Polat, M. (2022). Comparison of performance measures obtained from foreign language tests according to item response theory vs classical test theory. *International Online Journal of Education and Teaching (IOJET), 9*(1). 471-485.

Received : 19.09.2021
Revised version received : 15.12.2021
Accepted : 17.12.2021

COMPARISON OF PERFORMANCE MEASURES OBTAINED FROM FOREIGN LANGUAGE TESTS ACCORDING TO ITEM RESPONSE THEORY VS CLASSICAL TEST THEORY

Research article

Murat Polat http://orcid.org/0000-0001-5851-2322

Anadolu University

mpolat@anadolu.edu.tr

Biodata:

Murat Polat received his Ph.D. from Osmangazi University, Graduate School of Educational Sciences. Currently, he is working as a language instructor at Anadolu University, School of Foreign Languages. His research interests include language assessment, IRT and educational statistics.

COMPARISON OF PERFORMANCE MEASURES OBTAINED FROM FOREIGN LANGUAGE TESTS ACCORDING TO ITEM RESPONSE THEORY VS CLASSICAL TEST THEORY

Murat Polat

mpolat@anadolu.edu.tr

Abstract

Foreign language testing is a multi-dimensional phenomenon and obtaining objective and error-free scores on learners' language skills is often problematic. While assessing foreign language performance on high-stakes tests, using different testing approaches including Classical Test Theory (CTT), Generalizability Theory (GT) and/or Item Response Theory (IRT) may help both to obtain results closer to true scores on students' proficiency levels and to minimize the amount of measurement error on test results, depending on item parameters, testing objectives and the amount of time and resources for valid and reliable evaluation. In this study, two popular testing theories the CTT and IRT were compared in testing language proficiency. Multi-dimensionality of two multiple-choice language tests taken by 2032 low-int and intermediate level language students in the spring term of 2018-2019 academic year was examined via CTT and IRT. In the first step of the analyses, the dimensionality test results revealed that test results were multidimensional. As a result of the NOHARM test, carried out to analyze which IRT model the data fit finest, it was determined that the test data fit the 3-parameter-logistic model. Eventually, it was determined that the correlation coefficients between foreign language proficiency estimations based on the CTT and IRT varied between 0.806 and 0.891. Thus, it was concluded that the two assessment theories (CTT focuses on measurement errors while the IRT focuses on individual traits) could let test designers obtain valid and reliable measurement scores, while the latter approach was observed to be slightly better at testing language achievement and/or proficiency.

Keywords: IRT, CTT, achievement tests, language proficiency, multi-dimensionality

1. Introduction

Foreign language testing might be considered as a distinct type of educational assessment since various cognitive and metacognitive processes involve in performance measurements considering the student's effort to express him/herself in a foreign language on one hand and the need to show his/her academic skills like reading, writing or interpreting on the other hand. Similar to most educational testing studies, estimating student achievement closer to its true value or the "quest for the true score" has been an ongoing craving for the language testing researchers for so long. As is known, language learners' real abilities (actual scores) are latent variables and cannot be measured directly in any language test (Anastasi & Urbina, 2002). For this reason, simulative settings (test questions) that are believed to be the valid indicators of students' real abilities are often designed considering a number of particular learning outcomes and language learners' reactions in these simulative settings are assumed to be the indicators of their language skills by the test designers.

One of the processes that must be completed with minimum errors in order to elicit the information obtained from these simulative settings to reflect the real ability of language testees is the degree to which the scores obtained from the test items that make up the simulative

situation represent the intended abilities to be measured. There are several ways to test whether the scores from the exam questions represent the intended skill. Of these, the most widely used is the dimensionality test to examine the number of dimensions among test questions. Measuring the exam to find out if it is unidimensional or multi-dimensional is a basic indicator of how well the exam can measure **(**Linden & Hambleton, 1997). The most widely used of these ways is to examine the dimensions of the test questions. Thinking of the test itself as a single dimension or a model consisting of a number of dimensions is a fundamental indicator of how well the test can fulfil the purpose of foreign language measurement.

While the discussion so far has been about the validity of foreign language test questions, how much these questions (if they represent) can measure students' real language ability is another important issue worth studying **(**Donlon, 1984). For example, let's consider the different measurement dimensions that an item of a reading skill test may contain. It is an important question how much of the score obtained from an item measuring reading for inference is related to a language learner's reasoning skills and how much it is related to reading skills. Moreover, when it comes to psychological characteristics, it is necessary to ask the question of how much of the test-takers' responding behaviour is the reason of the external stimuli (such as the test question).

The answer to whether the correct answer to a reading test question has the same probability for each respondent, whether it contains random error, and whether it requires the same amount of mental effort also determines if there is a change in the amount of diagnostic output that each question can trigger in each individual's mind. Thus, a language test-designer might wonder if a reading test question measures the same amount of information for each student. That's why, this question is synonymous with the question of how much of the real talent is measured by each question in a language test. To sum up, in this study, dimensionality of the questions in two separate language tests developed by a foreign language school of a state university in Turkey for lower-intermediate and intermediate level students was examined and whether the ability estimations obtained by different methods were similar to each other was tested. Thus, it was aimed to show the observed advantages and disadvantages of different scoring methods in foreign language tests according to different testing theories.

1.1. Statement of the Problem

Primarily, there are three theories for estimating student ability (measured performance score) in educational measurement and evaluation. The first and most widely used one is the CTT, the second and the most striking one in the last decades has been the IRT and the last but not the least one is Generalizability Theory (GT). Test scores in language schools (for instance the school where the data was taken from) are generally analysed based on the CTT and students' language skill estimations are made mostly by the use of those achievement scores gathered from the use of the CTT. However, taking only one testing approach into consideration would not be sufficient to determine the language learners' true language performance levels since various test models including productive and descriptive skills are used by means of cognitive and meta cognitive measurement processes.

It is evident that there might be more than one dimension in language tests and this possibility should be kept in mind and remind us the facts that not only the test item parameters but also the language level and some other critical variables such as the location and order of the items in the test booklet, and which skill the item measures before or after another particular language skill should also be considered (Hambleton & Jones, 1993). That is why, the current research aimed to determine how dimensionality of language scores obtained from two

different tests (end-course and proficiency exams) designed to test language learners of two different language levels (low-int and intermediate) test students' language skills in a language teaching program where more than 3000 students enrolled in two Turkish state universities, administered in the 2018-19 academic year.

Besides, it was aimed to investigate if the results obtained from the language tests' analyses were more compatible with the IRT or CTT model. Also, the analysis covering the estimation of the scores obtained from the language tests according to the CTT and the IRT would be would be studied to see if those methods show a significant difference in language learners' proficiency scores. Finally, with the help of the findings, it would be determined whether each item in the language tests used in the study has the same score weighting, and whether the objected measurement approaches (CTT or IRT) would make a significant difference in the success ranking of students without an in-depth analysis including some item parameters.

In line with the above research objectives, answers to the following questions were sought in the study:

1. How many dimensions does the data set obtained from the language tests have?

2. Which item response theory fits better with ability measure predictions according to item response theory?

3. Is there a significant correlation between the ability measures analysed according to the CTT and the IRT?

4. Do the ability measures obtained based on the CTT and IRT differ significantly according to the booklet type?

5. Do the ability measures obtained based on the CTT and IRT differ significantly according to the participants' language levels?

6. Does the booklet type have a significant effect on students' language scores?

2. Literature Review

2.1. Classical test Theory (CTT)

Classical Test Theory (CTT), which is the most common measurement approach and popular especially in exams taken by a small number of students, assumes that while measuring the students' abilities, the final test score can be found by an equation in which there is absolutely some measurement error (Steiger, 2000).

Accordingly: $X= T+E$

In this equation, X represents the calculated or observed score with the help of achievement tests, T stands for the actual score, and E represents the amount of random error in the measurement. Based on this equation, the CTT claims some assumptions **(**Traub,1997). The most important of these are:

1. The standard normal distribution is the distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

2. The correlation of random error with true score is zero.

3. The correlation between the error sums of the measurement results of two different features is equal to zero.

It should be underlined that above three assumptions are based on the concept of measurement error. The CTT urges that defining the error is critical and studying on its detection will bring the tester closer to the true score. Thus, in the CTT approach, the error is seen as the actual difference between the true score and the observed score. Fan (1998) stated that the CTT focuses on random errors rather than fixed and systematic errors, the direction and amount of which can be determined, namely the third type of error included in the theory. The existence of random error has led to the emergence of the concept of reliability of measurements in the CTT. Thus, Hambleton and Swaminathan (1985) concluded that the random error involved in the measurement processes is naturally in the model in accordance with the normal distribution.

In the CTT, there are two item statistics, the former is item difficulty whereas the latter is item discrimination (Taylor, 2009). Crocker and Algina (1986) define item difficulty (pj) as the percentage of correct answers in the group to which an item was asked and item discrimination (rjx) as the power of the item distinguish those who have the target skill and those who do not. While item difficulty is the ratio of those who answered the item correctly to all respondents, item discrimination is obtained by calculating the correlation between item scores and the scores obtained from the whole test.

For item discrimination in multiple-choice tests, the "lower-upper group method" can also be conducted by subtracting the response rate in the unsuccessful group from the rate of successful group considering the test results **(**Bachman, 2004). These two basic statistical features of the item are affected by the sample group parameters to which the test is applied (Ebel, 1979). That is, the difficulty and discrimination values obtained by analysing an item set from one group may differ from the difficulty and discrimination values obtained by applying the same item set to another group. This situation does not make it possible to predict the values to be obtained in subsequent applications by utilizing the parameters of the subjected items. That is, in the CTT item statistics are group dependent.

2.2. Item response Theory (IRT)

Item Response Theory (IRT) was developed in the late 1960s as an alternative to the testing assumptions of the CTT, especially in test development studies, and have found more and more applications up to now (Linden & Hambleton, 1997). Lord developed the two-parameter normal ogive model in this first study. Later, Birnbaum had efforts to develop user friendly logistic models for the normal ogive model, but Rasch developed one of the first logistic models based on the IRT, which would later be named after him (Hambleton & Swaminathan, 1985). Soon, the IRT started to be more and more popular, both because it was easy to adapt the tests to different ability groups, and because you can predict the ability measures of individuals based on the items without being dependent on the test (Ellis & Ross, 2014).

Henning et al. (1985) stated that the IRT is based on two basic premises:

• A person's performance on a test item can be predicted through a factor called implicit trait.

• The relationship between a person's performance on an item and the feature that enables him to respond to that item (item determination) can be shown with a curve called the item characteristic curve.

The mathematical term called the "item characteristic curve" is used to explain the success of a person's ability on an item **(**Van der Linden & Hambleton, 1997). This curve expresses the regression of the ability to be measured with the item on the probability of responding correctly

to that item. The shape of this regression curve depends on the number of parameters to be used to define the regression curve. Three parameters, item difficulty, item discrimination and chance success are used to determine the item characteristic curve and depending on which of these parameters are used, different models with one, two or three parameters have been determined in the item response theory (Crocker & Algina, 1986).

The most significant innovations brought by the IRT are the ease of estimating item parameters independently of a particular group sample (Hambleton & Dirir, 2003), estimating ability measures independently of the item sample (Chalmers, 2012), matching with the skills to be measured with the test, revealing the skills that have not yet been acquired and that have already been acquired on an individual basis (Embretson & Reise, 2000), and providing each individual with the appropriate item for his/her ability level (Kane, M. (2013).

In addition, some of the superior features the IRT also offers to researchers can be listed as follows:

• Being able to define different point values for each question, considering item characteristics (such as item difficulty and discrimination).

• Being able to predict at which cognitive level each question will be answered correctly.

• Being able to subtract the probability of giving a correct answer by chance for each question from the score estimated for that particular item.

• To be able to report student scores closer to the levels at evenly spaced scales.

• Calculating performance measures more accurately than the CTT.

• Estimating students' ability regardless of which ability group they are calculated in (item parameters are not affected by the group performance.

• Enabling individualized test applications in with computers (Çetin, 2019; Çokluk et al. 2016; Hofmans et al., 2009; Uzun et al., 2010).

2.3. Related Literature Findings

Studies related to the comparison of the CTT and IRT have revealed care-taking findings. To start with, their publication dates were considered a milestone and the findings were presented from the earliest to the most recent. Initially, Young (1991) compared the IRT and the classical cumulative score calculation method in terms (CTT) of estimating the abilities of male and female students and found that the scoring system based on the IRT was more predictive for both males and females than the CTT. Next, Gelbal (1994) compared the ability parameters estimated by the IRT Rasch model and the CTT on the findings of an achievement tests developed for primary school 5th grade Turkish and Mathematics lessons in his study.

According to the results, a positive and high relationship was found between the ability estimations obtained with both theories, and he concluded that this relationship increased as the students' ability level increased. However, in a study conducted by Fan (1998), it was reported that there was no significant difference between the characteristic measures of the IRT and the CTT, and one single theory could not be preferred over the others in various assessment and evaluation cases.

On the other hand, in a Monte Carlo study, individual parameters that emerged according to the IRT and CTT were compared and their results were reported to be very close to 1.00 (MacDonald & Paunonen, 2002). However, considering the calculations made according to the IRT, it was noticed that the item parameters revealed a great agreement. Therefore, the IRT

was recommended over the CTT when it comes to item selection for a test. What is more, in a study which estimated scores obtained from the driver's license test according to the CTT and IRT, even if they showed a strong correlation, the IRT's 3 parameter-logistics model allowed for more realistic estimations because of the in-depth item parameters (Wiberg (2004). Next, in her study, Ozkan (2012) compared the success scores of the Turkish and Mathematics subtests of the Student Achievement Examination predicted by the CTT and unidimensional and multidimensional IRT. As a result, it was found that the ability parameters obtained within the scope of multidimensional IRT models contained less error and provided more sensitive measurements than the CTT. Finally, Akyildiz and Sahin (2017) compared an achievement test of open education faculty on the basis of the CTT and IRT. They reported that the scores fit better to the multi-dimensional logistics model better.

To sum up, literature review findings has shown that it is not possible to talk about a definite result on the correlations between the IRT and the CTT in terms of superiority. While some studies reported a high correlation, some others showed a contrast (which was rarely significant) between those two assessment approaches. However, it was summoned that almost all researchers recommended the use of the IRT when it comes to item selection and reveal the achievement difference better compare to the CTT.

This study could be considered highly important for language performance measurement since the item response theory-based ability and parameter estimations on learners' language skills were made the first time for a Turkish language school's end-course and proficiency exams. Thus, the language school's confidence in the multiple-choice language tests would differ seeing the fact that for both testing approaches test results were similar and this similarity is highly significant. Moreover, seeing the results of the current research, they can adopt alternative test development ideas and compare those with the ones they use. Finally, this study is useful for researchers as it studied the multidimensionality and revealed an analysis in language skills estimation for large-scale testing.

3. Methodology

This study was designed as an exploratory research to identify which assessment model best fits the test data gathered from language learners. The answers given by the students to the language proficiency questions in two different tests were examined via different measurement techniques. Exploratory studies investigate such issues which are vague and confusing for scientists. They are mostly conducted to gain deeper understanding and knowledge about the existing research problems, whilst, they are not supposed to provide a definite and steady answer to the research matters. Within this aim in mind, the researcher uses the conventional method and tests its effectiveness comparing it with an alternative method; thus, the CTT was used to test the language learners skills with two multiple choice language proficiency tests while the same analysis was conducted on the basis of IRT to compare its results with the CTT.

3.1. Participants

For data collection, a particular sampling method was not used in this study. A total of low-intermediate and intermediate 2032 learners' language achievement and proficiency tests' scores were used in the Spring Term of 2018-2019 academic year. The research data used in the analyses was taken from a language school which serves for two separate state universities in Turkey.

3.2. Data Collection Process

In the initial phase of data collection, the objective of the study was announced and necessary official permission was taken from the language school administration. Next, exam scores of the participant group were taken from the school's database. The data was recorded in two different memory cards by the researcher to ensure its safety. Later, students' test scores were coded and grouped with the SPSS 26 data analysis program.

First, the data of 2097 students were taken from the database and examined carefully. 65 students' test data was excluded from the data set since the score sets had some missing values. Thus, 2032 students' test data was considered appropriate and identified as two separate score sets (end-course test and proficiency test) and the original form of each item was re-coded in a 1-0 matrix (to illustrate, the test item 34's answer was A but the student response was either C or D) with the answer keys suitable for each test's booklet codes. A similar re-coding application on achievement scores was reported by Akyildiz, and Sahin, (2017). They re-coded the students' answers of an achievement test conducted at open education faculty as 1-0 in a matrix for their comparative study. Finally, the accuracy and reliability of the data was tested and reported by comparing and contrasting the total scores of each participant with the total scores listed in the original database score set.

3.3. Data Analysis

After the exclusion of missing values from the data set, the analyses of language test scores according to the CTT and IRT were conducted. The data-analysis model used in the study was organized by taking the example of Akyildiz and Sahin's (2017) research. First, dimensionality analyses were performed on the language score data. The dimensions of the tests were determined with two different statistical tests. First, the dimensionality was tested with the DIMTEST method, which is a nonparametric approach developed by Stout in 1987. Next, dimensionality analysis was conducted with the NOHARM test, which is another advanced dimensionality test investigating curvilinear relationships (Yavuz & Dogan, 2015). In order to find an answer to the research question considering the dimensionality of the test, the scores were analysed to find out the appropriate IRT model by using the IRTPRO. The reason for using the test was to determine how many parameters the data related to the language scores fit the IRT model (Zwick & Velicer, 1986).

In the next stage, in order to find an answer to compare the CTT and IRT models, raw scores of students were analysed by using the end-course and proficiency exam data. Then, ability estimations were driven using the IRT model, in which the model fit was higher, with the use of the IRTPRO software. Whether or not these ability estimations differ significantly was analysed by calculating the correlation coefficients of two different testing models. To be able to find an answer to the last research question, the relationships between the item parameter estimations obtained according to both the CTT and the IRT were examined through correlation coefficients test regarding the test booklet type. Thus, it was investigated whether the booklet type had a significant effect on the item parameters in both testing models. The item parameters (a) item discrimination, (b) item difficulty and (c) probability of answering the question by chance, which were most highlighted based on IRT, were obtained respectively from the data set.

4. Findings

In this exploratory study, it was aimed to investigate what kind of similarities or differences would emerge in the scoring of foreign language tests using the CTT and IRT methods. It was also aimed to find out which of these measurement models would be more efficient in assessing foreign language skills via multiple-choice tests in different language levels. In order to answer the first research question of the present study (How many dimensions did the research data have?), the DIMTEST and NOHARM tests were conducted.

The DIMTEST checks if two groups of items show a significant difference from each other in terms of test correlation when the items of the test are divided into two groups. Thus, the first group of items analysed in the study was the group of items with a higher correlation than the other group. The second group of items included the remaining items. The DIMTEST reveals whether there is a significant difference between these two item groups. As a result of the DIMTEST analysis, a T value and the significance of this T value are underlined. If the T value is significant, it is decided that there is a significant difference between the two groups of items, so it is understood that the measurement tool has more than one dimension.

The values obtained as a result of the DIMTEST analysis were shown in Table 1.

Table 1. DIMTEST results

DIMTEST	T	p
End Course	0,3201	0,2365
Proficiency	4,0036	0,0000

The results of the DIMTEST analysis presented in Table 1 led us to the conclude that the end-course test results were unidimensional while the results of the proficiency test were multidimensional. Hence, it was understood that the proficiency exam data set should be analysed with the IRT. It should also be noted that, although the DIMTEST is supposed to be a better dimensionality analysis method than other classical dimensionality approaches such as Exploratory Factor Analysis (EFA), it basically analyses probable relations between items and student abilities within a linear approach. For this reason, by the use of NOHARM test, which enables more appropriate analysis to the main objective of the IRT, dimensionality analyses were repeated considering the fact that correlation among language skills and students' answers to different questions might have been curvilinear.

Nevertheless, thinking of the probability that the study data could be one-dimensional, two-dimensional or multidimensional, the data was re-analysed accordingly with the use of NOHARM analysis. Tanaka's GFI and RMSR (Root Mean Score of Residuals) values were calculated by the researcher against each dimensionality possibility prediction. In such calculations, the GFI (Goodness of Fit) value is expected to be close to 1. Although there is no specific criterion reported in the literature among the indices with normal distribution such as Tanaka's GFI indices (Tanaka, 1987), which is obtained for different dimensionality cases, Cheung and Rensvold (2002) state that if there is a difference of 0.01 or more for the fit indices, the difference will be significant. Next, it was suggested that the NOHARM analyses on the data set could be repeated for up to 2 factors, and the analyses were continued until taking the 2-factor results since no different results were obtained for the 3-factor and subsequent factors

(Raykov & Marcoulides, 2006). Accordingly, the fit indices obtained as a result of NOHARM were given in Table 2.

Table 2. NOHARM Dimensionality Analysis

		With Prediction Correction		Without Prediction Correction	
Test	Factor	RMSR	TANAKA'S GFI	RMSR	TANAKA'S GFI
End-Course	1	0,007	0,896	0,007	0,889
	2	0,003	0,994	0,003	0,995
Proficiency	1	0,007	0,883	0,008	0,942
	2	0,005	0,976	0,004	0,988

As can be seen from Table 2, the data obtained from foreign language tests were more in line with a multidimensional model since they were close to 1. For this reason, the exam data was considered multidimensional for the analyses to be made in the following steps.

The second research question was "Which item response model would best fit the language skill measures to be predicted according to the item response theory? ". At this stage, the end-course and proficiency exams were analysed according to one-parameter, two-parameter and three-parameter IRT models, respectively, not only to determine the appropriate model, but also do find the best fit measurement model which could be determined by comparing the fit values (RMSEA) obtained from each IRT model. The fit values calculated as a result of the analyses and the proposed model to be used in the language school for the language exams were presented in Table 3.

Table 3. IRT Data-Model Fit Coefficients

Test	RMSEA (1 parameter)	RMSEA (2 parameters)	RMSEA (3 parameters)
End-Course	0,06	0,04	0,03*
Proficiency	0,06	0,05	0,04*

*Best fit

Table 3 shows the model fit coefficients (RMSEA) obtained from the data of end-course and proficiency exams. It should be noted that the RMSEA value must be at least 0.05 in order to be able to say that the data fits well to the model (Meade et al.,2008). Similarly, the model-data fit increases as the RMSEA value is close to 0.00. Checking the results presented in Table 3, it was interpreted that the most appropriate model for the ability measures to be estimated according to the IRT for the end-course and proficiency exams was the two-dimensional 3PL (Parameter Logistic) IRT model.

The third question of the study was to test whether there was a high and significant relationship between the ability measures estimated according to the CTT and the IRT. After determining the dimensionality and the most suitable IRT model for the end-course and

proficiency exams of the language program, students' language skill estimations were analysed using the IRTPRO software, based on the IRT model. In the next step, the Pearson Product-moment Correlation Coefficient between the students' ability measures obtained using the IRT and the CTT were calculated using the SPSS 26 package program.

For each dimension of the foreign language exams, the correlation between the ability measures obtained from the IRT and CTT was examined and two separate correlation values were taken for the end-course and proficiency exams. The results of this test were presented in Table 4.

Table 4. IRT and CTT correlation coefficients according to language test scores

Test	Correlation Coefficients	
	1st Dimension	2nd Dimension
End-Course	0.902*	0.899*
Proficiency	0.861*	0.882*

*p<.01

In Table 4, it is seen that there was a high and significant relationship between the student ability measures obtained by analysing the exam scores on the basis of the CTT and IRT **(**p<.01**)**. It was also seen that student achievement scores measured according to CTT and IRT for the 1st dimension in the end-course exam, showed a high and positive correlation 0.902 and for the second dimension, this correlation was 0.899. As for the proficiency exam' 1st dimension, it was found that the student achievement scores measured according to the IRT and CTT showed a correlation up to 0.861, and for the second dimension this correlation was 0.882.

The fourth question of the study was to find out whether the ability measures and item parameters obtained from the data set according to the CTT and ITT differ significantly according to the booklet type. In this stage, correlation coefficients between the ability measures obtained from the exams according to the IRT and CTT were analysed separately for booklet A and B. The correlation coefficients were shown in Table 5.

Table 5. IRT and CTT correlation coefficients according to different test booklets

Test	Correlation Coefficients			
	Booklet A		Booklet B	
	1st Dimension	2nd Dimension	1st Dimension	2nd Dimension
End-Course	0.887*	0.878*	0.885*	0.870*
Proficiency	0.858*	0.813*	0.861*	0.846*

*p<.01

When the results in Table 5 were examined, it was found that the correlation coefficients calculated between the ability measures according to the CTT and IRT for each booklet in end-course exam was multidimensional and varied between 0.870 and 0.887, and all these correlation coefficients were statistically significant (p<.01). Accordingly, depending on the booklet, it was concluded that the ability measures of the students in the end-course test showed insignificant changes depending on the estimation method.

Talking of the proficiency exam, it was concluded that the correlation coefficients calculated between the ability measures according to the CTT and IRT for each language test booklet was determined to be multidimensional and varied between 0.813 and 0.861, and these correlation coefficient values were statistically significant (p<.01). It should also be added that the relatively higher correlation value between the first and the second dimensions of the test was observed in booklet B.

Table 6. Correlation coefficients between the IRT and CTT according to language levels

Test	Correlation Coefficients			
	Low-Int		Intermediate	
	1st Dimension	2nd Dimension	1st Dimension	2nd Dimension
End-Course	0.877*	0.869*	0.861*	0.850*
Proficiency	0.853*	0.831*	0.853*	0.831*

*p<.01

The last but not the least, findings in Table 6 revealed that the correlation coefficients calculated between the ability measures according to the CTT and IRT considering the participants' language levels showed that the end-course test was multidimensional and the correlation level varied between 0.850 and 0.877, and all these correlation coefficients were statistically significant (p<.01). Accordingly, depending on the learners' language levels, it was concluded that the ability measures of the intermediate students in the end-course test showed minor changes depending on the estimation method. Talking of the proficiency exam, it was found that the correlation coefficients calculated between the ability measures according to the CTT and IRT considering the participants' language levels showed that the test was multidimensional and the correlation level varied between 0.831 and 0.853, and all these correlation coefficients were statistically significant (p<.01). The relatively higher correlation coefficient for both the first and the second dimensions was observed in low-int language level.

5. Discussion & Conclusion

When the findings of the study were examined as a whole, it was concluded that the proficiency test to find out if the student is compatible in English language, which were taken by more than two thousand foreign language learners, was multidimensional. This finding is highly important for decision makers since it should be considered by language school test teams and IRT based assessment and evaluation models are ought to be developed to test students' language skills better. Bachman (2004) underlines the fact that proficiency tests are critical exams in which serious decisions are taken on learners' actual language levels. Possible error in those tests could affect the individuals' whole lives and lead to other problems no matter if the error causes the student pass or fail. Therefore, this finding is critical especially for the high-stakes language exam designers. Next, the analyses revealed a positive relationship between the total scores of end-course and proficiency exams, which were calculated by assigning 1 point to each item according to the CTT, and the language skill measures estimated by the IRT approach in which scores assigned to each question with different values, considering the student's ability, item difficulty, item discrimination quality, and the probability of the question being answered correctly by chance. This finding is in line with the findings of MacDonald and Paunonen (2002). Seeing this fact, they suggested the use of the IRT especially in forming the item sets since the IRT enables the identification of dimensionality.

Moreover, high correlation coefficients between the IRT and CTT were found indicating positive and significantly high score similarity. This similarity was a bit higher than the 1st dimension compared to the test items considering the 2nd dimension in the language tests. This result revealed that there is a slight difference between the measurement results between the IRT and CTT. Özkan (2012) reported a similar finding in her study. compared the success scores of the Turkish and Mathematics subtests of the Student Achievement Examination predicted by the CTT and unidimensional and multidimensional IRT. Nevertheless, it can also be concluded that both methods can be used by language schools to have reliable and valid ability estimations in terms of language testing when the similar scores are considered, but as the findings of the research suggest the use of the IRT can be recommended to have more valid and reliable test results.

Next, the research findings revealed that the ability measures estimated according to the CTT and the IRT showed high correlation coefficients between 0.801 and 0.894. Although these relationships were high, they do not mean that the way an institution measures learners' success according to a particular theory can easily be replaced with another. To underline this finding, Akyildiz and Şahin (2017) compared the CTT and IRT results and reported that the scores fit better to the multi-dimensional logistics model better; whereas, these ability estimates were not interchangeable and administrator should not prefer the IRT or CTT interchangeably. For this reason, it can be concluded that only one of these two approaches should be determined by institutions and all language exams should be planned, administered and evaluated in accordance with the pre-determined testing approach.

Additionally, it was found that the type of booklet could affect the exam results, which were considered as the dependent variable in the study. Provided that grammar questions, which were supposed to have less cognitive load, were asked right after the listening section (answered jointly in both tests) while the questions were placed in the booklet, the average scores of the students was found to be a bit higher compared to the scores of the students who answered the reading questions after listening section. Akyildiz and Şahin (2017) reported the same finding as the correlations between the estimations of the students' ability measures according to the CTT and IRT showed slight differences compared to the booklets, and they advised that the difference observed between the booklets should be examined. This finding; therefore, could help language test designers to make important interpretations on which skills be measured former and which skills be measured latter in high-stakes language tests by which language assessment and evaluation practices could direct students' academic lives.

Finally, it was observed that language learners' proficiency levels could have a minor effect on the exam results. This finding could be interpreted as the sign of the high validity and reliability degrees of both language tests since there was a positive and significantly high correlation among test scores, testing theories and the students' language levels considering the mean scores. The small difference between the dimensions in booklets (particularly in intermediate level booklet) could be the sign of possible item difficulty which is supposed to be higher in intermediate level end-course test. Ebel, R. L. (1979) reported a similar case in his book and suggested the test designers prepare more complex items to language learners as their language levels increase in time since language learning process is not static, the language measurement should not be static either. Language test contents and the way they test should revolve and develop in parallel with the student's continuous cognitive development and language learning.

References

Akyildiz, M. & Sahin, M. D. (2017). Açık öğretimde kullanılan sınavlardan Klasik Test Kuramına ve Madde Tepki Kuramına göre elde edilen yetenek ölçülerinin karşılaştırılması. *Açık öğretim Uygulamaları ve Araştırmaları Dergisi, 3 (4),* 141-159. Retrieved from https://dergipark.org.tr/tr/pub/auad/issue/34247/378491

Anastasi, A. & Urbina, S. (2002). *Psychological testing*. Prentice Hall: New York.

Bachman, L. (2004). *Statistical analyses for language assessment*. New York, NY: Cambridge University Press.

Chalmers, R. P. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48(6*).

Cheung, G., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* 1, 233-255.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory,* (1st ed.). Belmont, CA: Wadsworth Group/ Thomson Learning.

Çetin, B. (2019). *Eğitimde Ölçme ve Değerlendirme,* Anı Yayıncılık.

Çokluk, Ö., Şekercioğlu G. & Büyüköztürk, Ş. (2016). *Multivariate statistics, SPSS and lyrical applications for social sciences* (4th Edition). Ankara: Pegem Academy.

Donlon, G. (1984). *The college board technical handbook for the Scholastic Aptitude Test and achievement tests.* New York: College Entrance Examination Board.

Ebel, R. L. (1979). *Essentials of educational measurement,* 1st Edition. Upper Saddle River, NJ: Prentice Hall.

Ellis, D., & Ross, S. (2014). *Item response theory in language testing.* In A. Kunnan (Ed.), Companion to language assessment, Vol. 3. Hoboken, NJ: Wiley Blackwell.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item person statistics. Educational and Psychological Measurement, 58(3), 367-381.

Gelbal, S. (1994). p madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 10,* 85-94.

Hambleton, R., & Dirir, M. (2003). *Classical and modern item analysis.* In Fernández-Ballesteros (Ed.), Encyclopaedia of psychological assessment (pp. 188–192). London, UK: Sage Publications.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement*: Issues and Practice, 12(3), 3847.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*: Principles and applications. Boston: Kluwer.

Hofmans, J., Theuns, P., & Van Acker, F. (2009). Combining quality and quantity. A psychometric evaluation of the self-anchoring scale. *Qual. Quant. 43*, 703–716.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2(2),* 141–154.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1*), 1–73. doi:10.1111/jedm.12000.

Linden, W. J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement, 62(6*), 921-943.

Meade, A.W., Johnson, E.C. & Braddy, P.W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568–592.

Özkan, Ö. Y. (2014). Öğrenci başarılarının belirlenmesi sınavından klasik test kuramı, tek ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması. *International Journal of Human Sciences,* 11(1), 20-44.

Raykov, T. & Marcoulides, G.A. (2006). *A First Course in Structural Equation Modeling*. Lawrence Erlbaum Ass., London.

Steiger, J. H. (2000). *Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser*. Structural Equation Modeling, 7, 149-162.

Tanaka, J.S. (1987). How big is big enough? Sample size and goodness-of-fit in structural equation models with latent variables. *Child Development, 58,* 134-146.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21-36. doi:10.1017/S026719050909003

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and practice*, 8-1 4.

Uzun, N. B., Gelbal, S., & Öğretmen, T. (2010). *Kastamonu Üniversitesi Kastamonu Eğitim Dergisi. 18* (2), 531-544

Van der Linden, W., & Hambleton, R. (1997). *Handbook of modern item response theory.* New York, NY: Springer-Verlag.

Yavuz, G., & Doğan, N. (2015). Test Boyutluluğunun Analizinde Kullanılan Programların İncelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 6(2);* 293-312

Young, J. W. (1991). Gender bias in predicting college academic performance: a new approach using IRT. *Journal of Educational Measurement*. 28(1), 37- 47.

Wiberg, M. (2004). *Classical Test Theory vs. Item Response Theory. An evaluation of the theory test in the Swedish driving-license test*. Umba University.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.