ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(2); 98-112

# An Analysis of Parameter Invariance according to Different Sample Sizes and Dimensions in Parametric and Nonparametric Item Response Theory *

Çiğdem REYHANLIOĞLU ** Nuri DOĞAN ***

**Abstract**

The aim of this study is to analyze the invariance of the estimated parameters in regard to Item Response Theory (IRT) applications within different dimensionality and sample characteristics. To realize this aim, students' answers in test booklet 'A' of the first stage of TEOG, Secondary Education Placement Test applied by the Ministry of Education in Turkey, in 2015 were used as the data. The population size of the study was determined as 63,871. The groups of 50, 100, 200, 500, 1000 and 5000 people who were randomly selected from the population comprised the sample of the study. One-dimensional mathematics subtest and artificially formed two-dimensional test results were used within IRT applications. As a result of the findings obtained from the study within the analysis of the one-dimensional test in regard to one-dimensional nonparametric IRT (ODNPIRT), item parameter invariance was obtained from the sample size 200. When it was analyzed in regard to one-dimensional parametric IRT (ODPIRT), it was found that at least a sample size of 1000 has to be used for item parameter estimations nearly as high as the population value (parameter). It was found that invariance for item parameters was not obtained by analyzing the two-parameter test in regard to ODPIRT, ODNPIRT and Multidimensional IRT (MDIRT).

*Key Words:* Item Response Theory applications, sample size, parameter invariance.

## INTRODUCTION

In the literature, in a situation where a test is multidimensional, it is discussed whether the score of the whole test can be used to determine the performance of individuals or how. Therefore, in a situation where a test is not one-dimensional, test theories based on the one-dimensional assumption may be insufficient in the analysis of data obtained from multi-dimensional tests. In this case, the models used in the process of estimating the ability and item parameters belonging to individuals obtained from a multi-dimensional test are significant (Meara, Robin & Sireci, 2000). These models must be models capable of analyzing multi-dimensional data that do not require a one-dimensional assumption. On the other hand, it is quite common to use parametric methods that require many assumptions when assessment theories are analyzed. However, situations where parametric conditions cannot occur, and one-dimensionality is not achieved are frequently encountered in educational practices. The absence of consistent parametric conditions in education and psychology and the necessity of using tests with multiple sub-dimensions to make decisions about individuals has made it necessary to develop applications other than parametric and one-dimensional models. In other words, nonparametric models have been developed in cases where parametric conditions are not achieved, and multi-dimensional models have been developed for cases where one-dimensionality cannot be achieved. However, before these developed models can be used, their functionality must be demonstrated experimentally. For this purpose, the results obtained from these models developed within the framework of theories hast to be compared with the results obtained from existing theories.

* This study is based on Çiğdem REYHANLIOĞLU's doctoral dissertation titled An Analysis of Parameter Invariance according to Different Sample Sizes and Dimensions in Parametric and Nonparametric Item Response Theory.

** PhD., Gaziantep College Foundation Private Schools, Gaziantep-Turkey, dr.cigdemreyhanlioglu@gmail.com, ORCID ID: 0000-0002-4685-0495

*** Prof. PhD., Hacettepe University, Ankara-Turkey, e-posta: nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

To cite this article:
Reyhanlıoğlu, Ç., & Doğan, N. (2020). An analysis of parameter invariance according to different sample sizes and dimensions in parametric and nonparametric item response theory. *Journal of Measurement and Evaluation in Education and Psychology, 11*(2), 98-112. doi: 10.21031/epod.584977

*Geliş Tarihi: 01.07.2020*
*Kabul Tarihi: 07.04.2020*

The development and popularity of parametric IRT models is undoubtedly a significant step in the development and use of modern test theory. Significant findings have been obtained that large samples are needed to use these models. This is a significant limitation in terms of the applicability of the parametric models in schools and other field studies. Due to this limitation, in field studies and especially in schools, the item parameters are estimated based on the sample characteristics, whereas the success levels of individuals can be estimated based on the item parameters. However, it is considered that obtaining the invariance characteristic, which is the most important characteristic of IRT and helps estimate the item and ability parameters independent of each other, is considered to be significant in terms of student ability levels and decisions made based on test and item parameters. By obtaining the invariability of IRT, it is possible to estimate the item parameters of the tests applied in schools independent of ability distributions of students, whereas it is also possible to estimate the ability distributions independent of the item and test parameters (Price, 2017). Thus, the decisions made depending on the tests applied in schools are expected to be more accurate thanks to the invariance characteristic. It is of great importance to spread and use nonparametric IRT models, especially in schools in order to apply to small study groups. Nonparametric IRT models have significant advantages over parametric models in order to provide a solution to the necessity of using large samples. Another important advantage of nonparametric IRT models compared to parametric models is that the relationship between the responses to the items and the latent variable measured by the items has fewer assumptions. This is because the item characteristic curves of nonparametric IRT models do not have a predefined parametric form (Sodano & Tracey, 2011). Accordingly, it can be stated that nonparametric models are more useful than parametric models. However, although nonparametric models have significant advantages, there is a need to collect evidence that they work as well as parametric models in order to use these models, and the results obtained from this study will provide significant evidence in this regard. This study compared the results of a newly widespread theory in Turkey, one-dimensional nonparametric IRT, with the results obtained from one-dimensional parametric IRT and multidimensional IRT, which is used in the analysis of multidimensional tests.

### *The Aim of the Study*

Depending on the literature research in Turkey, although there are various studies on comparing the nonparametric one-dimensional and multidimensional IRT applications on multicategory data (Koğar, 2018; Şengül-Avşar, 2018; Şengül-Avşar, 2017), there is a limited number of studies carried out with dichotomous data. These studies were carried out by Koğar (2014) and Mor-Dirlik (2017). This study is significant in terms of being one of the first research carried out on actual dichotomous data in Turkey. Within this context in this study, it was aimed to test the functionality of parametric and nonparametric IRT models over a dichotomous one-dimensional data set. However, one-dimensional parametric and nonparametric IRT models of a multidimensional data set were compared with the results obtained from the multidimensional IRT model. Thus, it is aimed to demonstrate the sufficiency of one-dimensional parametric and nonparametric models in the analysis of data in a situation where the one-dimensional hypothesis fails.

Reliable data is expected to have a repeatable structure. Therefore, the same causes are expected to give the same results under the same conditions. Data that is not duplicated and based on a cause-effect relationship may not be considered scientific. Scientific data consists of the results obtained by repeating the research in different environments. Also, scientific data is cumulative. A study carried out by a scientist is expected to support previous studies, and in cases where it does not support previous studies, it is expected to reveal the causes clearly. Depending on the literature research in Turkey and around the world, many studies that compare the results obtained from IRT applications in terms of different variables such as the number of dimensions in scales (Smits, Timmerman & Meijer 2012) and ability level distributions (Syu, 2013) were found. One of these variables is the sample size (Koğar, 2014; Köse, 2010; Sünbül, 2011). The sample size is an important factor to consider when deciding which of the parametric and nonparametric models serve the same purpose. Whether the results obtained in this study differ or not were analyzed according to the sample size as well as theoretical applications. In this context, this study is also significant because it is one of the studies in which the results obtained from IRT applications are compared according to the sample size. One of the aims of the study is to determine a sample size that can be a criterion for the usage preference of parametric and nonparametric IRT models in the literature. In line with this purpose, it was aimed to reveal whether the ability levels obtained from the sub-tests of the TEOG application in different dimensions differ within the scope of the multidimensional IRT and the parametric and nonparametric IRT, considering the sample sizes. Within this purpose, the problem statement to be addressed in this research is structured as "To what extent is the invariance of the parameters ensured for the parametric and nonparametric IRT according to the variables of different dimensionality and sample size?" This study seeks answers for the sub-problems stated below.

1. In one-dimensional tests, what are the estimated standard error of the means for the item parameters calculated regarding the parametric and nonparametric IRT, when the sample size is 50, 100, 200, 500, 1000, and 5000 and estimated from the population?
2. In multidimensional tests, what are the estimated standard error of the means for the item parameters calculated regarding one-dimensional and multidimensional IRT when the sample size is 50, 100, 200, 500, 1000, and 5000 and estimated from the population?
3. In multidimensional tests, what are the estimated standard error of the means for the item parameters calculated regarding the parametric and nonparametric IRT, when the sample size is 50, 100, 200, 500, 1000, and 5000 and is estimated from the population?

Different indicators are used for item difficulty and discrimination index for one-dimensional nonparametric IRT (ODNPIRT), one-dimensional parametric IRT (ODPIRT), and Multidimensional IRT (MDIRT) when interpreting the findings obtained from the results of sub-problems. For ODNPIRT, the Hi parameter was used when interpreting the discriminative power of the item, and the fact that the value of Hi is less than 0.30 indicates that the item is weak in terms of discrimination according to Sijtsma and Molenaar (2002). For ODNPIRT, p values, which are the classic difficulty parameter, were used as an indicator of item difficulty. For ODPIRT, parameter *a* is used as an indicator of discrimination, and parameter *b* is used as an indicator of item difficulty. Theoretically, *a* and *b* parameters take values in the range of $(-\infty, +\infty)$. Finally, for MDIRT, as in ODPIRT, parameter *a* is used as an indicator of discrimination. A different discrimination parameter is used for each dimension of the test in MDIRT. Since the integrated test used in this study is two-dimensional, two discrimination parameters are estimated. These are parameters $a_1$ and $a_2$. Parameter *d* is used as the indicator of item difficulty for MDIRT. Similarly, parameter *d* is interpreted as parameter *b* in ODPIRT.

## METHOD

This study is a descriptive study in terms of obtaining descriptive statistics about one-dimensional parametric and nonparametric IRT and multi-dimensional IRT models, and the existence and degree of the relationship between two or more variables. Descriptive research tries to explain "what" cases, objects, assets, institutions and various fields are (Kaptan, 1977).

### *Population and Sample*

The data used in the study was obtained from the students' answers in the test booklet 'A' of the first stage of the TEOG, Secondary Education Placement Test applied by the Ministry of Education in Turkey, in 2015. Booklet 'A' comprises of Turkish, Mathematics, Science, Religion Culture and Moral Knowledge, Revolution History, English, German and French subtests. In this study, students' answers in booklet A in all subtests comprise the population of the research. The population size is 63,871. As the results obtained from the study were compared regarding the sample sizes, the groups of 50, 100, 200, 500, 1000 and 5000 people who were randomly selected from the population comprised the sample of the study. While determining the sample sizes, field applications were taken into consideration especially for small samples. While determining the lower limit of the sample size, the average number of 8th-grade students studying in private institutions in Gaziantep during the 2017-2018 academic year was considered. According to the information received from the R&D unit affiliated to Gaziantep Provincial Directorate of National Education, there are a total of 1188 grade 8 students studying in 24 private middle schools in Gaziantep. The average of the number of students per school is 49.5. Therefore, the lower limit of the sample size was determined as 50. Other small sample sizes (100, 200, and 500) were determined to be doubles of 50. In determining the large sample sizes, the results of the studies conducted by Hullin, Lissak and Drasgow (1982), Goldman and Raju (1986) and Thissen and Wainer (1982) were taken into consideration.

Samples of 50, 100, 200, 500, 1000 and 5000 randomly selected from the population were selected only once without replication. Although not performing replication in the selection of the sample was considered as a limitation of the study, this method was used to overcome the problems encountered in case of replication for large samples. As an example, when selecting a sample of 5000 people if 50-100 replications are performed, there may be many identical individuals in each sample, and these individuals can expand parameter invariance. In other words, they automatically cause invariance and / or cause estimations to be biased. Therefore, no replication was performed in this study.

### *Data Collection Tools*

The data used in the study was obtained from the students' answers in 'A' test booklet of the first stage of TEOG, which comprises of Turkish, Mathematics, Science, Religion Culture and Moral Knowledge, Revolution History, English, German and French subtests.

### *Process*

In accordance with the purpose of the study, it is aimed to use the results of one-dimensional and two-dimensional subtests in the booklet A used in the first stage of TEOG in the analysis process. For this purpose, it was examined whether all subtests in TEOG were suitable for factor analysis according to the results of KMO and Bartlett's Test of Sphericity. KMO values of all subtests in TEOG were above 0.90. Bartlett's Test of Sphericity results were statistically significant for all subtests. In this case, the Bartlett test obtained for each subtest included in the TEOG is statistically significant, indicating that the data derived from the multivariate normal distribution and therefore the data has a suitable structure to apply factor analysis (Çokluk, Şekercioğlu & Büyüköztürk, 2010). In addition, the KMO value needs to be greater than 0.60 in terms of the suitability of the sample size to factor analysis (Tabachnick & Fidell, 2001). Factor analysis was carried out using the FACTOR 10.5.01 program, which provides dimensional results based on parallel analysis using the polychoric correlation matrix for each subtest in line with the obtained results. As a result of the dimensionality analysis, it was determined that all subtests of TEOG had a single dominant dimension. Two indicators were used to demonstrate the compatibility of the data of each subtest with the single factor model obtained. These are GFI and RMSR. The model was compatible with the data in GFI's proximity to 1. If RMSR is smaller than Kelly's criterion value (0.0316), it can be stated that the model complies well with the data (Harman, 1962). The GFI values obtained were in the range [0.998, 1] and the RMSR values were in the range [0.001, 0.023].

After determining that all the subtests in TEOG have a one-dimensional structure, a new test in the same length (20 items) with the test used in the two-dimensional and one-dimensional IRT analyzes was created to be used for MDIRT analysis by selecting items from two subtests. Factor analysis results obtained from the binary combinations of the subtests were taken into consideration in the selection of the subtests used in the creation of the test required for the MDIRT analysis. When combined, the subtests that best provide two dimensions are Science and Religion Culture and Moral Knowledge subtests. A two-dimensional integrated test was created by selecting 10 items that have high correlations within themselves and low correlations with other test items in Science and Religious Culture and Moral Knowledge subtests. The necessary evidence for the two-dimensional structure of the created test was obtained by factor analysis. The GFI value is 0.99 and the RMSR value is 0.01 among the two indicators used to demonstrate the compatibility of the existing data with the two-factor model obtained. As a result, it was concluded that the data of the integrated test comply with the two-factor structure.

### *Data Analysis*

The data analysis process started with testing the necessary assumptions of IRT. One-dimensionality assumption was tested for ODPIRT, and local independence was tested for MDIRT, whereas one-dimensionality and monotony assumptions were tested for ODNPIRT. Local independence assumption was not tested separately for ODPIRT and ODNPIRT. The reason for this is that obtaining one-dimensionality is enough to obtain the local independence assumption (Hambleton, Swaminathan & Rogers, 1985). Therefore, the local independence assumption was tested only for MDIRT.

One of the methods used to test local independence is to analyze the correlations between conditional items (Ferrara, Huyny & Baghi, 1997; as cited by Bulut, 2015). In this study, inter-item correlations between a certain ability range (high ability and low ability groups) were used to test local independence. 20% and 80% of the raw scores were used to identify the lower and the higher groups. The fact that the elements in the diagonal of the variance and covariance or correlation matrices obtained for individuals with limited ability levels are 0 or very close to 0 indicates that the local independence assumption is achieved (Hambleton, 1991; McDonald, 1981; as cited by Bulut, 2015). Based on this, the correlations between the items obtained from the responses of the individuals in the low and high ability groups to the items were obtained in this study. According to the results obtained, the correlations between the items were very low in both ability groups. Thus, it can be concluded that the local independence assumption was achieved for the integrated test.

Finally, the monotony assumption was tested for ODNPIRT. In this study, package Mokken developed by Van Der Ark (2007) for the R 3.0.2 software was used to test the monotony assumption. Before proceeding to interpret the results of the monotony assumption, it is useful to explain important symbols and abbreviations. (# AP) represents the number of active pairs for each item, (#VLM) represents the amount of the violation of latent monotony, (#VLM / #AP) represents the average amount of the monotony violation for each item pair, (maxVLM) represents the maximum value of the amount of the monotony violation, (TOT) shows the amount of total monotony violation, whereas (TOP / #AP) shows the amount of total monotony violation for each item pair. If all these values are significantly greater than 0, the latent monotony assumption is violated (Van der Ark, 2007). Another indicator that is significant for interpretation is the item scalability coefficient Hj. The fact that the Hjs belonging to each item is less than 0.30 indicates that the item is weak in terms of discrimination. Hj is interpreted as the item discrimination coefficient (Sijtsma & Molenaar, 2002). In line with the results, it was observed that the assumptions were not achieved by some items for Religion Culture and Moral Knowledge, English and Science subtests. All items in Turkish and Maths subtests were found to achieve the assumption. Especially for all the items in the Math test, (#AP), (#VLM) (#VLM/#AP), (maxVLM), (TOT) and (TOT/#AP) values were 0. Similarly, the scalability coefficient of all items in the math test was over 0.30.

Based on the results obtained, the results of the Math subtest, which was determined to have stronger one-dimensional evidence in ODPIRT and ODNPIRT analyses and which also met the monotony assumption for ODNPIRT, were used. The results of the two-factor integrated test created by selecting certain items from the Religion Culture and Moral Knowledge and Science tests were used for MDIRT, as stated earlier. After determining the tests used for one-dimensional and multi-dimensional data analysis, it is necessary to determine which models will be used within the scope of one and multi-dimensional IRT. 2 PLM and 3 PLM were applied to the results of the mathematics test for ODPIRT. The estimated number of parameters for 2 PLM is 40 and the -2 LL value obtained is 1419674.370. For the 3 PLM, the estimated number of parameters is 60 and the -2 LL value obtained is 1702461.230. The difference value obtained in 20 degrees of freedom is 282786.86. This result was significant when compared with the critical value of $\chi^2$ (31.410) at 20 (60-40) degrees of freedom.

Therefore, 2PLM with a low -2 LL value makes a significant difference compared to 3PLM; In other words, it can be said that 2PLM is a model that is more suitable to the data. In addition, Embretson and Reise (2000) recommend using 3PLM when working with multiple-choice test items, whereas they recommend using 1PLM or one of 2PLM when working with personality data. However, considering the -2LL values of the models, it was decided to apply 2PLM as the ODPIRT model.

For MDIRT, extended M2PLM and M3PLM were applied to the results of the integrated test. The estimated number of parameters for M2PLM is 40, and the -2 LL value obtained is 1214446.23. For M3PLM, the estimated number of parameters is 60, and the -2 LL value obtained is 1214490.81. The difference value obtained in 20 degrees of freedom is 44,.8. This result was significant when compared with the critical value of $\chi 2$ (31.410) at 20 (60-40) degrees of freedom. In this case, the preferred model for analysis is the extended M2PLM, which has a smaller value of -2 LL.

For ODNPIRT, it was decided to use Monotone Homogeneity Model chosen between the Double Monotony Model (DMM) and Monotone Homogeneity Model (MHM) proposed by Andries van der Ark (2007). This is because every data set that can be explained by DMM can be explained by a weaker Model (MHM) (Andries van der Ark, 2007).

In ODNPIRT analyses, package Mokken developed by Van Der Ark (2007) for the R 3.0.2 software was used. In ODPIRT and MDIRT analysis, FlexMIRT 3.5 software recommended by Cai (2017) was used. ODPIRT parameter estimation was made according to 2 PLM, and multi-dimensional IRT analyzes were performed according to the extended 2 PLM. In both ODPIRT and MDIRT analyses, error values were determined by estimating the error values of the EM algorithm of Cai (2008).

In this study, test scores were used instead of the item scores. Therefore, the average difficulty and discrimination coefficients were estimated for the total test scores. For models belonging to all three theories, the average difficulty and discrimination levels of the test were calculated separately for each sample drawn once from the population. The difference between the parameter averages estimated for each sample from the population value was analyzed. However, this difference was not statistically tested. The comments were made only within the framework of a large-small size relationship.

In order to analyze the parameter invariance, which is the focus of the study, standard error of the means (SEM) was used as an indicator of parameter invariance (Koğar, 2014; Sünbül, 2011). However, while interpreting the findings, it was examined how the parameter means were affected by the sample size as well as the SEMs for the estimation of the parameter. For this purpose, it was examined how SEMs of the parameter estimated for each sample differed from the population value. Just like the parameter means, this difference was not statistically tested. Since this is a descriptive study, the findings obtained were interpreted only within the frame of a large-small size relationship.

How the findings of the study were obtained can be explained by two phases. The first one is that the one-dimensional Mathematics test was modeled under one-dimensional parametric and nonparametric IRT, and item parameters were estimated according to this modeling. In the second phase, the analysis was performed with a one-dimensional model, according to ODPIRT and ODNPIRT based on the assumption that the two-dimensional integrated test is accepted as one-dimensional. In addition to these, the two-dimensional integrated test was modeled and analyzed according to MDIRT in accordance with its nature, and item parameters were estimated as a result of these analyzes.

### *Internal and External Validity of the Research*

The internal validity of the research is related to the degree to which the changes in the dependent variable are explained by independent variables. In this study, the internal validity has been ensured since the change in the item and ability estimations and the reliability levels of these estimations can be explained by the sample size and different applications of IRT (Fraenkel & Wallen, 2006).

The external validity of the research is related to the generalizability degree of the findings. Since the generalizability of the findings obtained from this study is limited with the sample size used, IRT applications and the subject area of the tests used, it is necessary to analyse the external validity of the research in this framework. Therefore, it is thought that the research results can be generalized within the framework of the sample sizes used, IRT applications and the subject area of the tests (Fraenkel & Wallen, 2006).

## RESULTS

### *Findings Related to the Solution of the First Sub-Problem*

For the solution of the first sub-problem, item parameters and standard error of the means (SEM) of the mathematics test estimated according to ODPIRT and ODNPIRT are demonstrated in Table 1.

Table 1. Item Parameters and Standard Error of the Means (SEM) of the Mathematics Test Estimated According to ODPIRT and ODNPIRT

| Sample size | ODNPIRT | | | | ODPIRT | | | |
|---|---|---|---|---|---|---|---|---|
| | H | $H_{SEM}$ | p | $p_{SEM}$ | a | $a_{SEM}$ | b | $b_{SEM}$ |
| 50 | 0.32 | 0.09 | 0.48 | 0.109 | 1.07 | 0.32 | 0.09 | 0.37 |
| 100 | 0.39 | 0.07 | 0.45 | 0.109 | 1.31 | 0.28 | 0.24 | 0.23 |
| 200 | 0.33 | 0.05 | 0.43 | 0.107 | 1.32 | 0.22 | 0.34 | 0.18 |
| 500 | 0.33 | 0.03 | 0.43 | 0.107 | 1.41 | 0.15 | 0.34 | 0.10 |
| 1000 | 0.33 | 0.02 | 0.43 | 0.107 | 1.48 | 0.11 | 0.30 | 0.07 |
| 5000 | 0.33 | 0.01 | 0.43 | 0.107 | 1.50 | 0.05 | 0.29 | 0.03 |
| Population | 0.33 | 0.00 | 0.43 | 0.107 | 1.52 | 0.01 | 0.30 | 0.01 |

The first one of the item parameters of the one-dimensional test estimated according to ODNPIRT is the H parameter means, which are the indicators of discrimination. When the H means given in the table are examined, it is seen that the population value is 0.33. The most distant H mean relative to the population value is estimated from the sample size 100 (H = 0.39). Even the most distant H mean to the population value is not much different from the population value itself. Therefore, it can be concluded that the H means estimated from the samples are close to the population value in size. However, starting from the sample size of 200, it is seen in the table that Hs have a stable structure to reflect the population value (H = 0.33). The population value of the SEM of the H is very close to 0. It is also seen in the table that the sample size that reflects the population value the least is the smallest sample 50 ($H_{SEM}$=0.09), and that the SEMs of the H parameter approximates to the population value as the sample size increases. In addition, regardless of the sample size, SEMs are estimated to be near the population value in size. According to these results, it can be concluded that it is not necessary to use large samples to estimate the H with SEMs close to the population value.

One of the item parameters of the one-dimensional test estimated according to ODNPIRT is the p means, which are the indicators of difficulty. The p mean that is most distant from the population value was estimated in the smallest sample (p = 0.48). However, as seen in Table 1, even the most distant p mean is not much different from the population value.

The population value of the SEM of the p was estimated to be 0.107. The most distant SEM relative to the population value was estimated from 50 and 100 sample sizes ($H_{SEM}$ = 0.109). In the 200 sample size, SEM got the lowest value ($p_{SEM}$ = 0.107), and it can be said that this value is equal to the population value. Just as with the SEMs of the H, it was determined that there was no big difference between the SEMs of the p estimated from the samples and the population value.

However, there was no change in SEM sizes in the samples between the sample size 200 and the population. As a result, it was observed that SEMs of item parameters and item parameter means estimated according to ODNPIRT have a stable structure to reflect the population as from the sample size 200. Accordingly, it was concluded for ODNPIRT that the parameter invariance was obtained from the sample size 200.

The first one of the item parameters of the one-dimensional test estimated according to ODPIRT is the a parameter means, which are the indicators of discrimination. It can be seen in Table 1 that the mean of parameter a estimated from the samples has values between 1.07 and 1.52 depending on the sample size. The population value of the a parameter mean was estimated at 1.52. The a parameter takes the most distant value to the population value at 50, which is the smallest sample size, and as the sample size increases, it is also seen in the table that the values of the a parameter mean tend to increase and approximate to the population value. The a parameter means are estimated to be very close to the population value as from the sample size of 1000.

For ODPIRT, the population value of the SEM of the a parameter was estimated to be 0,01. It is seen in the table that the most distant SEM is estimated from the sample size of 50 ($a_{SEM}$ = 0.32). As the sample size increased, the SEM of the parameter a showed a decreasing trend and approximated to the population value. As seen in the table, it can be concluded that the parameter a is estimated with a SEM much closer to the population value in terms of the sample size of 500.

One of the item parameters of the one-dimensional test estimated according to ODPIRT is the parameter b mean, which is the indicator of difficulty. The population value of the parameter b is estimated to be 0.30. The mean of the b parameter that is most distant from the population value was obtained from the sample size of 50 (b = 0.09). The sample size that reflects the population value best is 1000. According to the results obtained, it is not possible to say that the mean of the parameter b tends to increase or decrease regularly with the change in the sample size. However, it can be said that the changes in the mean of the parameter b are less as from the sample size of 1000. In this case, it can be concluded that the means of parameter b are close to the population as from the sample size of 1000.

The population value of the SEM of the b parameter was estimated to be 0.01 for ODPIRT. The SEM that is the most distant to the population value was estimated from the smallest sample size 50 ($b_{SEM}$ = 0.37). It can be seen in Table 1 that the SEM of b parameter is steadily approximating to the population value due to the increase in the sample size. As seen in the table, it can be concluded as from the sample size of 500 that the parameter b is estimated with a SEM that is closer to the population value.

As a result of the analysis of the item parameter estimations obtained through ODPIRT, it was concluded that the parameter invariance could not be ensured since the a and b parameters and SEMs belonging to the a and b parameters showed a continuous change.

### *Findings Related to the Solution of the Second Sub-Problem*

For the solution of the second sub-problem, the item parameters and standard error of the means (SEM) of the integrated test estimated, according to MDIRT and ODPIRT are demonstrated in Table 2.

Table 2. Parameters and Standard Error of the Means of the Integrated Test Estimated According to One and Multi-Dimensional IRT

| Sample Size | MDIRT | | | | | | ODPIRT | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $a_1$ | $a_{1\ SEM}$ | $a_2$ | $a_{2\ SEM}$ | d | $d_{SEM}$ | a | $a_{SEM}$ | b | $b_{SEM}$ |
| 50 | 0.98 | 1.47 | 0.51 | 0.75 | 1.12 | 0.99 | 1.23 | 0.54 | -0.82 | 0.48 |
| 100 | 0.60 | 0.70 | 0.44 | 0.45 | 0.77 | 0.53 | 1.24 | 0.75 | -0.74 | 0.47 |
| 200 | 0.68 | 0.57 | 0.42 | 0.30 | 1.02 | 0.43 | 1.48 | 0.52 | -0.79 | 0.36 |
| 500 | 0.68 | 0.43 | 0.40 | 0.20 | 0.87 | 0.36 | 1.68 | 0.69 | -0.55 | 0.19 |
| 1000 | 0.67 | 0.25 | 0.41 | 0.14 | 0.84 | 0.20 | 1.76 | 0.16 | -0.48 | 0.09 |
| 5000 | 0.67 | 0.11 | 0.42 | 0.06 | 0.88 | 0.08 | 1.94 | 0.07 | -0.49 | 0.03 |
| Population | 0.41 | 0.01 | 0.43 | 0.01 | -0.18 | 0.01 | 1.52 | 0.01 | 0.30 | 0.01 |

The first of the item parameters of the two-dimensional test estimated according to MDIRT are the parameters $a_1$ and $a_2$, which are indicators of discrimination. The population value of $a_1$ estimated from MDIRT was estimated to be 0.41. For $a_1$, the most distant estimation from the population value was obtained from the smallest sample size of 50. There is no sample size that reflects the population value of $a_1$ well. The population value of $a_2$ was estimated to be 0.43. The most distant estimation from the population value of $a_2$ was obtained from the smallest sample size and, the estimates of $a_2$ were obtained to be close to the population value as from the sample size of 100. According to the results obtained, it is not possible to say that $a_1$ and $a_2$ means have a regular increasing or decreasing tendency depending on the change in the sample size.

It can be seen in the table that the population value of SEM, which is the mean of the discrimination parameters for MDIRT, is estimated to be 0.01. The SEM estimate that is the most distant from the population value was obtained from the smallest sample size ($a_{1\ SEM}$=1.47; $a_{2\ SEM}$=0.75). With the increase in sample size, the SEM of both discrimination parameter means tended to decrease, and they approximated to the population value. It can be seen in Table 2 that $a_1$ is estimated with SEM that is close to the population value as from the sample size of 5000, whereas the $a_2$ mean is estimated with SEM that is close to the population value as from the sample size of 1000. However, SEM, which is the mean of both discrimination parameters, shows a continuous change from the samples to the population.

As a result, it was concluded that the parameter invariance could not be achieved since the average values of the discrimination parameters estimated from the MDIRT and their SEMs demonstrated a continuous change.

The population value of the d parameter estimated from the MDIRT was estimated as (d = -0.18). The mean of the d parameter that is most distant from the population value was obtained from the smallest sample size (d = 1.12). When the d parameter means obtained from the samples are examined, it is seen that they are different from the population value. According to the results obtained, it is seen that the mean of the d parameter does not have a regular increase or decrease tendency due to the change in the sample size. However, it can be seen in Table 2 that the variation between 500 and 5000 sample sizes is less than that of other sample sizes.

For MDIRT, the population value of the SEM for the mean of the d parameter was estimated to be 0.01. It is seen in the table that the SEM estimate that is the most distant to the population value is obtained from the smallest sample size and the SEM estimates approximate to the population value as the sample size increases. It is also seen in Table 2 that SEMs of the d parameter mean received different values throughout the samples.

Just like the mean of the discrimination parameters, it has been concluded that the parameter invariance cannot be achieved because the mean of the d parameter and the SEM of the d parameter mean demonstrated a continuous change from the samples to the population.

The first of the item parameters of the two-dimensional test estimated according to ODPIRT is a parameters, which are indicators of discrimination. It can be seen in Table 2 that the mean of a parameter estimated from the samples has values between 1.23 and 1.94 depending on the sample size. The population value of the mean of a parameter is 1.52, and the sample size that best reflects the population value is 200. In a more general sense, while the closest estimates to the population value were obtained in the medium-sized samples used in this study, the more distant estimates were obtained in the largest and smallest samples. It should also be noted that the population value of a parameter is the same as the population value obtained from the one-dimensional test. The point that makes the comparison meaningful here is that the compared values are obtained from the same population but from structures with two different dimensions. The conclusion is that in cases which one-dimensionality assumption is ensured and is not ensured, in other words, the mean of the population values of parameter an estimated from one and multi-dimensional tests does not change.

For ODPIRT, the population value of the SEM of the a parameter mean was estimated to be 0.01. The most distant SEM to the population value was obtained from the sample size 100 ($a_{SEM}$ = 0.75). It cannot be said that the SEM of a parameter mean shows a regular increase or decrease trend depending on the sample size. In some of the consecutive sample sizes, the population demonstrated a value that was more distant to the SEM of the a mean, while in others, it demonstrated a closer value. Modeling the multi-dimensional data structure under a one-dimensional model is the main reason for this irregularity. It is also seen in the table that the SEMs belonging to the a parameter mean get values that do not reflect the population up to the sample size of 1000, whereas they steadily decreased and approximated to the population value as from the sample size of 500. In addition, the SEMs belonging to the a parameter show a continuous change from the smallest sample size to the largest sample size.

Other item parameters of the two-dimensional test estimated according to ODPIRT is the b parameters, which are indicators of difficulty. The population value of the b parameter mean is estimated to be 0.30. The most distant estimation in terms of the population value was obtained when the sample size was 50 (b = -0.82). However, there is no sample size that reflects the population value of the mean of the b parameter well. The conclusion “in cases which one-dimensionality assumption is ensured and is not ensured, in other words, the mean of the population values of parameter a estimated from one and multi-dimensional tests does not change” also applies to parameter b. In other words, the population values of the estimated b parameters, whether one-dimensional or multi-dimensional, are the same. According to the results obtained, it is seen that there is no clear increase or decrease tendency due to the change of the sample size of the mean of the parameter b.

For ODPIRT, the population value of the SEM of the a parameter mean was estimated to be 0.01. The most distant SEM of the b parameter to the population value was obtained from the sample size 50 ($b_{SEM}$ = 0.48). As the sample size increased, SEMs approximated to the population value. It is also seen in Table 2 that the b parameter is estimated by a SEM that is very close to the population value as from the sample size of 500, SEM ($b_{SEM}$ = 0.19). Just like the a parameter, SEMs belonging to the b parameter got values in different sample sizes.

As a result of the analysis of the SEMs belonging to the item parameters and the item parameters estimated from ODPIRT, it was concluded that the parameter invariance could not be achieved since the a and b parameters and SEMs of the a and b parameters demonstrated a continuous change.

### *Findings Related to the Solution of the Third Sub-Problem*

For the solution of the third sub-problem, the item parameters and standard error of the means (SEMs) of the integrated test estimated according to ODPIRT and ODNPIRT are shown in Table 3.

Table 3. The Item Parameters and Standard Error of the Means (Sems) of the Integrated Test Estimated According to ODPIRT and ODNPIRT

| Sample Size | ODPIRT | | | | ODNPIRT | | | |
|---|---|---|---|---|---|---|---|---|
| | a | $a_{SEM}$ | b | $b_{SEM}$ | H | $H_{SEM}$ | p | $p_{SEM}$ |
| 50 | 1,23 | 0,54 | -0,82 | 0,48 | 0,46 | 0,10 | 0,65 | 0,097 |
| 100 | 1,24 | 0,75 | -0,74 | 0,47 | 0,38 | 0,07 | 0,65 | 0,096 |
| 200 | 1,48 | 0,52 | -0,79 | 0,36 | 0,39 | 0,05 | 0,68 | 0,096 |
| 500 | 1,68 | 0,69 | -0,55 | 0,19 | 0,38 | 0,03 | 0,65 | 0,096 |
| 1000 | 1,76 | 0,16 | -0,48 | 0,09 | 0,40 | 0,02 | 0,65 | 0,096 |
| 5000 | 1,94 | 0,07 | -0,49 | 0,03 | 0,40 | 0,01 | 0,66 | 0,095 |
| Population | 1,52 | 0,01 | 0,30 | 0,01 | 0,33 | 0,00 | 0,43 | 0,107 |

The results obtained by analyzing the integrated test according to ODPIRT, were explained in detail in the solution of the second sub-problem. The first parameter obtained by analyzing the integrated test according to ODNPIRT is the H parameter.

It can be seen in the table that the H mean of the population is 0.33 (H = 0.33). The Hi mean that is the most distant from the population value was estimated from the sample size 50. (H = 0.46). In cases which one-dimesionality assumption is ensured and is not ensured, in other words, the mean of the population values of parameter H estimated from one and multi-dimensional tests does not change. The sample sizes that reflect the population best in terms of the H mean are 100 and 500 sample sizes.

For the SEM of the H, a population value close to 0 is estimated ($H_{SEM}$= 0.00). The most distant SEM in terms of population value was estimated at 50, which is the smallest sample size ($H_{SEM}$ = 0.10). It is seen that SEMs belonging to the H decrease steadily as the sample size increases and approximate to the population value. In addition, it was determined that SEMs obtained from the samples did not differ greatly from the population value. In this case, when working with small sample sizes over multi-dimensional data, it can be said that it is possible to estimate the discrimination parameter with SEMs close to the population value. In addition, when one and multi-dimensional tests were analyzed according to ODNPIRT, it was observed that SEMs belonging to H parameters obtained were very close to each other. In this case, it can be concluded that in cases where one-dimensionality is violated, the H means are estimated with similar SEM values. However, SEMs show a continuous change in size throughout the samples.

As seen in Table 3, it can be concluded that invariance cannot be obtained for the H because the H and its SEMs differ with each other throughout the samples. It was explained in the findings of the first sub-problem that invariability was ensured as from a small sample size in the case that one-dimensionality was ensured. The violation of one-dimensionality, which is the most important assumption for using one-dimensional models, affected ODNPIRT results in terms of invariance.

Another item parameter of the two-dimensional test estimated according to ODNPIRT is the p parameters, which are indicators of difficulty. It is seen in Table 3 that the test is relatively easy in the samples compared to the p means. However, it is seen that there is no big difference between the population value of the p mean and its value that is most distant from the population in terms of size. The population value of the p was estimated to be 0.43 (p = 0.43). In terms of size, the most distant p-value in the population was estimated as from the sample size 200. As in the population value of the H, the population value of the p was found to be the same as the estimated population value when the one-dimensional test was analyzed according to ODNPIRT. In that case, it can be concluded that in cases which one-dimensionality assumption is ensured and is not ensured, in other words, the mean of the population values of the p estimated from one and multi-dimensional tests does not change.

As seen in Table 3, the population value of SEM for p was obtained as 0.107 ($p_{SEM}$=0.107). The most distant SEM value of the population value was estimated from the sample size 5000 ($p_{SEM}$=0.095). When the one-dimensional test is analyzed according to ODNPIRT, it is a significant finding that the p means obtained from the population are estimated with the same SEM. Accordingly, it can be concluded that in cases where one-dimensionality is ensured and violated, SEMs of the p means are the same. It can be seen in Table 3 that there is no big difference between the population value of the SEMs of the p means and its most distant value to the population value. For this reason, it can be said that the change in sample size does not cause big differences on SEMs in which p means are estimated. In this case, if one-dimensionality cannot be achieved and large samples are not reached, it can be concluded that it is possible to estimate the p parameter with SEMs that is close to the population value. The SEMs belonging to the p mean have different values in the largest and smallest sample sizes but did not change in other sample sizes.

In addition, since the p mean values and SEMs had different values in the largest and smallest sample sizes, it was concluded that the invariance could not be achieved for the p mean. In case that one-dimensionality is achieved, it was explained in the findings of the first sub-problem that invariability is achieved as from a small sample size. Similar results were presented for the H mean.

If the results obtained from ODPIRT and ODNPIRT are compared, it can be seen in table 3 that H means are estimated with SEMs close to the population value even in small sample sizes according to a parameter means. According to the results obtained from Table 3, it is necessary to work with a minimum sample size of 1000 or even sample size 5000 in order to estimate the a parameter mean with a SEM close to the population value. It can be concluded that ODNPIRT is more advantageous than ODPIRT since the average discrimination parameters can be estimated with a SEM close to the population value from a smaller sample. A similar result was obtained when the one-dimensional test was analyzed according to parametric and nonparametric IRT. So whether the one-dimensionality assumption is achieved or not, large samples must be used for ODPIRT to estimate the average discrimination parameter with a SEM close to the population value. There is no such limit for ODNPIRT. Similar comments are valid for the average difficulty parameters obtained from parametric and nonparametric IRT. It is necessary to work with a minimum sample size of 1000 to estimate the b parameter mean with a SEM that is close to the population value. However, it is seen in Table 3 that it is possible to work with smaller sample sizes to estimate the p parameter mean with a SEM close to the population value.

## DISCUSSION and CONCLUSION

In accordance with the results obtained from the findings of the first sub-problem, it was seen that the one-dimensional test was analyzed according to ODNPIRT and the item parameter invariance was achieved as from the sample size 200. In this case, it was concluded that ODNPIRT application could be used for the data obtained from small samples for the purpose of estimating the average item parameter when large sample sizes cannot be achieved. This result is significant in terms of enabling the item parameter estimations of the tests applied at different levels of education, especially in school applications.

By analyzing the one-dimensional test according to ODPIRT, it was observed that as the sample size increased for ODPIRT, the SEM of the a parameter mean decreased and approximated to the population value. Thissen and Wainer (1982) stated that for parameter estimation, a sample size of 10.000 and more are needed, while Goldman and Raju (1986) stated that at least 1000-people samples are needed for the correct estimation of the parameter a. The case that the sample size exceeds 5000 and the mean value of the a parameter and the standard error mean of the a parameter mean continue to change supports the findings of Thissen and Wainer (1982).

The SEM belonging to the b parameter mean obtained by analyzing the one-dimensional test according to ODPIRT, tended to decrease as the sample size increased and approximated to the population value. The b parameter mean did not show a regular increase or decrease trend depending on the sample size. Sünbül (2011) expressed this situation as “the insignificance of sample size over parameter b”. However, Hulin, Lissak and Drasgow (1982) revealed that there were no significant changes in the estimated parameters for samples larger than 1000 for 2 PLM. Therefore, the findings obtained are consistent with those of Hulin, Lissak and Drasgow (1982).

As a result, the fact that the item parameters obtained from small sample sizes are stable to reflect the population, in other words, ensuring parameter invariance as from a small sample size is a proof that ONDPIRT has significant advantages over ODPIRT. Therefore, instead of ODPIRT, ODNPIRT applications can be preferred, especially in school applications.

Based on the results obtained from the findings of the second sub-problem, it was concluded that the mean of the $a_1$ and $a_2$ parameters obtained by analyzing the integrated test according to the MDIRT did not show a regular increase or decrease trend depending on the sample size. Ackermann (2005) concluded that as the sample size increases, the discrimination power of the items increases. In this respect, the result obtained from this study differs from the result obtained by Ackermann (2005).

According to the results of the MDIRT analysis, it is seen that there is no clear increase or decrease tendency due to the change in the sample size of the d parameter mean. It was concluded that the parameter invariance could not be achieved since the d parameter mean and the SEM of the d parameter mean also showed a continuous change from the samples towards the population.

When the integrated test was analyzed according to the MDIRT, both of the mean of a parameters obtained for each dimension demonstrated values in the range (0,1). The average discrimination parameters estimated according to ODPIRT demonstrated values above 1. This result differs from the results found by Ansley and Forsyth (1985) in their study where they obtained the estimation of a parameter from two theories with values that are close to each other. However, according to the results obtained by analyzing the multi-dimensional test according to ODPIRT, it was concluded that the parameter invariance could not be achieved since the a and b parameter means and SEMs belonging to the a and b means demonstrated different values in all samples.

As a result, whether a multidimensional data is modeled by ODPIRT or MDIRT, it is necessary to use the sample size 5000 at the lowest to make an average parameter estimation with SEM that is close to the population value. If 5000 or larger samples are used, the SEM approximates to the population value. However, although the SEM of the parameter means approximates to the population value, the parameter means obtained by analyzing the multi-dimensional data according to the one-dimensional model and the parameter means obtained when analyzed according to the multi-dimensional model differed from each other regardless of the sample size. In other words, parameter invariance was not achieved. This result is consistent with the findings of Drasgow and Parsons (1983), who stated that the results of parameter estimations differ when the one-dimensionality assumption is not achieved. In this study, it is considered that replication was not performed in the sample selection as the reason for the fact that the invariance could not be achieved, and the parameter means did not show a clear increase or decrease depending on the sample size.

In line with the results obtained from the findings of the third sub-problem, it was concluded that the SEMs belonging to H mean obtained by analyzing the integrated test according to ODNPIRT decreased steadily as the sample size increased and approximated to the population value. Koğar (2014) revealed that as the sample size increases, the standard error of the H coefficients decreases. This result was consistent with the findings of Koğar (2014). In addition, regardless of the sample size, SEMs have taken very close values to the population in terms of size. In this case, it was concluded that violating one-dimensionality did not have a significant effect on SEM values which Hi and *p* means were estimated.

The most important effect of violating the one-dimensionality assumption on parameter estimates was that when the one-dimensionality assumption was achieved, the parameter invariance was achieved, and when the assumption was not achieved, the parameter invariance could not be achieved. For this reason, Koğar’s (2014) finding that invariance is achieved for *p* values as the sample size increases could not be observed in this study. This result obtained by Koğar (2014) was obtained for *p* means of one-dimensional data as stated in the findings of the first sub-problem. In other words, as the sample size increased, parameter invariance was achieved for the *p* means of one-dimensional data.

When the results obtained from ODPIRT and ODNPIRT are compared, it was concluded that the parameter means estimated from ODNPIRT can be estimated with a SEM close to the population value. In order to estimate the parameter means obtained from ODPIRT with a SEM close to the population value, it is necessary to work with a minimum sample size of 1000. The result that estimating item parameters from a smaller sample with SEM close to the population value is proof that ODNPIRT is advantageous over ODPIRT. However, it is a significant finding that the fact whether the one-dimensionality assumption is achieved does not change this result.

### *Suggestions*

In this study, 50, 100, 200, 500, 1000, and 5000 sample sizes were studied to analyze the effect of sample size. A similar study can be carried out with different sample sizes. In this study, 2PLM for ODPIRT, MHM for ODNPIRT, and M2PLM for MDIRT were used. A similar study can be carried out with different models whose results belong to the theories studied.

The parameter values estimated on the models used in this study are the parameter mean values of the entire test. A similar study can be carried out for parameter estimates of items in a test. In this study, standard error mean of parameter estimation was used in order to research the effect of sample size. In other studies, different indicators can be used to research the effect of sample size.

In this study, the effect of chance success was not taken into consideration. Other studies can be carried out on real scores that are free of chance success. In this study, the effect of sample size on parameter invariance was studied. A similar study can be carried out with different independent variables. In this study, each sample size was selected only once in the sample selection. In other studies, the effect of sample size on parameter invariance can be studied by selecting samples through replication.

## REFERENCES

Ackerman, T. A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (p. 3–25). Lawrence Erlbaum Associates Publishers.

Andries van der Ark, L. (2007). Mokken Scale Analysis in R. *Journal of Statistical Software*, *20*(11).1-19.

Ansley, T.N. and Forsyth, R.A. (1985). An Examination of the characteristics of Unidimensional IRT estimates derived from two dimensional data. *Applied Psychological Measurement, 9*(1), 37-48.

Şengül Avşar A. (2018). Kategori sayısının psikometrik özellikler üzerine etkisinin mokken homojenlik modeli’ne göre incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 9*(1), 49-63.

Şengül Avşar A. , Tavşancıl E. (2017). Examination of polytomous items’ psychometric properties according to nonparametric item response theory models in different test conditions, *Kuram ve Uygulamada Egitim Bilimleri*, 17, 493-514.

Cai L. (2017). *flexMIRT Version 3.51: Flexible multilevel multidimensional item analysis and test scoring* (Computer software). Chapel Hill, NC: Vector Psychometric Group.

Cai, L. (2008). SEM of Another Flavour: Two new applications of the supplemented em algorithm. *British Journal of Mathematical and Statistical Psychology,* 61, 309–329.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik teknikleri.* Ankara: Pegem Akademi.

Drasgow, F. & Parsons,K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*(2), 189-199.

Embretson, S. E., & Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Fraenkel, J.R. & Wallen, N.E. ( 2006). *How to design and evaluate research in education* (Sixth edition). Boston: McGraw-Hill Pub.

Goldman, S.H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational And Psychological Measurement*, *46*(1),11-21.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Academic Puslishers Group.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo Study. *Applied Psychological Measurement, 6*(3), 249-260.

Harman, H. H. (1962). *Modern Factor Analysis* (2. edition). University of Chicago Press, Chicago.

Kaptan, S. (1977). *Bilimsel Araştırma Teknikleri,* Ankara: Tekışık Matbaası ve Rehber Yayınevi.

Koğar H. (2018). Examining invariant item ordering using mokken scale analysis for polytomously scored items. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *9*(4), 312-325.

Koğar, H. (2014). *Madde tepki kuramının farklı uygulamalarından elde edilen parametrelerin ve model uyumlarının örneklem büyüklüğü ve test uzunluğu açısından karşılaştırılması.* (Doktora tezi, Hacettepe Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Meara, K., Robin, F. & Sireci, S.G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35*(2), 229–259.

Mor-Dirlik, E. (2017). *Parametrik ve parametrik olmayan madde tepki kuramı modellerinden çeşitli faktörlere göre elde edilen madde ve yetenek kestirimlerinin karşılaştırılması*. (Doktora tezi, Ankara Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Price, L. R. (2017). *Psychometric methods: Theory and practice*. New York, NY: The Guilford Press.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Thousand Oaks, CA: Sage.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory mokken scale analysis as a dimensionality assessment tool: why scalability does not ımply unidimensionality. *Applied Psychological Measurement, 36*(6), 516-539.

Sodano, S. M., and Tracey, T. J. G. (2011). A brief inventory of interpersonal problems- circumplex using non-parametric item response theory: introducing the IIP-C- IRT. *Journal of Personality Assessment, 93*(1), 62-75. doi:10.1080/00223891.2010.528482

Syu, J. J. (2013). *Applying person fit-in faking detection-the simulation and practice of non parametric item response theory*. (Doctoral Dissertation, National Chengchi University).
Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi.* (Doktora tezi, Mersin Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Tabachnick, G. B. ve Fidell, S. L. (2001). *Using multivariate statistics* (4th Edition), Boston MA: Allyn&Bacon.

Thissen, D., & Wainer, H. (1982). Some standart errors in item response theory. *Psychometrika, 47*(4), 397-412.