International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 602–616
https://dx.doi.org/10.21449/ijate.554212
Published at http://www.ijate.net http://dergipark.org.tr *Research Article*

Investigating Differential Item Functioning of Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model

Özge Altıntaş [1,*], **Ömer Kutlu** [1]

[1]Department of Educational Measurement and Evaluation, Ankara University, Turkey

ARTICLE HISTORY

Received: 15 April 2019
Revised: 26 July 2019
Accepted: 10 November 2019

KEYWORDS

Differential Item Functioning,
Rrecursive Partitioning Analysis,
Rasch Tree Method,
Examinations for International Students

Abstract: This study aims to determine whether items in the Ankara University Examination for Foreign Students Basic Learning Skills Test function differently according to country and gender using the Recursive Partitioning Analysis in the Rasch Model. The variables used in the recursive partitioning of the data are country and gender. The population of the study is composed of 2476 individuals. Since the study includes comparisons across countries, the country is accepted as a criterion in determining the sample group. Thus, the sample of the study consists of 615 individuals selected from Azerbaijan, Bulgaria, and Syria. To investigate differential item functioning (DIF) of the items of the test, the Rasch tree method was used. As a result of the analysis, DIF has been detected in 16 items at the 0.001 significance level. However, these items have been identified to have similar difficulty parameters in all countries. Finally, items have not shown DIF according to gender.

1. INTRODUCTION

The constant social, economic, political and cultural changes have encouraged societies to explain the world in which they live based on the sovereign power of knowledge. It is possible to see the products of the mind in the transition to agriculture, emergence of cities, and birth of urban-state civilization. Interest, curiosity and needs have led our ancestors to learn about the natural world in terms of food, heating and protection, and use it for their own benefit. Since the 1600s, resources that help disseminate information; e.g., the printing press, and those that increase production, such as steam, coal, and electricity have entered the life of societies (Asimov, 2004).

The effort to understand human and social life in the new century has further revealed the value of scientific knowledge and scientific research. Scientific information has contributed to the rapid development of information technologies. With the emergence of new technologies,

CONTACT: Özge Altıntaş ✉ oaltintas@ankara.edu.tr Department of Educational Measurement and Evaluation, Ankara University, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

social life has undergone a change and has been reshaped, thus the need for continuous regeneration of information required to sustain life has grown.

Since scientific knowledge is a shared value, incomplete, erroneous or even inaccurate information has a negative effect on the changes of societies. Toffler (1980) stated that change is not linear; it is forward, backward or lateral. Transformation of change into social development; i.e., forward change, depends on the production of accurate knowledge, which requires the application of principles of reason and logic to understand and explain natural and social events. This is closely related to raising individuals that value scientific knowledge, seek information, and know how to obtain it.

Today, many societies consider that economic, cultural and political development is only possible through raising qualified and educated individuals. Therefore, they attach importance to all the educational stages, especially preschool education, and the construction educational institutions in accordance with the requirements of the era. The increase in the number of educated individuals leads to an increase in individuals who want to receive higher education. Thus, societies not only open new higher education institutions and programs but also cooperate with foreign educational institutions to ensure that their citizens receive a better quality of education (Altıntaş, 2016).

Many graduates of secondary education programs want to study in various higher education programs of universities outside their country in order to obtain a higher quality university education. Those countries that accept international students try to determine the application requirements for those candidates. Such education programs in Western countries that accept a large number of overseas students require internationally recognized high school diploma degrees and/or scores from internationally valid tests, such as the Scholastic Aptitude/Assessment Test (SAT), the American College Testing (ACT), and the Thinking Skills Assessment (TSA). These tests are prepared in English, a language which is widely used in the world, and the test items include verbal, numerical and formal expressions. The items mainly measure high-level thinking processes, such as reasoning, critical thinking, problem solving, and abstract thinking (Zwick & Sklar, 2005). In terms of the measured characteristics, these tests aim to assess high-level mental processes; e.g., using, interpreting and generalizing knowledge, making differentiations, and establishing and evaluating relations between different components (Kutlu & Karakaya, 2007).

Selection and placement tests implemented in educational processes are used in the transition of students from the present learning step to a higher educational stage and in deciding whether the student can move to the upper level (Cronbach, 1990). The selection and placement tests currently used in Turkey differ from those implemented from 1961 to 1980 in that they aim to measure the academic ability of the students and consist of items based on secondary education programs (Oral, 1985). These tests measure the students' ability to use the basic knowledge and skills acquired in school programs, and in 1981, this new approach was also adopted for the development of Examination for Foreign Students (YOS) tests, used in the selection and placement of foreign students.

In 2010, the Turkish Council of Higher Education (YOK) examined the process of foreign student selection with a view to increasing the competitive power of Turkey in the international arena. In parallel to the studies concerning student placement in the higher education system in the country, beginning from the 2010-2011 academic year, YOK abandoned the YOS system and decided that universities were to determine the principles and conditions to be applied for the admission of foreign students and receive YOK's approval before implementing them (YOK, 2013).

In accordance with this decision, universities began to determine their own programs and quota for foreign students with the approval of YOK. Within the framework of these principles, through the decree of the academic senate, the Rectorate of Ankara University accepted SAT 1, Abitur, the International Baccalaureate Diploma, and the scores in Ankara University Examination for Foreign Students (AYOS) as the criteria to be considered in the selection and placement of foreign students to study in the university (Ankara University, 2013). Since the scores obtained from the tests used in these exams differ, the scores of diplomas are converted by the university's Student Affairs Office to comply with the scoring of AYOS. However, the lack of equivalence in the scoring of placement tests constitutes a measurement problem (Altıntaş, 2016).

The Student Selection and Placement Center (OSYM) tried to maintain the goal of selecting and placing students in higher education in the 1980s, at which time this center was responsible for determining the framework of foreign student examinations. The items included in the tests that aim to select and place foreign students in higher education programs in Turkey require the student to use thinking processes, such as comprehension, application, and analysis (Toker, 1997; USYM, 1978, 1980a, 1980b).

These items are based on the relation between figures, numbers, and letters, which are independent of language. In addition, they are associated with mental processes; e.g., analytical thinking, reasoning, and abstract thinking that develop in individuals over a long period of time. The reason why such tests are developed in a language-independent manner; i.e., containing very limited verbal language and mostly utilize figures, numbers and letters, is that word use, relationships between words, and verbal instructions are not suitable for those individuals who do not have an adequate knowledge of the target language (Resing, 2005).

Since 2011, AYOS has been conducted in accordance with the aims of YOS tests and the general purposes of developing student selection and placement tests in higher education. In this exam, the Basic Learning Skills Test (TOBT) consisting of 100 items is used. The items are prepared independently of verbal expression, language, and the content of the curricula of the schools. The first 60 items of test are based on the relationships between shapes, numbers, and letters, the items measuring psychological properties, such as analytical thinking, reasoning, abstract and spatial thinking; the remaining 40 items consist of the items that measure numerical thinking skills that require the use of mathematics and geometry information (ANKUDEM, 2011). In other words, the test is a "non-verbal or verbal neutral" measurement tool. This is mostly because the foreign students preferring to study in higher education programs in Turkey are coming from different cultures and therefore, they do not know Turkish or another foreign language well.

The first 60 items in TOBT mainly aim to measure students' analytical thinking, abstract thinking, and reasoning. Analytical thinking is the process of breaking things down into their constituent components in order to understand the whole and examining the relationships between these components. Reasoning refers to the process of making inferences and reaching a conclusion based on the information given (Bruner, 1957 as cited in Lohman & Lakin, 2011). The remaining 40 items measuring students' ability to use numerical skills based on basic mathematics and geometry predominantly require the individuals to establish connections based on shapes, numbers and letters, make logical inferences, and engage in abstract thinking and reasoning.

Since the beginning of the 1900s, many researchers have attempted to measure skills, such as analytical thinking, reasoning, and abstract thinking through intelligence tests based on the relationships between shapes, numbers and letters. Looking at the process related to psychological measures, the tools used to measure the mental abilities of individuals are now known as academic aptitude tests. These tests also contribute to the estimation of school

achievement by identifying individuals' abilities (Anastasi, 1979). Walsh and Betz (1995) stated that aptitude tests used in education can also help predict future educational achievement. The high correlation between the scores obtained from the academic aptitude tests and the academic achievement scores indicates that the students with higher aptitude scores may have higher school grades than those with lower aptitude scores (Sternberg, 1997).

Certain psychometric properties are sought in the tests designed to measure academic aptitude, one of which is the tests of having predictive power. Thus, one of the most important features of an academic aptitude test prepared for student selection and placement is predictive validity. The predictive power of a test means that the feature measured by the test in a given period is related to a particular feature in the future. Prediction is a matter of making estimations, and predictive validity is often used in the selection of tests for educational placement and recruitment for employment (Turgut & Baykul, 1992).

Tests developed for student selection and placement should be able to predict achievement scores to be obtained from future tests based on the scores in the currently applied tests (Thornell & McCoy, 1985). One of the factors that influence the predictive power of a test is the items representing the psychological characteristics that are measured. For a test to have high predictive power, the items contained should measure the mental characteristics that develop in individuals over a long period, rather than psychological features that develop in a shorter time (Anastasi & Urbina, 1997; Cronbach, 1990).

Psychological measurement tools are affected by demographic properties, such as age and gender, and cultural properties; e.g., ethnicity and country, whether developed in verbal or non-verbal language (Messick, 1989). This situation makes the accuracy of the decisions given based on the scores obtained from the measurement tools questionable. Especially in the selection and placement tests, the items which are important for students need to have equal response possibility for those who have the same ability level. Otherwise, individuals with the same ability level will be advantageous or disadvantageous compared to each other.

An essential part of the test development and item preparation process in the education and psychology fields is to divide the tests into different groups and determine whether the measurement results are the same for each group. In particular, the items included in selection and placement tests, which have an important impact on the lives of individuals, must create the possibility of an equal response from individuals at the same level of ability and skills. In other words, the test items should not be biased toward certain subgroups. Otherwise, individuals with the same skill level taking the test will gain an advantage or be disadvantaged in relation to each other. Since AYOS is developed for selection and placement purposes, it is necessary to examine the psychometric properties of these tests to ensure that such bias is not present.

Karakaya and Kutlu (2012) state that the determination of bias of the items in the tests is one of the important studies to increase the validity and reliability of the test. Therefore, undertaking an investigation of item-level bias is very important in the test development process, especially in the process of item writing, in establishing the preliminary evidence on the validity.

The concept of validity is addressed and defined from different perspectives in the literature due to the variety of validation methods. According to Anastasi (1979), validity refers to what a measurement tool measures and the extent of which the measurement process is undertaken correctly. A measurement tool applies only to a specific purpose under certain conditions; it cannot be asserted that the same measurement instrument applies to other purposes or conditions. Messick (1989) defined validity as a degree of all assessments supporting the accuracy and appropriateness of implications related to the scores of a measurement tool or other measurement cases based on theoretical information and empirical evidence. In this

respect, validity is not only a feature of measurement or evaluation; rather, it is concerned with the meaning of the scores obtained from the measurement tool. Linn and Gronlund (2000), including all these definitions, described validity as the most fundamental and important factor in determining the accuracy and appropriateness of implications and judgments about the results obtained from measurement tools.

The issue of validity, discussed in detail in the Measurement Standards in Education and Psychology, prepared in 2014 by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), has found wide coverage in *Standard 1.0* to *Standard 1.25*. In the same source, validity is defined as the degree of evidence and theories supporting the interpretability of the scores obtained from tests. According to this definition, validity describes an ongoing process, rather than merely results. In this process, additional information is always available to achieve a better understanding of the implications of the test. Making inferences about validity is similar to undertaking scientific inferences. Therefore, validity studies are conducted by presenting supporting information about test scores. A prerequisite for this, as emphasized in *Standard 1.1*, is to confirm that the construct(s) intended to be measured in a test are clearly defined and are not extensively linked to other constructs.

The *Test Validation* section of *Educational Measurement*, written by Cronbach (1971) and edited by Thorndike, emphasizes the importance of psychological constructs in educational measurement. According this source, whenever a test developer[†] asks, "What does this measurement tool really measure?", s/he wants to obtain information about construct validity. In this respect, a test being able to measure the construct of interest despite the presence of disrupting effects emerges as a condition to which testers pay great attention.

Whether the test is appropriate or powerful enough to measure the intended construct is determined by obtaining evidence on construct validity because the power of a test to measure a construct is an important indicator of the extent to which the test serves its purpose (Linn, 1989). Therefore, considering that the items in a test are also developed in accordance with the purpose of that test, it can be assumed that investigation of construct validity begins with the test development process.

Cronbach and Meehl (1955) emphasized that construct validity was important for tests that measure many psychological features, such as interest, attitude, ability, and success. By the beginning of 1900s, studies starting with Charles Edward Spearman, led psychometrists to investigate the real value of the observed characteristics. Since then, the importance of the reliability and validity values of measurement instruments has been acknowledged, and the related coefficients have become the most important criteria for these tools. Achieving a certain reliability and validity value for the scores obtained from measurement tools is considered as a requirement but not sufficient alone for the measured characteristics. This understanding has resulted in deepening the studies on the structural features of measurement instruments. For example, the tests used in education have been enriched in terms of psychometric features measured, and national standard achievement tests and more recently large-scale tests that can be used at the international level have been developed. The widespread use of tests has raised the question of whether the items contained in these tests can measure the intended constructs independently of individuals and their associated groups (Cohen et al., 1988; Crocker & Algina, 1986).

After 1950s, many researchers began investigating the reasons for the conflicting results obtained from intelligence tests. The 1960s mark the discussion of issues related to the fair use

[†] The word 'educator' is used in the original source. However, 'test developer' was used here in accordance with the context.

of tests and item bias, which was mainly a result of the human rights movement initiated in the United States of America in 1964. This movement led to the signing and enforcement of certain laws on equality and equal opportunities. The fundamental changes brought about by the human rights movement also attracted the attention of test developers to the use of tests that may have an adverse impact[‡] on recruitment and education (Osterlind & Everson, 2009). After the 1970s, psychometrists concentrated their research on this issue of 'bias', for the development of tools that would make more objective and sensitive measurements (Reynolds & Suzuki, 2013).

In the following years, the related studies were not limited to a specific culture, but were deepened by extensive research on cross-cultural comparisons. However, during these investigations that focused on bias in tests developed to measure psychological features, the items included in tests, and individuals (gender, age, etc.) or groups (culture, ethnic origin, etc.) responding to these items, researchers commonly faced four important problems and limitations: (1) lack of a consensus on the definition of psychological constructs or characteristics to be measured, (2) failure to select a sample that represents the groups to be compared, (3) inability to standardize the conditions of the test application (absence of standardization), and (4) lack of rules regarding the translation, adaptation and scoring of test booklets (Hambleton, 2002).

Psychological measurement tools are developed to obtain information about the psychological characteristics of individuals living in a particular culture. Thus, a measurement tool developed in a culture possesses features specific to that culture (Öner, 1987). Culture is a very important factor in the test development process since it can affect the scores obtained from tests and the psychometric characteristics related to these scores (Hambleton, Merenda, & Charles, 2005). Wicherts (2007) noted that the results of the measurement might differ according to individual characteristics, but it would be wrong to attribute these differences to individual characteristics alone since they might also result from the measurement tool. For example, assuming that a girl and a boy have a similar level of knowledge in mathematics, if there is a systematic difference between the scores of these students in a mathematics test developed to measure the related construct, it can be stated that the test has a gender bias.

The first known studies on bias date back to the 1900s, when it was determined that the scores of socioeconomically disadvantaged children who had taken the intelligence test developed by Alfred Binet were related to what they had learned at home or school, rather than the mental characteristics of the individuals, which led to the removal of certain items from this test (Camilli & Shepard, 1994).

According to Crocker and Algina (1986), bias research has two main objectives. First refers to whether the test scores are affected by different variance sources in different subgroups taking the test, and second is whether the test scores are affected by the same variance sources for all subgroups. If the test scores are judged to be affected by the same variance sources in all subgroups, it should also be investigated whether there are unrelated sources that provide unfair advantage to certain subgroups.

There are two basic statistical approaches to the investigation: external methods, in which an external measure independent of the test is used, and internal methods, in which psychometric features of the items included in the test are used as criteria. In external methods, analysis of bias is undertaken by comparing the differences in averages of the total scores obtained from a test for different subgroups to those obtained from a different test considered to measure the same construct. In cases where an external measure is not available, internal methods can be used to determine bias by examining the psychometric features of the items included in the test (the total test score obtained from the items and skill level) (Shepard, Camilli, & Averill, 1981).

[‡] A term used to refer to evidence supporting unlawful discrimination claims.

However, item analysis methods, such as item-total score correlations performed in accordance with the Classical Test Theory (CTT) or variance analysis that compares the items and total test scores in a similar manner do not provide sufficient information about bias based on the averages of groups and differences between these averages. Bias investigations undertaken using such methods may have deficiencies since the psychometric features of items and the total scores obtained from the test are affected by the skill distribution of the sample in CTT. Therefore, the differences in individuals' test performance or average scores in test items should not be interpreted as evidence of a bias in direct comparison groups.

In the third section of the Measurement Standards in Education and Psychology[§], AERA, APA and NCME (2014) present 20 standards related to the fairness of the test scores; i.e., they should be free from bias. In all these standards, it is emphasized that the test or item scores should have the same meaning for all individuals (within subgroups) that have taken the test, and that the test scores should be comparable. *Standards 3.2* and *3.13* explain in detail the necessity to take into account the different characteristics of the test respondents, such as language and culture, in accordance with the purpose of the test. If there are thought to be differences in the test and item performance of respondents in terms of the measured characteristic according to ethnicity, language, culture, gender and age groups, these differences should be investigated in as much depth as possible.

In other words, the tests to be administered to the individuals from different groups and the items to be included in these tests should be designed in such a way as to reduce the situations that may lead to bias. Therefore, the evidence related to whether or not the measurement instrument investigated in this study, TOBT, caused a bias for/against individuals in different subgroups was examined by investigating whether the function of the TOBT items differed between these individuals. Accordingly, the aim of this research was to investigate whether the items contained in the Basic Learning Skills Test in the AYOS-TOBT 2017 showed differential item functioning (DIF) according to three countries (Azerbaijan, Bulgaria, and Syria) and gender.

2. METHOD

2.1. Research Design

This research is a survey type taken from descriptive research models. The study uses a descriptive research model, since it aims to investigate whether the items of AYOS-TOBT show DIF in terms of country and gender variables and describe the current situation (Karasar, 2015). Descriptive research is describing and interpreting the factors that are the subjects of the study; however, this goes beyond gathering and classifying the data. Research process also includes collecting, classifying, describing, analyzing and inferring results results from the data (Best, 1970).

2.2. Population and Sample

The population of the study is 2476 individuals (N_{female} = 1184 approx. 48%, N_{male} = 1292 approx. 52%) from 75 countries who took AYOS 2017. Based on the purpose of the research, a purposive sampling method was used to select the sample in order to conduct an in-depth research and obtain rich information. In this study, criterion sampling was used from purposive sampling methods (Büyüköztürk et al., 2015). Since the study includes comparisons across different countries, culture was accepted as a criterion in determining the sample group. Thus, it was represented in both the number of students and different cultures, and the sample of the study consisted of 615 individuals selected from Azerbaijan, Bulgaria, and Syria. The distribution of the sample according to country and gender is given in Table 1.

[§] Fairness in Testing

Table 1. Distribution of sample by country and gender

Gender / Country	Female		Male		Total	
	n	%	n	%	n	%
Azerbaijan	51	38.35	82	61.65	133	21.63
Bulgaria	125	58.69	88	41.31	213	34.63
Syria	107	39.78	162	60.22	269	43.74
Total	283	46.02	332	53.98	615	100.00

The numbers of individuals taking the exam from Bulgaria (213, approx. 35%) and Syria (269, approx. 44%) were close to each other than Azerbaijan (133, approx. 22%). Besides, the number of individuals in the sample was close to each other in terms of gender (n_{female} = 283 approx. 46%, n_{male} = 332 approx. 54%) (Table 1).

2.3. Data collection

The research data was composed of the students' responses to AYOS-TOBT 2017, simultaneously implemented in three different exam centers located in Ankara/Turkey, Cologne/Germany and Baku/Azerbaijan in a single session. The responses obtained from two different booklets (A and B) were reordered according to Booklet A, and the data were prepared for analysis by converting it to the 1-0 scoring matrix and merging.

In this study, in order to limit the number of items examined, the first 60 items of the test were selected since they were considered to be more similar to the characteristics measured.

2.4. Data Analysis

In this study, bias analysis performed at the item level in terms of different subgroups was undertaken using a DIF analysis within the scope of the Rasch Model. This approach of model-based recursive partitioning (MBRP), proposed by Zeileis, Hothorn, and Hornik (2008), includes tests for both predefined groups and all possible groups without complicating the interpretation process. This allows for the determination of parameter imbalances. Similar to implicit class or mixed models, the main idea underlying this approach is based on identifying the groups in which the model parameters are differentiated is the sequential testing of all groups by investigating all possible sources that may cause DIF. In recursive partitioning, groups are defined not by an implicit factor as in implicit class models, but through the combinations of the observed common variables, based on an intuitive approach. Thus, MBRP offers intuitive, yet easy-to-interpret alternatives to implicit class or mixed models.

The Rasch tree model, a very new method for determining DIF, is based on MBRP, in which tests for structural change adapted from econometrics are used. MBRP is highly correlated with classification and regression tree methods, in which a common variable field is recursively partitioned to determine the group of a categorical or continuous response variable with different values. MBRP has a semi-parametric approach including the parameters of a parametric model varying between groups instead of values for a single response variable. Such parameters may be those of the Rasch model, which vary between groups or constant and slope parameters of a linear regression model (Strobl, Kopf, & Zeileis, 2015).

In MBRP analysis, the purpose is to divide the data matrix into subgroups (classes) with a homogenous structure. Each of these subgroups is called a node. Subgroups are primarily defined by common variables (such as age and gender). Then, these nodes are broken down as in classification and regression trees, like the branches of a tree, until they have a homogenous structure within. This is known as a Rasch tree, and the branches of this tree contain critical values (leaves) for common variables. This process continues until each node has the lowest variance value and the variance between the nodes is the highest (Kopf, Augustin, & Strobl, 2010).

In this regard, a Rasch tree with each leaf containing a node associated with a suitable model (e.g., maximum likelihood model or linear regression model) is created to achieve model-data fitness. Here, the basic idea is that each node is associated with a single model. Below is the successive steps (algorithm) used to create a Rasch tree (Strobl, Kopf, & Zeileis, 2015):

1. Estimate the item parameters jointly for all subjects in the current sample, starting with the full sample.
2. Assess the stability of the item parameters with respect to each available covariate.
3. If there is significant instability, split the sample along the covariate with the strongest instability and at the cutpoint leading to the highest improvement of the model fit.
4. [Repeat Steps 1-3 recursively in the resulting sub-samples until there are no more significant instabilities (or the sub-sample becomes too small)].

The Rasch tree employs the model-based recursive partitioning algorithm to detect groups that display different item parameters in the Rasch model (Kopf, 2013). This analysis was performed using the 0.12-1 version of the add-on package PsychoTree (Zeileis et al., 2011) in R program, which is open source statistical software (R Core Team, 2013). The Psychotree package was used to identify the items that showed DIF. The data was analyzed using the Rasch tree method (RTM) included in Psychotree, and the items showing DIF according to country and gender were determined. The significance level for the determination criteria of DIF, usually set to 5%, serves as the most important stopping criterion (Strobl, Kopf, & Zeileis, 2015). In this study, in addition to the 0.05 level of significance, a lower value, 0.01, was used as the DIF criterion in RTM.

3. FINDINGS and DISCUSSION

Table 2 shows the parameter instability tests conducted to determine whether the TOBT items indicated DIF at the level of significance according to the country and gender.

Table 2. Parameter instability tests: Test statistics and their corresponding p values by country and gender

Cov.	Par. Inst.	Node 1	Node 2
Country	statistic	.000	.000
	p value	.000*	.000*
Gender	statistic	90.2769556	85.3219230
	p value	0.2794151	0.4997366

*$p < .001$

As shown in Table 2, the country was accepted as a covariant, since the instability statistics obtained according to the country variable was significant. Gender was not considered a covariant since the value obtained for the gender variable was not significant. The Rasch tree showing this situation is given in Figure 1.

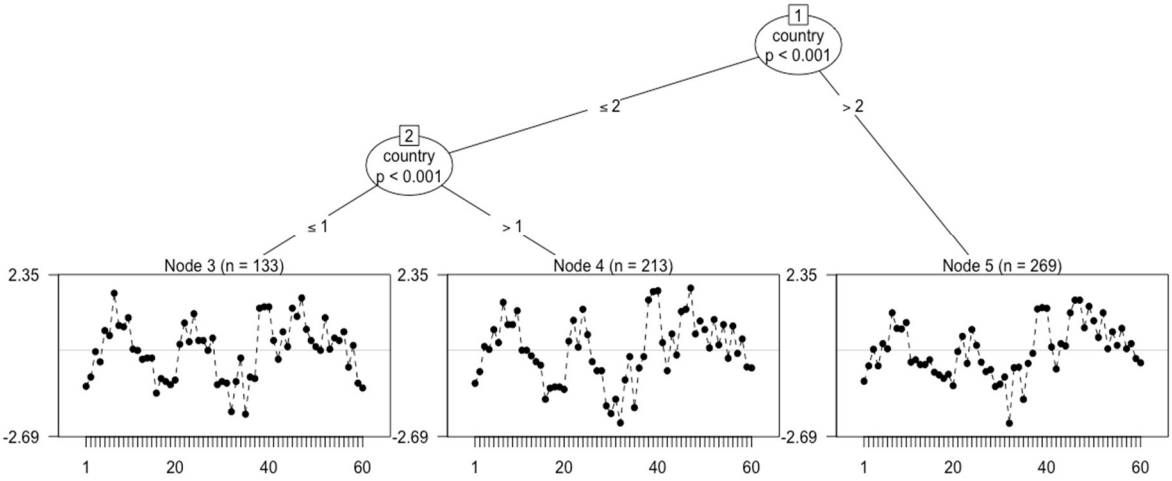

Figure 1. Rasch tree by country

As shown in Figure 1, the estimates of the difficulty parameters of 60 items in TOBT were between 2.35 and -2.69. The high level of these values means that the items were difficult while those with low values were easy (Strobl, Kopf, & Zeileis, 2015). Accordingly, it can be stated that the difficulty levels of the items were close to the average value of zero. Figure 1 also includes the significance values for each partition. Hence, there was a variation in terms of the countries in which the partition took place.

The Rasch tree obtained also provides information on which countries the difference appears. Thus, Syria differs according to the other two countries (Azerbaijan and Bulgaria) in the first partition and Azerbaijan differs from Bulgaria according to the second partition. When the Rasch tree is examined in detail, the 16 items (3, 4, 5, 6, 11, 12, 21, 23, 26, 27, 28, 44, 51, 53, 55, 57) in the TOBT included DIF in terms of countries. Distribution of difficulty parametres of the 16 items showing DIF is shown comparatively in Table 3.

Table 3. Distribution of difficulty parameters of items showing DIF according to the countries

Item no.	Countries		
	Azerbaijan	Bulgaria	Syria
3	-0.046	0.118	0.027
4	-0.364	0.022	-0.488
5	0.612	0.642	0.204
6	0.457	0.235	0.045
11	0.032	-0.002	-0.373
12	-0.007	-0.002	-0.298
21	0.187	0.281	-0.044
23	0.264	0.094	-0.411
26	0.302	-0.358	-0.373
27	-0.007	-0.640	-0.667
28	0.379	-0.640	-0.606
44	0.110	-0.150	0.134
51	-0.007	0.070	0.397
53	0.032	0.165	0.045
55	0.302	-0.252	0.151
57	-0.530	-0.100	0.045

In Table 3 items 3, 51 and 57 are in favour of Azerbaijani; items 44 and 55 are in favour of Bulgarian; items 5, 6, 11, 12, 21 and 23 are in favour of Syrian students. According to this, items 3, 51 and 57 are easier only for Azerbaijani students, items 5, 6, 11, 12, 21 and 23 are

easier only for Syrian students and items 44, 51 and 55 are easier only for Bulgarian students. Another finding is that some items are in favour of two countries. For instance items 4 and 53 are in favour of both Azebaijani and Syrian students whereas items 26, 27 and 28 are in favour of Bulgarian and Syrian students.

When items showing DIF are analyzed in terms of their cognitive properties they require the following skills:

- Items 3,4,5 and 6 require the ability to reach a conclusion by constructing meaning between related parts
- Items 11 and 12 require the ability to find the part of a meaningful whole
- Items 21 nd 23 require the ability to predict the whole that the parts construct
- Items 26,27 an 28 require the ability to predict the parts from the whole by three dimensional-thinking
- Items 44, 51, 53, 55 and 57 require the ability to reach a conclusion by combining the given parts using the knowledge of arithmetic operation, letter and symbol according to a rule.

Similarly, Maller (2001) investigated the DIF of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III). The WISC-III national standardization sample (N = 2200) was used to determine DIF in six WISC-III subtests. After fitting two parameter logistic and graded response models to the data, the items were tested for DIF using the DIF detection method based on item response theory likelihood ratio. Of the 151 items studied, 52 were found to function differently across the groups.

Carman and Taylor (2010) examined the relationship between the Naglieri Nonverbal Ability Test (NNAT), ethnicity and gender, as well as the socioeconomic status and NNAT performance. Correlations and multiple regression were used to examine the relationships between ethnicity, SES, and NNAT performance in a large kindergarten sample. The results suggest a significant relationship between ethnicity, SES, and NNAT performance.

As presented in Table 2, the instability statistics value in terms of gender was not significant. Therefore, it can be stated that TOBT did not contain DIF in terms of gender. Since there is no significant difference between females and males ($p > 0.001$), a Rasch tree cannot be produced. In other words, no partition occurs, which is shown in Figure 2.

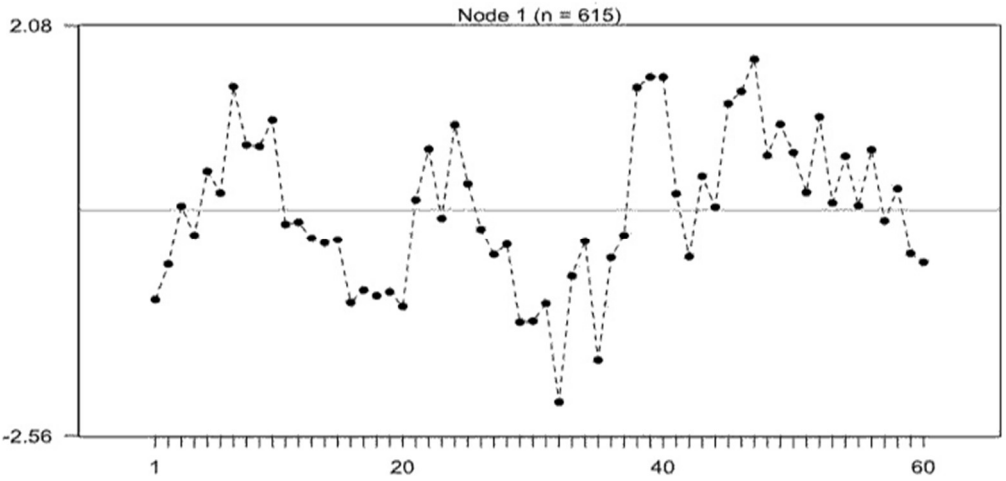

Figure 2. Rasch tree by gender

When Figure 2 is examined, it can be seen that there is no partition between females and males due to the absence of variation in response to the items in TOBT. This indicates that no item in TOBT shows DIF by gender.

Toivainen et al. (2017) used a large longitudinal twin sample to estimate sex differences in non-verbal and verbal abilities over time, using a variety of measures. Their study also investigated the influence of prenatal testosterone on these differences by comparing females with male co-twins to females with female co-twins. The sample size used in that study varied from 14187 participants at age 4 to 4959 participants at age 16. One-way ANCOVAs were used to establish significant group differences, either between sexes or between sex-by-zygosity twin groups. In all analyses, age was used as a covariate to account for the possible effect of age differences. The results showed negligible sex differences in non-verbal and verbal ability across development.

Empirical research consistently finds that standardized cognitive tests are not biased in terms of predictive and construct validity. Furthermore, continued claims of test bias, which appear in academic journals, the popular media, and some psychology textbooks, are not empirically justified. These claims of bias should be met with skepticism and evaluated critically according to established scientific principles (Brown, Reynolds, & Whitaker, 1999).

4. CONCLUSION

This study investigated whether recursively partitioned manifest variables can reveal DIF patterns in a non-verbal test using a Rasch tree approach. As a result of the research, 16 items in AYOS-TOBT 2017 indicated DIF in terms of countries. Concerning the 16 items showing DIF as a whole, it is noteworthy that the level of difficulty of the items according to the countries was close to the average value of zero. This situation showed that there was no significant variation in terms of countries.

In this study, whether the items in test showed DIF in terms of countries and gender was determined using quantitative analysis methods. Further research can be undertaken on the items which show DIF. Thus, comprehensive information can be obtained about the reasons why these items show DIF.

Acknowledgement

This article is the extended version of the paper titled “Investigating Differential Item Functioning of the Ankara University Examination for Foreign Students by Recursive Partitioning Analysis in the Rasch Model” and presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology (C-MEEP) in 5 - 8 September 2018 in Prizren, Kosovo.

ORCID

Özge Altıntaş https://orcid.org/0000-0001-5779-855X
Ömer Kutlu https://orcid.org/0000-0003-4364-5629

5. REFERENCES

Altıntaş, Ö. (2016). *Ankara Üniversitesi Yabancı Uyruklu Öğrenci Seçme Testinin Ölçme Değişmezliğinin Örtük Sınıf ve Rasch Modeline Göre İncelenmesi [Investigating The Measurement Invariance of Ankara University Foreign Student Selection Test by Latent Class and Rasch Model]*. PhD diss., Ankara University.

AERA., APA., & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Anastasi, A. (1979). *Fields of Applied Psychology*. 2nd ed. New York: McGraw-Hill Book Company.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing*. 7th ed. New Jersey: Prentice-Hall International, Inc.

Ankara University. (2013). *Ankara Üniversitesine Yurtdışından Öğrenci Kabulüne İlişkin Yönerge. Ankara Üniversitesi Senato Kararı Örneği [Instruction on Admission of Foreign Students to Ankara University. Ankara University Senate Decision Example]*. Decision Date: 23.06.2013. Number of Meetings: 365. Number of Decisions: 3100.

ANKUDEM. (2011). *Ankara University Examination for Foreign Student Selection and Placement Exam (AYOS) Project Final Report*. Project No: 11Y5250001. Ankara: Ankara University Scientific Research Projects Scientific Coordinator.

Asimov, I. 2004. *Asimov's Chronology of Science and Discovery: Updated and Illustrated*. Norwalk: The Easton Press.

Best, J. W. (1970). *Research in Education*. 2nd ed. New Jersey: Prentice-Hall Inc.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since Bias in Mental Testing. *School Psychology Quarterly*, *14*(3), 208-238. DOI: http://dx.doi.org/10.1037/h0089007

Büyüköztürk, Ş., Akgün, Ö. E., Demirel, F., Karadeniz, Ş., & Kılıç Çakmak, E. (2015). *Bilimsel Araştırma Yöntemleri (Scientific Research Methods).* Ankara: Pegem Academy Publishing Co.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. California: Sage Publication, Inc.

Carman, C. A., & Taylor, D. K. (2010). Socioeconomic Status Effects on Using the Naglieri Nonverbal Ability Test (NNAT) to Identify the Gifted/Talented. *Gifted Child Quarterly*, *54*(2), 75–84. DOI: https://doi.org/10.1177/0016986209355976

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik M. E. (1988). *Psychological Testing: An Introduction to Tests and Measurement*. California: Mayfield Publishing Company.

Crocker, L., & Algina, J. (1986). *Introduction to Classical Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1971). Test Validation. In *Educational Measurement*. 2nd ed., edited by R. L. Thorndike, 443-507. Washington: American Council on Education.

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. 5th ed. New York: Harper and Collins Publishers, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Classics in the History of Psychology. *Psychological Bulletin*, *52*, 281-302. Retrieved from https://psychclassics.yorku.ca/Cronbach/construct.htm

Hambleton, R. K. (2002). Adapting Achievement Tests into Multiple Languages for International Assessments. In *Methodological Advances in Cross-National Surveys of Educational Achievement*, edited by A. C. Porter and A. Gamoran, 58-79. Washington: National Academy Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D., eds. (2005). *Adapting Educational and Psychological Tests for Cross-cultural Assessment.* New Jersey: Lawrence Erlbaum Associates, Inc.

Karakaya, İ., & Kutlu, Ö. (2012). An Investigation of Item Bias in Turkish Sub Tests in Level Determination Exam. *Education and Science, 37*(165), 2-15. Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1342/433

Karasar, N. (2015). *Bilimsel Araştırma Yöntemi [Scientific Research Method]*. 28th ed. Ankara: Nobel Academy Publishing Co.

Kopf, J. (2013). *Model-based Recursive Partitioning Meets Item Response Theory: New Statistical Methods for the Detection of Differential Item Functioning and Appropriate Anchor Selection*. PhD diss., Munich: Ludwig-Maximilians-University Department of Statistics. Retrieved from https://edoc.ub.uni-muenchen.de/16434/1/Kopf_Julia.pdf

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. Retrieved from https://pdfs.semanticscholar.org/c8ee/da16de9cb040066e5cb64aa37ceb58493286.pdf

Kutlu, Ö., & Karakaya, İ. (2007). Orta Öğretim Kurumları Öğrenci Seçme ve Yerleştirme Sınavının Faktör Yapılarına İlişkin Bir Araştırma. [A Research on the Factor Structure of Secondary Education Institutions' Student Selection and Placement Test]. *Elementary Education Online, 6*(3), 397-410. Retrieved from http://dergipark.gov.tr/download/article-file/90996

Linn, R. L., ed. (1989). *Educational Measurement*. 3rd ed. New Jersey: American Council on Education and Macmillan Publishing Company.

Linn, R. L., & Gronlund N. E. (2000). *Measurement and Assessment in Teaching*. 8th ed. New Jersey: Prentice-Hall International, Inc.

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

Maller, S. J. (2001). Differential Item Functioning in the WISC-III: Item Parameters for Boys and Girls in the National Standardization Sample. *Educational and Psychological Measurement, 61*(5), 793–817. DOI: https://doi.org/10.1177/00131640121971527

Messick, S. (1989). Validity. In *Educational Measurement*. 3rd ed., edited by R. L. Linn, 13-103. New Jersey: American Council on Education and Macmillan Publishing Company.

Oral, T. (1985). *Lise Başarı Ölçüleri ile ÖSYS Puanları Arasındaki Uyum [Concordance between High School Success Metrics and University Entrance Exam (ÖSYS) Scores]*. PhD diss., Hacettepe University.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. California: Sage Publications, Inc.

Öner, N. (1987). Kültürlerarası Ölçek Uyarlamasında Bir Yöntembilim Modeli [A Methodology Model in Intercultural Scale Adaptation]. *Turkish Journal of Psychology*, *6*(21), 80-83.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Resing, W. C. M. (2005). Intelligence Testing. In *Encyclopedia of Social Measurement*, edited by K. Kempf-Leonard, 307-315. San Diego: Elsevier Academic Press.

Reynolds, C. R., & Suzuki, L. (2013). Bias in Psychological Assessment: An Empirical Review and Recommendations. In *Handbook of Psychology Vol. 10: Assessment Psychology*. 2nd ed., edited by J. R. Graham, J. A. Naglieri, & I. B. Weiner, 82-113. New Jersey: John Wiley and Sons, Inc.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of Procedures for Detecting Test-Item Bias with Both Internal and External Ability Criteria. *Journal of Educational and Behavioral Statistics*, *6*(4), 317-375. DOI: https://doi.org/10.3102/10769986006004317

Sternberg, R. J. (1997). The Concept of Intelligence and Its Role in Lifelong Learning and Success. *American Psychologist*, *52*(10), 1030-1037. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.826.5234&rep=rep1&type=pdf

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316. Retrieved from https://eeecon.uibk.ac.at/~zeileis/papers/Strobl+Kopf+Zeileis-2015.pdf

Thornell, J. G., & McCoy, A. (1985). The Predictive Validity of The Graduate Record Examination for Subgroups of Students in Different Academic Disciplines. *Educational*

and Psychological Measurement, *45*(2), 415-419. DOI: http://dx.doi.org/10.1177/0013164485045002229

Toffler, A. (1980). *The Third Wave*. New York: McGraw-Hill Book Company.

Toker, F. (1997). *Türkiye'de Yükseköğretime Giriş [Entrance to Higher Education in Turkey]*. Ankara: ÖSYM Publications.

Toivainen, T., Papageorgiou, K. A., Tostoc, M. G., & Kovas Y. (2017). Sex Differences in Non-verbal and Verbal Abilities in Childhood and Adolescence. *Intelligence*, 64, 81-88. DOI: https://doi.org/10.1016/j.intell.2017.07.007

Turgut, M. F. & Baykul, Y. (1992-1). *Ölçekleme Teknikleri [Scaling Methods]*. Ankara: ÖSYM Publications.

USYM. (1978). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sınavının Temel İlkeleri [Basic Principles of Two-Stage University Selection and Placement Exam]*. Ankara: ÖSYM Publications.

USYM. (1980a). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sistemi [Two-Stage University Selection and Placement System]*. KY-02-80-0001. Ankara: ÖSYM Publications.

USYM. (1980b). *İki Aşamalı Üniversitelerarası Seçme ve Yerleştirme Sistemi [Two-Stage University Selection and Placement System]*. KY-02-80-0002. Ankara: ÖSYM Publications.

Walsh, W. B., & Betz, N. E. (1995). *Tests and Assessment*. 3rd ed. Englewood Cliffs. New Jersey: Prentice-Hall International, Inc.

Wicherts, J. M. (2007). *Group Differences in Intelligence Test Performance*. PhD diss., University of Amsterdam. Retrieved from https://pure.uva.nl/ws/files/4175964/46967_Wicherts.pdf

YOK. (2013). *Yurtdışından Öğrenci Kabulüne İlişkin Esaslar [Principles for the Acceptance of Students From Abroad]*. The decision of the General Council of Higher Education. Decision Date: 01.02.2013.

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

Zwick, R., & Sklar, J. C. (2005). Predicting College Grades and Degree Completion Using High School Grades and SAT Scores: The Role of Student Ethnicity and First Language. *American Educational Research Journal*, *42*, 439-464. DOI: https://doi.org/10.3102/00028312042003439