimation related to 60 items, obtained according to gender, which is included in TOBT. Accordingly, estimation values referring to item difficulty parameters are between 4.10 and -3.28. The evaluations carried out about the circumstances which the values are high and low are similar to the ones related to countries.

In this context, on the significance level (*p>0.01*), a differentiation in terms of female and male students exposed to partitioning isn't observed. In spite of the fact that the Rasch tree which was obtained indicates that some items (for example 23, 42, 45, 51, and 57) found in TOBT show DIF regarding to gender, p value which is not on the significance level in respect to partitioning demonstrates that items do not include DIF in terms of gender.

Apinyapibal, Lawthong, and Kanjanawasee (2015) emphasized that on the account of Rasch tree obtained from Recursive Partitioning Analysis, both predefined and latent groups are able to be determined in terms of DIF and item groups can be identified by means of results shown in a diagram which is easily comprehensible and an output of the method.

Discussion and Conclusion

The examination of the overall results of the study shows that the 60 items in the TOBT test included in AYOS 2013 had a three-factor structure. The analysis of the distribution of items according to the factors revealed that there were 12 items in the factor related to establishing part-to-whole relations, 14 items in the factor on making inferences, and 34 items in the analytical thinking factor. The reason for the higher number of items in the last factor was that it contained items requiring the respondents to combine or separate components or identify the relations between them. Analytical thinking is defined as the process of abstractly breaking down a concept into its constituent parts in order to understand the whole and examining the relations between the parts (Bruner, 1957, as cited in Lohman & Lakin, 2011). Considering the construct that TOBT intends to measure as a whole, the test is based on figures, numbers and letter relations, and the items in the test constructed according to the relationships between these components. In this respect, TOBT exhibits a structure similar to other international tests developed for the same purpose. The content validity studies on such tests indicate that they consist of cognitive processes, such as analytical thinking, critical thinking, drawing conclusions, problem solving, reasoning, and induction (Enright & Powers, 1991; Lord & Wild, 1985; McCallum, 2003).

LCFA was used to determine the factor structure of TOBT. This analysis was used because the structure of the test from which the data was obtained was at the rank-order level and the test items consisted of binary responses. In studies conducted with dichotomous data (1-0, yes-no, true-false) obtained from binary responses, the common practice is to perform techniques involved in FA through a tetrachoric correlation matrix. However, this method has many requirements, including a Gramian Correlation Matrix, which often distorts the results (Christoffersson, 1975). It is observed that FA, which is only suitable for continuous variables, is often used in scale types that include variables at classification and rank-order levels. This, in many respects, causes the violation of the linearity hypothesis by interpreting the model as a linear model while in reality, it is not (Magidson & Vermunt, 2003b). LCFA should also be preferred in terms of providing an easier approach to explaining and interpreting the data structure. In addition, LCFA models are more suitable for multidimensionality (Moors, 2003). Therefore, in the current study, FA was considered not to be appropriate for categorical variables; instead, as a dimensionality reduction method more suitable for the categorical data structure, latent class models were utilized.

Using SLCA, whether TOBT had measurement invariance according to country and gender was examined at different levels of homogeneity in hierarchical order from the heterogeneous to the homogeneous model. The fact that the homogeneous model was found to best fit the data in the measurement invariance measurements for country and gender indicates that TOBT had measurement invariance in terms of both variables. This suggests that the three-factor structure of the test was the same for the three countries and two genders; in other words, the test measured the same psychological traits in terms of both variables. This is expected since the aim of TOBT is to perform measurements independent of language considering both the intended use of the test and the target group consisting of individuals from different cultures. Although ensuring this equivalence seems to be difficult due to the greatly varying profiles of different cultures, there are studies in the literature suggesting that such tests can successfully measure constructs independent of cultural characteristics (Cattell, 1979; Raven, 2000).

In the investigation of the measurement invariance of TOBT, rather than FA-based methods, approaches based on latent class analysis were chosen. The statistical methods used to examine measurement invariance are based on either the structural equation model or MGCFA that utilizes mean and covariance models (Jöreskog, 1971; Sörbom, 1974). However, these techniques do not always provide an appropriate approach for the analysis of the categorical data structure due to various factors, such as the requirement of a certain type of correlation matrix and normal distribution assumption as is the case in FA methods. In contrast, latent class model-based analysis provides robust evidence for the determination of all kinds of measurement biases by providing a highly flexible approach (Moors, 2003).

As the model selection criteria employed in the analysis of measurement invariance, the BIC value adjusted for sample size was taken into account in the formulation for country-level comparisons. This is because BIC and Consistent Akaike Information Criterion (CAIC) are not suitable for use in small samples (Güngör Culha, 2012). Another reason is that the BIC value tends to result in a simpler model, which facilitates the interpretation of estimation values. In addition, studies indicate that BIC provides more consistent results with a sample size of about 1000 (Dias, 2006; Kankaras, Vermunt, & Moors, 2011; Moors & Wennekers, 2003; Morren et al., 2011). Although the use of BIC and AIC values is generally accepted as an alternative way to select a model, there is no consensus concerning which existing criteria; e.g., AIC3, CAIC, and Sample-Size-Adjusted Bayesian Information Criterion (SABIC), are most suitable for this purpose. In the literature, there is discussion by many researchers on this issue (Bauer & Curran, 2003; Dias, 2006; Lin, 2006; Nylund, Asparouhov, & Muthen, 2007; Vrieze, 2012; Yang & Yang, 2007, as cited in Güngör, Korkmaz, & Sazak, 2015).

The items in the test were studied to identify whether they present DIF according to country and gender based on Recursive Partitioning Analysis in Rasch model. As a result of the analysis of DIF in terms of two variables, it was observed that 14 items regarding to country variable show DIF.

Considering 14 items which indicate DIF as a whole, it is observed that the difficulty and easiness level of the items according to country is close to zero which is the average value. 46 items, except for these, have approximate distribution and values.

Westers and Kelderman (1992), by emphasising assessment instruments used in education and psychology including items showing DIF on the significance level are not fair for some subgroups, state that such items required to be revised or taken out of assessment instrument by detecting. However, taking out the items showing DIF of the test directly could cause some problems particularly when the test has few items or in the circumstances content validity related to measured psychological structure is damaged. Taking such conditions into account, following the items are determined by means of experimental techniques, whether relevant items are advantageous or disadvantageous for the subgroups taken the test ought to be put forward by experts (Hambleton, Swaminathan, & Rogers, 1991).

In the studies carried out according to gender, it is concluded that there is no differentiation regarding to female and male students on the significance level, in other words; the items found in test do not involve DIF according to gender.

Suggestions

This study examined whether TOBT had measurement invariance; whether the items included in the test show different functions according to the country and gender using quantitative analysis methods. Accordingly, based on the items including DIF, the reasons for different functions of the items included in the test can be extensively determined in future studies examining functional differences at item level or with respondents and item developers.

The measurement invariance and differential item functioning analyses carried out within the scope of this study only took into consideration the country and gender variables. Future research can be undertaken to investigate other variables, such as age, type of school, region, and minority status. In addition, our data consisted of the responses of the individuals to TOBT that participated in AYOS from Germany, Azerbaijan, and Iran. Further comparative studies can be conducted with data obtained from other groups of individuals with similar characteristics taking the same test in AYOS.

Lastly, this study was designed based on estimates using response patterns obtained from a test application. Tests to be examined by other studies can be trialed on similar groups before their actual application to confirm the absence of bias, which is important for validity.

Acknowledgements

The data of this research was obtained from ANKUDEM.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

ANKUDEM. (2011). *Ankara Üniversitesi Yabancı Uyruklu Öğrenci Seçme ve Yerleştirme Sınavı (AYÖS) projesi kesin raporu*. Proje No: 11Y5250001. Ankara: Ankara Üniversitesi Bilimsel Araştırma Projeleri Ofisi.

ANKUDEM. (2012). *AYÖS 2012 Temel Öğrenme Becerileri Testi üzerine bir çalışma*. Kurum içi Rapor. Ankara: Ankara Üniversitesi Ölçme ve Değerlendirme Uygulama ve Araştırma Merkezi.

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A comparative analysis of the efficacy of differential item functioning detection for dichotomously scored items among logistic regression, SIBTEST and Rasch tree methods. *Procedia-Social and Behavioral Sciences, 191*(2), 21-25.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton: Taylor and Francis Group.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley-Interscience Publication.

Cattell, R. B. (1979). Are culture fair intelligence tests possible and necessary? *Journal of Research and Development in Education, 12*(2), 3-13.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79*(388), 762-771.

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. *Sociological Methodology, 15*, 81-110.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis with application in the social, behavioral, & health sciences*. New Jersey: John Wiley and Sons, Inc.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5-32.

Dias, J. G. (2006). Latent class analysis and model selection. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, & W. Gaul (Eds.), *From data and information analysis to knowledge engineering* (pp. 95-102). Berlin: Springer Heidelberg.

Enright, M. K., & Powers, D. E. (1991). *Validating the GRE analytical ability measure against faculty ratings of analytical reasoning skills*. New Jersey: Lawrence Erlbaum Associates, Inc.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Güngör Culha, D. (2012). *Örtük sınıf analizlerinde ölçme eşdeğerliğinin incelenmesi* (Unpublished doctoral dissertation). Ege University, İzmir.

Güngör D., Korkmaz, M., & Sazak, H. S. (2015). Örtük sınıf analiziyle yapılan ölçme eşdeğerliği çalışmalarında model seçimi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(1), 90-105.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications, Inc.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Kankaras, M., & Moors, G. B. D. (2009). Measurement equivalence in solidarity attitudes in Europe: Insights from a multiple-group latent-class factor approach. *International Sociology, 24*, 557-579.

Kankaras, M., & Vermunt, J. K. (2014). Simultaneous latent class analysis across groups. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 5969-5974). Dordrecht: Springer.

Kankaras, M., Moors, G. B. D., & Vermunt, J. K. (2011). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 359-384). New York: Taylor and Francis Group.

Kankaras, M., Vermunt, J. K., & Moors, G. B. D. (2011). Measurement equivalence of ordinal items. A comparison of factor analytic, item response theory, & latent class approaches. *Sociological Methods and Research, 40*(2), 279-310.

Kopf, J., Augustin, T., & Strobl, C. (2010). *The potential of model-based recursive partitioning in the social sciences: Revisiting Ockham's Razor* (Technical report number 88). Munich: University of Munich, Department of Statistics.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Lin, H. T. (2006). A comparison of model selection indices for nested latent class models. *Monte Carlo Methods and Applications, 12*(3), 239-259.

Lohman, D. F., & Lakin, J. M. (2011). Reasoning and intelligence. In R. J. Sternberg, & S. B. Kaufman (Eds.), *Handbook of intelligence* (pp. 419-441). New York: Cambridge University Press.

Lord, F. M., & Wild, C. L. (1985). Contribution of verbal item types in the GRE general test to accuracy of measurement of the verbal scores. GRE Board Professional Report GREB No. 84-6P, ETS Research Report No. 85-29. New Jersey: Educational Testing Service. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.1909&rep=rep1&type=pdf.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology, 31*(1), 223-264.

Magidson, J., & Vermunt, J. K. (2003a). *A nontechnical introduction to latent class models*. White Paper. Statistical Innovations.

Magidson, J., & Vermunt, J. K. (2003b). Comparing latent class factor analysis with the traditional approach in data mining. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 373-383). Boca Raton: Chapman and Hall/CRC.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp.175-198). Thousand Oaks: Sage Publications, Inc.

McCallum, R. S. (2003). Context for nonverbal assessment of intelligence and related abilities. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 3-18). New York: Kluwer Academic/Plenum Publishers.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage University Papers Series. Quantitative Applications in the Social Sciences. No. 07-064. Newbury Park: Sage Publications, Inc.

McCutcheon, A. L. (2002). Basic concepts and procedures in single- and multiple-group latent class analysis. In J. A. Hagenaars, & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56-86). Cambridge: Cambridge University Press.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research: Applications of Item Response Theory, 13*, 123-144.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3th ed., pp. 13-103). New Jersey: American Council on Education and Macmillan Publishing Company.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York: Taylor and Francis Group.

Moors, G. B. D. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity, 37*(3), 277-302.

Moors, G. B. D., & Wennekers, C. H. W. (2003). Comparing moral values in western European countries between 1981 and 1999. A multigroup latent-class factor approach. *International Journal of Comparative Sociology, 44*(2), 155-172.

Morren, M., Gelissen, J., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology, 41*(1), 13-47.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Strobl, C., Kopf, J., & Zeileis, A. (2015). A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316.

Strobl, C., Wickelmaier, F., & Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics, 36*(2), 135-153.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, & recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Vermunt, J. K., & Magidson, J. (2000). Graphical displays for latent class cluster and latent class factor models. In W. Jansen, & J. G. Bethlehem (Eds.), *Proceedings in Computational Statistics 2000* (pp. 121-122).

Vermunt, J. K., & Magidson, J. (2005a). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In L. A. Van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 41-62). Mahwah: Lawrence Erlbaum Associates, Inc.

Vermunt, J. K., & Magidson, J. (2005b). *Latent GOLD 4.0 user's guide*. Massachusetts: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013a). *Latent GOLD 5.0 upgrade manual*. Belmont: Statistical Innovations Inc.

Vermunt J. K., & Magidson J. (2013b). *Technical guide for Latent GOLD 5.0: Basic, advanced, & syntax*. Belmont: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2013c). *LG-Syntax user's guide: Manual for Latent GOLD 5.0 syntax module*. Belmont: Statistical Innovations Inc.

Westers, P., & Kelderman, H. (1992). Examining differential item functioning due to item difficulty and alternative attractiveness. *Psychometrika, 57*(1), 107-118.

Wolcott, H. F. (1994). *Transforming qualitative data: Description, analysis, & interpretation*. Thousand Oaks, California: Sage Publications, Inc.

YOK. (2013). *Yurtdışından öğrenci kabulüne ilişkin esaslar*. Karar Tarihi: 01.02.2013. Retrieved from http://www.yok.gov.tr/documents/10279/19575784/Yurtdisindan_Ogrenci_Kabulune_Iliskin_Esaslar161115.pdf.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*(2), 492-514.

Zeileis, A., Strobl, C., Wickelmaier, F., & Kopf, J. (2011). *Psychotree: Recursive partitioning based on psychometric models*. R package version 0.12-1. Retrieved from http://CRAN.R-project.org/package=psychotree.