ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(3); 254-266

Comparison of Testlet Effect on Parameter Estimates Using Different Item Response Theory Models

Esin YILMAZ KOĞAR *

Abstract

In this study, the testlet effect was calculated for each testlet in the PISA 2018 reading literacy test, and it was examined whether this effect caused a difference in item and ability parameters. The data set was analyzed with a two-parameter logistic item response theory model and a two-parameter logistic testlet model. The results show that variances of testlet effects range from .100 to .432. When the item and ability parameter estimation results of the models were compared, it was determined that the item and ability parameters estimated from the two approaches were highly correlated with each other. It can be said that the item slope and item intercept parameters estimated from different models remained unaffected. However, when the local dependency assumption is not met, it was observed that the standard error values of the two-parameter model for the ability parameter were underestimated. The implications for the analysis and evaluation of the tests based on testlet are discussed. In conclusion, in this study, it was concluded that the testlet effect caused a difference in parameter estimates, but the local dependence among the items was negligible because of the small testlet effects.

Key Words: Local item dependency, item response theory, testlet response theory, testlet effects, PISA.

INTRODUCTION

A testlet is defined as a cluster of items that share a common stimulus (Wainer & Kiely, 1987). This common stimulus can be presented as a passage, scenario, table, or figure. Testlets are widely used in testing for several reasons such as ensuring the effective use of the time required for the test application, reducing the context effect that may arise from the content of the items in the test, eliminating the concerns that a single independent item may be too atomistic (measuring a very specific or narrow concept) because of its nature (Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007). However, if different items are collected in the same testlet, these items may be related to each other beyond the effect of the latent trait that is tried to be measured. This situation, known as local item dependency (LID), leads to the violation of the local independence assumption of standard item response theory (IRT) models. For example, the performance of students in a reading comprehension test may be affected by their interest in or knowledge of reading passages, as well as their reading skills (Yen, 1993). Therefore, items in the same set of items may be locally dependent.

The local item dependency (LID) between testlet items is called the testlet effect (Wainer & Kiely, 1987). Bradlow, Wainer, and Wang (1999) proposed a new model by adding this effect as a parameter to the 2-parameter logistic model (Birnbaum, 1968, 2PLM). In this model, which is called the testlet response theory (TRT) model, there is a random-effects parameter, γ, that considers account the dependencies between the items in the same testlet. In the standard 2PL IRT model, there are item difficulty and item discrimination parameters, and it is assumed that there is no local dependence between items. In the TRT model, calculations are made by including item difficulty and item discrimination parameters, as well as a random effect parameter. The 2PL TRT model, which is developed in the standard 2PL IRT model, can be written as (Li, Bolt, & Fu, 2006; Ip, 2010);

* Asst. Prof. Dr., Niğde Ömer Halisdemir University, Faculty of Education, Niğde-Turkey, esinyilmazz@gmail.com, ORCID ID: 0000-0001-6755-9018

To cite this article:
Yılmaz Koğar, E. (2021). Assessing testlet effect on parameter estimates obtained from item response theory models. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 254-266. doi: 10.21031/epod.948227

Received: 5.06.2021
Accepted: 21.07.2021

$$P\left(Y_{ij} = 1 \middle| \theta_j, \gamma_{jd(i)}\right) = \frac{\exp(a_i\,(\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i\,(\theta_j - b_i - \gamma_{jd(i)}))} \qquad (1)$$

where P(Y_{ij} = 1) is the probability that examinee *j* answers item *i* correctly, θ_j is the ability of examinee *j*, a_i denotes the discrimination parameter of item i, b_i is the difficulty of item *i*. The testlet effect $\gamma_{jd(i)}$ for examinee *j* is such that his or her response to item *i* is nested within testlet *d(i)*, and this testlet effect is assumed to be independent of the latent trait θ.

It has been thought that the use of standard IRT models for these tests may be insufficient since the LID assumption has been violated in the tests involving testlets. Therefore TRT models have become a frequently used model in research to testlet effect (DeMars, 2006; Eckes, 2014; Geramipour, 2021; Min & He, 2014; Özdemir, 2017; Paap & Veldkamp, 2012; Wainer & Wang, 2000; Yılmaz Kogar & Kelecioglu, 2017). Glas, Wainer, and Bradlow (2000) examined in their simulation study that when the testlet effect was ignored and the standard IRT model was used, the mean absolute errors of discrimination and difficulty parameter estimation were poorly predicted. Wainer and Wang (2000), in their study based on TOEFL results, determined that the testlet model developed by adding the γ parameter, expressed as the random testlet effect, to the standard 3PL IRT model, gave better results in parameter estimation. Özdemir (2017) conducted a study in which he analyzed the English Proficiency Test data with the TRT model, the dichotomous and polytomous IRT models. In this study, he compared item and ability parameter estimations and determined that the results differed, especially for item parameters. Studies in the literature show that the use of standard IRT models when LID is present can lead to problems such as biased item parameter estimates, overestimation of the accuracy of ability estimates, overestimation of test reliability and test information, and underestimation of standard errors for ability parameter (Sireci, Thissen, & Wainer, 1991; Wainer et al., 2007; Yen & Fitzpatrick, 2006). Based on the results of these studies, it can be said that serious problems may be encountered for the psychometric properties of the tests when LID is ignored. This may lead to incorrect results regarding the interpretation and use of test scores.

Testlets, which are based on a common stimulus and group of items, are used in many large-scale tests because of the previously specified advantages. One of these tests is the PISA (Program for International Student Assessment) applied on the international platform by OECD (Organisation for Economic Co-operation and Development). This application, which evaluates the knowledge and skills of 15-year-old students every three years, focused on reading literacy skills in 2018. Testlets are used in tests that measure language skills, such as reading comprehension. However, in such items, some students have a special interest or better prior background knowledge in a passage than other students, in this situation, they are likely to perform better on the items related to this passage than on other items of the same difficulty level, or they tend to perform better than other students with the same general ability level (Li, 2017, p.1). Therefore, testlets lead to the emergence of additional variance sources, such as content knowledge in an item response function (Chen & Thissen, 1997). However, it is still not commonly enough to perform analyzes through the models that take this effect. The current study is aimed to fill this gap.

PISA applications, which are very important to national and international platforms, are classified as low-stake tests because the important personal decisions associated with the test performance of the participants are not taken. However, the role of these applications in the educational policies of countries is great. IRT approach is used for item and ability estimates in PISA; these models are not special IRT models developed for testlets. In this respect, it is a condition that the results obtained from the standard IRT models and the results obtained from TRT models will change all interpretations. Because it is desirable to be estimated by the least amount of error to achieve a high degree of accuracy. If the LID is a large effect on the estimates of the testlets, this may be compromised.

This study is aimed to calculate the LID magnitude caused by testlets and to compare the effect of this magnitude on parameter estimates and test precision. The following research questions have been established to address these situations:

1. What is the LID level of testlets included in the PISA 2018 reading literacy test?

2. Do the person and item parameters obtained with the standard IRT model and TRT model differ?

By determining the level of testlet variances obtained through the real data sets with these research questions, it is aimed to make an inference about the situations in which the use of TRT models proposed in the literature may be necessary. Also, this study aims to help researchers, especially those used to standard IRT models, to better understand and interpret testlet models because TRT models are less known and less used models than standard IRT models.

METHOD

Participants

PISA application is carried out on 15-year-old students enrolled in formal education. Schools and students participating in the PISA research are determined by the OECD randomly. There are more than 600,000 students from 79 countries and economies participating in the PISA 2018 application (Organisation for Economic Co-operation and Development, 2019). In this study, countries participating in PISA 2018 application as computer-based administrations were preferred. The data of these countries were examined in terms of the same test design and testlets, and analyses were carried out on 3105 students, who were suitable for the study.

Data Sources and Measures

In PISA 2018 application, the main domain is reading literacy. In PISA 2018, a multistage adaptive test (MSAT) design was used to measure reading skills. The MSAT design for the PISA 2018 main survey consisted of three stages (Core stage, Stage 1 and, Stage 2) and 245 items. Different designs were created by applying these stages in different orders. In this design, between 33 and 40 items were applied to each student, depending on which test was taken at each stage. The data used in this study were obtained from design A (Core> Stage 1> Stage 2) applied to 75% of the students. From 64 different ways defined for design A, the selected path is RC1 for the core stage, R15H for stage 1, and R21H for stage 2. For detailed information, it is recommended to consult the report of Yamamoto, Shin, and Khorramdel (2019).

The items in the reading literacy test are in a format that includes constructed response or selected response. However, this study focused on only multiple-choice and dichotomous items because the models used in the study were developed for the items scored dichotomously. The data in PISA applications are open to everyone's use. However, the items are not shared because the items in cognitive instruments are used in other years. For this reason, only data coding was considered for the testlet decision regarding the items. The “label” section of the reading literacy test has been examined in the SPSS format and assumed that the items in the same label are testlets. After this review, 39 items comprising seven testlets were used in the study. The reason for the use of PISA data in the study is that it provides a real set of data in testing and applies to many people.

The data of the study were accessed at https://www.oecd.org/pisa/data/

Data Analysis

Two different measurement models were used in the study: (a) standard 2PL IRT model, (b) 2PL testlet response model. The reason 2PL models are used in the study is that when 3PL is used in TRT models, convergence problems can be experienced for parameter estimation (Eckes, 2014). In this study, the item and ability parameters estimate obtained from the standard IRT model and TRT model was compared with the corresponding standard errors. Root Mean Square Error (RMSE) was examined to compare the capability parameters estimated from different IRT models. RMSE values are calculated by taking the square root of the mean square of the standard errors of the ability parameters. Besides, to better understand the degree of agreement between the estimates, correlations related to the estimates of the two models were calculated, and statistics based on mean differences were used (Mean Difference-MD, Mean Absolute Difference-MAD, Root-Mean-Square Difference-RMSD).

Analyses were performed using the mirt package (Chalmers, Pritikin, Robitzsch, & Zoltak, 2015) included in the R software. mirt is a package developed for multidimensional IRT models. Therefore, it includes slope and intercept parameters as item parameters. For the unidimensional 2PL model, the slope parameter is the same as the discrimination parameter (a_i), while the intercept parameter (di) is calculated over the discrimination and difficulty parameter (b_i) ($d_i = -a_i b_i$). In this study, the intercept parameter transformation is used instead of the difficulty parameter. The item intercept parameter is interpreted as item easiness and is the opposite of the item difficulty parameter. In general, a high value means that the item is easy (Reckase, 2009). The item slope parameter is interpreted as the item discrimination parameter. Higher values indicate that the item is more distinctive (Baker, 2001).

It is also assumed that the population ability distribution in the pack follows a normal distribution. Therefore, there is a normal distribution with mean and standard deviation equal to 0 and 1, respectively, for model identification purposes in IRT calibrations (Paek & Cole, 2020). In this way, parameter estimates obtained from different IRT models are provided to be on the same scale (Li, Li, & Wang, 2010). Also, the calculation of IRT scale scores was performed using the EAP (expected a posteriori) method.

RESULTS

The current study, firstly, analysis results based on the TRT model are presented and focus on testlet effect variance as an indicator of LID for each testlet. Then, the item parameter estimates obtained from the TRT model and the standard IRT model were compared, and the RMSE values showing the precision of these estimates were calculated for each model. Then, various statistics based on correlation values and mean differences are given to examine the fit between models. The same operations were done for the estimations regarding the ability parameter.

The Testlet Effects

The testlet effect variance shows the degree of local dependency among items included in a particular testlet. When the testlet effect variance is zero, there is no local dependence between items. The more this variance exceeds zero, the higher the degree of LID. However, there are different approaches to interpret this value. In simulation studies, it is generally stated that variances below .25 can be considered negligibly small (Glas et al., 2000; Wang & Wilson, 2005). For the testlet effect variance, values of .50 and above are considered to be more important (Wang & Wilson, 2005; Wainer et al., 2007). Table 1 shows the magnitudes of γ and standard errors of testlet effects.

Table 1. Testlet Statistics

Testlet	Number of Items	Testlet Variance	Standard Error
Testlet 1	4	.173	.099
Testlet 2	5	.432	.142
Testlet 3	6	.088	.077
Testlet 4	3	.157	.123
Testlet 5	2	.200	.235
Testlet 6	7	.100	.044
Testlet 7	6	.365	.070

As shown in Table 1, some testlets have much higher LID than others. The variance of the testlet effect for testlet 2 (the code of the testlet is "South Pole") is .489, which is much greater than for other testlets. However, it is seen that all testlet effect variances are less than .50. Looking at the estimations for standard errors, it can be said that these values are not very high, and therefore each testlet effect variance is estimated precisely.

Item Parameter Estimates

The standard IRT model which ignores LID and TRT model item parameters and RMSE values are showed in Table 2.

Table 2. Summary Statistics for Estimated Item Parameters

Model	Slope					Intercept				
	Mean	SD	Min	Max	RMSE	Mean	SD	Min	Max	Mean
IRT	.87	.31	.37	1.57	.08	1.27	1.49	-1.83	5.30	.09
TRT	.87	.33	.35	1.71	.09	1.32	1.58	-1.88	6.01	.13

Note: SD = standard deviation, Min = minimum, Max = maximum, RMSE = Root Mean Square Error.

The summary statistics are shown in Table 2 show to a very high correspondence between the item parameters estimated by the standard IRT and TRT models. Especially item slope parameters were estimated with extreme precision by both models but item intercept parameters, the precision was somewhat lower but still very high. Besides, when the RMSE values are examined, it is seen that the values obtained from the TRT model are higher.

Correlation values and mean differences calculated to determine the amount of agreement of item parameters obtained from different models are given in Table 3.

Table 3. Correlations and Mean Differences for Item Parameter Estimates from Different Models

Parameter	Correlation	MD	MAD	RMSD
Slope	.996	-.009	.032	.034
Intercept	.998	-.095	.103	.303

Note: MD = mean differences, MAD = mean absolute differences, RMSD = root mean square differences.

Table 3 presents the correlations and difference-based statistics for item slope and intercept estimates, respectively. When the correlation values in this table are examined, it is seen that the item parameters obtained from both models are highly correlated. Mean differences between the item parameters obtained from the two models were also calculated to see if one model produced higher or lower parameters than the other model. It can be seen that the average differences for both parameters are very small. However, when looking at the RMSD values, it can be said that the item parameters are affected by the testlet structure. It is seen that testlet structure in the test can produce biased results especially for the intercept parameter.

The relationships of the estimations on item parameters obtained from the IRT model and TRT model are shown in Figure 1.

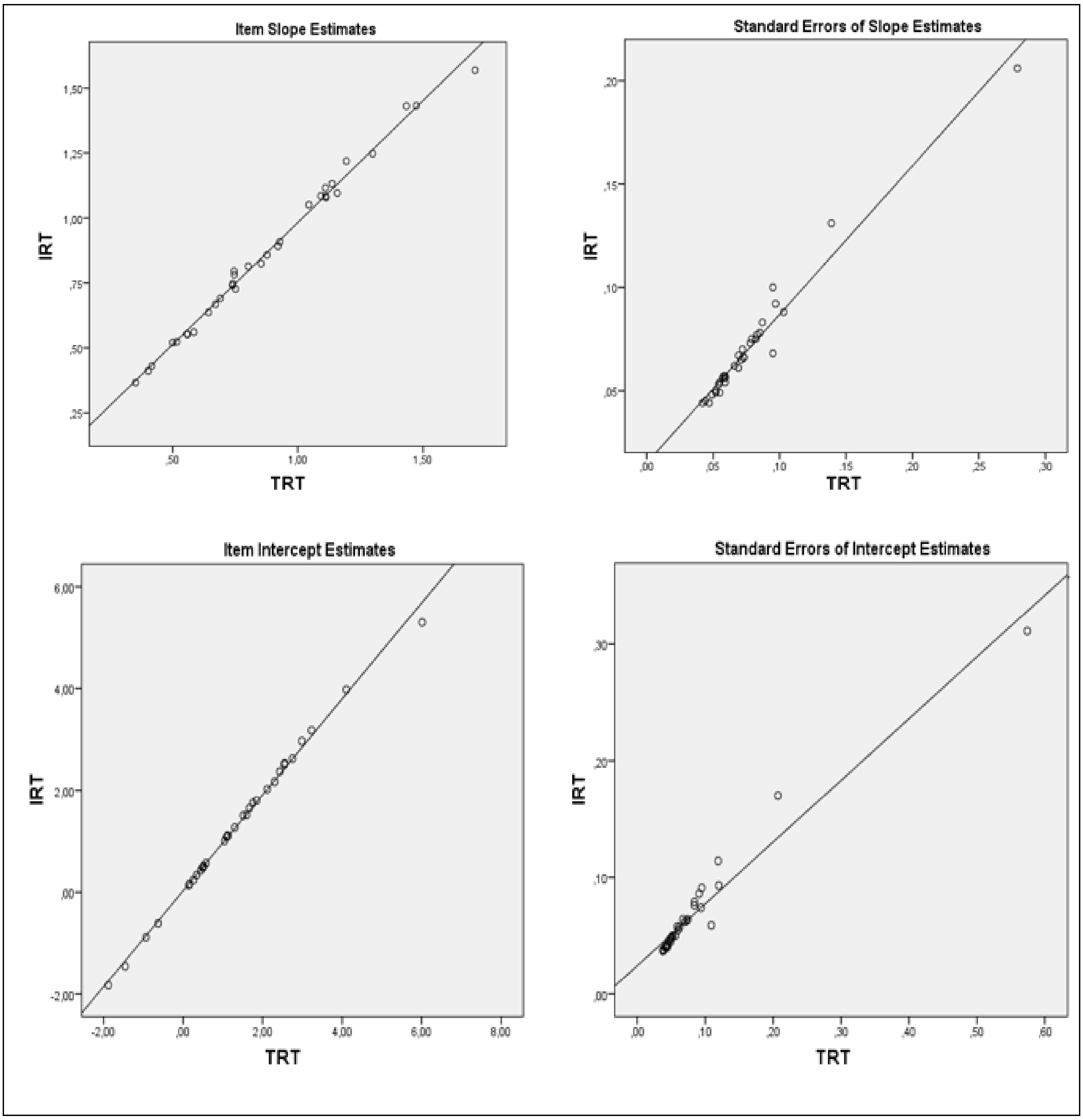

Figure 1. Item Slope and Item Intercept Estimates Under the Standard 2PL Model and Testlet Response Model

When Figure 1 is analyzed, it can be said that item parameter estimates are similar in both models. However, while the standard errors related to the item slope parameters are still similar, there is a slight difference in the standard errors for the item intercept parameter. The standard errors estimated from the standard IRT model for item slope parameters vary between .04 and .21, while the standard errors estimated from the TRT model vary between .04 and .28. The standard errors estimated from the standard IRT model for item intercept parameters vary between .04 and .31, while the standard errors estimated from the TRT model vary between .04 and .57. Therefore, it can be said that the standard IRT model underestimated the measurement error.

Person Ability Estimates

Descriptive statistics for the ability parameters obtained from two IRT models and the RMSE values for the accuracy of this estimate are given in Table 4.

Table 4. Summary Statistics for Person Ability Estimates

Model	Minimum	Maximum	SD	RMSE
IRT	-3.04	2.29	.87	.49
TRT	-2.70	2.15	.81	.58

Note: The mean of the ability distribution was fixed at 0 for estimation purposes for the two models, SD = standard deviation; RMSE = root mean square error.

When looking at the minimum and maximum values and standard deviation values for the ability estimation in Table 4, the estimates from the 2PL IRT model showed a somewhat larger variation than the estimates from the testlet model. When the RMSE values are examined, the higher measurement precision was obtained from the 2PL IRT model compared to the TRT model. In addition to these values, correlation and mean differences were calculated to show the fit between the ability parameters estimated from the two models. It was determined that there is a high correlation between ability parameters obtained from independent items and the TRT model (r = .996). The value found for MAD is .098, and the value found for RMSD is .123. For this reason, it can be said that the ability parameters estimated from both models are similar. Figure 2 shows the scatter plots of ability estimates obtained from both models and the standard errors of these estimates.

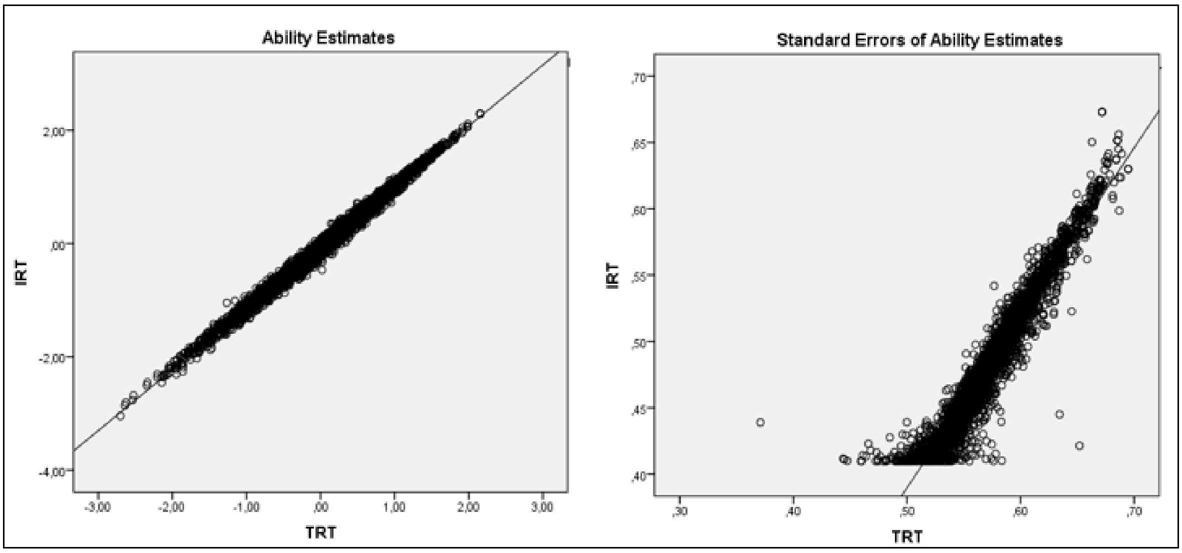

Figure 2. Person Ability Estimates and Associated Standard Errors under the Standard IRT Model and the Testlet Response Model

On the left of Figure 2, the distribution of ability estimates of different models and on the right side, the distribution graphs of the standard errors of the relevant parameter are shown. It can be said that the estimates of the two models are almost the same according to the scatter plot of the ability parameters obtained from the standard IRT model and TRT model. However, when the graph regarding the standard errors is examined, it is seen that the standard IRT model estimates the errors less. While the standard errors estimated from the IRT model ranged from .41 to .67, the standard errors estimated from the TRT model ranged from .37 to .69. Therefore, it can be said that the standard IRT model underestimated the measurement error.

DISCUSSION and CONCLUSION

The aim of this study is to calculate LID magnitude resulting from the testlets in the PISA 2018 reading literacy test and to compare the effect of this size on parameter estimates and test accuracy. For this purpose, item and ability parameter estimations were performed using the IRT model with local independency assumption and the TRT model.

First, the LID status among the items was examined by calculating the testlet effect variance. It was determined that the testlet effects found for the seven testlets were lower than .50. Therefore, it can be said that there is no strong testlet effect in the data set. In studies conducted on real data in the literature, it has been observed that testlet effect variances are lower than .50 (Baghaei & Ravand, 2016, Chang & Wang, 2020; Eckes, 2014).

Then, the item parameters estimated on the standard IRT model and TRT models were compared. The results obtained show that the item parameter estimates are similar. In general, the RMSDs between the item parameters estimated from the two models were low. It was also determined that the slope parameter gives more similar results than the intercept parameter. However, this result differs from the results of the study conducted by Min and He (2014). Comparing the item parameters of different IRT models, the researchers stated that the slope parameter was estimated more suspiciously than the intercept parameter. However, in this study, the bifactor model, another model used in testlets, was chosen as the basic model, and this model was compared with other models. In the present study, the bifactor model was excluded. The difference observed may be due to comparison with different models.

Correlations between item parameter estimates obtained from both models are quite high. DeMars (2006), in his research with PISA 2000 data, used both mathematics and reading literacy data to examine the ability estimations of the independent item model and testlet effect model and stated that the correlations between these estimations were close to 1. A similar result has been observed in other studies (Baghaei & Ravand, 2016; Eckes, 2014; Eckes & Baghaei, 2015; Yılmaz Kogar & Kelecioglu, 2017).

For the last stage of the research, the estimates regarding the ability parameters were examined. Although the ability parameter results obtained from the standard IRT and TRT models are similar, it is seen that the results of the standard IRT model differ more. However, considering the correlation for this parameter and the statistics based on the mean differences of these estimates, it was determined that the IRT and TRT models show high correlation and are quite compatible with each other with small RMSD values. This finding is in line with the findings of the studies conducted by Eckes (2014) and Özdemir (2017). Besides, standard errors related to the ability parameter are estimated higher in the TRT model. In the literature, it is stated that if the item team effect is ignored, the standard error for the ability parameter is underestimated (Chang & Wang, 2010; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000).

Conclusion and Suggestions

Testlets allow more than one item to be asked based on the same stimulus, allowing more than one information to be collected from a stimulus, thus improving the efficiency of the test (information per unit time) (Wainer et al., 2000). Therefore, the use of such items in tests is inevitable. However, it is also necessary to deal with the violation of the local independence assumption of testlet items. To this end, it is important to determine in which cases breaking this assumption will affect the results.

The current study was determined that the results obtained from the standard IRT model and the TRT model are quite close to each other. This result is similar to the studies conducted on the real data set (Baghaei & Ravand, 2016; Demars, 2006; Eckes, 2014; Eckes & Baghaei, 2015; Özdemir, 2017; Yılmaz Kogar & Kelecioglu, 2017). The reason why the result is this way is probably the small variance of the testlet in the data set used in this study because Glas et al. (2000) stated that the testlet effect variances lower than .50 had a negligible effect on the results. They also stated that in this case, standard IRT models, such as 2PL or 3PL could be used without compromising the quality of the parameter estimates. However, even in studies with a high testlet effect, correlations between standard IRT models and TRT models were high (Baghaei & Ravand, 2016; Özdemir, 2017). Beside, it was observed that there were partial variations in RMSE and standard errors obtained from the parameters. According to DeMars (2006), although the complex model results in slightly higher RMSE than the less complex model, this is not a bias. Differences in standard errors were observed, especially in the ability parameter. Such differences can lead to negative consequences when it comes to high-risk decisions (Baghaei & Ravand, 2016). Besides, this can cause serious problems when using computer adaptive tests, which are test termination criteria, the standard error of ability estimates.

As a result, when there is a very strong dependency between the items in the tests, standard IRT models will not give appropriate results for testlet as they neglect this addiction because the studies conducted show that neglecting the assumption of local independence violation causes overestimation of reliability or knowledge and underestimation of standard error of ability estimation (Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). However, researchers who have difficulty using more complex models when the testlet effect is low can use standard IRT models since when the testlet effect is low, it can be said that these models do not make very different predictions from the TRT models. Researchers working with testlets are primarily recommended to examine the testlet variance. Then, if the testlet effect is low, it can be said that standard IRT models can be used for parameter estimates. If there is a high testlet effect, TRT models are required.

Limitations

Despite the contribution of this research to the field, it has several limitations that require further research. Since real data was used in the study, the results of the current situation were examined and the testlet effect variance was estimated to be low. With different studies it can be examined how high these effects can be based on real data. Also, instead of determining only this effect, studies can be conducted to determine the source of the variance created by this effect. For this purpose, the characteristic features of the testlet can be examined using real data, where each item in the test can be accessed. However, since not all the items could be accessed in PISA applications, the characteristics of the testlet items could not be examined in this study. Also, only dichotomous items were used in the study. In future research, the regulations that will consider account the polytomous items can be made.

In the current study, the 2PL TRT model, one model dealing with testlets, was used. TRT models are a limited form of bifactor models. For this reason, the testlet effect can also be handled with bifactor models. In the future, similar studies can be done using the bifactor model and models containing more parameters.

REFERENCES

Baghaei, P. & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*(1), 85-104.

Baker, F. B. (2001). *The basics of item response theory*. College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment.* Paper presented at 4th IEA International Research Conference, Gothenburg, Sweden. Retrieved from http://www.iea-irc.org/fileadmin/IRC_2010_papers/PIRLS/Chang_Wang.pdf

Chalmers, P., Pritikin, J., Robitzsch, A., & Zoltak, M. (2015). Package ‘mirt’. Retrieved January 10, 2021, from https://mran.microsoft.com/snapshot/2014-12-27/web/packages/mirt/mirt.pdf

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61.

Eckes, T. & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education, 28*(2), 85-98.

Geramipour, M. (2021). Rasch testlet model and bifactor analysis: how do they assess the dimensionality of large-scale Iranian EFL reading comprehension tests?. *Language Testing in Asia, 11*(1), 1-23. https://doi.org/10.1186/s40468-021-00118-5

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-287). Boston, MA: Kluwer-Nijhoff.

Ip, E. H. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*(7), 467-482.

Li, F. (2017). An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory. *ETS Research Report Series, (1)*, 1-25. https://doi.org/10.1002/ets2.12151

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21.

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (ETS RR-10–21). Princeton, NJ: Educational Testing Service.

Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477.

Organisation for Economic Co-operation and Development (2019). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing. https://doi.org/10.1787/b25efab8-en

Özdemir, B. (2017). Examining testlet effects in english proficiency test: A Bayesian testlet response theory approach. In I. Koleva & G. Duman (Eds.), *Educational Research and Practice*, (pp. 425-437). Sofia: ST. Kliment Ohridski University Press.

Paap, M. C., & Veldkamp, B. P. (2012). *Minimizing the testlet effect: Identifying critical testlet features by means of tree-based regression*. Psychometrics in Practice at RCEC, 63. Retrieved January 12, 2021, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1001.1923&rep=rep1&type=pdf#page=71

Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. London: Routledge.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). New York, NY: Springer.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247.

Wainer, H., Bradlow, E. T., & Du. Z. (2000). Testlet response theory. An analog for the 3PL useful in testlet-based adaptive testing. In W. J. van der Linden & G. A. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer, Dordrecht. https://doi.org/10.1007/0-306-47531-6_13

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296–318.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 16-27.

Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.

Yılmaz Kogar, E., & Kelecioglu, H. (2017). Examination of different item response theory models on tests composed of testlets. *Journal of Education and Learning, 6*(4), 113-126.

Parametre Tahminleri Üzerindeki Madde Takımı Etkisinin Farklı Madde Tepki Kuramı Modelleri Kullanılarak Karşılaştırılması

Giriş

Madde takımı (testlet), ortak bir uyaranı paylaşan maddeler kümesi olarak tanımlanır (Wainer ve Kiely, 1987). Bu ortak uyaran bir metin, senaryo, tablo ya da şekil olarak sunulabilir. Madde takımları, test uygulaması için gerekli zamanın etkili kullanılmasını sağlaması, testteki maddelerin içeriğinden kaynaklı oluşabilecek içerik etkisini azaltması, tek bir bağımsız maddenin doğası gereği fazla atomistik (çok özel veya dar bir kavramı ölçme) olabileceğine dair endişeleri ortadan kaldırması gibi çeşitli nedenlerle testlerde oldukça kullanılmaktadır (Wainer, Bradlow ve Du, 2000; Wainer, Bradlow ve Wang, 2007). Ancak farklı maddelerin aynı madde takımında toplanması durumunda bu maddeler, ölçülmeye çalışılan gizil özelliğin etkisinin ötesinde birbirleriyle ilişkili olabilir. Yerel madde bağımlılığı olarak bilinen bu durum standart madde tepki kuramı (MTK) modellerinin yerel bağımsızlık varsayımının ihlâl edilmesine yol açar. Örneğin okuduğunu anlama becerisinin ölçüldüğü bir testte yer alan maddelerde öğrencilerin performansı, okuma becerisinin yanı sıra okuma parçası içeriğine olan ilgisinden veya bilgisinden etkilenebilir (Yen, 1993). Bu nedenle de aynı madde takımında yer alan maddeler yerel bağımlı olabilir.

Madde takımlarından kaynaklanan yerel madde bağımlılığına madde takımı etkisi denir (Wainer ve Kiely, 1987). Bradlow, Wainer ve Wang (1999), 2 parametreli lojistik modele (Birnbaum, 1968, 2PLM) bu etkiyi de bir parametre olarak eklemiş ve yeni bir model önermişlerdir. Madde takımı tepki kuramı (MTTK) olarak isimlendirilen bu modelde, aynı madde takımında yer alan maddeler arasındaki bağımlılıkları da hesaba katan bir rastgele etkiler parametresi, γ, bulunur. Standart 2PL MTK modelinde madde güçlük ve madde ayırt edicilik parametreleri bulunmakta ve maddeler arasında yerel bağımlılık olmadığı varsayılmaktadır. MTTK modelinde ise madde güçlük ve madde ayırt edicilik parametrelerinin yanı sıra bir rastgele etki parametresi de dâhil edilerek hesaplamalar yapılır.

Madde takımı etkisini göz önüne alan MTTK modelleri araştırmalarda sıklıkla kullanılan bir model hâline gelmiştir (DeMars, 2006; Eckes, 2014; Min ve He, 2014; Paap ve Veldkamp, 2012; Wainer ve Wang, 2000). Glas, Wainer ve Bradlow (2000) yaptıkları simülasyon çalışmasında, madde takımı etkisinin görmezden gelindiği ve standart MTK modelinin kullanıldığı durumda, ayırt edicilik ve güçlük parametre kestiriminin ortalama mutlak hatasının kötü tahmin edildiğini belirlemişlerdir. Wainer ve Wang (2000) TOEFL sonuçları üzerinden yürüttükleri çalışmada Standart 3PL MTK modeline tesadüfi madde takımı etkisi olarak ifade edilen γ parametresinin eklenmesiyle geliştirilen madde takımı modelinin parametre kestirimlerinde daha iyi sonuç verdiğini belirlemişlerdir. Alanyazında yer alan araştırmalar, yerel madde bağımlılığı mevcutken standart MTK modellerinin kullanılmasının yanlı madde parametre kestirimlerine, yetenek kestirimlerinin kesinliğinin fazla tahmin edilmesine, test güvenirliğinin ve test bilgilerinin fazla tahmin edilmesi ve yetenek parametresine ilişkin standart hataların olduğundan az tahmin edilmesi gibi sorunlara yol açabildiğini göstermektedir (Sireci, Thissen ve Wainer, 1991; Wainer vd., 2007; Yen ve Fitzpatrick, 2006). Bu araştırmaların sonuçlarına dayanarak yerel madde bağımlılığı göz ardı edildiğinde testlerin psikometrik özellikleri için ciddi sorunlarla karşılaşılabileceği söylenebilir. Bu durum ise test puanlarının yorumlanması ve kullanılmasıyla ilgili yanlış sonuçlar doğurabilir.

Birçok geniş ölçekli testte, ortak bir uyarana dayanan ve madde takımı olarak adlandırılan madde grupları kullanılmaktadır. Özellikle okuduğunu anlama becerileri için geliştirilen testlerde madde takımlarına oldukça yer verilir. Ancak bu madde takımlarının neden olduğu madde takımı etkisi, bir madde cevap fonksiyonunda ek bir varyans kaynağı oluşturur. Buna karşın bu etkiyi göz önüne alan modeller üzerinden analizler gerçekleştirmek hâlâ yeterince yaygın değildir. Bu çalışma ile bu boşluğun doldurulmasına katkı sağlamak hedeflenmektedir. Bu çalışmada; madde takımlarından kaynaklı oluşan yerel madde bağımlılığı büyüklüğünü hesaplamak, bu büyüklüğün parametre tahminleri ve test kesinliği üzerindeki etkisini karşılaştırmak amaçlanmaktadır. Bu durumları ele almak için aşağıdaki araştırma soruları oluşturulmuştur:

1. PISA 2018 okuma becerileri testinde yer alan madde takımlarının yerel madde bağımlılığı derecesi nedir?

2. Standart 2-PL MTK modeliyle elde edilen kişi ve madde parametreleri ile 2-PL MTTK modeliyle elde edilen kişi ve madde parametreleri farklılaşmakta mıdır?

Yöntem

PISA uygulaması, örgün öğretimde kayıtlı olan 15 yaş grubu öğrencilerin katıldığı bir uygulamadır. PISA araştırmasına katılacak okul ve öğrenciler, OECD tarafından seçkisiz yöntemle belirlenmektedir. PISA 2018 uygulamasına toplam 79 ülke ve ekonomiden katılan 600.000'den fazla öğrenci bulunmaktadır (Organisation for Economic Co-operation and Development, 2019). Bu çalışmada PISA 2018 uygulamasına bilgisayar tabanlı değerlendirme şeklinde katılan ülkeler tercih edilmiştir. Bu ülkelerin verileri test düzeninin ve madde takımlarının aynı olması bakımından incelenmiş ve araştırmanın amacına uygun olan 3105 öğrenci üzerinden analizler gerçekleştirilmiştir.

PISA 2018 uygulamasında ağırlıklı alan okuma becerileridir (reading literacy). PISA 2018'de okuma becerilerini ölçebilmek için çok aşamalı uyarlanmış test (multistage adaptive test-MSAT) deseni kullanılmıştır. Bu deseni içeren uygulamada okuma becerileri alanı için toplam 245 madde bulunmaktadır. Maddeler; temel, 1. aşama ve 2. aşama olacak şekilde üç aşamada yer alacak şekilde yapılandırılmıştır. Bu aşamaların farklı sıralarda uygulanmasıyla farklı düzenler oluşturulmuştur. Bu desende her öğrenciye her aşamada hangi testin alındığına bağlı olarak 33 ile 40 arasında madde uygulanmıştır. Bu çalışmada kullanılan veriler, öğrencilerin %75'ine uygulanan A düzeninden (Core>Stage 1>Stage 2) elde edilmiştir. A düzeni için tanımlanan 64 farklı yoldan ise seçilen yol temel aşama için RC1, 1. aşama için R15H ve 2. aşama için R21H şeklindedir. Ayrıntılı bilgi için Yamamoto, Shin ve Khorramdel'in (2019) raporuna bakılması önerilir.

Okuma becerileri testinde yer alan maddeler seçme gerektiren ya da öğrencinin cevabı kendisinin yapılandırmasını gerektiren formattadır. Ancak bu çalışmada yalnızca çoktan seçmeli ve ikili puanlanan maddeler üzerine odaklanılmıştır. Çalışmada farklı sayıda madde içeren 7 madde takımının oluşturduğu toplam 39 madde kullanılmıştır.

Çalışmada iki farklı ölçme modeli kullanılmıştır: (a) 2PL Madde takımı tepki modeli (Wainer et al.,2007), (b) standart 2PL MTK modeli (Birnbaum, 1968). Çalışmada 2PL modellerinin kullanılmasının nedeni, MTTK modellerinde 3PL kullanıldığında parametre kestirimleri için yakınsama problemi yaşanabilmesidir (Eckes, 2014). Bu çalışmada standart 2PL MTK ve 2PL MTTK modellerinden elde edilen madde ve yetenek parametreleri kestirimleri ile bunlara karşılık gelen standart hatalar karşılaştırılmıştır. Farklı MTK modellerinden kestirilen yetenek parametrelerini karşılaştırmak için hataların ortalama karekökü (RMSE) incelenmiştir. RMSE değerleri yetenek parametrelerinin standart hatalarının karesinin ortalamasının karekökü alınarak hesaplanmıştır. Ayrıca kestirimler arasındaki uyuşma derecesini daha iyi anlamak için iki modelin kestirimlerine ilişkin korelasyonlar hesaplanmış ve ortalama farklılıklarına dayalı istatistikler kullanılmıştır (MD, MAD, RMSD). RMSD, iki modelden kestirilen parametrelerine ilişkin hatalar farkının karesinin ortalaması alınarak elde edilmiştir. Analizler R programında mirt paketi (Chalmers vd., 2015) üzerinden gerçekleştirilmiştir.

Sonuç ve Tartışma

Bu çalışmanın amacı; PISA 2018 okuma becerileri testindeki madde takımlarından kaynaklı oluşan yerel madde bağımlılığı büyüklüğünü hesaplamak, bu büyüklüğün parametre tahminleri ve test kesinliği üzerindeki etkisini karşılaştırmaktır. Bu amaçla madde ve yetenek parametresi kestirimleri yerel bağımsızlık varsayımı bulunan MTK modeli ile MTTK modeli kullanılarak gerçekleştirilmiştir.

İlk olarak madde takımı etki varyansı hesaplanarak maddeler arasındaki yerel madde bağımlılığı durumu incelenmiştir. Yedi madde takımı için bulunan madde takımı etkisi düşük düzeydedir. Bu nedenle veri setinde güçlü bir madde takımı etkisinin olmadığı söylenebilir. Literatürde gerçek veriler üzerinden yapılan çalışmalarda da madde takımı varyanslarının .50'den düşük olduğu gözlenmiştir (Baghaei ve Ravand, 2016, Chang ve Wang, 2020; Eckes, 2014).

Daha sonra MTK ve MTTK modeli üzerinden kestirilen madde parametreleri karşılaştırılmıştır. Elde edilen sonuçlar madde parametre kestirimlerinin benzer olduğunu göstermektedir. Genel olarak, iki modelden tahmin edilen madde parametreleri arasındaki RMSD'lerin küçük olduğu belirlenmiştir.

Ayrıca a parametresinin, d parametresine göre daha benzer sonuçlar verdiği belirlenmiştir. Her iki modelde elde edilen madde parametre kestirimleri arasındaki korelasyonlar ise oldukça yüksektir. DeMars (2006) PISA 2000 verisiyle yaptığı araştırmada hem matematik hem okuma verileri için MTTK modeli ile standart MTK'nin yetenek kestirimlerinin korelasyonlarının 1'e yakın olduğunu belirtmiştir.

Yetenek parametrelerine ilişkin kestirimler incelendiğinde her iki modelden elde edilen sonuçlar benzer olsa da standart MTK modeli sonuçlarının daha çok farklılaştığı görülmektedir. Ancak bu parametre için korelasyon ve bu kestirimlerin ortalama farklılıklarına dayalı istatistikler göz önüne alındığında, MTK ve TRMTTKT modellerinin yüksek korelasyon gösterdiği ve küçük RMSD değeriyle birbirine oldukça uyumlu olduğu belirlenmiştir. Bu bulgu Eckes (2014) ve Özdemir (2017) tarafından yapılan çalışmaların bulgularıyla paralellik göstermektedir. Ayrıca yetenek parametresine ilişkin standart hatalar MTTK modelinde daha yüksek kestirilmiştir. Literatürde de madde takımı etkisinin göz ardı edildiğinde yetenek parametresinin standart hatasının olduğundan düşük kestirildiği belirtilmektedir (Chang & Wang, 2010; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000).

Sonuç olarak bu çalışmada standart MTK modeli ile MTTK modelinden elde edilen sonuçların birbirine oldukça yakın olduğu belirlenmiştir. Bu sonuç gerçek veri seti üzerinden yapılan çalışmalarda da bu şekildedir (Baghaei ve Ravand, 2016; Demars, 2006; Eckes, 2014; Eckes ve Baghaei, 2015; Özdemir, 2017; Yılmaz Kogar ve Kelecioglu, 2017). Bu çalışmada bu sonucun nedeni büyük olasılıkla çalışmada kullanılan veri setinde bulunan madde takımlarının madde takımı varyanslarının düşük olmasıdır. Çünkü Glas vd. (2000) 0.50'ten düşük madde takımı etki parametrelerinin sonuçlar üzerinde göz ardı edilebilir bir etki yaptığını belirtmişlerdir. Ayrıca bu durumda 2PL veya 3PL gibi modellerin parametre tahmininin kalitesinden ödün vermeden kullanılabileceğini ifade etmişlerdir. Ancak madde takımı etkisinin yüksek olduğu belirlenen çalışmalar da bile standart MTK modelleri ve MTTK modelleri arasındaki korelasyonlar yüksek bulunmuştur (Baghaei ve Ravand, 2016; Özdemir, 2017). Ancak parametrelerden elde edilen RMSE ve standart hatalarda kısmen farklılaşmalar olduğu görülmüştür. DeMars (2006) belirttiği gibi karmaşık model daha az karmaşık modele göre biraz daha yüksek RMSE'ye yol açmıştır. Standart hatalardaki farklılıklar ise özellikle yetenek parametresinde gözlenmiştir. Bu tür farklılıklar yüksek riskli kararlar söz konusu olduğunda olumsuz sonuçlara yol açabilir (Baghaei ve Ravand, 2016). Ayrıca bu durum, test sonlandırma kriteri kişi tahminlerinin standart hatası olan bilgisayar uyarlamalı testler kullanıldığında da ciddi sorunlara yol açabilir.

Bu araştırmanın alana katkısı olmasına rağmen, daha fazla araştırma gerektiren bazı sınırlılıkları vardır. Çalışmada gerçek veriler kullanıldığı için mevcut durumun sonuçları incelenmiş ve madde takımı etki varyansının düşük olduğu kestirilmiştir. Farklı çalışmalarla bu etkilerin ne kadar yüksek olabileceği gerçek verilere dayanılarak incelenebilir. Ayrıca sadece bu etkiyi belirlemek yerine, bu etkinin yarattığı varyansın kaynağını belirlemeye yönelik çalışmalar yapılabilir. Bu amaçla, testteki her bir maddeye ulaşılabilen gerçek veriler kullanılarak madde takımının karakteristik özellikleri incelenebilir.

Mevcut çalışmada madde takımlarını ele alan modellerden biri olan 2PL MTTK modeli kullanılmıştır. MTTK modelleri, bifaktör modelinin sınırlı bir şeklidir. Bu nedenle madde takımı etkisi bifaktör modeliyle de ele alınabilir. İleride yapılacak çalışmalarda bifaktör modeli ve daha fazla parametre içeren modeller kullanılarak benzer çalışmalar yapılabilir.