International Journal of Educational Studies and Policy (IJESP)

Volume: 5, Issue: 2, November 2024

Item Response Theory Assumptions: A Comprehensive Review of Studies with Document Analysis

Mahmut Sami Yiğiter[1], Erdem Boduroğlu[2]

ABSTRACT

Item Response Theory (IRT), over its nearly 100-year history, has become one of the most popular methodologies for modeling response patterns in measures in education, psychology and health. Due to its advantages, IRT is particularly popular in large-scale assessments. A pre-condition for the validity of the estimations obtained from IRT is that the data meet the model assumptions. The purpose of this study is to examine the testing of model assumptions in studies using IRT models. For this purpose, 107 studies in the National Thesis Center of the Council of Higher Education that use the IRT model on real data were examined. The studies were analyzed according to sample size, unidimensionality, local independence, overall model fit, item fit and non-speedness test criteria. According to the results, it was observed that the unidimensionality assumption was tested at a high level (89%) and Factor Analytic approaches were predominantly used. Local independence assumption was not tested in 36% of the studies, unidimensionality was cited as evidence in 40% of the studies and tested in 24% of the studies. Overall model fit was tested at a moderate level (51%) and Log-Likelihood and information criteria were used. Item fit and Non-Speedness testing were tested at a low level (26% and 9%). IRT assumptions should be considered as a whole and all assumptions should be tested from an evidence-based perspective.

Keywords: Item response theory, assumption, unidimensionality, local independence, overall model fit, Item fit, non-speedness test, factor analysis.

DOI: https://doi.org/10.5281/zenodo.14016086

Article Info:

Received: 03.06.2024 **Accepted: 12.10.2024** **Article Type: Research Article**

Cite as: Yiğiter, M. S. & Boduroğlu, E. (2024). Item response theory assumptions: A comprehensive review of studies with document analysis. *International Journal of Educational Studies and Policy*, *5*(2), 119-138.

[1] Corresponding Author: Mahmut Sami Yiğiter, Social Sciences University of Ankara, mahmutsamiyigiter@gmail.com, ORCID: 0000-0002-2896-0201

[2] Erdem Boduroğlu, Ministry of National Education, erdemboduroglu@gmail.com, ORCID: 0000-0001-8318-4914

* This article was presented as an oral presentation at the VII. International Congress on Measurement and Evaluation in Education and Psychology.

Introduction

Many models have been developed throughout the history to place scores obtained from educational and psychological measurements on a scale. Classical Test Theory (CTT) and Item Response Theory (IRT) are the most widely used, known and important among these models. Classical Test Theory models are often referred to as "weak models". The reason for this is that the assumptions of the models can be easily met with the test data. On the other hand, the test data are less likely to meet the assumptions since they are strict in Item Response Theory models and therefore they are called as “strong models” (Hambleton and Jones, 1993). Classical Test Theory is a theory based on observed score (X), true score (T) and error score (E). This theory has a simple linear equation expressed as X=T+E, consisting of the sum of the observed test scores (X), the unobservable and often latent true score (T) and the error score (E) (Novick, 1966). Since there are two unknown variables in the equation (T and E), the equation cannot be solved unless there are some assumptions. These assumptions of the CTT are (a) the true scores and error scores are uncorrelated, (b) the expected value of the error score is equal to zero, (c) the true scores and error scores in parallel tests are uncorrelated (Lord & Novick, 1968). The true score in the basic equation of the theory is the difference between the observed test score and the error score. The true score of the examinee is also defined as the expected score from the parallel forms. CTT models are focused on modelling at the test score level. These models relate the true score to the total score obtained from the test, not to the scores obtained from the items. The biggest advantage of CTT is that its assumptions are easily met and item parameters can be easily calculated (Fan, 1998). However, CTT has some limitations. Lord (1953) states that the true score in the CTT varies according to the difficulty of the test. For example, while an examinee will score low on a difficult test, he/she will score high on an easy test. While the examinee's ability level has not changed, the fact that the examinee's true score takes different values indicates that the examinee's true score in CTT is dependent on the test or the group he is in. Thus, different methods and models have been sought to overcome the limitations of CTT.

Item Response Theory (IRT) is one of the most popular methodologies used to model response patterns from measurements (Boduroğlu & Anil, 2023). IRT studies started with the modelling of latent variables with the work of Thorndike, Thurstone and Symond in the early 20th century until its foundations were laid with the studies of Lazarsfeld and Lord in the 1950s. In the 1960s, IRT developed with Rasch's studies on the "Rasch model" and Birnbaum's studies on the "logistical model" (Baker, 2001; Himelfarb, 2019). It remained as a theory with no practical application until the 1970s, as computational technology did not enable data analysis with IRT models. With the development of computers in computing technology, IRT applications and research have become widespread. Over time, more complex models have been developed on logistic models (De Boeck and Wilson, 2004; Reckase, 1997; Rijmen et al., 2003). In the 21st century, IRT has found a wide and fundamental application area, especially in large scale educational assessment. Today, IRT is used in the social sciences and behavioral sciences as well as education, psychology and medical sciences (Reise & Waller, 2009; Thissen & Steinberg, 2020; Zanon et al., 2016; Mutluer & Çakan, 2023).

IRT is a powerful scaling method used to determine the characteristics of the items and examinees based on the responses of examinees to the items in the test (Embretson & Reise, 2000; Selçuk & Demir, 2024; Sözer & Kahraman, 2021). In IRT, there is a parameter called ability, denoted by theta, which corresponds to the true score of the individual in CTT. In addition, IRT provides useful information about the contribution of the items in the measurement of the latent construct, its quality and at which points of the ability scale it performs the best measurement. One of the important features of the IRT is that it places both examinees and items on the same scale. An examinee may have a high or low ability level, and an item may have high or low difficulty and be on the same scale. Having a common scale for examinees and items makes it possible to evaluate the amount of information that items provide in terms of latent structure and to match items in accordance with the ability level of the individual taking the test (Van der Linden & Glas, 2010). Another advantage of IRT is that both item parameters and ability parameters can be estimated without being dependent on the group or the test. That is, (a) examinees' ability parameters are independent of the test items they take, (b) item parameters are independent of examinees' ability distributions (Hambleton et al., 1991). Thanks to its advantages, IRT is actively used in large-scale tests, computerized adaptive testing, test equating, differential item functioning, cognitive diagnostic model and scale development applications (Aybek, 2023; Ayva Yörü, 2024; Doğan & Atar, 2024; Kılıç et al., 2023; Saatcioglu & Sen, 2023; Şahin, Yildirim & Boztunc Öztürk, 2023; Yiğiter & Doğan, 2023).

As previously mentioned, IRT models are strong due in part to the fact that their underlying assumptions are challenging to meet. To take advantage of the benefits of the above mentioned IRT, the assumptions of the model need to be tested and met. Estimation made without meeting assumptions will contain systematic error, and the validity of the obtained item and ability parameters will become doubtful (Hambleton & Swaminathan, 1985; Reckase, 2009). On the other hand, IRT needs large samples for the estimation of item and ability parameters. The minimum sample size differs according to the IRT model used for accurate estimation of parameters in IRT applications. As the IRT model used gets more complex, larger samples are required (Sireci, 1991). The assumptions of the Item Response Theory have been discussed in many different sources. Trabin and Weiss (1983) discussed the assumptions of IRT under three headings: (a) unidimensionality, (b) local independence, (c) item characteristic curve graph. Hambleton and Swaminathan (1985) stated that there are four assumptions: (a) dimensionality, (b) local independence, (c) item characteristic curve fit, (d) non-speedness test. Crocker and Algina (1986) express that there are two assumptions: (a) unidimensionality and (b) local independence. According to Embretson and Reise (2000), IRT has two basic assumptions: (a) item characteristic curve have a specified form and (b) local independence. Demars (2010), on the other hand, discussed this under the headings of (a) unidimensionality, (b) local independence and (c) fit. Stone and Zhu (2015) lists five different assumptions: (a) dimensionality, (b) local independence, (c) form of the IRT Model (Overall Model Fit), (d) non-speedness test and (e) Model Fit (Item and Person Fit).

In the following sections of the study, firstly, sample size in IRT is discussed. Then, the assumptions of IRT are described and the methods used to test these assumptions are explained under separate headings in line with the main sources and books in the literature.

Sample Size

It is difficult to accurately determine the sample size required for an accurate estimation of the item and individual parameters to be obtained from a test. In particular, as the IRT model used becomes more complex, both larger sample sizes and longer tests are needed to obtain accurate estimations (Hambleton, 1989). As an outgrowth, IRT models are used in large scale assessment. In addition, scaling and estimation can be made with IRT in tests which consist of fewer items and are applied to groups with a certain sample size (Emretson & Reise, 2000). In the literature, there are many studies on the sample size which is required to obtain accurate and stable parameter estimations with IRT. Lord (1968) discusses that a sample size of more than 1000 is needed to estimate the item discrimination parameter accurately. It is stated that the Rasch or 1PL model with fixed item discrimination can be used with a sample size of 100 or 200 (DeMars, 2010). In models where item discrimination is estimated, it is seen that a larger sample size is required. Ree and Jensen (1980) state that a sample size of 500 or more is required in order to estimate the item discrimination and difficulty parameters accurately. According to Hulin et al. (1982), a sample size of 500 or more is required for the 2PL model, and a sample size of 1000 or more for the 3PL model. Swaminathan and Gifford (1983) state that a sample size of 1000 gives good results for the 3PL model. Harwell and Janosky (1991) suggested that the sample size should be more than 250 in order to estimate the parameters correctly. Demars (2010) says that the sample size should not be less than 500 for the 2PL and 3PL models.

Unidimensionality

Unidimensionality means that there is only one type of ability that affects a test taker's performance in a test subject (Lord & Novick, 1968). In other words, it is a single feature that keeps the items in the measurement tool together. Unidimensionality appears as the basic assumption of unidimensional IRT models. In order for an item group to be considered as a single dimension, these items must have a common characteristic and this item group must have a common variance that explains the variability among examinees. If unidimensionality is violated, the multidimensional structure of the latent trait space will not match one-to-one with the unidimensional IRT model. When unidimensionality is violated, scores obtained with the unidimensional IRT may be biased. On the other hand, it is very difficult to achieve pure unidimensionality in practice. Because examinees cannot be expected to act in line with only one trait while answering the items. Also, the measured trait may be a multidimensional construct. Multidimensional IRT models have been developed for multidimensional tests. These models assume that more than one trait underlies performance. Multidimensional IRT models can be used if more than one trait determines the examinee's performance (Reckase, 2009; Kartal & Mor Dirlik, 2021).

It is seen that Factor Analytical methods are generally used to test the unidimensionality assumption (Ziegler & Hagemann, 2015). Factor Analytical methods try to explain the relationships between responses to test items with fewer factors (Stone & Zhu, 2015). For this reason, with these methods, factor analysis is applied to the data obtained from the items and a dominant factor is sought (Erkus et al., 2017). Factor Analysis is categorized as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). While EFA helps us to determine the possible factor structure underlying the observed variables based on examinees responses, CFA allows testing the hypothesis that the determined relationships exist (Suhr, 2006). Apart from factor analysis, many different methods have been developed to test dimensionality. Horn's Parallel Analysis (Watkins, 2006), Velicer's Map Test (Zwick & Velicer, 1986), DIMTEST (Stout, 1987; Nandakumar & Stout, 1993), hierarchical cluster analysis (HCA/CCPROX) (Roussos, Stout, & Marden, 1998) and DETECT index (Zhang & Stout, 1999) are other methods used to determine dimensionality. Principal Component Analysis on standardized residuals is another method used to test dimensionality (Chou & Wang, 2010). As noted above, Factor Analysis functions as a dimension reduction. Decision Tree Induction, Support Vector Machine (SVM), Naive Bayes Classifier and Random Forest Classifier methods from cluster analysis and Machine Learning (ML) algorithms as dimension reduction methods also appear in the literature as methods used in examining dimensionality (Hasan et al., 2021; Dogan & Basokcu, 2010).

Local Indepencence

Local independence means that an examinee's probability of answering an item is independent of their response behavior to other items. In other words, when the ability that affects test performance is kept constant, examinees' responses to items are statistically independent from each other (Hambleton & Swaminathan, 1985). Local independence comes from the basic rule of the probability function on which the IRT is based. Failure to meet this assumption causes the estimations obtained from statistical calculations to be incorrect (Looney & Spray, 1992). For example, in the estimation of ability with the Maximum Likelihood Function, the estimation of ability is estimated by multiplying the probabilities obtained from the responses of the examinees to the items. Likelihood functions calculate probability results by treating items as if they are independent of each other at a given ability level. In order for the probability of two events to occur at the same time to be equal to the product of the probabilities of the two cases, the cases must be independent of each other. It is stated that if the local independence assumption is not met, the test information and item discrimination parameters are overestimated (Chen & Thissen 1997; Embretson & Reise, 2000; Sireci et al., 1991). Junker (1991), on the other hand, states that ability parameters are strongly biased in case of local dependency. In addition, there are studies in the literature showing that the item difficulty parameters of local dependency are incorrectly estimated (Eckes, 2011; Min & He, 2014). There can be many factors affecting the assumption of local independence in educational tests. Response to an item in items with a common root may affect the responses to other items with the same common root (Chen & Thissen (1997). Also, cheating behavior, fatigue (Yen, 1993), students' different practice situations or the test being a speed test (Embretson & Reise, 1999). 2000) affects local independence. In cases where local independence is violated, three solutions are distinguished. First, one of the two items with local dependence between them can be excluded. Second, by creating an item group from local dependent items, this item group can be scaled with IRT models under multi-category models (Yen, 1993). Third, Testlet Response Theory models, which make predictions by considering grouped items, can be used, (Wainer et al., 2007).

A wide variety of methods have been developed to test local independence. Local independence is usually examined through the relationship between the items over the residual matrix calculated by the difference between the observed matrix and the produced matrix from the model. Analyses such as Yen's Q_3 (1984), G^2 squared (or χ^2) (Chen & Thissen, 1997), correlation between residuals (Linacre, 2009), JSI (Edwards et al., 2018) are the methods used to test local independence. It is also stated that local independence can be examined by categorizing the ability scale into different ability ranges and examining the correlation or covariance between the items (McDonald, 1981; Tucker et al., 1986).

Overall Model Fit

As with any modelling study, it is necessary to measure the misfit between the model and the data to determine which IRT model to use. Estimates and inferences made with an inappropriate IRT model will be invalid (Maydeu-Olivares, 2006). Evaluation of the overall fit of the model and the data can be done by comparing the total observed score distribution with the expected score distribution by the model. Evaluation of residuals from differences between observed and expected correct response rates at all skill levels provides precious information for overall model data fit. Model-data fit can be mentioned if the residuals are small and randomly distributed. Embretson and Reise (2000) state that residuals approaching zero for a model can be taken as a measure of model-data fit. In the evaluation of the general model-data fit, information criterion values that exhibit the χ^2 distribution are generally used. The prominent ones are Log-Likelihood Test [-2*LL], Akaike Information Criteria [AIC], Consistent AIC [CAIC], Bayesian Information Criteria [BIC], sample size–adjusted BIC [SABIC], Hannan-Quinn Criterion (HQ) values (Antoniou et al., 2022; Hambleton & Swaminathan, 1985). The low statistics calculated for the models indicate a better model-data fit. The differences between these values obtained from the models allow comparison of the models with the chi-square statistics at the relevant degrees of freedom. Another test's goodness-of-fit statistic is M_2 (Maydeu-Olivares & Joe, 2006).

In addition, goodness of fit indexes (χ^2/sd, GFI, AGFI, CFI, NFI, TLI, SRMR, RMSEA, et al.) are used in model selection in order to determine which model fits the data better in IRT (Chalmers, 2012). (For more information on general model data fit, see Demars, 2010; De Ayala, 2013; Maydeu-Olivares, 2006).

Item Fit

In order to make inferences from a data set scaled with IRT, it is important to meet the fit of the items to the model. In terms of the accuracy of the estimations to be obtained from the model, the items in the test should fit with the model. Estimates from models with non-fit items will lead to biased estimation of ability parameters, unfair ranking of examinees, and incorrectly equalized scores (Wainer & Thissen, 1987; Yen, 1981). The Item Characteristic Curve (ICC) can be used to evaluate the fit of the items to the model. The indicator of the item's fit with the model is the similar distribution of the estimated ICC and the observed values across the throughout ability scale. In other words, the small difference between the ICC and the observed values will indicate the fit of the item to the model. The difference between the ICC and the observed values is called the residual. The fact that the residuals approach zero is a sign of good item fit (Embretson & Reise, 2000). Visual inspection of the residuals on the ICC is helpful, but it also draws criticism for the subjectivity of the evaluation. For this reason, many item fit indices have been developed. These indices are divided into two groups as Traditional and Alternative item fit indices. Traditional indices divide examinees into specific groups and examine the differences between the expected and observed mean values of these groups. Alternative item fit indices have been developed since it was stated in the traditional indices that if the item is misfit, the estimation made is also incorrect, and therefore the expected scores produced from the model will be incorrect. Traditional item fit indices can be listed as follows: OUTFIT, INFIT indices (for Rasch and 1PL models) (Wright & Panchapakesan, 1969), Bock's χ^2 index (Bock, 1972), Yen's Q_1 index (Yen, 1981), G^2 index (McKinley and Mills, 1985). Alternative item fit indices can be listed as S-χ^2 (Orlando & Thissen, 2000), scaling corrected fit statistics (χ^{2*}) (Stone, 2000), and adjusted chi-square/degrees of freedom ratio (χ^2/df ratio) (Drasgov et al., 1985).

Non-Speedness Test

It is an assumption of IRT models that the test is not performed under accelerated conditions. That is, if an examinee did not answer some test items, it must not be because he did not reach the test items or the time period has expired. This situation must be due to the insufficient level of talent of the examinee. This assumption is sometimes referred to in the unidimensionality assumption. When speed affects test performance; test performance is affected by at least two dimensions - measured talent and speed. This situation has a disruptive effect on unidimensionality. The non-speedness test assumption assumes that examinees should have enough time to answer the items they think they can answer. Many different methods have been proposed to test the non-speedness assumption. A few methods in the literature are as follows: The first is to examine the relationship between the scores obtained by applying the same test form to the same group under a certain time limit and without a time limit. The second is the ratio of the variance of the number of items that each examinee left blank to the variance of the number of items they answered incorrectly. If this ratio is close to zero, it is stated that the test is a non-speedness test, and if it is close to one, it is a speed test (Gulliksen, 1950). The third is the examination of the marking rates of the items. It is expected that the percentage of those who reach the items and give correct or incorrect answers is high. The fourth is the evaluation of the percentage of "unreachable" items. In order to determine that the test is a non-speedness test, 80% of the examinees must complete the test by reaching all the items, and each examinee taking the test must reach at least 75% of the items (Swineford, 1956).

Aim and Significance of the Research

In this study, it is aimed to determine the status of examining the IRT assumptions of the studies using the IRT model in the literature. For this aim, master's theses and doctoral dissertations written using the IRT model in Türkiye are addressed. The methods by which the researchers tested the IRT assumptions were examined in detail. It is known that IRT models, which have been widely used recently, have many strengths. However, it should be taken into account that the estimated parameters and interpretations will be erroneous in cases where IRT assumptions are violated. When the literature is examined, it is seen that there is no study that deals with the examining of assumptions in detail. It is thought that this study will contribute to the field in terms of revealing the general framework of the IRT assumptions in the literature and presenting an awareness of testing of these assumptions.

Method

This study is a descriptive research as it reports the existing characteristics of the studies conducted using IRT models in terms of IRT assumptions. Descriptive research aims to report the characteristics of the situation examined in the research as it exists (Fraenkel & Wallen, 2011; Koyuncu & Kılıç, 2021). At the same time, document analysis method was used in this research, which was created by gathering information from the studies in the literature. Document analysis is a systematic process that enables the analysis of the information and content in the written elements considered for the purpose of the research (Ary, Jacobs, & Sorensen, 2010). In this study, document analysis method was preferred to examine the master's and doctoral theses written on Item Response Theory between 1993 and 2023 in Türkiye.

Many different and complex IRT models have been developed, such as Multidimensional IRT Models, Mixture IRT Models, Nonparametric IRT Models, Explanatory Item Response Models, etc. Studies conducted on these different IRT models were excluded from the context of this study due to the different scaling methods and assumptions. Therefore, in this study, only master's and doctoral theses prepared using unidimensional IRT models were analysed. This situation is a limitation of this study.

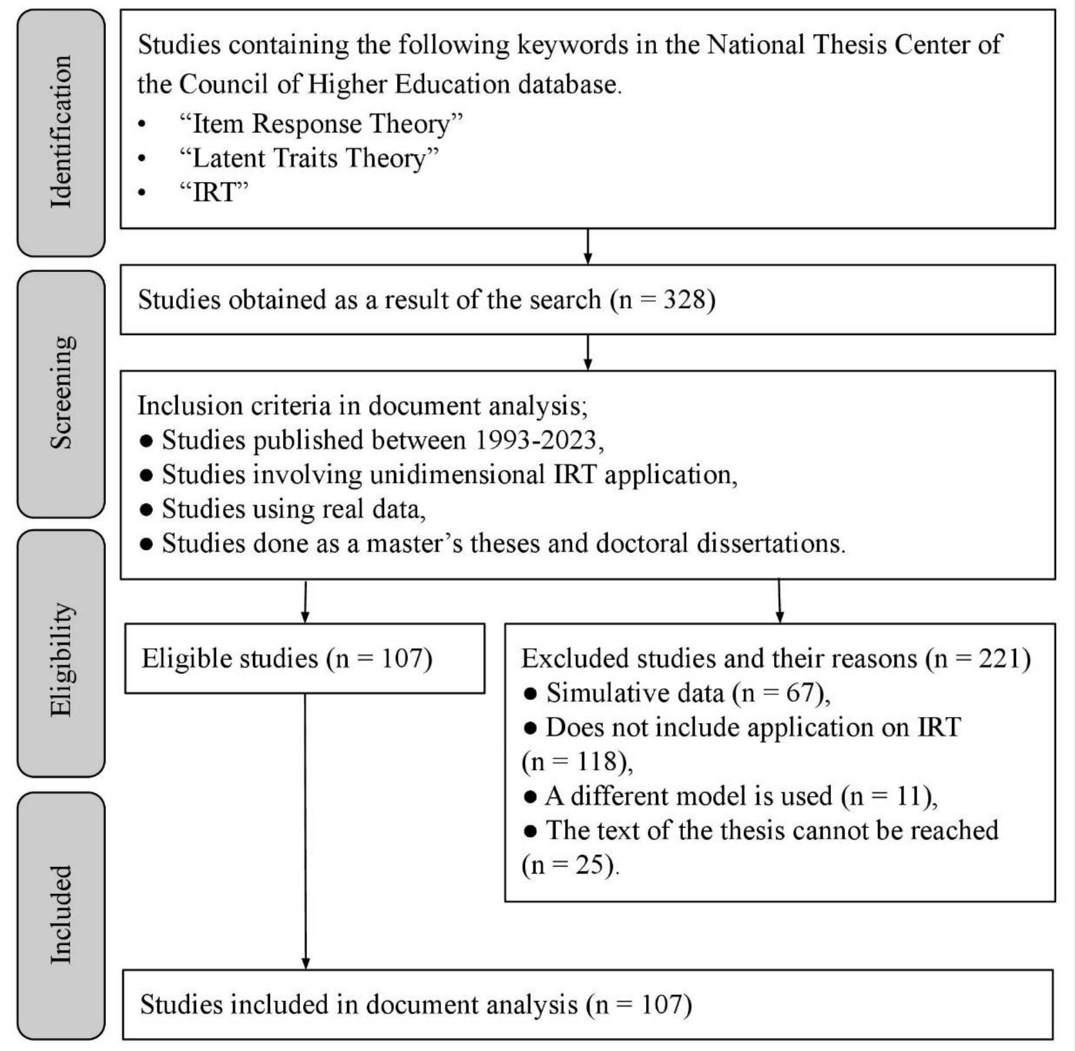

Identification
Studies containing the following keywords in the National Thesis Center of the Council of Higher Education database.
• "Item Response Theory"
• "Latent Traits Theory"
• "IRT"
Screening
Studies obtained as a result of the search (n = 328)
Inclusion criteria in document analysis;
● Studies published between 1993-2023,
● Studies involving unidimensional IRT application,
● Studies using real data,
● Studies done as a master's theses and doctoral dissertations.
Eligibility
Eligible studies (n = 107)
Excluded studies and their reasons (n = 221)
● Simulative data (n = 67),
● Does not include application on IRT (n = 118),
● A different model is used (n = 11),
● The text of the thesis cannot be reached (n = 25).
Included
Studies included in document analysis (n = 107)

Figure 1. Literature review flow chart

Within the scope of the research, the database of the National Thesis Center of the Council of Higher Education was searched. Details of the literature review are presented in Figure 1.

107 studies that are in line with the inclusion criteria were examined within the scope of the research. In determining the criteria to be investigated, the researchers examined the books and articles that were the main sources in the development and dissemination of the IRT. Aligning with the literature review, unidimensional IRT assumptions are discussed under five headings in Figure 2.

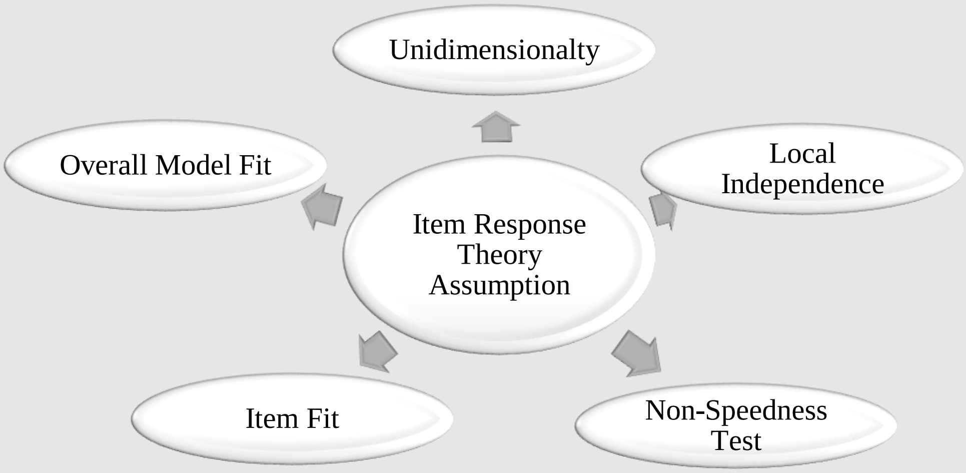


Figure 2. Unidimensional IRT assumptions

As noted in the Sample Size section, the sample must be of a certain size in IRT. For that reason, it was considered to analyze the sample size as a criterion. Then, 107 studies were analyzed according to the assumptions in Figure 1 (1) unidimensionality, (2) local independence, (3) overall model fit, (4) item fit and (5) non-speedness test.

After the included studies were reviewed within the scope determined by the researchers, the data were analysed using descriptive analysis. Results were reported using frequency and percentage.

Ethics

The ethics application for the study was made on 20/06/2021 and the research was carried out with the approval of Social Sciences University of Ankara Ethics Commission dated 06/08/2021 and numbered 14020.

Results

Distribution of Studies by Years

Table 1 shows the distribution of the studies in the study group by years.

Table 1. Distribution of IRT studies by years

Year	Number of Studies	Year	Number of Studies	Year	Number of Studies	Year	Number of Studies
1993	1	2001	0	2009	4	2017	6
1994	2	2002	1	2010	1	2018	11
1995	1	2003	1	2011	4	2019	10
1996	0	2004	0	2012	4	2020	5
1997	0	2005	2	2013	6	2021	8
1998	0	2006	3	2014	7	2022	2
1999	1	2007	0	2015	9	2023	6
2000	0	2008	5	2016	7	Total	107

When Table 1 is examined, it is seen that the studies written using IRT have increased significantly in the last 10 years.

Sample Size

The sample sizes of the studies were investigated in four classes as [0,200], [201,500], [500,1000], [1001+]. In order to evaluate the sample size of the polytomous IRT models, the dichotomous IRT model was coded as the corresponding models (Partial Credit Model -> Rasch or 1PL; Generalized Partial Credit Model, Graded Response Model -> 2PL) (Brzezińska, 2016).

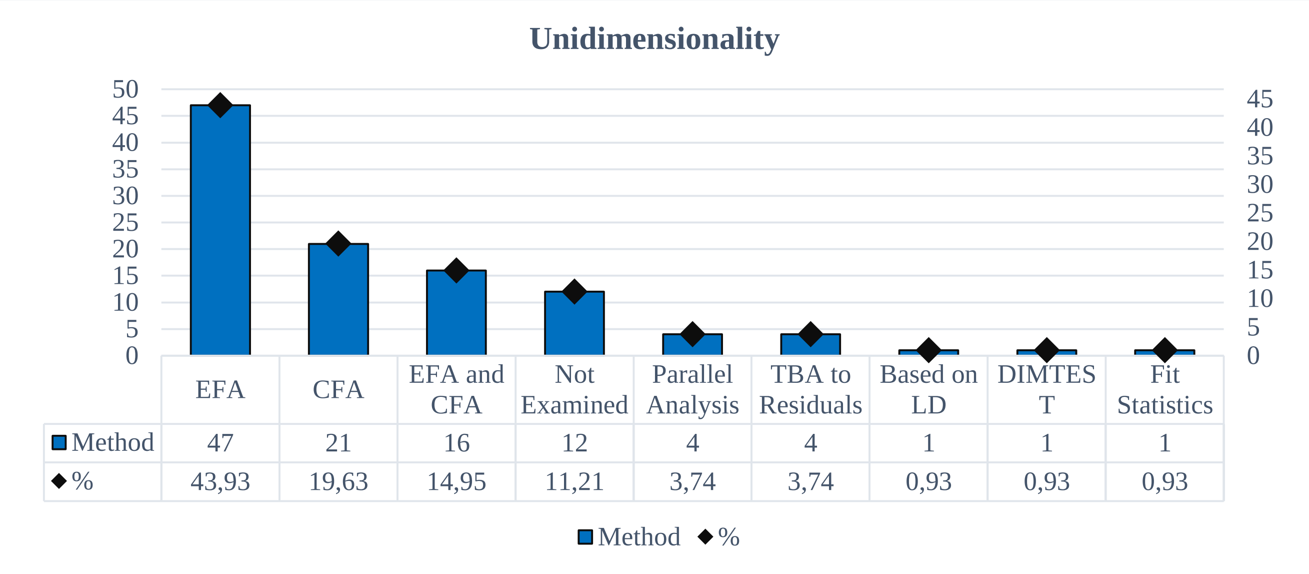
Table 2. Sample size

Model	Sample Size				Total
	[0,200]	[201,500]	[501,1000]	[1001+]	
Rasch or 1PL	3	9	3	16	31
2PL	3	10	14	25	52
3PL	0	1	3	20	24
Total	6	20	20	61	107

When Table 2 is examined, it is seen that there are few studies with low sample size. While there are no studies with a sample size of 200 or less in the 3PL model, there are some studies with low sample sizes in the Rasch-1PL and 2PL models.

Unidimensionalty

The distribution of the methods used in testing the unidimensionality assumption in 107 studies examined in the research is presented in the figure below.



	EFA	CFA	EFA and CFA	Not Examined	Parallel Analysis	TBA to Residuals	Based on LD	DIMTEST	Fit Statistics
■ Method	47	21	16	12	4	4	1	1	1
◆ %	43,93	19,63	14,95	11,21	3,74	3,74	0,93	0,93	0,93

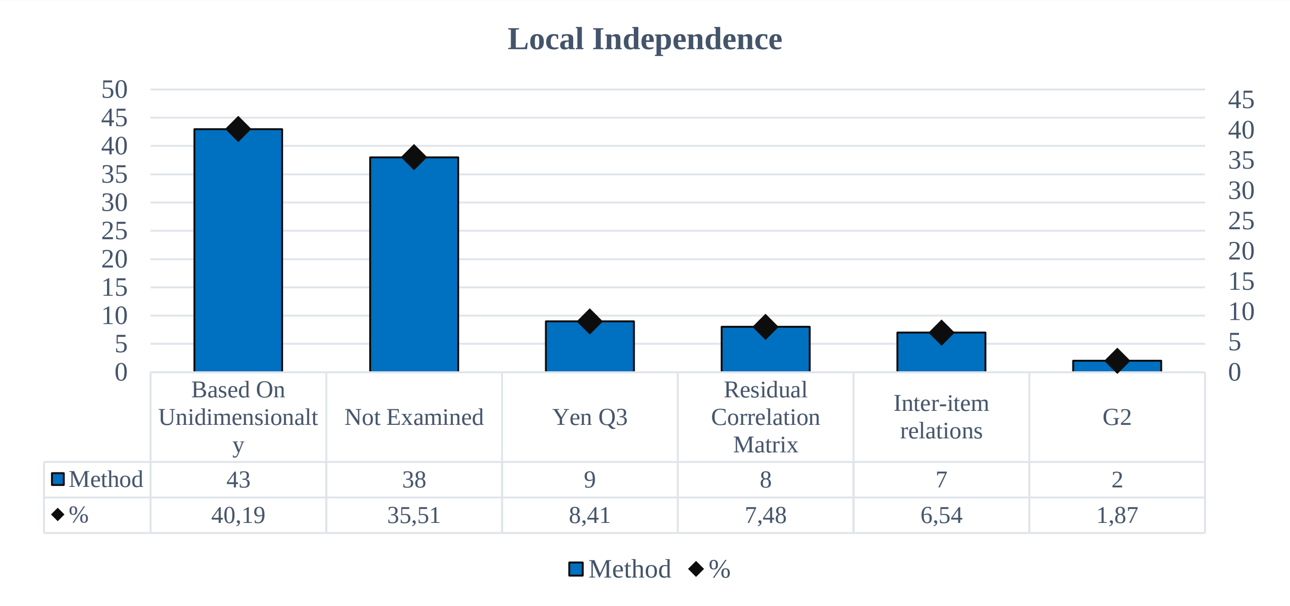
**EFA: Explanatory Factor Analysis, CFA: Confirmatory Factor Analysis, PCA: Principal Component Analysis, LD: Local Dependence*.

Figure 3. Methods used in testing unidimensionality assumption

When the results are analyzed, it is seen that unidimensionality is tested at a high level (n=95, 88.79%). EFA (n=47,%=43.93), CFA (n=21, %=19.63) and EFA and CFA (n=16, %=14.95) are the most used methods for testing unidimensionality.

Local Independence

The distribution of the methods used in testing the local independence assumption in the studies examined in the research is presented in the Figure 4 below.



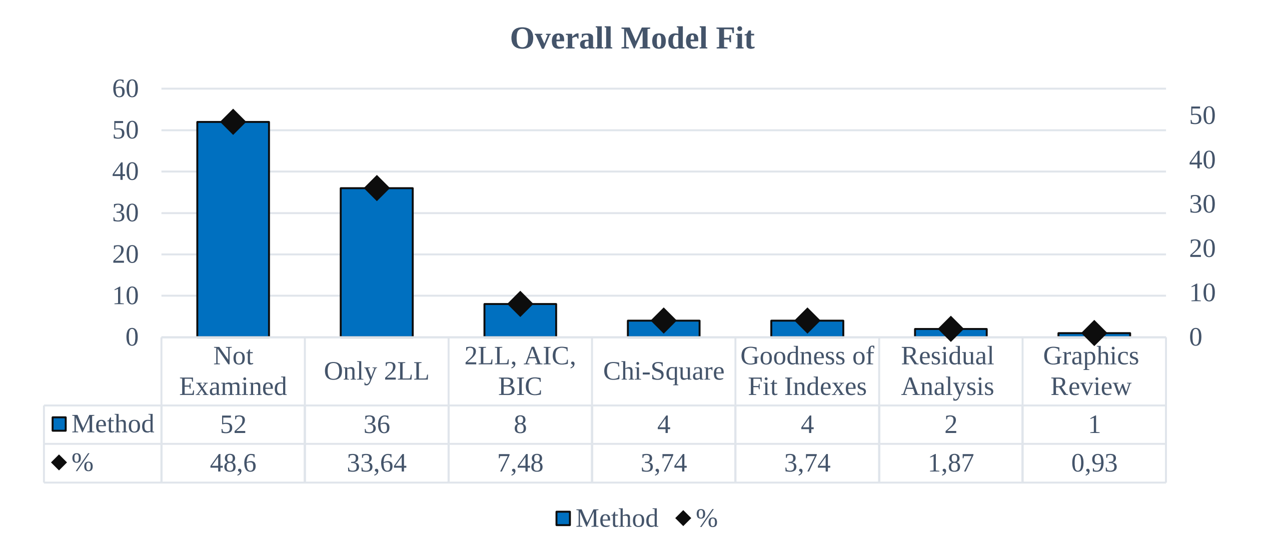
	Based On Unidimensionalty	Not Examined	Yen Q3	Residual Correlation Matrix	Inter-item relations	G2
■ Method	43	38	9	8	7	2
◆ %	40,19	35,51	8,41	7,48	6,54	1,87

Figure 4. Testing the local ındependence assumption

When the results are analyzed, it is seen that the local independence is mostly handled on the basis of unidimensionalism and there is no additional testing (n=43,%=40.19). Yen Q_3 (n=9, %=8.41), Residual Correlation Matrix (n=8, %=7.48), Inter-item relations (n=7, %=6.54) and G^2 (n=2, %=1.87) methods are used to test local independence. In many studies, local independence was not examined (n=38, %=35.51).

Overall Model Fit

The distribution of the methods used in testing the overall model fit assumption in the studies investigated in the research is presented in Figure 5.

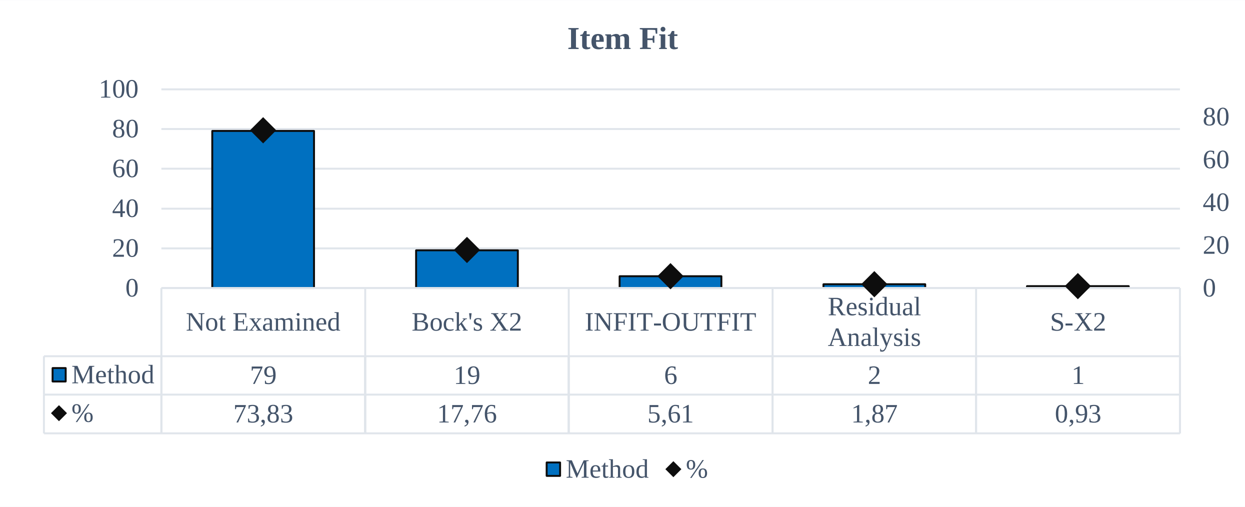


	Not Examined	Only 2LL	2LL, AIC, BIC	Chi-Square	Goodness of Fit Indexes	Residual Analysis	Graphics Review
■ Method	52	36	8	4	4	2	1
◆ %	48,6	33,64	7,48	3,74	3,74	1,87	0,93

Figure 5. Testing of overall model fit

When the results are analyzed, it is seen that the overall model fit was tested at a moderate level (n=55, %51.4). In the testing of overall model fit, it is seen that Log Likelihood (2LL) (n=36, %=33,64) value is mostly analyzed. Information Criteria values (2LL, AIC, BIC) (n=8, %=7.48), Chi-Square (n=4, %=3.74%), goodness of fit indexes (n=4 , %=3.74), Residual Analysis (n=2, %=1.87) and Graphical Review (n=1, %=0.99) are other methods used to test the model data fit.

Item Fit



	Not Examined	Bock's X2	INFIT-OUTFIT	Residual Analysis	S-X2
■ Method	79	19	6	2	1
◆ %	73,83	17,76	5,61	1,87	0,93

Figure 6. Testing of ıtem Fit

The distribution of the methods used in testing the item fit of the studies included in the study is presented in Figure 6.

When the results are analyzed, it is seen that the item fit is tested at a low level (n=28, %=26,16). It is seen that Bock's χ^2 (n=19, %=17.76) statistics are mostly used in the testing of item fit. INFIT-OUTFIT (n=6, %=5.61), Residual Analysis (n=2, %=1.87), and S-χ^2 (n=1, %=0.93) methods are other methods used to test item fit.

Non-Speedness Test

The distribution of the methods used in testing non-speedness test assumption is presented in Figure 7.

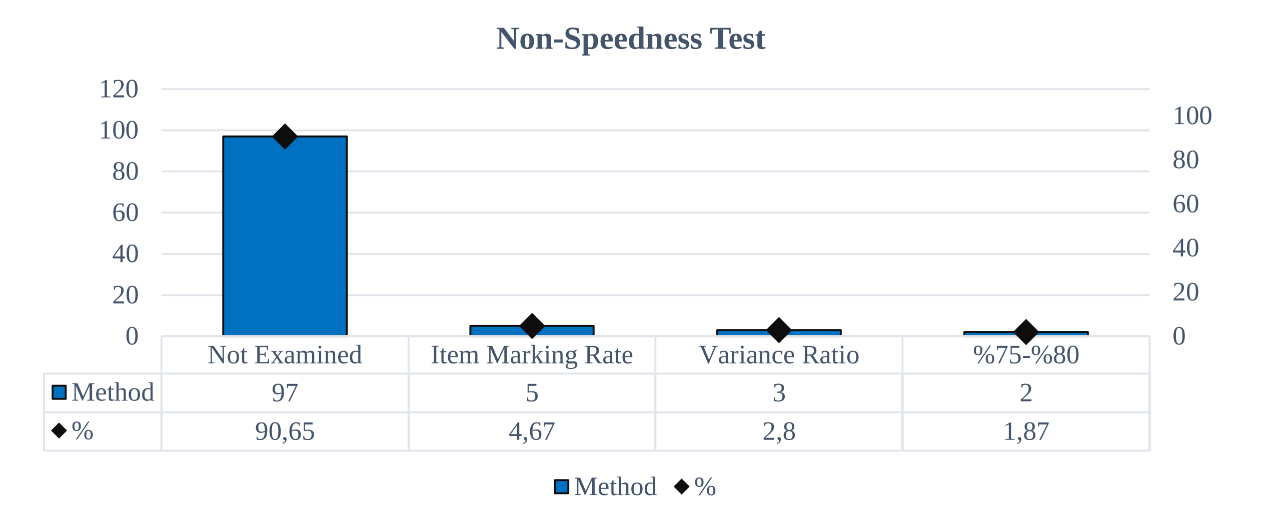


Figure 7. Testing of non-speeded test assumption

When the results are analyzed, it is seen that non-speedness test assumption is tested at a low level (n=10, %9.34). Item Marking Rate (n=5, %=4.67), variance ratio (n=3, %=2.80) and 75%-80% (n=2, %=1.87) methods were used in the testing of non-speedness test.

Discussion and Conclusion

Item Response Theory has been widely used by researchers in recent years thanks to the advantages it offers. In this study; the studies carried out with the IRT model in the last 30 years in Türkiye have been examined and it has been observed that approximately 93% of these studies have been carried out from 2005 to the present. This situation can be interpreted as an indication of the increasing interest in the IRT in recent years. It is known that model assumptions must be met in order to benefit from the strengths of IRT and to interpret test scores correctly. The studies examined in this study were limited to unidimensional IRT models. In this study, the testing of the dimensionality assumption was discussed primarily. Due to the difficulties in the unidimensional test development process, researchers are trying to obtain a dominant factor or component. Although there are many different methods to test the unidimensionality assumption, it is stated that factor analysis methods provide more effective results and are widely used (Erkus, 2006; Lumsden, 1961, 1976; Ziegler & Hagemann, 2015). As a result of the investigations made in this study, it was seen that 78% of the researchers resorted to factor analytical (EFA or CFA) methods in testing the unidimensionality assumption. It is stated that the unidimensionality assumption should be checked in studies using IRT models (Crocker & Algina, 1986; Demars, 2010;

Hambleton & Swaminathan, 1985; Lord & Novick, 1968). In approximately 11% of studies, it was observed that the unidimensionality assumption was not examined.

When the testing of the local independence assumption was analyzed, it was seen that approximately 40% of the researchers did not make a further test, referring to the fact that when the unidimensionality assumption is met, the local independence assumption will also be met. This situation is generally based on Lord's (1980) view that the correlation between the responses of individuals to the items for a given ability level in a one-dimensional test will be zero. So, when the unidimensionality assumption is met, the local independence assumption will also be met. Similarly, since Hambleton & Swaminathan (1985) stated that these two assumptions are equivalent when ϑ ability level is unidimensional, the researchers did not perform a local independence test other than unidimensionality. DeMars (2010), on the other hand, stated that in cases where the dependence between item pairs is at limited levels, it may not emerge as a separate dimension, for that reason, local independence may not be determined with unidimensionality tests and local independence should be tested with different methods. Local independence was tested in 24% of the studies reviewed. Yen's Q_3 test, Residual Correlation Matrix, Inter-item relations and G^2 methods were used to test local independence. In 36% of the studies, the assumption of local independence was not examined. Considering that the estimations obtained from the statistical calculations will be incorrect if the local independence assumption is not met, it is an important problem that there are many studies that do not examine this assumption.

The benefits of IRT for applications such as test development, item bank creation, differential item function (DIF), computerized adaptive testing (CAT), and test synchronization may not be realized unless a fit IRT model is used for a given dataset. The success of IRT applications requires a satisfactory fit between the model and data. The most critical problem caused by model-data misfit may be that parameter invariance, which is the hallmark of IRT, is no longer valid (Rupp & Zumbo, 2006; Shepard, Camilli, & Williams, 1984). Similarly, the items in the test should be fit with the model, that is, the values observed with the estimated ICC should exhibit a similar distribution across the throughout ability scale. When examining the overal model fit in the theses in this research, it was seen that this assumption was not tested in approximately 49% of the studies, and only the LogLikelihood (2LL) value was examined in 34% of the studies. Information criteria, Chi-Square, Goodness of Fit Indexes, Chart Review and Residual Analysis are other methods used to test the overall model fit.

It is seen that in 26% of studies, item fit is tested. Bock's χ^2, INFIT-OUTFIT, Residual Analysis, S-χ^2 indices are the methods used to test item fit. In approximately 74% of the studies, item fit was not examined. The fact that overall model fit and item fit were not tested at a high rate in the studies discussed makes the validity of the results obtained from the model questionable.

In IRT models, the failure of examinees to respond to test items should occur not because of their inability to reach the test items, but because of their limited abilities. In other words, the measurement tool in which IRT models are used should not be a speed test. There are many methods developed to test this assumption. However, it was observed that the assumption of non-speedness test was tested in only 9% of the studies examined. In 91% of the studies, this assumption was not examined. Evidence that the data included in the study was obtained from a measurement tool, which is a non-speedness test, should be presented.

As a result of the examinations carried out, it was seen that there is no certain standard for examining the assumptions in the studies. It is also thought that this situation is caused by the differences between the basic books in the literature in their handling of IRT assumptions. At the same time, it was concluded that the researchers did not test many IRT assumptions with the necessary rigor. Assumptions must be met in order to benefit from the advantages provided by the IRT. However, in many studies, estimations were made without testing these assumptions. In this case, it should be considered that the measurement results obtained and the decisions made based on these results may be incorrect. Another point is that thanks to the many packages, programs and software developed with the advances in computer technologies, it has become easier to test the assumptions and access to many methods that can be used. It is thought that the use of alternative statistical methods in testing the assumptions in studies to be carried out with IRT models will contribute to the field.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Ethics

The ethics application for the study was made on 20/06/2021 and the research was carried out with the approval of Social Sciences University of Ankara Ethics Commission dated 06/08/2021 and numbered 14020.

References

Antoniou, F., Alkhadim, G., Mouzaki, A., & Simos, P. (2022). A Psychometric Analysis of Raven's Colored Progressive Matrices: Evaluating Guessing and Carelessness Using the 4PL Item Response Theory Model. *Journal of Intelligence 10*(1),6 MDPI AG. https://doi.org/10.3390/jintelligence10010006

Ary, D., Jacobs, L. C., Sorensen, C., & Razavieh, A. (2010). *Introduction to research in education (Eight).* Belmont: wadsworth Cengage Learning.

Aybek, E. C. (2023). The relation of item difficulty between Classical Test Theory and Item Response Theory: Computerized Adaptive Test perspective. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, *14*(2), 118–127. https://doi.org/10.21031/epod.1209284

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker..

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Boduroğlu, E., & Anil, D. (2023). Examining group differences in mathematics achievement: Explanatory item response model application. *OPUS Toplum Araştırmaları Dergisi*, *20*(53), 385–395. https://doi.org/10.26466/opusjsr.1226914

Brzezińska, J. (2016). A polytomous item response theory models using R / Politomiczne modele teorii odpowiedzi na pozycje testowe w programie R. *Ekonometria*. https://doi.org/10.15611/ekt.2016.2.04

Chalmers, R., P. (2012). mirt: A multidimensional ıtem response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. doi: https://doi.org/10.18637/jss.v048.i06

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in ıtem response models with principal component analysis on standardized residuals. In educational and psychological measurement (Vol. 70, Issue 5, pp. 717–731). SAGE Publications. https://doi.org/10.1177/0013164410379322

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

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: a generalized linear and nonlinear approach.* New York: Springer.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dogan, N., & Basokcu, T. O. (2010). İstatistik tutum ölçegi için uygulanan faktör analizi ve asamalı kümeleme analizi sonuçlarının karsilastirilmasi. *Journal of Measurement and Evaluation in Education and Psychology, 1*(2), 65-71.

Doğan, Ö., & Atar, B. (2024). Comparing differential item functioning based on multilevel mixture item response theory, mixture item response theory and manifest groups. *Egitimde ve Psikolojide Olcme ve Değerlendirme Dergisi*, *15*(2), 120–137. https://doi.org/10.21031/epod.1457880

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *The British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Frankfurt am Main: Peter Lang.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23*(1), 138–149. https://doi.org/10.1037/met0000121

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Erkuş, A. (2006). S*ınıf öğretmenleri için ölçme ve değerlendirme: kavramlar ve uygulamalar*. Ekinoks Yayınları, Ankara.

Erkuş, A., Sünbül, Ö., Sünbül, S. Ö., Yormaz, S., & Aşiret, S. (2017). *psikolojide ölçme ve ölçek geliştirme-II ölçme araçlarının psikometrik nitelikleri ve ölçme kuramları.* Pegem Yayınları, Ankara.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their ıtem/person statistics. *Educational and Psychological Measurement, 58*(3). 357–381. https://doi.org/10.1177/0013164498058003001

Gökçen Ayva Yörü, F. (2024). Thematic and metadological analysis of doctoral dissertations on measurement and. *International Journal of Education Technology and Scientific Researches*. https://doi.org/10.35826/ijetsar.721

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika, 15*(3), 259-269.

Hambleton, R. K. (1989). *Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement*. Washington, DC: American Council on Education and Macmillan.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., & Swaminathan, H. (1985). *A look at psychometrics in the Netherlands.*

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions of item response theory. *Item response theory,* 15-31. Springer, Dordrecht.

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

Harwell, M. R., & Janosky, J. E. (1991). An Empirical Study of the Effects of Small Datasets and Varying Prior Variances on Item Parameter Estimation in BILOG. *Applied Psychological Measurement, 15*(3), 279–291. https://doi.org/10.1177/014662169101500308

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining, 2*(1), 20-30. https://doi.org/10.30880/jscdm.2021.02.01.003

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *The Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/jce-18-22

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

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

Kartal, S., & Mor Dirlik, E. (2021). Examining the dimensionality and monotonicity of an attitude dataset based on the item response theory models. *International Journal of Assessment Tools in Education*, *8*(2), 296–309. https://doi.org/10.21449/ijate.728362

Kılıç, A. F., Koyuncu, İ., & Uysal, İ. (2023). Scale development based on item response theory: A systematic review. *International Journal of Psychology and Educational Studies*, 10(1), 209–223. https://doi.org/10.52380/ijpes.2023.10.1.982

Koyuncu, İ., & Kılıç, A. F. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *TED EĞİTİM VE BİLİM*. https://doi.org/10.15390/eb.2019.7665

Linacre JM. (2009). Local independence and residual covariance: a study of olympic figure skating ratings. *Journal of Applied Measurement, 10*(2), 157-69.

Looney, M. A., & Spray, J. A. (1992). Effects of violating local independence on IRT parameter estimation for the binomial trials model. *Research Quarterly For Exercise and Sport*, *63*(4), 356-359.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational And Psychological Measurement, 13*(4), 517-549.

Lord, F. M. (1968). An Analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.*

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin*, *58*(2), 122–131. https://doi.org/10.1037/h0048679

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of mathematical and statistical Psychology, 34*(1), 100-117.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453–477. https://doi.org/10.1177/0265532214527277

Mutluer, C., & Çakan, M. (2023). Comparison of test equating methods based on Classical Test Theory and Item Response Theory. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *36*(3), 866–906. https://doi.org/10.19171/uefad.1325587

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*(1), 41–68. https://doi.org/10.3102/10769986018001041

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal Of Mathematical Psychology, 3*(1), 1-18.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer, New York, NY.

Ree, M. J., & Jensen, H. E. (1980). E*ffects of sample size on linear equating of item characteristic curve parameters*. D. J. Weiss (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology, 5*(1), 27-48.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63–84. https://doi.org/10.1177/0013164404273942

Saatcioglu, F. M., & Sen, S. (2023). The analysis of TIMSS 2015 data with confirmatory mixture item response theory: A multidimensional approach. *International Journal of Testing*, *23*(4), 257–275. https://doi.org/10.1080/15305058.2023.2214648

Selçuk, E., & Demir, E. (2024). Comparison of item response theory ability and item parameters according to classical and Bayesian estimation methods. *International Journal of Assessment Tools in Education, 11*(2), 213–248. https://doi.org/10.21449/ijate.1290831

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(2), 93–128. https://doi.org/10.3102/10769986009002093

Sireci, S. G. (1991, June). *Sample independent item parameters?An investigation of the stability of IRT item parameters estimated from small data sets*. Paper presented at the annual Conference of Northeastern Educational Research Association, New York, NY.

Sözer, E., & Kahraman, N. (2021). Investigation of psychometric properties of likert items with same categories using polytomous item response theory models. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, *12*(2), 129–146. https://doi.org/10.21031/epod.819927

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*(1), 58–75. https://doi.org/10.1111/j.1745-3984.2000.tb01076.x

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. SAS Institute Inc.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Suhr, D. (2006). Exploratory or Confirmatory Factor Analysis? *Statistics and Data Analysis*, 1–17.

Swaminathan, H., & Gifford, J. A. (1983). *Esimation of parameters in the three-parameter latent trait model*. In D. J. Weiss (Ed.), New horizons in testing, (pp. 9-30). New York: Academic Press.

Swineford F. (1956). Technical Manual for Users of Test Analyses, *Statistical Report*, 56-42. Princeton, NJ: Educational Testing Service.

Şahin, M. G., Yildirim, Y., & Boztunc Öztürk, N. (2023). Examining the achievement test development process in the educational studies. *Participatory Educational Research*, *10*(1), 251–274. https://doi.org/10.17275/per.23.14.10.1

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation*, *1*(1). https://doi.org/10.59863/gpml7603

Trabin, T. E., & Weiss, D. J. (1983). *The person response curve: Fit of individuals to item response theory models*. New horizons in testing (pp. 83-108). Academic press.

Tucker, L. R., Humphreys, L. G., & Roznowski, M. A. (1986). C*omparative accuracy of five ındices of dimensionality of binary ıtems*.

van der Linden, W. J. (2010). *Elements of adaptive testing*. C. A. Glas (Ed.). New York, NY: Springer.

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

Watkins, M. W. (2006). Determining parallel analysis criteria. *Journal of Modern Applied Statistical Methods, 5*(2), 344-346.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational And Psychological Measurement, 29*(1), 23-48.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yiğiter, M. S., & Doğan, N. (2023). Comparison of different computerized adaptive testing approaches with shadow test under different test length and ability estimation method conditions. *Egitimde ve Psikolojide Olcme ve Degerlendirme Dergisi*, *14*(4), 396–412. https://doi.org/10.21031/epod.1202599

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(1). https://doi.org/10.1186/s41155-016-0040-x

Ziegler, M., & Hagemann, D. (2015). Testing the Unidimensionality of Items. *European Journal of Psychological Assessment, 31(4)*, 231–237. https://doi.org/10.1027/1015-5759/a000309

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432.