The Comparison of Item Parameters Estimated From Parametric and Nonparametric Item Response Theory Models in Case of The Violance of Local Independence Assumption

Ezgi Mor Dirlik [i]
Kastamonu University

Abstract

Item response theory(IRT) has so many advantages than its precedent Classical Test Theory(CTT) such as non-changing item parameters, ability parameter estimations free from the items. However, in order to get these advantages, some assumptions should be met and they are; unidimensionality, normality and local independence. However, it is not always so easy to be met these assumptions by datasets. Especially when the normality of data is not provided, another approach for IRT can be applied, which is Non-Parametric Item Response Theory (NIRT). NIRT provides more flexible methods to scale datasets and it is used when the assumptions of Parametric Item Response Theory (PIRT) are not met at a satisfactory level. The assumption of local independence, is one of the situations in which NIRT can be used more effectively than PIRT. In this study, by using a real dataset, taken from TIMSS 2011, the effect of local dependence on the item parameters was investigated. With this goal, a dataset composed of 1,000 students was formed randomly from the TIMSS 2011 eight grade Mathematic test. Firstly, the item parameters were calculated from data set according to the two approaches without any manipulation. After that, two items were arranged as local dependent by changing the response patterns completely the same and the item parameters have been estimated from each sample by using R program, *ltm* and *mokken* packages. Two sets of item parameters estimated from data set were compared and the differences of the parameters were analyzed with statistical test. By this way, the effect of local independence has been analyzed on the item parameters have been decided.

Keywords: Local Independence, Item Parameters, Parametric Item Response Theory, Non-Parametric Item Response Theory.

DOI: 10.29329/ijpe.2019.203.17

---

[i] **Esra Uçak,** Assist. Prof. Dr., Kastamonu University, Measurement and Evaluatıon.

Correspondence: ezgimor@gmail.com

INTRODUCTION

Theories and models in the educational testing field try to explain and anticipate relationships among test items and examinee abilities. The theories provide a framework for deeper understanding among these variables. One of the important mission of these theories and models is to find out the errors in measurement and control them. Classical Test Theory(CTT) and Item Response Theory(IRT) are the most known theories in educational assessment. The former's much of concern is on the estimation and controlling of errors in the process of testing. It assumes a linear relation among test score (X-observed score), true score (T, which is a hypothetical value) and error(E) and the relationship between these components is shown with a basic equation as $X=T+E$. In this equation, the error is assumed to be randomly distributed and its mean is accepted as 0 across the population of examinees. The other feature of errors is that they are to be uncorrelated with error scores on parallel test administration. Thanks to this assumption, reliability in CTT is defined as the correlation between the parallel tests. In summary, for CTT, all features have relied on this linear relationship and by means of this reasonable linear model, this theory has long been used to guide test development, evaluation and interpretation of the test scores (Crocker and Algina, 1986; Gulliksen, 1950; Lord and Novick, 1968).

Despite the simplicity and wide usage in educational psychological testing, CTT has several important shortcomings. One of them is that the item parameters (difficulty and discrimination) depend on the ability of the sample of examinees. Also the mean and standard deviation of the sample affect the parameters' values, for this reason, the item parameters are not stable from sample to sample. The sample dependent parameters are only valid when the samples of examinees are similar with regards to ability distribution. Consequently, for different samples, item parameters are required to be calculated again, and this issue creates the test equation problems. The other important drawback of the theory is another dependence issue that examinees' scores are largely a function of the difficulty level of items administered to them. Because of that, test scores of examinees cannot be compared when several forms of tests varying in difficulty are used. In other words, it is not possible to compare test scores of examinees unless the forms are not parallel to each other. Developing parallel forms is another challenging issue, too and as before mentioned, the reliability of the test scores is based on the parallel test assumption. Another shortcoming of CTT is that the most preferred classical test model requires the assumption of equal errors of measurement for all participants. According to the Lord (1980), violation of this assumption is the rule most of the time. For instance, guessing creates violation of this assumption and low- ability examinees use their chance in response process more while the average and high-ability examinees prefer guessing less. So the error of measurement changes according to the group of examinees and this is a violation of the assumption. However, these violations may not be a threat for all over the theory, another model can be used, but this feature does not address the other drawbacks of the theory.

In an effort to bring solutions to the shortcomings of CTT, another test theory known as Item Response Theory (IRT) has been developed (Hambleton and Swaminathan, 1985; Lord, 1980). In this theory, models have been developed for use with discrete or continuous item responses which are dichotomously or polytomously scored. Also if the assumptions of models which are local independence, normality of the test scores, unidimensionality, are met, most of the CTT drawbacks' can be solved in this theory. Group independent item parameters, examinee ability estimates independent from item difficulty levels can be obtained thanks to the IRT models. Thanks to IRT models, it is so easy to develop computer adaptive tests, test equation and detecting the differentiating items. However, in order to get these advantages of IRT models, model requirements should be met by the data sets. Common requirements of general unidimensional parametric IRT models are; unidimensionality, local independence, and model-data fit. In unidimensionality, it is assumed that the test principally measures a single or unitary trait. This assumption can be analyzed several techniques and the most known is the factor analysis methods. The other important assumption is local independence. According to this assumption, one's answering probability of an item should be totally independent of his/her response to another item in the same test. Local independence is in the focus of this study, so it is required to be analyzed in detail here.

Local Independence Assumption

Local independence has been considered to be the principal axiom of test theory (Lazarsfeld, 1958; Lord and Novick, 1968; Jannarone, 1988). This concept was first discussed in its general form by Lazarsfeld(1958), and later Anderson(1956) and McDonald(1962) studied on it (cited in Henning, 1989) . It is widely represented as the basic assumption underlying the latent trait models and IRT models require this assumption, too. The standard unidimensional IRT models such as one, two and three-parameter models, require the local independence assumption (Embretson & Reise, 2000; Hambleton and Swaminathan, 1985; Lord & Novick, 1968). According to this assumption, an examinee's answering probability of an item should be totally independent of his/her response to another item in the same test. In these models, the probabilities that an examinee will provide a specific response to an item are a function of two components and they are just like that:

1. The examinee's ability; and,

2. Item parameters affecting examinee's answers to the items such as difficulty and discrimination parameters, pseudo-chance parameter.

Thanks to these two components, unidimensional IRT models define one's success probability only using his/her ability and item characteristics'. It means that the response to any item is unrelated to any other item for the specific latent trait level. Local independence means that one's answering an item only depends on his or her ability even the items are highly correlated with each other. If a dataset meets this assumption, when the trait level is controlled, local independence implies that there is no relationship remains between the items (Embretson & Reise, 2000). When this assumption is violated, items become local dependent and this situation will affect the estimations of item and ability parameters. This violation may create substantial consequences such as misleading item discrimination parameters. Many studies have shown that statistical analysis conducted on local dependent data sets is misleading and create different results from the real estimations (Chen and Thissen 1997; Chen and Wang, 2007; Junker, 1991, Sireci, Thissen and Wainer, 1991). The results of Tuerlinckx and De Boeck,(1998) research showed that if negative local dependence is not included into model estimations, the item discrimination parameters of the locally dependent items are underestimated. Moreover, they showed that the value of item discrimination parameter depends on the difficulty parameter of the item it interacts with. Because of the underestimated discrimination parameters of the items, test information function is calculated less than its real values, so the standard error of measurement is underestimated, too. In addition to the deviations from the item parameters, Junker (1991) showed that in case of local dependence, ability parameters are estimated biased strongly.

As for the reasons for occurring the local dependence between the items, there are several potential causes (Yen, 1993). According to the Yen (1993), the violations can be grouped as two major causes; the first group of causes is not related with item content, they can bring on from external reasons, such as, assistance from a teacher, fatigue which means that items at the end tend to be more challenging. Practice effect and speediness can be the other external causes of local dependence and this type of causes has been named as "surface local dependence". The other type of causes of local dependence has been named as" underlying local dependence" by Chen and Thissen (1997). In this type of causes, violation from the local independence can generally occur due to the item content. Item chaining can be one of these reasons and in this situation, items are organized as steps, so their answers will affect each other. In this group of causes, the reasons are assumed as minor dimensions besides the unique essential latent dimension which is tested.

As noted earlier, local independence is an essential requirement for an accurate item and ability parameters, so this assumption should be checked via analysis before starting estimation of abilities or item parameters in IRT models. There are several statistics that can be used in order to detect the local dependence in datasets. Chen and Thissen(1997) proposed four statistics to be used so as to detect item pair dependence. These statistics are the *Pearson's* X^2, *the* G^2 *statistic*, the *Standardized Phi Coefficient Difference*, and *the Standardized Log-Odds Ratio Difference*. With these

statistics, the covariance is examined by using two-way contingency tables and these tables are composed on expected and observed values. For this reason, when the observed value is 0, the standardized phi coefficient difference and the standardized log-odds ratio difference may not be calculated. Yen (1984) proposed *Q_3 coefficient* in order to detect local item dependence and this coefficient is based on pairwise index of correlation of the residuals from the IRT model. Chen and Thissen (1997) compared Yen's *Q_3 coefficient,* the *Pearson's X^2* and the *G^2 statistic* and they found that *X^2* and *G^2 statistic* indices appear less powerful than Yen's *Q_3 coefficient* for detecting underlying local dependence. As for surface local dependence, it was determined that all of them appear equally powerful. In addition to these coefficients, *DETECT,* which is a statistical tool developed to estimate features of multidimensional latent space, can also be used in an attempt to analyze underlying local dependence. It works like Yen's *Q_3 coefficient* and can be used to explore the homogenous items subsets as a separate dimension (Balazs and De Boeck, 2006; Stout, 2000).

After detecting local dependence, researchers should handle this matter before going on estimation. There are several ways to cope with local dependence issue. The first one is creating "testlets" with the dependent items (Wainer and Kiely, 1987). In this method, instead of item scoring, testlet scoring is used and the scores in a testlet are summed and each score represents a category of polytomous item, and this method is used in the Graded Response Model of Samejima(1969), the Partial Credit Model and the Rating Scale Model(Andrich, 1985; Wright and Masters, 1982; cited in Monseur, Baye, Lafontaine and Quittre, 2011). The second way of modeling local dependence is including this effect to the IRT model estimations. According to this approach, the response patterns of testlet are modeled by including additional fixed item interaction parameters in addition to item parameters. By this way, total item information is preserved and local dependence is viewed as an item characteristic. In the third approach, it can be named as random-effects models, local dependence is modeled as a variable of examinee and depends on the abilities of examinees. There are lots of random-effects models and Bayesian random effects model, the random weights linear logistic model, the random-effects two-facet model are some of them which are used commonly with this goal (Monseur, Baye, Lafontaine and Quittre, 2011).

In addition to the models developed for handling local dependence, using Non-Parametric Item Response Models is another new way of solving this issue. Non-parametric Item Response Theory (NIRT) is a non-parametric approach of parametric IRT and has some advantages as parametric models. One of these advantages is that the non-parametric models can be applied more datasets than the parametric ones thanks to the limited assumptions. In NIRT models, it is available to analyze the datasets composed of fewer people, like 50 or 100 and fewer items such as 10-15 items than the parametric models. Also the other advantage is that this approach allows making ability and item estimations by lightening the assumptions of IRT. It can be applied when the normality is not met by the dataset and it is possible to get the more accurate item and ability estimations when the local independence assumption is not met totally (Junker, 1991, Sijtsma and Molenaar , 2002; Sijtsma and Meijer, 2007). So when the local independence is violated for few numbers of item, non-parametric item response theory models can be preferred instead of modeling the dependence with different techniques. NIRT is a newly developed approach, so it is under research for many issues and local independence is one of the matters that have to be investigated according to this approach.

Local independence is one of the important underlying assumptions of not only all item response models but also all latent trait models such as factor analysis, latent trait analysis, latent class analysis and latent profile analysis (Vermunt and Magidson, 2004). It requires that the response to an item on a test not be influenced by the response to any other items. This assumption is often taken for granted and it is paid little or no attention to determine if the process of responding to one item affect the response to other item/s. Also this assumption can be easily violated when several items are embedded in the same passage, or when items composes of multiple parts. According to Ackerman (1987), this assumption is violated whenever the response process of one item provides the necessary cognitive schema to trigger a response to a subsequent item. Many techniques and modelling have proposed in order to investigate this assumption and NIRT models may be the new alternatives to deal with the violation of this assumption. As stated before, NIRT is a relatively new domain in

psychometrics literature and new studies should be conducted in order to reveal the real performance of NIRT approach especially in the issues in which PIRT models do not allow to scale the tests and violation of local independence assumption is one of these issues. For this reason, in the concept of this study, it is aimed to investigate the item parameter estimations of NIRT and PIRT in case that the local independence assumption is violated for the data set. The implicit objective of the study is to demonstrate the usability of NIRT models as alternatives to PIRT models especially for item parameter estimations in cognitive assessment, which is generally used in typical performance and health sciences.

Research Questions

The research questions guiding the study are as follows:

1. How are the item parameters estimated from PIRT and NIRT change in case of local dependence violation?

2. Is there any significant relationship between the item parameters estimated from PIRT and NIRT both local dependent and independent data sets?

METHOD

This study was designed as a fundamental research because the main purpose of the study is to analyze the differences and similarities of parametric and non-parametric IRT model item parameter estimations in case of local independence violation. In accordance with this goal, the NIRT models' features have been explored when the local independence assumption is not met by the data set. Comparing the estimations from two different approaches composes the basic goal of the study, hence it is thought that the results of the study will be beneficial in determining the features and usages of models. For these reasons, this research has been classified as a fundamental study.

Composing Data Set

When the studies on the NIRT approach are analyzed, it is seen that most of them have been aimed to figure out the features of this new approach and have used the simulative data sets in order to analyze the features of the approach. The ones that used real data sets are about much more health science, psychology. Hence there is not enough study analyzing the suitability of NIRT approach on educational settings. For this reason, in the concept of the study, it was aimed to use a real data set in order to create an example of NIRT usage and explore the NIRT features in educational settings.

In order to achieve the goal of the study, initial data set was composed by using from the whole data set of Trends in International Mathematics and Science Study (TIMSS) 2011. While composing the data set, general IRT assumptions, such as unidimensionality, local dependence, and normality were taken into consideration. All of Mathematics and Science booklets were examined at 8^{th}-grade level and the one which meets the unidimensionality assumption at most was selected to be analyzed. The reason of choosing only the 8^{th}-grade booklets for dimensionality analysis is that at this grade level, there are more items than the 4^{th}-grade level booklets, and as stated before, parametric IRT models require as many items as possible in order to get an accurate estimation of parameters. According to the unidimensionality analysis results, it has been found that the booklets of Mathematics met unidimensionality assumption higher level than the Science booklets. After checking the level of unidimensionality in all of the 8^{th}-grade Mathematics booklets, it was found that this assumption was met at most in the 11^{th} booklet of Mathematics composed of 30 dichotomous items and the analyses were gone through on the items of this booklet.

In addition to test length and unidimensionality assumption, the sample size is another regarded factor while composing the study's data set. Like longer tests, parametric IRT approach

needs a larger sample size than the non-parametric one to get an accurate estimation (Embretson and Reise, 2000; Junker, 2001). In order to decide the required sample size, previous studies results were investigated and it was found that the sample composed of 1000 people is suitable for both parametric and non-parametric IRT models. In selecting 1000 people from the whole TIMSS 2011 data set, the success rating of the countries was taken into account and the top 20 countries according to the Mathematics success at 8^{th} grade level in TIMSS 2011 were included into data set. The reason of incorporating the highest Mathematics score countries to the data set is that previous research have shown that parametric IRT models may provide lower standard error of measurement when the examinees' abilities are at high levels.

After composing a basic data set by using TIMSS 2011 database, the whole data set was transformed into two data sets, by changing the response pattern of an item in order to create local dependent test. One of the items were selected randomly from the test and its response pattern was made same with the item following it. Hence a new data set was created violating the local independence assumption. The estimates from the original and local dependent data sets were compared.

Data Analysis

After composing data set, the item parameters were estimated according to the parametric and non-parametric approach. For the parameter estimations from both approaches, R Studio program was used. In R Studio, non-parametric analysis was done with "mokken" package and parametric analysis were done with "ltm" and "irtoys" packages. The data analyses were made in two phases; 1. checking the assumptions and finding the suitable model for data, and 2. item parameters estimation according to the both approaches.

At the first phase of the study, the general IRT assumptions were tested. While checking the unidimensionality assumption, explanatory and confirmatory factor analysis was used. In order to apply explanatory factor analysis, tetrachoric correlation matrix was composed by using STATA program. After obtaining the matrix, the explanatory factor analysis was conducted via SPSS 22.0. While determining the dimensional structure of the test, the scree-plot and explained variance ratios were taken into account. It was found that the dominant factor of the test explains 58% of total variance and other two factors of which eigenvalues are higher than 1 explain little amount of total variance, 9% and 5% respectively. Also the item loadings changes between 0.52 and 0.86 as a unidimensional structure, hence the test was accepted as unidimensional. After explanatory factor analysis, the decided factor structure was tested by using confirmatory factor analysis via LISREL program. The model was composed as unidimensional and it was found that all of the items have significant t-values for the proposed model. The standardized factor loadings of the items change between 0.56 and 0.84. As for the model- level fit statistics, several indexes were analyzed. The first one is X^2/df which was found as 3,42, and this value indicates a moderate level fit (Kline, 2005). Then the RMSEA value was analyzed and it was found as 0.06 which shows a good fit (Jöreskog and Sörbom, 1993). The other fit indexes analyzed are GFI, NNFI and AGFI. All of them are found higher than 0.90 and these values mean that model-data fit is at good level (Brown, 2006). Considering the results of explanatory and confirmatory factor analyses, it was decided that the test has a unidimensional structure and the examination of the other assumptions of IRT was continued.

The data analysis process was continued with the other IRT assumption; local independence was investigated in detail with three methods. Yen's Q_3 *coefficient* and G^2 *coefficient* were calculated and no violation of local independence was detected for the data set. Also the DETECT was applied in order to define any other dimension which may affect students' performance, but it was found that the test is composed of a homogenous structure. Hence it was decided that the test meets the local independence assumption.

Subsequently checking the assumptions, item parameters were estimated according to the parametric and non-parametric models. These parameters were accepted as the precise ones. Then the

data set was manipulated by changing the response pattern of two items. One item was selected randomly from the dataset then the following item's response pattern was totally changed with the first one. In this way, these items were made artificially local dependent and the assumption of local independence is violated with these items. It means that if the examinee answers the first item true, the second one is also true and if the first answer is false, the second one is false, too. The modified data set was accepted as a new and local dependent data set and again the assumptions of IRT were checked. At this time, as planned, it was found that the local independence assumption is violated and the second item parameters were estimated through this data set.

In the last phase of the data analysis process, item parameters estimated according to two approaches from the both data sets were analyzed by using statistical tests and correlation analyses. The mean differences between the parameters were analyzed by Related Sample T-test and Wilcoxon Rank Order Test. Also the relationship between the parameters were investigated by calculating both parametric and non-parametric correlation coefficients.

FINDINGS

The first analysis in the research is the investigation of the model data fit of the composed data sets to the parametric and non-parametric IRT models. For parametric IRT approach, one, two and three-parameter logistic models, were tested. In model-data fit analysis, item level chi-square and -2 loglikelihood values were taken into consideration. The model-data fit statistics are presented in Table 1 below.

Table 1. Model-Data Fit Statistics for PIRT Models

Model Comparison	**Original Data Set**			**Modified Data Set**		
	-2LL	df	p	-2LL	Df	p
1 PLM-2PLM	100.33	29	<0,001	167,65	29	<0,001
2 PLM-3PLM	57,64	30	>0,001	56,49	30	>0,001
1 PLM-3 PLM	62,49	59	<0,001	142,14	59	<0,001

When the results of model data fit statistics given in Table 1 were analyzed, it is clear that the highest improvement of the model-data fit occurred between one and two parameter logistic model for both data sets. Also while the difference between the likelihood values between one and two parameter models were significant, the values computed for three parameter logistic model were not significant. Hence for both data sets, it was decided that the model-data fit was provided at most for two parameter logistic model, so the parameter estimations were conducted according to this model.

Like PIRT, model-data fit analyses were conducted for NIRT. The process of determining the suitable model in the non-parametric approach is different from the parametric one. There are several steps to be followed and in first, the item popularities, which is the percent of correct answer, and the item difficulty value in Classical Test Theory, are calculated for each item. Then item scalability coefficients, which is shown as *H* coefficients, are calculated and they give information about item discrimination levels. After this step, the assumptions of the models are checked with specific methods. While the unidimensionality assumption is tested with Automated Item Selection Procedure(AISP), monotonicity of the item characteristic curves can be analyzed with rest-group methods, and both of these methods are specific to NIRT models. When these two assumptions are met for the data set, it is possible to scale the data set with Monotone Homogeneity Model, which is the basic model of the NIRT. In addition to these assumptions, invariant item ordering is the last assumption of NIRT's strict model, Double Homogenity Model and in the concept of this assumption, item popularities are expected to be invariant across different ability groups. The invariant item ordering makes possible to estimate abilities of examinees at interval level and this feature is only provided with the Rasch model in parametric IRT models. This feature is the strongest point of this model and can be checked with several techniques, which are composing p matrixes, rest-group analysis and H^T coefficients (Junker, 2000; Meijer, Tenderio and Wanders, 2015; Van Schuur, 2011).

These steps were followed in order to define the applicable model for the datasets and it was found that while the unidimensionality and monotonicity assumptions were met for the data sets, invariant item ordering was not met, so the Monotone Homogeneity Model was used in order to estimate item parameters from both data sets. After determining the applicable models for the data sets according to the approaches, item parameters were estimated and the descriptive statistics of the parameters are given in Table 2.

Table 2. Descriptive Statistics Calculated from the Item Parameters Being Estimated with PIRT and NIRT

IRT Approach	Data Sets	Parameters	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
PIRT	LocalDependent	b	-1,17	1,20	-,121	,665	,457	-,736
	LocalIndependent	b	-1,18	1,17	-,093	,654	,313	-,795
	LocalDependent	a	,79	2,70	1,558	,406	,658	1,388
	LocalIndependent	a	,00	2,78	1,495	,495	-,259	3,090
NIRT	LocalDependent	b	,18	,73	,527	,160	-,485	-,797
	LocalIndependent	b	,18	,73	,515	,157	-,379	-,882
	LocalDependent	a	,23	,67	,385	,081	1,815	5,122
	LocalIndependent	a	,24	,67	,386	,080	1,765	5,068

Table 1 includes the descriptive statistics of the estimated item parameters. The item difficulty and discrimination parameters are shown as b and a sequentially. When the values of PIRT are examined, it is clear that b parameter values are so close to each other in the datasets. Especially the minimum and maximum values of b parameters are nearly the same. Also, the mean values are so similar to each other. In the NIRT approach, both parameters values are nearly the same and the descriptive statistics of b are closer than the ones estimated from the PIRT approach. As for item discrimination, a parameters, nearly the same values are obtained from the non-parametric approach. However, a parameters estimated from the parametric approach are so different from each other. In the local independent dataset, the minimum value of a parameter is calculated as ,00 while in the local dependent, it is estimated as ,79. This finding shows that violation of the local independence assumption increases the values of a parameters and it affects a parameters rather than b parameters. Though, the mean value of a parameter is so close in the both data sets, so this increase doesn't affect the other items and doesn't create any change in the mean value. But there is a big change in the skewness and kurtosis values in the local dependent data set, therefore it can be said that the form of the distribution of a parameters is changing in case there is a violation of local independence assumption. In short, it is clear that in case of local dependency, the item discrimination parameters estimated from PIRT are open to change while the difficulty parameters don't change so much.

After the first analyses with the descriptive statistics, in order to investigate the differences among the parameters deeply, the significance of the mean values was tested. Before conducting the significance tests, the assumptions of the tests are tested. Firstly, the normality of the parameters is tested by using the skewness, kurtosis values, histograms and normality tests. According to the results of normality analyses, it was found that, T-test is applicable for b parameters in all the data sets and two approaches. As for a parameters, Related Sample T-test was used for the parameter estimations obtained from PIRT and Wilcoxon Ranked Order Test was used for the parameters taken from NIRT. According to the test results, it was found that there is no significant difference between the parameters' means calculated from two approaches. Also, the differences between the means calculated from the local dependent and independent data sets were found insignificant, too. After the difference of the means analyses, the data analyses process was retained with the investigation of the relationships among the parameters estimated from different datasets and two approaches. The relationships between the parameters were investigated by calculating the correlation coefficients. Due to the normality of the b parameters, Pearson Moment Correlation Coefficient was calculated for item difficulty parameters. As for item discrimination parameters, the Spearman Brown Correlation Coefficient was used and the results are given Table 3 and 4 below.

Table 3. Pearson Correlation Coefficients Between the Item Difficulty Parameters Estimated from PIRT and NIRT

		PIRT		NIRT	
IRT approach		LocalDependent	LocalIndependent	LocalDependent	LocalIndependent
PIRT	LocalDependent	1			
	LocalIndependent	,933**	1		
NIRT	LocalDependent	-,981**	-,906**	1	
	LocalIndependent	-,909**	-,981**	,919**	1

**. p< 0.01

Table 4. Spearman-Brown Correlation Coefficient Between the Item Discrimination Parameters Estimated from PIRT and NIRT

		PIRT		NIRT	
IRT approach		LocalDependent	LocalIndependent	LocalDependent	LocalIndependent
PIRT	LocalDependent	1			
	LocalIndependent	,793**	1		
NIRT	LocalDependent	,921**	,766**	1	
	LocalIndependent	,918**	,791**	,968**	1

**. p< 0.01

In Table 3, there are Pearson Correlation Coefficients calculated from the item difficulty parameters estimated from PIRT and NIRT. When the correlation coefficients calculated in PIRT analyzed, it can be seen that the item difficulties estimated from the local dependent and independent data sets are so close to each other and the correlation between them is calculated as ,933, which is so high value for a correlation coefficient of which maximum value is 1.00. As for the non-parametric approach, the correlation coefficient of b parameters calculated from different sets was found so high again. Considering the relationship between the parameters estimated from two different approaches, it can be seen that the b parameters estimated from the same data set are highly correlated with each other. For example, the correlation coefficient of b parameters estimated from the local dependent data sets is calculated as -,981 which shows a negative but nearly a perfect correlation. In a short, the correlation coefficients in Table 3, show that b parameters are highly related with each other even if the local independence assumption is violated or not.

In Table 4, the Spearman-Brown Correlation Coefficients calculated between the item discrimination parameters are taken place. Starting with the parametric approach, one can say that there is a high-level correlation between a parameters in case of local dependence. However, the correlation coefficient is calculated as ,793 and it is lower than the one calculated for b parameters shown in Table 3. In respect to the non-parametric approach, the correlation coefficient calculated from the local dependent and independent data set is,968 and it was found as significant at 0,01 level. Considering this finding, it is possible to say that the item discrimination parameters estimated from NIRT are affected less than the parameters estimated from PIRT, in case of local independence violation. Comparing coefficients calculated from two approaches, it was found that a parameters estimated from the local dependent data set according to the PIRT are correlated higher than the ones estimated from the NIRT. It shows that whether or not there is a violation of the local independence assumption, item discrimination parameters estimated according to the NIRT are open to change less than the PIRT. In summary, it is clear that, item discrimination parameter is more sensitive to the violations of local independence assumption than the item difficulty parameters in the parametric approach and in PIRT, there is much more change in the values of this parameter than NIRT.

DISCUSSION

In the concept of this study, the effects of the local independence assumption on the item parameter estimations are analyzed by using two different IRT approaches, which are parametric and non-parametric approaches. First of all, two data sets were composed and one of them is prepared as a

local dependent. Then the item parameters, difficulty, and discrimination, are estimated both parametric and non-parametric IRT models. After the estimations of parameters, the descriptive statistics of them, the relationships and differences among them are investigated by using tests. When the findings are examined, all in all, it is clear that there are differences in the parameters estimated from the approaches but these differences are not at the significant level. The reason for the differences between the parameters being insignificance may be the limited number of local dependent items in the dataset, which may affect the whole data set at low-level. In different researches, the number of items violating the assumptions may be increased and the effects of the increment in the numbers of violating items can be analyzed.

Also, the other finding is that item difficulty parameters estimated the different approaches are highly and negatively correlated with each other whether local independence assumption is met or not. The reason for the negativity of this coefficient is the difference in the definition of item difficulty in the parametric and non-parametric approach. In PIRT, the b values change in -2 and +2 and the higher value means the harder item. However, the situation is totally contrasted in non-parametric IRT. In this approach, item difficulty is the popularity of the item, like in Classical Test Theory. It is calculated as the percent of the correct, so increase in the p value means easier items. Due to the stated difference in the definition the item difficulty parameters, the direction of the correlation coefficient is negative.

Despite of the high and significant correlation between item difficulties estimated from both approaches, it is not valid for item discrimination parameters, and as stated in the previous researches, item discrimination parameters are highly affected from local independence violation in parametric approach. Moreover, the other finding of the study is that item discrimination parameters are inflated artificially in case of local independence violation in PIRT approach. As for NIRT, just like in item difficulty parameters, item discrimination parameters are not influenced by the violation of the assumption. Hence as in the related studies, it is found that NIRT models do not require the IRT assumption as the same level as PIRT. As stated in the basic sources about NIRT, such as Sijtsma and Molenaar (2002), Van Schuur (2011) and Junker (2001), etc., non-parametric IRT is free from traditional IRT assumptions. This finding is consistent with the other researches and it can be said that in case of the possibility of local independence violation, one should prefer NIRT rather than PIRT especially in the estimation of item discrimination parameters, because the item discrimination parameters estimated from NIRT is more accurate than the ones estimated from PIRT in case of local dependence. Moreover, the high correlation coefficients computed between the parameters estimated from original data sets can be interpreted that the usage of the NIRT models should not be limited only when the requirements of PIRT are not provided. The findings of the study have made clear that the item parameters obtained from both parametric and non-parametric approaches are so similar to each other and this finding is consistent with the similar studies (Mor Dirlik, 2017; Meijer, Sijtsma and Smidt, 1990). Hence it is clear that when the basic goal is to scale and determine the item parameters, especially in the small samples and few items, NIRT models will provide so similar item parameters with the PIRT models, so they can be preferred. As for recommendations for the new researches, the number of items can be changed and this effect can be analyzed in detail. Also, in the other researches, the change in the ability parameters can be investigated in case of local independence violation.

REFERENCES

Balazs, K., & De Boeck, P. (2006). Detecting local item dependence stemming from minor dimensions: Interuniversity Attraction Pole statistics network [technical report].

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. (First Edition). NY: Guilford Publications, Inc.

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.

Chen, C. T., & Wang, W. C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement*, *31*(5), 388-411.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Gulliksen. H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hambleton, R. K., & Swaminathan, H. (1985). 1985: Item response theory: principles and applications. Boston, MA: Kluwer-Nijhoff.

Henning, G. (1989). Meanings and implications of the principle of local independence. *Language Testing*, *6*(1), 95-108.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, 56(2), 255-278.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In Boomsma A., et al. (Eds.) Essays on Item Response Theory. New York: Springer-Verlag.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Kline, R. B. (2005). *Principles and practice of structural equation modelling*. (Second Edition). NY:Guilford Publications, Inc.

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

Lord, F. M.& Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. (2014). The use of nonparametric item response theory to explore data quality. In *Handbook of Item Response Theory Modeling* (pp. 103-128). Routledge.

Meijer, R. R.. Sijtsma, K and Smidt, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. Applied Psychological Measurement. 14. 283–298

Mor Dirlik, E. (2017). The Comparison of Item and Ability Estimations Calculated From The Parametric And Non-Parametric Item Response Theory According To The Several Factors. Unpublished Doctoral Dissertion. Institute of Educational Sciences, Ankara University.

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

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics, vol. 26: Psychometrics (pp. 719–746). Amsterdam: Elsevier, North Holland.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage.

Sijtsma K, Molenaar IW (2002). Introduction to Nonparametric Item Response Theory. Sage, Thousand Oaks, CA.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55,293-325.

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology*, 82(3), 448.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer, New York, NY.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis* (Vol. 169). Sage.

Vermunt, J. K., & Magidson, J. (2004). Local independence. *Encyclopedia of social sciences research methods*, 732-733.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.