






Scale Development Based on Item Response Theory: A Systematic Review

Abdullah Faruk KILIÇ¹, İlhan KOYUNCU², İbrahim UYSAL³

¹ Faculty of Education, Adıyaman University, Adıyaman, Türkiye  0000-0003-3129-1763

² Faculty of Education, Adıyaman University, Adıyaman, Türkiye  0000-0002-0009-5279

³ Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Türkiye  0000-0002-6767-0362

ARTICLE INFO

Article History

Received 01.07.2022

Received in revised form
08.09.2022

Accepted 04.10.2022

Article Type: Research
Article

ABSTRACT

Classical test theory (CTT) and item response theory (IRT) are two fundamental approaches used in scale development research. Although CTT is the preferred methodology in scale development research in the Republic of Turkey, the IRT methodology has started to gain traction in recent years. In this study, we used the systematic review methodology to review scale development research in Turkey that used the IRT procedures. Results revealed that most of the studies were conducted in 2017. Most of the studies have a sufficient sample size. Exploratory factor analysis was used more than confirmatory factor analysis. Some studies failed to choose the proper IRT model, report and interpret item parameters, check IRT assumptions, and report and interpret information functions. We recommend some practical considerations to researchers in terms of IRT assumptions.

Keywords:

Classical test theory, item response theory, scale development, systematic review

1. Introduction

While measuring in education and psychology, latent structures that cannot be directly observed (academic success, anxiety, etc.) are mainly studied. Making inferences about latent structures is possible by answering the items that are indicators of the relevant construct. Tests or scales are created for this purpose by combining items that measure the same construct. Two fundamental theories are used in the development of tests and scales. These are Classical Test Theory (CTT) and Item Response Theory (IRT). Both approaches assume that the measured construct is continuous. In CTT, the extent to which an individual has the measured characteristic can be found by averaging an infinite number of measurements. The observed score, which shows the performance of the individual in the test, consists of the differences between the individuals (true scores) and the error arising from the interaction between the respondents and the items. Since scale-independent measurements could not be obtained in CTT, the concept of reliability became prominent. Reliability shows the level of errorless in each measure, ranging from 0 to 1 (Bock & Gibbons, 2021).

CTT assumes that true scores in one measurement tool are independent of errors; errors in one measurement tool are independent of true scores in another measurement tool; and errors in one measurement tool are independent of errors in another measurement tool (de Ayala, 2009). Since the above-mentioned weak assumptions are likely to be met in the CTT, it may be preferred when developing a scale or test in the literature. However, it is not possible to examine the model-data fit in CTT. In addition, CTT requires normal distribution for polytomous items. In practice, it is not always possible to ensure normality (Hambleton et al., 2011).

¹Corresponding author's address: Adıyaman University, Faculty of Education, Block A., Room 42, Center/Adıyaman /Türkiye
e-mail: abdullahfarukkilic@gmail.com

Citation: 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://dx.doi.org/10.52380/ijpes.2023.10.1.982>

Although scales are developed based on CTT or IRT, similar processes are used in the scale development process. These are scale conceptualization, scale construction, scale tryout, item analysis, and scale revision (Cohen & Swerdlik, 2018). While the observed scores of the people are the focus in the CTT, the items are the focus in the IRT (de Ayala, 2009). In other words, while the scores that individuals get from each item in the CTT are summed, inferences are made about the traits of the individuals based on correct or incorrect answers specific to the items in the IRT (Baker, 2001). This situation differentiates the item selection process in CTT and IRT. Ensuring the invariance of item statistics is considered an essential factor in the emergence of IRT. In addition, the advantages of IRT in computerized adaptive test applications, coping with missing data, and test equating have enabled the theory to create a wide area of use (Hambleton et al., 2011).

The traits of individuals answering different test forms can be compared through IRT (Crocker & Algina, 2008). Unlike CTT, when IRT is used, person and item characteristics become independent of one another. Moreover, person and item statistics can be placed on the same scale through IRT (de Ayala, 2009). Although IRT has solid mathematical models, it also requires strong assumptions (Hambleton et al., 2011). One of the IRT assumptions is that the measured construct is unidimensional, while another is local independence (Crocker & Algina, 2008). Unidimensionality means that the participants have only one dominant trait that explains test performance. Unidimensionality can be studied through techniques such as Stout's Test of Essential Unidimensionality (DIMTEST), residual analysis in unidimensional models, eigenvalues, or parallel analysis. Local independence means that the participants' answers to one item do not create clues to other items or affect their answers in a good or bad way. Yen's (1984) Q3 statistics are frequently used to analyze local independence (DeMars, 2010; Hambleton & Swaminathan, 1985). In addition to those mentioned above, model-data fit can also be considered an assumption. In studies based on IRT, it is seen that the authors mainly examine the item fit but do not focus enough on the model-data fit. Many techniques can be used to ensure model-data fit. These are X^2 and G^2 analysis, $-2\log$ likelihood value, AIC, and BIC values (DeMars, 2010). It should be noted that the inferences made will be incorrect if the assumptions mentioned above are not met.

IRT provides researchers with the probability of answering the item correctly for each individual at each ability level. The item characteristic curve is drawn through these probability and ability levels. In the item characteristic curve, the y-axis shows the probability of answering the item correctly, and the x-axis shows the ability. The item characteristic curve makes inferences about item difficulty and discrimination. The slope at the midpoint of the item characteristic curve indicates item discrimination, while the point where the item is on the ability scale indicates item difficulty. The Joint maximum likelihood (JML), conditional maximum likelihood (CML), and marginal maximum likelihood (MML) methods are used to estimate item parameters. Although the item and ability parameters are estimated together in the JML method, there may be cases where the estimations are inconsistent. Although the CML method solves this problem by determining the likelihood function of the answer vector over the total score, it is only applicable for 1PLM. On the other hand, MML adds a likelihood function to the ability distribution, mostly assuming a normal distribution, thus separating the item parameter estimation from the ability parameter estimation. Bayesian methods with marginal distribution can be used in cases where the above methods based on maximum likelihood do not yield logical values or cannot make parameter estimations. On the other hand, Markov chain Monte Carlo (MCMC) methods use Bayesian methods in more advanced IRT models (Meijer & Tendeiro, 2018).

The answers given to the items in the IRT are used to estimate the abilities of the individuals by using the item characteristic curves. Abilities theoretically range from minus infinity to plus infinity but are generally in the range of -3 to 3. Typically, the average of the abilities is 0, the standard deviation is 1, and an increase in the value means that the ability has increased (Baker, 2001; Hambleton et al., 2011). There are three approaches used in the estimation of ability parameters. These are maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP). It determines the point where the ML function is maximum and defines it as theta. The Newton-Raphson method is mainly used to determine the maximum point. A common variant of ML is MAP, which is a Bayesian approach. An additional curve representing the population is used in the theta estimation with MAP. Another variant, EAP, uses the weighted average value in the function instead of the Bayes modified curve (Thompson, 2009).

IRT models were developed for items scored in two categories as 1/0 (e.g., multiple-choice items in academic achievement tests) and items scored in more than two categories (e.g., open-ended items, Likert-type scales) (Reckase, 2009). Although normal ogive models were predominant in the early studies in IRT, they were later

replaced by logistic models. Logistic models scale using a constant ($D=1.7$) to eliminate scale differences (Hambleton et al., 2011).

Three logistic models are frequently used for dichotomous items. These are one-, two-, and three-parameter logistic models that use a cumulative logistic distribution function based on an item characteristic curve (Crocker & Algina, 2008). While the two-parameter model (Birnbaum, 1957, 1958a, 1958b) includes item discrimination (a parameter) and difficulty parameters (b parameter), item discrimination is considered equal in one-parameter models (Rasch, 1960). It can be stated that the b parameters vary between -3 and 3 , and the increase in the b parameter indicates that the item is becoming more difficult. The aforementioned parameter of an item that can be qualified as good varies between 0.40 and 2.50 (Hambleton et al., 2011). In the two-parameter model (2PLM), the ability level with a correct answer probability of 0.50 is defined as item difficulty. The equation of the three-parameter models (Lord & Novick, 1968) includes, in addition to 2PLM, another parameter related to the probability of giving the correct answer by chance. The third parameter (c) shows the probability of people with a very low ability to correctly answer the item. The reason for considering the third parameter is that individuals who know one or more of the options are wrong can give the correct answer by choosing from the remaining options after eliminating them (Bock & Gibbons, 2021). The three-parameter model (3PLM) is preferred in situations with a probability of giving a correct answer by chance, such as multiple-choice and true-false items. If the probability of giving a correct answer by chance is negligible, then it would be more appropriate to prefer 2PLM. This is because complex models can produce more inaccurate results. 2PLM is often used in short answer items scored in two categories. If the item discrimination change is negligible in the 2PLM, then a one-parameter model (1PLM) should be used. Although 1PLM can be used in narrow psychological testing, it can be problematic in educational testing. A special case of 1PLM is the Rasch model. In the Rasch model, the discrimination parameter is assumed to be 1 for all items. It is known that the Rasch model was developed on longitudinal intelligence tests (Crocker & Algina, 2008; Hambleton et al., 2011).

Polytomous IRT models establish item-person interaction for items containing various categories. The categories mentioned may belong to a Likert-type scale [such as strongly disagree (1), disagree (2), neutral (3), agree (4), and strongly agree (5)] or maybe scoring categories of a composition. The critical point here is that the scoring should not be in the form of true/false ($1/0$). Since items scored polytomous provide more psychometric information about structures, it can be understood why they have become popular in recent years (Ostini & Nering, 2011). However, polytomous IRT models are more complex and have been created with different perspectives besides providing more psychometric information. In addition, it has not yet been developed as much as dichotomous IRT models (Hambleton et al., 2011). In particular, the category threshold points associated with the parameters reveal differences between the models. Although they have the same numbers and types of parameters, the models differ from each other (Ostini & Nering, 2011). Therefore, the Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), the Graded Response Model (Samejima, 1969, 1972), the Rasch Rating Scale Model (Andrich, 1978, 1988), and the Nominal Response Model (Bock, 1972), there are many polytomous IRT models.

Threshold points are critical in understanding polytomous IRT models. When a person's ability exceeds the threshold, they are scored in a different category. If it does not pass the next threshold point, this failure is included in the probability of responding to the following category (Ostini & Nering, 2011). Although there are many models for polytomous IRT models, the two most commonly used models are GRM and GPCM. Samejima's (1969) GRM is an extension of 2PLM for ordinal polytomous items. In GRM, each category has the same discrimination value, so the model is called homogeneous. In GRM, the b parameter is the ability level at which the probability of responding to a particular category is 0.50 or higher. There are as many threshold parameters as one less number of categories in the GRM. This is because the threshold parameter, where the probability of responding to the categories is 1.00 , cannot be estimated (Hambleton et al., 2011).

Similarly, the GPCM model is an extension of the 2PLM, and unlike the GRM model, it directly calculates the probability of responding to a category through the ability function. In GRM, however, indirect estimation is used in two stages. In GPCM, as in GRM, the difficulty parameters are not ordered from smallest to largest. PCM is an extension of 1PLM, and the discrimination of all items is fixed in the model. The model can be used if the model-data fit is achieved (Hambleton et al., 2011).

In scale development studies, factor analytical techniques are used when deciding whether to remove items from the scale or keep them on the scale. It can be seen that exploratory factor analysis is mainly used in scale development studies based on CTT (Kaya-Uyanık et al., 2017). However, factor analysis generally has a more suitable structure than IRT (Whittaker & Worthington, 2016). Two-parameter normal ogive models and unidimensional factor analysis models are equivalent to each other in data scores with two categories. In addition, equivalence was obtained in polytomous models, ordinal and non-ordinal categorical data (Takane & de Leeuw, 1987). There is a relationship between factor analysis and IRT; both are important for scale development. In addition, the concept of reliability is classically taught and examined in CTT. When the items are removed from the scale, the change in the reliability coefficient is reviewed. In IRT, on the other hand, item and test information functions can be seen as an alternative to classical reliability and standard error. The standard error of the measurement is used to determine the accuracy of the measurement process for any ability level in IRT. In addition, each item's effect on the test's overall accuracy is evaluated. Thus, by combining the items in different tests in different ways, an idea about reliability can be obtained without applying the test (DeMars, 2010; Hambleton & Swaminathan, 1985). For reliability in IRT, there are also concepts of marginal reliability and test reliability. While the marginal reliability coefficient is based on expected information functions, IRT test reliability uses item response functions (Andersson & Xin, 2018).

It can be stated that reliability, item, and ability parameters can be estimated accurately when the sample size is sufficient in IRT (DeMars, 2010). When the literature is examined, it is seen that large-scale development studies are carried out with IRT. However, it is thought that it is essential to investigate the scale development studies in IRT since IRT has strong assumptions and researchers do not have as much knowledge of IRT as CTT. Whittaker and Worthington (2016) state in their research that IRT may not be used due to the lack of quantitative knowledge of counselors, difficulty in accessing IRT software, or insufficient understanding that confirmatory factor analysis and IRT require unique models and estimation techniques. As the mentioned research supports the argument presented in the current study, the current study gains importance in terms of providing information about how to scale development with IRT should be.

In a study examining scale development studies based on CTT in Turkey (Kaya-Uyanık et al., 2017), the articles were reviewed using 17 criteria. The results showed that only one of the 57 articles met all the requirements. Kaya-Uyanık et al. (2017) stated that exploratory and confirmatory factor analysis are generally not used together and that the number of items in the item pool is usually not 2-3 times the number of items in the final scale. Information about the scoring of the scales is often not included. Çüm and Koç (2013) examined 29 large-scale development studies. Çüm and Koç (2013) discovered that on average, 67% of the scale development steps (such as the purpose of scale development, drawing the theoretical framework, locating behavioral indicators, deciding on the scale development technique, determining the response category, pre-pilot and pilot studies, performing statistical analysis, and conducting validity and reliability studies) were followed, and 40% of the scale development principles (such as obtaining permission) were followed. As can be understood from the studies above, it is seen that scale development does not comply with the criteria presented in the literature, even in a theory where assumptions such as CTT are not challenging to provide. More problems will likely arise in scale development for a theory with solid assumptions such as IRT. In addition, although many evaluations of scale development studies based on CTT in Turkey, an assessment based on IRT has not been found in the literature. It is thought that the reasons mentioned add originality to this research.

In this context, we mainly sought answers to the following research questions: Did studies use IRT methodology in scale development research in the Republic of Turkey?

- Conduct exploratory and confirmatory factor analysis?
- Check for irt assumptions?
- Estimate and interpret item parameters correctly?
- Examine item and test information functions?

2. Methodology

In this study, we systematically reviewed scale development research in Turkey that used the IRT methodology. Researchers select and compile studies that contribute to a particular area based on certain selection criteria in a systematic review. They extract the related meta-data, choose the appropriate research

method, analyze and report the emerging meta-data, and provide evidence for making clear inferences about what is known and unknown (Denyer & Tranfield, 2009). While in systematic reviews generally, the aim is to analyze statistics obtained from reviewed studies with the meta-analysis technique (Petticrew & Roberts, 2008). However, sometimes narrative summaries can be used due to heterogeneity in the studies (Hemingway & Brereton, 2009). In this study, we used narrative summaries to report item response theory in scale development research.

2.1. Article Selection Procedure

To find IRT based scale development research conducted in Turkey, we searched Google Scholar and other national databases with the following keywords: "örtük özellikler teorisi" AND "ölçek geliştirme" (n=20), "örtük özellikler kuramı" AND "ölçek geliştirme" (n=23), "madde tepki kuramı" AND "ölçek geliştirme" (n=125), "madde yanıt kuramı" AND "ölçek geliştirme" (n=8), "örtük özellikler teorisi" AND "ölçek uyarlama" (n=4), "örtük özellikler kuramı" AND "ölçek uyarlama" (n=2), "madde tepki kuramı" AND "ölçek uyarlama" (n=42), "madde yanıt kuramı" AND "ölçek uyarlama" (n=2). We used three inclusion criteria: i) published articles in Turkey, ii) scale developed studies, and iii) scales developed with CTT but re-analyzed with IRT. We used three exclusion criteria: i) unpublished studies; ii) thesis or dissertation publications; and iii) comparison studies (e.g., item and ability estimation comparisons). So, we reached the total of 226 document. But we excluded 207 documents via exclusion criteria. In addition, some of the search results were duplicated papers. As of June 27, 2022, we examined 19 scale development studies published in Turkey and used IRT procedures (see Appendix 1). Seven of these studies used IRT to re-analyze scales developed with CTT (Aydın et al., 2017; Demirtaşlı et al., 2016; İlhan & Kınay, 2018; Karaca, 2017, 2019; Tatar & Ayhan, 2022; Toraman et al., 2022). The distribution of the reviewed studies by years is given in Figure 1.

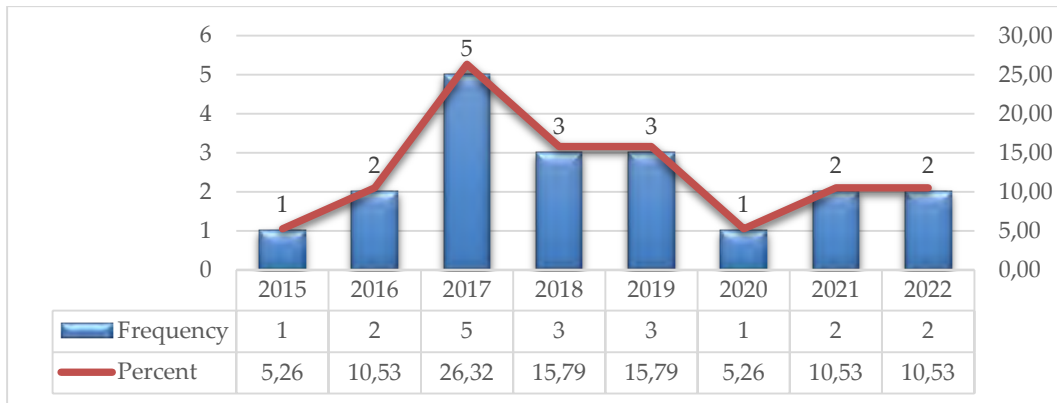


Figure 1. Distribution of Studies by Years

The first study that used IRT for scale development was published in 2015. Approximately 85% of the studies were carried out in 2017 and after (Figure 1). Three reviewed studies used Multilog, four used the *mirt* package in R software, four used Parscale, two used IRTPro, one used Xcalibre, one used FACETS, and one used the *eRm* package to obtain parameter estimates. Three of the reviewed studies did not report the software of choice.

2.2. Data Analysis

We used various metrics to categorize and evaluate selected studies as follows: sample size, EFA and CFA reporting and evaluation, dimensionality and local independence reporting and evaluation, IRT model type, reporting and evaluation of model-and item-fit statistics, estimation method for item parameters, reporting and evaluation of item parameters, reporting of marginal and CTT reliability coefficients, reporting and evaluation of item and test information functions, the software used, and other article-specific contents. We used Microsoft Excel software to describe and visualize some of these criteria. Also, article-specific content was reported and evaluated individually.

2.3. Reliability and Validity of the Study

To standardize the article examination process, we developed an "article evaluation form." First, we determined which properties should be examined in the articles. We used scale development procedures (Downing, 2006) to determine properties. Second, we created an article evaluation form. We received opinions

from two educational measurement and evaluation experts on the prepared form. In line with their opinion, we have made some modifications. Third, we select four articles randomly, and two of us evaluate that articles. We had precisely the same results evaluating those articles. Then, the remaining 15 articles were evaluated by two of us. In the evaluation process, we always keep in touch to ensure consistency.

Internal validity is related to the extent to which the research results correspond to reality (Merriam & Tisdell, 2016). We used an article evaluation form to examine articles, and our results can be said to meet the actual situations. So, we can say that the internal validity is sufficient. External validity is related to the generalizability of the results to other studies (Merriam & Tisdell, 2016). In terms of external validity, we generalize to all Turkey samples because we examine all of the published articles in Turkey. Our sample is the population, in reality.

2.4. Ethical

In this study, we followed all the rules of the “Higher Education Institutions Scientific Research and Publication Ethics Directive.”

Ethical Review Board Name: Adiyaman University Ethics Committee.

Date of Decision on Ethics Evaluation: April 26, 2021

Ethics Assessment Document Issue Number: 97

3. Findings

The findings are presented in the same order as the research questions posited in the introduction.

3.1. Examination of Studies in Terms of Exploratory and Confirmatory Factor Analysis

The sample size for each study is reported in Figure 2. Sample sizes of the reviewed studies ranged from 166 to 2223 (Mean = 779, Median = 519, SD = 596).

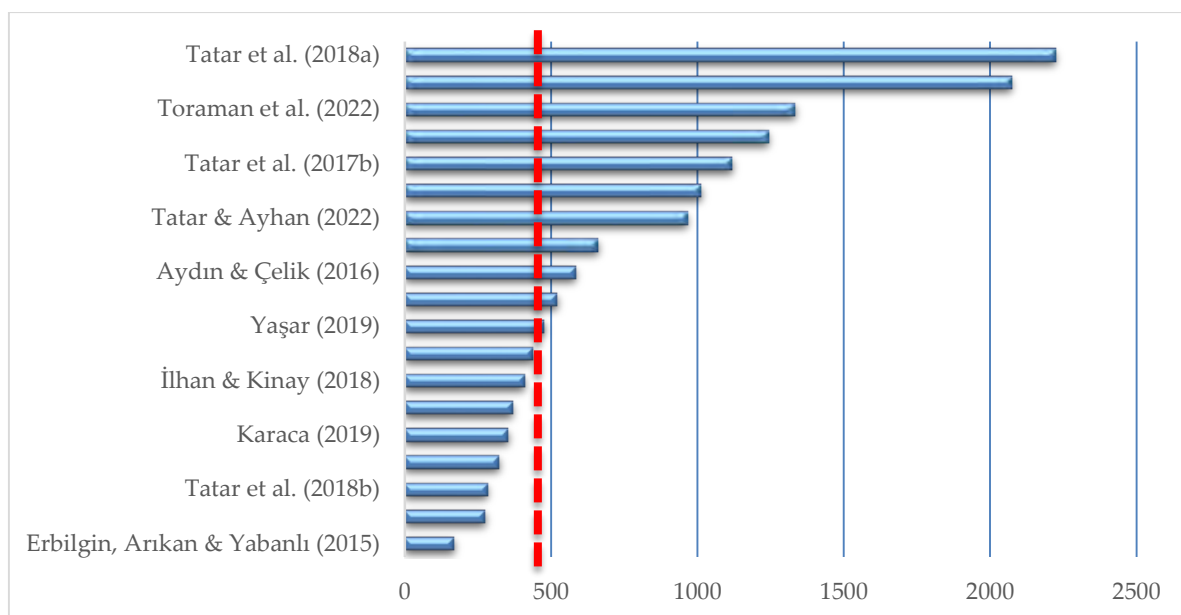


Figure 2. Sample Sizes of the Studies

Out of 19 studies, ten of them had a sample size of 500 or larger (see Figure 2). EFA was performed in 18 (94.73%) and CFA was performed in 12 (63.15%) of them.

Figure 3 presents the distribution of dimensional structure and the associated Total Explained Variance (TEV) statistics for the reviewed studies. Thirteen of the reviewed studies concluded that the factor structure was unidimensional (average TEV: 43.41%), whereas the remaining studies concluded that the factor structure was 2-dimensional (two articles, average TEV: 41.82%), 3-dimensional (two articles, average TEV: 59.25%), and 5-dimensional (one article, TEV: 44.42%). The total explained variance was not reported for the article, which concluded with a 4-dimensional factor structure.

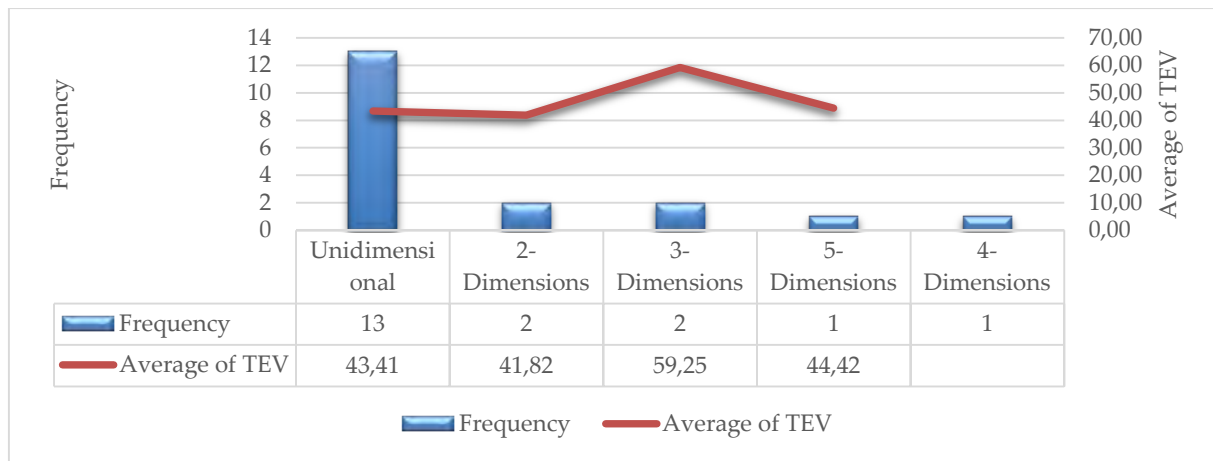


Figure 3. The Number of Dimensions and the Average Explained Variance of the Studies

3.2. Examination of the Studies in Terms of IRT Assumptions

Table 1 summarizes whether reviewed studies reported or checked IRT assumptions. Out of 14 studies, the unidimensionality assumption was checked in eight (42.10%), and the local independence assumption was checked in only six (36.84%). However, two out of 13 studies that did not check for local independence stated that local independence was also satisfied since unidimensionality was held. Eight (42.11%) of the reviewed studies examined model-data fit, whereas only 26.32% examined item-fit statistics. Six of the reviewed studies (31.58%) used GRM, nine of them (47.37%) used 2PLM, one of them (5.26%) used the Rasch model, one of them (5.26%) used the PCM model, one of them (5.26%) used the GPCM model, and one of them did not report any IRT model. The studies that used the 2PLM model had polytomous items.

Table 1. IRT Assumption Check and IRT Models Used in the Reviewed Studies

| Reporting Status | Assumptions of IRT | | | | Model-Data Fit | | Item Fit | | Utilized Models | f | % |
|------------------|--------------------|-------|--------------------|-------|----------------|-------|----------|-------|-----------------|---|-------|
| | Uni-dimensionality | | Local Independence | | f | % | f | % | | | |
| | f | % | f | % | | | | | | | |
| Yes | 8 | 42.11 | 6 | 31.58 | 8 | 42.11 | 5 | 26.32 | 2PLM | 9 | 47.37 |
| No | 11 | 57.89 | 13 | 68.42 | 11 | 57.89 | 14 | 73.68 | GPCM | 1 | 5.26 |
| | | | | | | | | | PCM | 1 | 5.26 |
| | | | | | | | | | GRM | 6 | 31.58 |
| | | | | | | | | | Rasch | 1 | 5.26 |
| | | | | | | | | | - | 1 | 5.26 |

Note: 2PLM = 2 Parameter Logistic Model, GPCM = Generalized Partial Credit Model, PCM = Partial Credit Model, GRM = Graded Response Model, - = not reported

3.3. Examination of the Studies in terms of Item Parameters

Table 2 summarizes whether reviewed studies reported parameter estimation methods and interpreted item parameters (notably, the discrimination or a parameter). Out of 19 studies, one (5.26%) reported the item parameter estimation method, and fifteen (78.95%) reported item parameters (three of them reported only a parameter) except for one of those reported item parameters that is also interpreted as a parameter.

Table 2. Reporting and Interpretation of Parameters

| Reporting Status | Parameter Estimation Method | | Item Parameters | | Interpretation of a parameter | |
|------------------|-----------------------------|-------|-----------------|-------|-------------------------------|-------|
| | f | % | f | % | f | % |
| Yes | 1 | 5.26 | 15 | 78.95 | 14 | 73.68 |
| No | 18 | 94.73 | 4 | 21.05 | 5 | 26.32 |

3.4. Examination of the Studies in terms of Item and Test Information Functions

Table 3 summarizes whether reviewed studies reported item and test information functions, the targeted ability level, test trimming based on IRT parameters (removing ill-functioning items), and reliability coefficients. Out of 19 reviewed studies, eight (42.11%) of them reported item information functions (though two did not provide any plots), nine (47.37%) of them reported the test information function, nine (47.37%) of them reported the target ability level for which the scale was most beneficial, eight (42.11%) of them reported marginal reliability coefficients obtained from the IRT procedures (Min = 0.85, Max = 0.97, Average = 0.92, Median = 0.93), and 12 (85.71%) of them reported CTT reliability coefficients (Min = 0.81, Max = 0.97, Average = 0.91, Median = 0.93).

Table 3. Information Functions, Item Selection, and Reliability

| Reporting Status | Item-Information Function | | Test-Information Function | | Appropriate Ability Levels | | Drop Out Items with IRT | | Marginal Reliability | | CTT Reliability | |
|------------------|---------------------------|-------|---------------------------|-------|----------------------------|-------|-------------------------|-------|----------------------|-------|-----------------|-------|
| | f | % | f | % | f | % | f | % | f | % | f | % |
| Yes | 8 | 42.11 | 9 | 47.37 | 9 | 47.37 | 8 | 42.11 | 9 | 47.37 | 16 | 84.21 |
| No | 11 | 57.89 | 10 | 52.63 | 10 | 52.63 | 11 | 57.89 | 10 | 52.63 | 3 | 15.79 |

4. Conclusion and Discussion

This study aims to review scale development research that used the IRT methodology in the Republic of Turkey. Using Google Scholar and national search engines, we accessed 19 studies. Studies are reviewed, summarized, and evaluated mainly based on their sample size, assumption check, IRT model type, reported fit statistics, and reliability coefficients.

The average sample size for the 19 reviewed studies was 779 (median = 519, SD = 596). However, in almost half of them (47.37%), the sample size was smaller than 500. Although these sample sizes are suitable for EFA and CFA (Comrey, 1988; Gorsuch, 1974), it is recommended to conduct an IRT analysis with a sample size of at least 500 (de Ayala, 2009; DeMars, 2010; Zanon et al., 2016). When the sample size is small, parameter estimates will be inaccurate. Moreover, ensuring the invariance of item and ability parameters will become difficult, which is stated to be an advantage of IRT over CTT.

Most (94.73%) of the reviewed studies conducted EFA, and twelve (63.15%) conducted CFA. Factor analysis is a frequently used method to examine the dimensional structure of the scales (Brown, 2015) and the unidimensionality assumption for the IRT analysis. Nonetheless, in some of the reviewed studies, factor analysis was performed after estimating IRT parameters (Tatar et al., 2018; Tatar, Namlı, et al., 2017; Tatar, Özüdoğru, et al., 2017). Since factor analysis is used to check the unidimensionality assumption for the IRT analysis, it should be performed before obtaining IRT item parameters. Failure to validate the unidimensional factor structure and ignoring it affect the item parameter estimates (unless multidimensional IRT is used). For example, if the inter-factor correlation is relatively low for a scale with a two-factorial structure, the item discriminations obtained from CTT and IRT will be underestimated. An item will load more on one factor but less on the other. In this case, an item with high discrimination may accidentally be excluded from the scale. As a matter of fact, in about 32% of the reviewed studies, the unidimensional factor structure was not validated. Unidimensionality and local independence are two fundamental assumptions of the unidimensional IRT (de Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991). However, only about 68% of the reviewed studies sought evidence for unidimensionality, and only about 30% checked for local independence. When these two assumptions are not met, item parameter estimates and examinees' scores will be biased (DeMars, 2010).

Some reviewed studies stated that since the unidimensionality assumption was met, the local independence assumption need not be checked (Karaca, 2017, 2019), but this practice is frowned upon (DeMars, 2010). DeMars (2010, p. 49) emphasized that the local independence and unidimensionality assumptions should be tested separately. Two items that may disrupt local independence (e.g., similar wording) can pass the unidimensionality test. Thus, unidimensionality and local independence assumptions should be tested separately. The Q3 statistic proposed by Yen (1984) can be used to test the local independence assumption,

and EFA or CFA can be used to test the unidimensionality assumption. One may consider a multidimensional IRT model when the unidimensionality assumption is violated.

Most (74%) of the reviewed studies did not check item fit statistics. For polytomous IRT models, $S-\chi^2$ (Orlando & Thissen, 2000), G^2 (McKinley & Mills, 1985), Bock's χ^2 (Bock, 1972), or Yen's Q statistic (Yen, 1981) can be used to check item fit. In their report, LaHuis et al. (2011) reported that $S-\chi^2$ (Orlando & Thissen, 2000) and Stone's χ^2 (Stone, 2000) gave less erroneous results in the graded response model. Accordingly, the correct item fit statistics should be selected according to the IRT model. Nearly half of the reviewed studies had polytomous items (Aydın et al., 2017; Aydın & Çelik, 2016; İlhan & Kınay, 2018; Tatar, 2020; Tatar et al., 2018; Tatar, Namlı, et al., 2017; Tatar, Saltukoglu, et al., 2017; Tatar, Özüdoğru, et al., 2017; Tatar & Ayhan, 2022), however, they used 2PLM or Rasch models. The IRT model should be selected based on the number of response categories (and whether those response categories are ordinal). Rasch, 1PLM, 2PLM, 3LM, or 4 PLM can be used for items with binary responses; RSM, GRM, PCM, or GPCM can be used for items with ordered polytomous responses, and NRM can be used for items with un-ordered polytomous responses. Also, multidimensional IRT models can be used with a multidimensional factor structure. The most appropriate IRT model for items with the same categories and structure could be decided by looking at the model-data fit. For example, for items with ordered polytomous responses, one of the RSM, GRM, PCM, or GPCM can be used (de Ayala, 2009). Among the reviewed studies, only Yaşar and Aybek (2019) compared GRM and GPCM and decided on the GRM based on superior model-data fit.

After choosing the appropriate IRT model, item parameters should be estimated. The item parameter estimation method was not reported in most of the reviewed studies (94.73%). Although Marginal Maximum Likelihood is frequently used in item parameter estimation, Joint Maximum Likelihood, Conditional Maximum Likelihood, and Bayesian Estimation methods are also used (Hambleton & Swaminathan, 1985). Along with the estimation method, item parameters should also be reported to guide researchers in removing ill-functioning items. The removal of items should be based on low discrimination and their location on the target ability level. At specific locations along the ability continuum, items may be clustered (removal of some may be justified) or sparse (adding some may be needed). Besides, items exhibiting Differential Item Functioning (DIF) can be modified or removed from the scale. Some of the reviewed studies removed items from the scale based on problematic IRT parameters (Demirtaşlı et al., 2016; Tatar et al., 2018; Tatar, Namlı, et al., 2017; Tatar, Saltukoglu, et al., 2017; Yaşar, 2019). In some of them, low-discriminating items were removed (Tatar et al., 2018; Tatar, Namlı, et al., 2017), and only Aybek and Gülleroğlu (2021) additionally considered the removal of DIF-exhibiting items. None of the reviewed studies with polytomous items reported threshold values.

Choosing items based on item information function and weaving them along the target ability level is an advantage of using IRT. Leveraging the results of IRT analysis maximizes a scale's validity beyond what could be accomplished via CTT analysis. About half (42.11%) of the reviewed studies reported item information functions, and about 53% reported test information functions. The item information function expresses how much information an item gives as a function of ability levels. For a binary item, the amount of information the item will provide is determined by the discrimination parameter (a). The ability level at which it will provide more information is determined by the difficulty parameter (b) (Hambleton et al., 1993). DeMars (2010) argued that item information functions obtained from polytomous items are not directly related to a parameter or category parameter (difficulties, steps, or thresholds). The test information function is the sum of item information functions (Hambleton et al., 1993). The test information function provides information on which location on the target ability continuum would produce more reliable test scores. In about half of the reviewed studies, it was not reported.

Besides, reporting standard errors along with item and test information functions is essential. When standard errors are reported, it may not be necessary to report the marginal reliability coefficient (DeMars, 2010). Moreover, per the Standards for Educational and Psychological Testing (AERA et al., 2014), reporting standard errors for different ability levels is recommended. Marginal reliability was not reported in nearly half (52.63%) of the reviewed studies. The majority of the reviewed studies (84.21%) reported CTT-based reliability estimates (CTT reliability > 0.70 for all).

5. Recommendations

To sum up, in this study, we reviewed IRT-based scale development research carried out in Türkiye. Although the IRT analysis processes in some of these studies were appropriate, some had essential shortcomings. For scholars embarking on scale development research using IRT methodology, we made several remarks as follows:

1. Ensuring a sufficient sample size: For IRT analysis, a sample size of at least 500 examinees is recommended (de Ayala, 2009; DeMars, 2010; Zanon et al., 2016). However, this may vary depending on the number of parameters and ability distributions in the model. Accordingly, DeMars's (2010, p. 34-37) recommendations regarding the sample size for different IRT models can be examined.
2. The unidimensionality assumption: NOHARM (Fraser & McDonald, 1988), DIMTEST (Nandakumar & Stout, 1993; Stout, 1987), DETECT (Zhang & Stout, 1999), or factor analysis methods (parallel analysis (Horn, 1965), scree plot (Cattell, 1966), MAP analysis (Velicer, 1976), can be used to test unidimensionality. Results from multiple methods should be compared and reported whenever possible.
3. Local independence assumption: Yen's Q3 statistics (Yen, 1984) can be used to check local independence. The *sirt* package (Robitzsch, 2017) in the R software (R Core Team, 2020) produces Q3 statistics when requested.
4. Selection of the appropriate IRT model: With binary items, Rasch, 1PLM, 2PLM, 3LM, or 4PLM can be used. One of the RSM, GRM, PCM, or GPCM can be used for ordinary polytomous items. For un-ordered polytomous items, NRM can be used. The corresponding multidimensional IRT model can be used for multidimensional structures. Along with the theory, model-data fit statistics should guide the model choice.
5. Examination of model and item data fit: Item fit statistics should be evaluated and reported along with the model-data fit statistics. DeMars (2010) argues that item fit is generally examined instead of model-data fit in IRT applications. For item fit, $S-\chi^2$ (Orlando & Thissen, 2000), G2 (McKinley & Mills, 1985), Bock's χ^2 (Bock, 1972), or Yen's Q statistic (Yen, 1981) can be used. In addition to these statistics, other methods in the *mirt* R package (Chalmers, 2012) can be used for this purpose.
6. Item and ability parameter estimation reporting and interpretation: Although Marginal Maximum Likelihood is frequently used in item parameter estimation, Joint Maximum Likelihood, Conditional Maximum Likelihood, and Bayesian Estimation methods are also widely used (Hambleton & Swaminathan, 1985). For ability parameters, one of the maximum likelihood, expected a-posterior (EAP) or maximum-a-posterior (MAP) estimation methods, could be used (DeMars, 2010). However, reporting the method and software used in the analysis is important because different methods or software may produce different results. On the other hand, item parameters should be re-estimated and reported with the final item set (after item trimming). Baker (2001) suggests the following classifications for the a parameter: if $a > 1.70$, discrimination is very high; if $1.35 < a < 1.69$, it is high; if $0.65 < a < 1.34$, it is moderate; if $0.35 < a < 0.64$, it is low; if $0.01 < a < 0.34$, it is very low. Item parameters for polytomous models should be reported accordingly.
7. Reporting and interpretation of item and test information functions: Plots for an item and test information functions should be examined along with standard errors. The ability at which test information is maximized should be compared to the target ability level. Items should be woven or equally spaced along with the target ability level.

6. References

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Andersson, B., & Xin, T. (2018). Large sample confidence intervals for Item Response Theory reliability coefficients. *Educational and Psychological Measurement*, 78(1), 32-45. <https://doi.org/10.1177/0013164417713570>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. <https://doi.org/10.1007/BF02293814>

- Andrich, D. (1988). *Rasch models for measurement*. Sage.
- Aydın, S., & Çelik, D. (2016). Preservice elementary mathematics teachers' beliefs about preparedness to teach mathematics: Scale adaptation and validation study. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 10(2), 469–469. <https://doi.org/10.17522/balikesirnef.280052>
- Aydın, S., Selçuk, G., Çakmak, A., & İlğan, A. (2017). Validity and reliability study of Turkish form of Epistemological Beliefs Scale. *Bartın University Journal of Faculty of Education*, 6(3), 1166–1188. <https://doi.org/10.14686/buefad.327723>
- Baker, F. (2001). *The basics of item response theory* (2nd ed.). <https://eric.ed.gov/?id=ED458219>
- Birnbaum, A. (1957). *Efficient design and use of mental ability for various decision-making problems* (Report No. 58-16). USAF School of Aviation Medicine.
- Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability* (Report No. 17). USAF School of Aviation Medicine.
- Birnbaum, A. (1958b). *On the estimation of mental ability* (Report No. 15). USAF School of Aviation Medicine.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51. <https://doi.org/10.1007/BF02291411>
- Bock, R. D., & Gibbons, R. D. (2021). *Item response theory*. Wiley.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). McGraw-Hill Education.
- Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56(5), 754–761. <https://doi.org/10.1037/0022-006X.56.5.754>
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press. <https://doi.org/10.1073/pnas.0703993104>
- Çüm, S., & Koç, N. (2013). The review of scale development and adaptation studies which have been published in psychology and education journals in Turkey. *Journal of Educational Sciences & Practices*, 12(24), 115–135.
- DeMars, C. (2010). *Item response theory*. Oxford University.
- Demirtaşlı, R. N., Yalçın, S., & Ayan, C. (2016). The development of IRT based attitude scale towards educational measurement course. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 133–144.
- Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. Buchanan & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 671–689). Sage.
- Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Lawrence Erlbaum Associates.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23(2), 267–269. https://doi.org/10.1207/s15327906mbr2302_9
- Gorsuch, R. L. (1974). *Factor analysis*. W. B. Saunders.

- Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, 30(2), 143–155. <https://doi.org/10.1111/j.1745-3984.1993.tb01071.x>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2011). IRT models for the analysis of polytomously scored data. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 21-42). Routledge.
- Hemingway, P., & Brereton, N. (2009). *What is a systematic review?* (2nd ed.). Hayward Medical Communications.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- İlhan, M., & Kinay, İ. (2018). The Turkish adaptation of teacher social stress scale-student related: The validity and reliability study based on the Classical Test Theory and Item Response Theory. In A. B. Bostancı & S. Koçak (Eds.), *2nd International Education Research and Teacher Education Congress* (pp. 599-608). ERTE.
- Karaca, E. (2017). The comparison of the reliability and validity of the scale of the problem perception of teaching profession based on the Classical Test Theory and Item Response Theory. *Academic Sight International Refereed Online Journal*, 64, 260–272.
- Karaca, E. (2019). Examining test and item scores of the commitment scale to the union via graded response model based on Item Response Theory. *Anadolu University Journal of Social Sciences (AUJSS)*, 10(1), 1–12. <https://doi.org/10.18037/ausbd.550234>
- Kaya-Uyanık, G., Güler, N., Taşdelen-Teker, G., & Demir, S. (2017). Investigation of scale development studies conducted in educational sciences published in Turkey by many-faceted Rasch model. *Journal of Measurement and Evaluation in Education and Psychology*, 8(2), 183-199. <https://doi.org/10.21031/epod.291367>
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9(1), 49–57. <https://doi.org/10.1177/014662168500900105>
- Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth & D. J. Hugh (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 413-433). Wiley. <https://doi.org/10.1002/9781118489772.ch15>
- Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation*. Jossey-Bass.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- 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/10.2307/1165182>
- 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>
- Ostini, R., & Nering, M. L. (2011). New perspectives and applications. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 3-20). Routledge.
- Petticrew, M., & Roberts, H. (2008). Systematic reviews – do they ‘work’ in informing decision-making around health inequalities? *Health Economics, Policy and Law*, 3(2), 197–211. <https://doi.org/10.1017/S1744133108004453>

- R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. <https://www.r-project.org/>
- Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.
- Robitzsch, A. (2017). *sirt: Supplementary item response theory models* [Computer software]. <https://cran.r-project.org/package=sirt>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34, 1–97. <https://doi.org/10.1007/BF03372160>
- Samejima, F. (1972). A general model for free-response data. *Psychometrika*, 37, 1-68.
- 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>
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589–617. <https://doi.org/10.1007/BF02294821>
- Takane, Y., & de Leeuw, J. (1987). On the relationship between Item Response Theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408. <https://doi.org/10.1007/BF02294363>
- Tatar, A., Namlı, M., Özüdoğru, M. T., Uysal, A. R., Yeşilkanat, G., Bahadır, E., Kalender, B., & Aydın, S. (2017). Development of a Mobbing Scale and examination of its psychometric properties. *Journal of Behavior at Work*, 2(2), 79–88. <https://doi.org/10.25203/idd.328590>
- Tatar, A., Özüdoğru, T., Uysal, R., & Uygur, G. (2017). Development of an Urban Life Satisfaction Scale and examination of its psychometric properties. *Reviwed Journal of Urban Culture and Management*, 10(4), 413–431.
- Tatar, A., Saltukoglu, G., & Ozmen, E. (2017). Development of a self-report stress scale using item response theory I: Item selection, formation of factor structure and examination of its psychometric properties. *Arch Neuropsychiatry*, 55, 161–170.
- Tatar, A., Tok, S., & Zekioğlu, A. (2018). A Stress Scale for Sports: Item selection, validity and reliability study. *Spormetre the Journal of Physical Education and Sport Sciences*, 16(3), 91–100.
- Thompson, N. A. (2009). *Ability estimation with Item Response Theory*. Assessment Systems Corporation.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327.
- Whittaker, T. A., & Worthington, R. L. (2016). Item Response Theory in scale development research: A critical analysis. *The Counseling Psychologist*, 44(2), 216–225. <https://doi.org/10.1177/0011000015626273>
- Yaşar, M. (2019). Development of a "Perceived Stress Scale" based on Classical Test Theory and graded response model. *International Journal of Assessment Tools in Education*, 6(3), 521–537. <https://doi.org/10.21449/ijate.626053>
- Yaşar, M., & Aybek, E. C. (2019). A Resilience Scale development for university students: Validity and reliability study based on Item Response Theory. *Elementary Education Online*, 18(4), 1687–1699. <https://doi.org/10.17051/ilkonline.2019.635031>
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245–262. <https://doi.org/10.1177/014662168100500212>
- 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. <https://doi.org/10.1177/014662168400800201>

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of Item Response Theory to psychological test development. *Psicologia: Reflexao e Critica*, 29, 18. <https://doi.org/10.1186/s41155-016-0040-x>

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64(2), 213–249. <https://doi.org/10.1007/BF02294536>

Appendix 1. List of Reviewed Articles

Aybek, E. C., & Gülleroğlu, H. D. (2021). Attitudes toward pirated content: A scale development study based on graded response model. *Eurasian Journal of Educational Research*, 21(91), 127–144.

Aydın, S., & Çelik, D. (2016). Preservice elementary mathematics teachers' beliefs about preparedness to teach mathematics: Scale adaptation and validation study. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 10(2), 469–469. <https://doi.org/10.17522/balikesirnef.280052>

Aydın, S., Selçuk, G., Çakmak, A., & İlğan, A. (2017). Validity and reliability study of Turkish form of Epistemological Beliefs Scale. *Bartın University Journal of Faculty of Education*, 6(3), 1166–1188. <https://doi.org/10.14686/buefad.327723>

Demirtaşlı, R. N., Yalçın, S., & Ayan, C. (2016). The development of IRT based attitude scale towards educational measurement course. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 133–144.

Erbilgin, E., Arıkan, S., & Yabanlı, H. (2015). Assessing line graph comprehension and construction skills. *Ahi Evran University Journal of Kırşehir Education Faculty*, 16(2), 43–61.

İlhan, M., & Kınay, İ. (2018). *The Turkish adaptation of teacher Social Stress Scale-Student Related: The validity and reliability study based on the Classical Test Theory and Item Response Theory*. In A. B. Bostancı & S. Koçak (Eds.), 2nd International Education Research and Teacher Education Congress (pp. 599-608). ERTE.

Karaca, E. (2019). Examining test and item scores of the commitment scale to the union via graded response model based on Item Response Theory. *Anadolu University Journal of Social Sciences (AUJSS)*, 10(1), 1–12. <https://doi.org/10.18037/ausbd.550234>

Tatar, A. (2020). Development of a Job Stress Scale-20 by using Item Response Theory: Study of reliability and validation. *Anatolian Journal of Psychiatry*, 21(Special Issue 1), 14–22. <https://doi.org/10.5455/apd.77173>

Tatar, A., & Ayhan, D. (2022). Evaluation of the Hospital Anxiety and Depression Scale in non-clinic Turkish sample using Item Response Theory analysis. *Cyprus Turkish Journal of Psychiatry and Psychology*, 4(1), 84–93. <https://doi.org/10.35365/ctjpp.22.1.09>

Tatar, A., Namlı, M., Özüdoğru, M. T., Uysal, A. R., Yeşilkanat, G., Bahadır, E., Kalender, B., & Aydın, S. (2017). Development of a Mobbing Scale and examination of its psychometric properties. *Journal of Behavior at Work*, 2(2), 79–88. <https://doi.org/10.25203/idd.328590>

Tatar, A., Özüdoğru, T., Uysal, R., & Uygur, G. (2017). Development of a Urban Life Satisfaction Scale and examination of its psychometric properties. *Reviwed Journal of Urban Culture and Management*, 10(4), 413–431.

Tatar, A., Saltukoğlu, G., Alioğlu, S., Çimen, S., Güven, H., & Ay, Ç. E. (2017). Measuring alexithymia via trait approach I: A Alexithymia Scale item selection and formation of factor structure. *Archives of Neuropsychiatry*, 54, 216–224. <https://doi.org/10.5152/npa.2017.12769>

Tatar, A., Saltukoglu, G., & Ozmen, E. (2017). Development of a self-report Stress Scale using Item Response Theory-I: Item selection, formation of factor structure and examination of its psychometric properties. *Arch Neuropsychiatry*, 55, 161–170.

Tatar, A., Tok, S., & Zekioğlu, A. (2018). A Stress Scale for Sports: Item selection, validity and reliability study. *Spormetre the Journal of Physical Education and Sport Sciences*, 16(3), 91–100.

- Toraman, Ç., Karadağ, E., & Polat, M. (2022). Validity and reliability evidence for the Scale of Distance Education Satisfaction of medical students based on Item Response Theory (IRT). *BMC Medical Education*, 22, 94. <https://doi.org/10.1186/s12909-022-03153-9>
- Tunç, E. B., & Şenel, S. (2021). Development of Test-Taking Strategies Scale: High school and undergraduate form. *International Journal of Contemporary Educational Research*, 8(4), 116–129. <https://doi.org/10.33200/ijcer.888368>
- Yaşar, M. (2019). Development of a “Perceived Stress Scale” based on Classical Test Theory and graded response model. *International Journal of Assessment Tools in Education*, 6(3), 521–537. <https://doi.org/10.21449/ijate.626053>
- Yaşar, M., & Aybek, E. C. (2019). A Resilience Scale development for university students: Validity and reliability study based on Item Response Theory. *Elementary Education Online*, 18(4), 1687–1699. <https://doi.org/10.17051/ilkonline.2019.635031>

Appendix 2. Article Evaluation Form

| Check List | Yes | No |
|--|-----|----|
| Exploratory Factor Analysis | | |
| Confirmatory Factor Analysis | | |
| Unidimensionality IRT Assumption Check | | |
| Local Independence IRT Assumption Check | | |
| Model-data Fit Check | | |
| Item Fit Checked? | | |
| Item Parameter Estimation Method Report | | |
| Item Parameters Report | | |
| A-parameter Examination | | |
| Item Information Function Examination | | |
| Test Information Function Examination | | |
| Targeted Ability Level Examination | | |
| Item Removal According to the IRT Parameters | | |
| Marginal Reliability Report | | |
| CTT Reliability Report | | |
| Ability Parameter Estimation Method Report | | |

| Questionnaire | Result |
|---|--------|
| Used software for IRT analysis: | |
| How many respondents are in the sample? | |
| How many dimensions are in the scale? | |
| How much is the total explained variance? | |
| What is the marginal reliability coefficient value, if reported? | |
| What is the Classical Test Theory reliability coefficient value, if reported? | |
| What is the IRT model used? | |