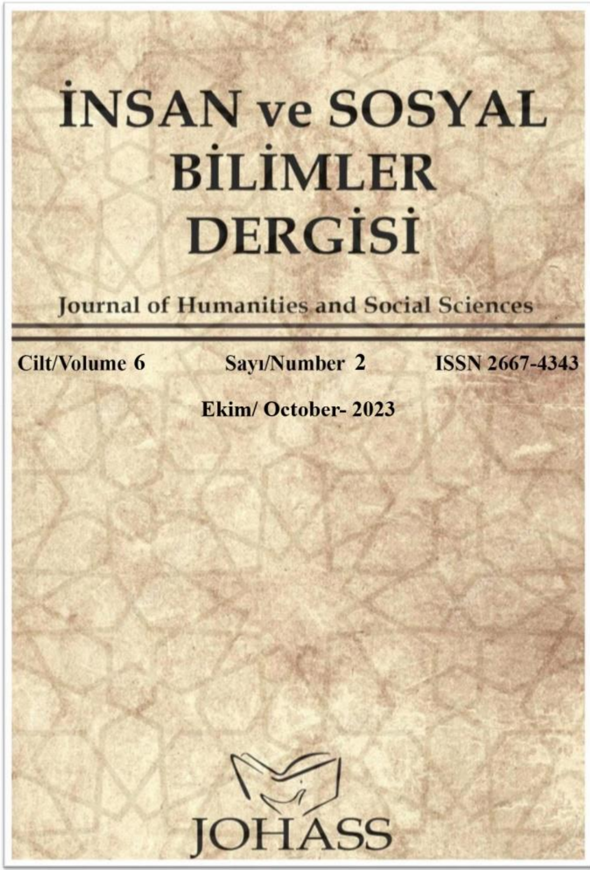
JOURNAL OF HUMAN AND SOCIAL SCIENCES (JOHASS)



https://dergipark.org.tr/tr/pub/johass

A Review of Measurement Tools Developed and Adapted Based on the Rasch Model

Emine Burcu TUNÇ[1]
Marmara University, Atatürk Faculty of Education, Measurement and Evaluation Department
Asst. Prof
burcupehlivantunc@gmail.com
Orcid ID: 0000-0002-8225-9299

Article Type: Research Article
Received: 30.09.2023
Revision received: 19.10.2023
Accepted: 25.10.2023
Published online: 27.10.2023

Citation: Tunç, E. B. (2023). A review of measurement tools developed and adapted based on the Rasch model. *Journal of Human and Social Sciences, 6*(2), 249-275.

A Review of Measurement Tools Developed and Adapted Based on the Rasch Model

Emine Burcu TUNÇ[1]

Marmara University, Atatürk Faculty of Education, Measurement and Evaluation Department

Abstract

It is often observed that the Rasch model is frequently used in determining the psychometric properties of measurement tools because the Rasch model has many advantages in the development and adaptation of measurement instruments. The aim of this study is to evaluate the theses included in the National Thesis Center, which examine the psychometric properties of measurement tools within the framework of the Rasch model, within the scope of the requirements of the Rasch model. In line with this purpose, the model of the research is a document analysis research within the scope of qualitative research. All theses containing the word Rasch in the thesis name and index were examined, and 24 theses in which the measurement tool was developed and adapted within the scope of the Rasch model were found. In order to examine these measurement tools, a coding list was created and the data was analyzed by applying categorical analysis which is one of the content analysis methods. According to the results obtained, it was revealed that in the majority of theses, information was given about unidimensionality, but in half of the theses, no information was given about the local independence assumption. There are studies that do not specify which model is used for polytomous items, and it was observed that model comparison was not performed. It was determined that item model fit was generally tested with different approaches in the theses, and item parameters were generally included. It is among the results that the Person separation index related to reliability was not reported in all studies, and sufficient information was not provided in some studies, even though Differential Item Functioning analyses were performed. In light of these results, it is seen that there is no common systematic approach in the development or adaptation of measurement tools within the framework of the Rasch model in the studies. Therefore, it is recommended that more detailed studies explaining this systematic approach should be conducted.

Keywords: Rasch model, measurement tool, scale development, scale adaptation

Research Article

Received: 30.09.2023
Revision received:.19.10.2023
Accepted: 25.10.2023
Published online: 27.10.2023

[1] Corresponding author:
Asst. Prof
burcupehlivantunc@gmail.com
Orcid ID: 0000-0002-8225-9299

Introduction

It is of utmost significance to assess the psychometric attributes, including validity and reliability, of the outcomes derived from the employed measurement tools during the measurement and evaluation process. Validity, which is one of the most important psychometric properties, is generally defined as the degree to which a measurement tool can assess the trait to be assessed without confusing it with other traits (Courville, 2004; Ebel & Frisbie, 1991; Murphy & Davidshofer, 2005) and it is meaningless to make any inferences on the results obtained from measurement tools that do not have validity (Hubley & Zumbo, 1996). Reliability is defined as the consistency between the scores of individuals taking two parallel instruments assessing the same characteristics; the consistency between the scores of the same individuals taking the same instrument at different times; the consistency between the scores of the same individuals obtained by dividing an instrument into two equivalent halves; and the internal consistency obtained depending on the covariance of the items in an instrument (Thorndike, 1982).

Different models are used in the evaluation of these psychometric properties of measurement tools and one of them is the Rasch model. In this study, the theses in which the psychometric properties of measurement instruments were studied within the scope of Rasch model were examined. When both national and international literature is examined, it is seen that the Rasch model is frequently used in determining the psychometric properties of measurement tools, because the Rasch model has many advantages in developing and adapting measurement tools. As Öztuna (2008) states the Rasch model has areas of use in different situations. These are the development of a new measurement tool, the evaluation of the psychometric properties of an existing measurement tool, the interpretation of measurement results obtained with ordinal results by converting them into interval scales, and the creation of item pools for computer adaptive tests.

In Classical Test Theory (CTT), item parameters are affected by the ability levels of individuals. When the same items are administered to individuals in different groups, different item parameters can be obtained, and therefore, it is seen that the obtained item parameters are group-dependent. However, as in all models within the scope of Item Response Theory (IRT), in the Rasch model, individuals' ability levels and item parameters are located along a common axis. Individuals' ability levels are estimated autonomously from the items in the measurement tool, and item parameters can be computed without being dependent on the ability levels of individuals within the group (Boone, 2016; DeMars, 2010; Embretson &

Reise, 2000; Engelhard, 2013; Hambleton & Swaminathan, 1985; Price, 2017; Wei *et al*., 2014). Moreover, considering the results obtained at the ranking scale level at the equal interval level in the CTT, the total score is taken and parametric statistics are used, which may lead to biased results (Brinthaupt & Kang, 2014). In the Rasch model, the results at the ordinal scale level are transformed into an equal interval logit scale and these limitations are overcome (Wright & Masters, 1982). In addition, while collecting the scores related to the responses given to the items in the CTT, the intervals between the options are considered equal and analyzed. However, it is known that the intervals between the options are not always equal (Elhan & Atakurt). These disadvantages are considered important in the preference of the Rasch model over the CTT.

The Rasch model was developed by Georg Rasch in the 1960s and started to be used to analyze the psychometric properties of dichotomous measurement instruments. It can be referred to as a 1-parameter logistic model of the IRT by researchers, and there are also researchers in the literature who advocate that it should be considered as a separate model from the IRT. While IRT uses a probabilistic distribution to determine ability levels, Rasch model uses a logistic technique. In addition, while the discrimination and chance parameters are held constant in the Rasch model, these parameters can change in the IRT. While an equation is created according to the data set in order to determine the psychometric properties in IRT, Rasch model requires the data set to fit the model (DeMars, 2010). In the two-category Rasch model, the likelihood of a correct response is represented as a logistic function of the disparity between an individual's ability and the item's difficulty, with both expressed in logit units (log-odds). In other words, it conceptualizes the raw scores obtained as the difference between item difficulty and an individual's ability and is obtained as the ratio of the probability of an individual agreeing with an item to the probability of disagreeing with it. When this probability ratio is transformed using logarithms, values from negative infinity to positive infinity are obtained and these values are called logits (Elhan & Atakurt, 2005; Hagquist *et al*., 2009; Pallant & Tennant, 2007; Tennant & Conaghan, 2007). According to Rasch, when an individual answers an item, there is a mathematical relationship that shows the probability of answering that item correctly. He argued that an individual with a higher level of ability than others should be more likely to answer an item correctly than others; he also argued that if there are two similar items, one of which is more difficult than the other, the easier item for any individual is more likely to be answered correctly (Bond & Fox, 2015).

Georg Rasch argued that there are two main causes that affect probabilities; one is the individual's ability, θ, and the other is the difficulty parameter of the item, β, and β and θ are additive. This means that they are in the same logit unit and range. This value is between -∞ and +∞, but in practice, it is evaluated between +3 and -3 (DeMars, 2010). For multi-category items, which is an extension of the Rasch model, the "Rating Scale Model (RSM)" was developed by David Andrich in 1978, and the "Partial Credit Model (PCM)" by Geofferey Masters in 1982 (Sumintono, 2017). In the RSM, the distance between thresholds is considered the same for all items. The analysis continues by estimating a single threshold for each item and adding other thresholds to this threshold value. The difficulty levels of the steps vary from item to item and the β value shows the average difficulty of a selected item according to the category thresholds. The PCM was developed for situations where partial scoring is important in the case of completing different stages in the analysis process or where the distances between response categories differ from item to item in Likert-type items. One of the important features of the model is that it is possible to score people with moderate θ (Koch & Dodd, 1989). Masters defines β parameters as "step difficulty". The reason for defining it as step difficulty is that the individual moves on to the next step after successfully completing one step. The item step difficulty parameter is also referred to as the category intersection parameter. As a result, the step difficulty parameter is defined as the amount of difficulty involved in selecting one response category from another response category. In PCM, there is one less step difficulty parameter than the number of item categories. For example, there are three step difficulty parameters for an item with four categories (Garrett, 2009). As in all Rasch models, items are assumed to have equal discrimination. Therefore, there is no item discrimination parameter in the model.

Unidimensionality, local independence and model-data fit are necessary assumptions for a Rasch model (DeMars, 2010). Unidimensionality is the presence of a single latent trait that adequately explains the common variance and the observed variables are a function of only one latent variable (de Ayala, 2009; Embretson & Reise, 2000). Meeting the unidimensionality assumption also indicates that there is no problem with local independence (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton *et al*., 1991; Lord, 1980; Morizot *et al*., 2007). Local independence means that the items are unrelated to each other. Although it is stated that if the unidimensionality assumption is met, the local independence assumption will also be met, it is recommended to examine the local independence assumption (DeMars, 2010). Violation of the local independence assumption may occur when the response to one item affects the other item and the measurement tool is multidimensional. The Q3 statistic, which is expressed as a correlation coefficient for the residual values between items, is a statistic that shows the dependency between item pairs. In order to test the local independence assumption, it is necessary to examine the relationship between all possible item pairs. Although a criterion of .20 is used in the evaluation of Yen's Q3 statistic (Christiensen *et al*., 2017), a criterion of .30 is generally considered (Riazi *et al*., 2014; Røe *et al*., 2014).

After testing the unidimensionality and local independence assumptions of the Rasch model, model-data fit should be tested with chi-square fit statistics. The chi-square fit statistic compares the difference between expected values and observed values between groups called class intervals, which represent different levels of ability along the trait to be measured (Tennant & Conaghan, 2007). The analysis programs used for the Rasch model usually report the fit statistics as two chi-square ratios, which are called the Infit MNSQ and Outfit MNSQ statistics (Wright & Linacre, 1994). The Infit value is sensitive to the individual's responses to items at a similar level of difficulty and provides centralized information. The Outfit value, on the other hand, is more sensitive to the unexpected responses of the individual to items that are more difficult or easier (Eckes, 2009). While Infit is more sensitive to responses to items that are close to the individual's ability level (Boone, 2016), Outfit is more sensitive to unexpected responses to items that are relatively easy or very difficult for individuals (Linacre, 2002). Infit and Outfit take values ranging from 0 to ∞, but the value indicating perfect fit is 1.00 (Eckes, 2009). However, it is difficult to find a perfect fit between the model and the data (Brentari & Golia, 2008). These two values are evaluated together and a value between 0.50 and 1.50 indicates that model-data fit is achieved (Linacre, 2015). Concordance statistics above 1.50 indicate that individuals gave extreme answers contrary to the item, that the answers given to the item were out of the expected or that the item was caused by the fact that the item did not belong to the structure formed by the other items. A concordance statistic of 0.50 and below indicates that the item is too compatible to be true, which means that individuals gave the same answers to the items (Elhan & Atakurt, 2005; Tennant & Conaghan, 2007; Maindal *et al.*, 2009; Mallinson, 2007). Infit and Outfit values can also be standardized to have an expected value of 0 and reported as standardized Infit (ZSTD Infit) and standardized Outfit (ZSTD Outfit) (Wright & Masters, 1982). When the model and data are compatible, the mean of the Z values is expected to be close to 0 and the standard deviation to be close to 1. In the studies, Z values greater than +2 and less than -2 are considered less compatible with the model than expected. Negative Z values indicate less differentiation than expected (all easy questions answered correctly, all difficult questions answered incorrectly and similar situations), while positive values indicate more differentiation than expected (such as more random answer patterns) (Bond & Fox, 2015).

Two reliability estimates can be obtained through the Rasch model: individual reliability and item reliability. Reliability indicates the repeatability of scores or predictions rather than their accuracy. The reliability coefficients obtained reflect the characteristics of the results rather than the measurement tool itself. High individual reliability means that individuals with a high level of ability are more likely to succeed than individuals with a low level of ability. Item reliability is a measure of the extent to which the item difficulty ranking obtained from the current sample can be repeated (Linacre, 2015). As with other reliability coefficients, it is known that the closer it is to 1.00, the higher the reliability. It is used to evaluate the appropriateness of the responses to the overall measurement tool (de Ayala, 2009). As with Cronbach's alpha internal consistency coefficient, it is recommended to take .70 as a criterion for the reliability index obtained from the Rasch model (Walker *et al.*, 2012). Along with reliability estimates, separation values are also estimated for individuals and items. Like reliability coefficients, separation coefficients are an indicator of the repeatability of item and individual parameters. The individual separation coefficient is used to categorize individuals and when this coefficient takes a value less than 2.00, it is interpreted that the test items are not sensitive enough to distinguish between low and high performing individuals and that more items are needed. The item discrimination coefficient is used to verify the hierarchy of items, and when this coefficient is less than 3.00, it means that the sample is not large enough to verify the item hierarchy (Linacre, 2015).

Differential Item Functioning (DIF) is one of the factors affecting model fit in Rasch model. DIF is the matching of individuals according to their abilities in terms of the variable to be measured and then statistically determining that these individuals in different groups have different probabilities of answering the item (Camilli & Shepard, 1994; Clauser & Mazor, 1998; Roever, 2005; Zumbo, 1999). If an item shows DIF, individuals in different groups with similar θ levels will not be equally likely to give a certain response to that item (Embretson & Reise, 2000). In other words, DIF occurs when different individuals with equal θ respond differently to a certain item (Tennant & Conaghan, 2007). There are two types of DIF: uniform and non-uniform DIF. When uniform DIF exists, the difference between the item characteristic curves for the focal and reference group is uniform (Finch & French, 2007; Jodoin & Gierl, 2001; Walker, 2011). Non-uniform DIF occurs when the difference between item characteristic curves is not constant (Walker *et al.*, 2001). As a result of statistical analysis, items are labeled in categories A (insignificant/insignificant DIF), B (moderate DIF) and C (high DIF) (Zieky, 1993).

In the Rasch model, testing the psychometric properties of the measurement tool is completed after the assumptions of unidimensionality and local independence are met, followed by model-data fit, reliability, and DIF analyses as described above. In recent years, there has been an increase in the number of scale development studies in particular, and this has led to low-quality studies. For this reason, studies discussing the psychometric properties of measurement tools are also increasing. Many of these studies examine measurement tools within the scope of the CTC (Acar Güvendir & Özer Özkan, 2015; Şengül Avşar & Barış Pekmezci, 2022; Barış Pekmezci & Ayan, 2020; Çüm & Koç, 2013; Delice & Ergene, 2015; Doğan 2009; Erkuş, 2007; Erol & Eskici, 2022; Fidan, 2021; Gül & Sözbilir, 2015; Güler & Ayan, 2020; Hinkin, 1995; Slavec & Drnovsek, 2012; Soycan & Babacan, 2019; Tavşancıl *et al.*, 2014; Tosun & Taşkesenligil, 2015; Worthington & Whittaker, 2006). In the studies conducted within the scope of IRT (Kılıç *et al.*, 2022), scale development articles were examined and suggestions were made especially on assumptions. There are many studies on why the Rasch model should be used. In this study, the theses in the National Thesis Center, in which only the psychometric properties of measurement tools were examined within the scope of the Rasch model, were evaluated within the scope of the requirements of the Rasch model.

Method

This section includes information on the research model, documents, data collection tool, and data analysis process.

Research Model

In this study, the psychometric properties of the measurement tools were examined within the scope of the requirements of the Rasch model. To this end, the model of the research is a document review study within the scope of qualitative research. Corbin & Strauss (2015) define document review as a research model in which both printed and electronic materials are systematically analyzed to obtain empirical information about a phenomenon. Document analysis aims to reach a synthesis that will reveal certain situations or views by finding and analyzing relevant documents (Bowen, 2009; Maxwell, 1996). O'Leary (2017) also explains document review as a research model that aims to collect, examine, question and analyze various written materials as a source of primary research data. In this study, within the scope of document review, theses containing measurement tools developed and adapted within the scope of the Rasch model were examined within the scope of the requirements of the Rasch model.

Documents

In this study, all the theses in the National Thesis Center Database of the Council of Higher Education that included the term 'Rasch' in their title and index were reviewed, and 24 theses (Appendix 1) in which the measurement tool was developed and adapted within the scope of the Rasch model were identified. In this context, no restriction was made and all theses were examined. Information about these theses is given in Table 1.

Table 1

Distribution of Thesis in Research According to Some Variables

	Year	Thesis	Development / Adaptation	Scope
1	2019	Specialist thesis	Adaptation	Department of Physical Medicine and Rehabilitation
2	2022	Doctoral thesis	Development	Department of Biostatistics
3	2018	Master thesis	Adaptation	Department of Teaching in Nursing
4	2015	Master thesis	Development	Department of Biostatistics
5	2019	Master thesis	Development	Department of Educational Sciences
6	2021	Specialist thesis	Adaptation	Department of Public Health
7	2013	Doctoral thesis	Development	Primary Education Department
8	2019	Master thesis	Development	Physiotherapy and Rehabilitation Program
9	2021	Master thesis	Adaptation	Occupational Therapy Program
10	2023	Doctoral thesis	Adaptation	Physiotherapy and Rehabilitation Program
11	2022	Specialist thesis	Adaptation	Child and Adolescent Mental Health and Diseases
12	2015	Master thesis	Adaptation	Internal Medicine Nursing
13	2020	Specialist thesis	Adaptation	Department of Public Health
14	2018	Master thesis	Development	Department of Physical Education and Sport
15	2019	Master thesis	Adaptation	Department of Nursing
16	2023	Master thesis	Development	Department of Biostatistics
17	2022	Doctoral thesis	Development	Department of Child Health and Diseases Nursing
18	2018	Master thesis	Adaptation	Department of Teaching in Nursing
19	2017	Master thesis	Adaptation	Department of Nursing
20	2022	Master thesis	Adaptation	Department of Speech and Language Therapy
21	2013	Master thesis	Adaptation	Department of Public Health
22	2019	Master thesis	Adaptation	Department of Mathematics and Science Education
23	2017	Master thesis	Adaptation	Department of Teaching in Nursing
24	2020	Master thesis	Adaptation	Department of Nutrition and Dietetics

As seen in Table 1, the theses examined are between 2013 and 2023. Sixteen of the theses are master's theses, four are specialization theses, four are doctoral theses, eight are measurement tool development studies and 16 are adaptation studies. When the fields are examined, it is seen that the measurement tools within the scope of Rasch are developed mostly in the field of health.

Data Collection Instrument

A coding list was developed to examine the measurement tools developed and adapted within the scope of the Rasch model. The coding list that has been developed consists of two main sections. The first section includes preliminary information about the theses (year, thesis type, field, sample size, number of items, number of dimensions, number of response categories, software used). The second section includes information about the requirements of the Rasch model in line with the main purpose of the study (unidimensionality and local independence assumption check, item data fit check, item parameter estimation method and item parameter reporting status, item and test information functions reporting, reliability and DIF analyses testing status). In this section, response categories of yes, no and partially were used for some categories and yes, no and partially for others. After the coding list was created, it was submitted to the opinions of three experts who are academicians in the field of measurement and evaluation. After the necessary arrangements were made, the final version of the form was decided.

Data Analysis

The data obtained within the scope of the research were analyzed by applying categorical analysis, which is one of the content analysis methods. Accordingly, the frequencies of each category were calculated. Tavşancıl & Aslan (2001) express that there are two approaches to following the category system in categorical analysis: theoretical categorization process and applied categorization process. In this study, categories were created based on the theoretical basis of the Rasch model. When the thesis review process started, there were changes in the categories created. Therefore, both deductive and inductive approaches were adopted. The findings were presented in the form of frequency/percentage tables. Two researchers coded seven theses independently of each other for the reliability of the coding on the form. The coding reliability of the data obtained from both coders was determined by the coding reliability formula (Coding reliability = Agreement / (Agreement + Disagreement)) proposed by Miles & Huberman (1994). As a result of the coding, the agreement between the codings was found to be 92%.

Findings

Information on the sample sizes, number of items, number of dimensions, number of categories and the statistical program used in the theses are given in Table 2.

Table 2

Sample Sizes, Number of İtems, Number of Dimensions, Number of Category and Software of the Studies

	Sample size	Number of items	Number of Dimensions	Number of Category	Software
1	179	10	2 Dimensions	5	RUMM 2020
2	308	21	Unidimensional	2	RUMM 2030
3	254	18	3 Dimensions	5	RUMM Version 5.3.
4	300	32	2 Dimensions	5	RUMM 2020
5	102	32	Unidimensional	2	Winsteps
6	110	9	2 Dimensions	5	Winsteps
7	502	16	2 Dimensions	2	-
8	370	44	3 Dimensions	2	RUMM 2020
9	101	25	7 Dimensions	4	-
10	100	10	Unidimensional	2	RUMM 2020
11	298	10	Unidimensional	8	RUMM 2030
12	130	22	Unidimensional	5	Winsteps
13	210	13	2 Dimensions	5	Winsteps
14	722	45	Unidimensional	2	-
15	367	33	2 Dimensions	2	SAS 9.4.
16	668	24	3 Dimensions	2	R
17	390	33	Unidimensional	2	Winsteps
18	296	16	2 Dimensions	5	RUMM Version 5.3.
19	499	39	6 Dimensions	4	RUMM Version 5.3.
20	71	24	4 Dimensions	7	Winsteps
21	150	25	4 Dimensions	5	RUMM 2020
22	250	20	4 Dimensions	3	Facets 3.65.0.
23	504	36	3 Dimensions	5	RUMM Version 5.3.
24	314	27	7 Dimensions	5	Winsteps

As seen in Table 2, the lowest sample size was 71 and the highest sample size was 722. The average sample size for 24 theses was 299.79. The number of items varied between 9 and 45, and the average number of items was 24. Seven of the measurement instruments were unidimensional, seven bi-dimensional, four three-dimensional, three four-dimensional, one six-dimensional and two seven-dimensional. Therefore, it was determined that the measurement tools were multidimensional in the majority of the studies. When the number of categories is analyzed, it is seen that the measurement tools have five-response categories in 10 studies and two-response categories in nine studies. In addition, there are measurement tools with three, four, seven and eight response categories. The programs used were RUMM, Winsteps, SAS, R and Facets, but it is seen that the RUMM program is mostly preferred. Three studies did not provide information on the program used. The results of testing the assumptions of the Rasch model are given in Table 3.

Table 3

Rasch Assumption Check

Reporting Status	Assumptions of Rasch			
	Unidimensionality assumption		Local independence assumption	
	f	%	f	%
Yes	17	%70.83	12	%50
No	7	%29.17	12	%50

As can be seen from Table 3, 17 studies provided information on the unidimensionality assumption. In 13 of these studies, Principal Component Analysis was used to meet the unidimensionality assumption. In two studies, it was stated that unidimensionality was also met since local independence was ensured. In two studies, it was stated that unidimensionality was accepted because the infit and outfit values were in the desired range, and in one study it was stated that the measurement tool had a unidimensional structure because the infit and outfit values were in the range of 0.70 and 1.30, and in the other study because they were in the range of 0.50 and 1.50. In seven studies, there was no information regarding the unidimensionality assumption. As can be remembered from Table 2, 17 of the measurement tools have a multidimensional structure. Therefore, the unidimensionality assumption should be tested separately for each dimension. However, only two of the studies specifically emphasized this information. Information on the variance explained by the items in the measurement tools was found in nine theses. In half of the theses, information on the assumption of local independence was given. The need to examine the relationship between all possible item pairs to check the assumption of local independence was tested with Yen's Q3 statistic. In six of the theses, the criterion of .30 was taken into consideration within the scope of this statistic. The assumption of local independence was interpreted by considering the criterion of .32 in four studies, .40 in one study and .50 in one study. In 12 studies, no information about local independence was given. The results of the Rasch model, item fit and item parameters are presented in Table 4.

Table 4

Utilized Rasch Models, Item Fit and Item Parameter

Utilized Rasch Models			Item Fit			Item Parameter		
	f	%	Reporting Status	f	%	Reporting Status	f	%
Dichotomous	9	%37.5	Yes	23	%95.83	Yes	20	%83.33
Partial Credit Model	6	%25.0	No	1	%4.16	No	4	%16.16

No information	9	%37.5

As seen in Table 4, the Dichotomous Rasch Model was used in nine of the theses and the Partial Credit Rasch Model was used in six of them. As can be recalled from Table 2, the measurement instruments had two response categories in nine of the theses; thus, the Dichotomous Rasch Model was preferred. No comparisons were made with other Rasch models that could be used for multiple response categories in any of the studies. The reason why the Partial Credit Model was used was not included in the studies comparatively. Nine studies did not provide information about the model used. Only one thesis did not provide information on item model fit. In ten theses, Infit values, which provide more central information, and Outfit values, which are more sensitive to unexpected responses, were given for all items in the measurement tool. These two values were evaluated together and it was interpreted that the items with values between 0.50 and 1.50 provided model fit. In five studies, standardized Infit and Outfit values were reported and it was stated that the items fit the model if they were in the range of ±2.5. In nine studies, since the chi-square values were higher than the Bonferroni corrected p value, it was stated that all items in the test fit the model. In five studies, overall goodness-of-fit statistics were given and it was stated that the mean of item fit statistic and individual fit statistic being close to 0.00 and standard deviation being close to 1.00 were the criteria for model-data fit. In only one of the theses examined, information on the estimation method was given and it was stated that the weighted likelihood estimation method was used. In 20 studies, it was determined that *b* values for items and standard errors for *b* values were calculated. In five of the instruments with multiple response categories, the threshold values of the items were given and it was checked whether the step transitions were regular. In one of these studies, it was determined that the threshold values of an item were not ordered and category merging was performed for the related item. Point Biserial values of the items were also included in two studies. Four studies did not include item parameters. Information on Item-Information Function, Test-Information Function and other maps are given in Table 5.

Table 5

Item-Information Function, Test-Information Function, Other Maps

Item-Information Function			**Test-Information Function**			**Other Maps**	
Reporting Status	f	%	Reporting Status	f	%		f

Yes	0	%0	Yes	1	%4.16	Person-Item Threshold Distribution	1
No	17	%70.83	No	23	%95.83	Person-Item Location Distribution	3
Partial	7	%29.17	Partial	0	%0	Person-Item Map	7

Item-Information Function is a mathematical function that describes the relationship between an individual's response to an item and his/her ability, usually logistically. Table 5 shows that seven studies included Item-Information Function for sample items rather than all items. In one study, expected and observed item characteristic curves were included, and the expected and observed probabilities were found to be compatible. Only one of the studies included the Test-Information Function. Two of the theses included Person-Item Threshold Distribution and three included Person-Item Location Distribution. Person-Item map was given in seven theses. The Person-Item map, which is also called Wright Maps, shows the distribution of item difficulties and the distribution of individuals' responses, and the left side of the graph shows the graph of individuals' ability estimates, while the right side shows the distribution of items according to their difficulties. The results related to reliability and Changing Item Function in the theses analyzed are given in Table 6.

Table 6

Reliability and Differential Item Functioning

Reliability					**Differential Item Functioning**		
Reporting Status	f	%		f	Reporting Status	f	%
Yes	22	%91.67	Person seperation index	18	Yes	10	%41.67
No	2	%8.33	Cronbach alfa	10	No	14	%58.33
			Test retest	6			
			KR-20	3			
			Split-half	1			

As seen in Table 6, 22 of the theses tested the reliability of the results obtained from the measurement tools. Two studies did not provide information on reliability. In 18 studies, the Person separation index value used within the scope of the Rasch model was given and the criterion of 0.70 was taken into account while interpreting. In 10 studies, Cronbach's alpha value, one of the reliability estimates based on the CTQ, was reported and in three studies only Cronbach's alpha value was given. Six studies reported test-retest reliability and three studies reported KR-20 internal consistency coefficient. In one of these studies, only KR-20 was reported as a reliability estimation. In one study, split-half reliability estimation was also included. In three of the theses, findings related to item reliability, indicating the extent to which the item difficulty ranking obtained from the current sample can be repeated within the context of the Rasch model, were also included. In addition, in four theses, information on the individual dissociation index used to separate individuals and the item dissociation index used to verify the hierarchy of items were also provided. When the Changing Item Function results were examined, it was found that 10 studies examined whether the items showed DIF or not, but in most of these studies, it was not explained that DIF determination method was used. One study reported that Mantel-Haenszel Chi-square DIF determination method was used, and three studies reported that DIF was determined by ANOVA. None of these studies commented on the size of the DIF and did not go through the item bias process. In 14 studies, DIF for items was not studied.

Discussion and Results

In this study, 24 theses in which the psychometric properties of measurement tools were examined within the scope of the Rasch model were reached and evaluated within the scope of the requirements of the Rasch model. Although all of the theses analyzed were published in the last decade, the majority of them are master's theses and unique to the field of health. However, 16 of them, the majority of studies, are adaptation studies.

When the sample sizes reached in the theses were analyzed, it was found out that the sample size was below 500 in 20 theses. Although there are researchers (de Ayala, 2009; DeMars, 2010) who state that the sample size should be at least 500 in IRT analyses, there are also different opinions on the appropriate sample size for parameter estimation (Hambleton & Swaminathan, 1985). It is stated that the Rasch model requires a smaller sample size than other IRT models and that the minimum sample size for a 20-item test can be 200 people within the scope of the Rasch model, and it was determined that the sample size was below 200 in eight of the theses examined. Unlike the findings of this study, Kılıç *et al*. (2022) state in their study in which they examined articles within the scope of IRT that more than half of the articles reached 500 for the sample size. In 17 theses, which constitute the majority, it was determined that the measurement tools were multidimensional and generally had five response and binary response categories. Although the RUMM program is generally preferred for Rasch analysis, there are also theses where program information is not provided.

When the assumptions of the Rasch model are analyzed, it is seen that most of the theses provide information on unidimensionality. However, there are also studies stating that unidimensionality is also ensured since local independence is ensured. However, it was also observed that there were studies stating that unidimensionality was accepted because the infit and outfit values were within the desired range. Brown (2015) states that factor analysis is the most commonly used method to check the unidimensionality assumption of measurement instruments. Unlike the findings of this study, Kılıç *et al*. (2022) state that the unidimensionality assumption was not met in more than half of the articles. In this study, information on the variance explained by the items in the measurement tools was found in ten theses. Azrilah *et al*. (2013) state that the data may be unidimensional if the percentage of variance explained for the Rasch model is at least 40% and the percentage of variance in the first opposite structure is less than 15%. Therefore, the reported variance explained is considered important. Half of the theses do not provide information on the local independence assumption. The residual correlation matrix was used and the criteria that were addressed differed from each other in all of the theses where information was provided. Although the .30 criterion is generally used, .32, .40 and .50 criteria are also used, and it is interpreted that there may be dependence between item pairs with values above these values. Marais (2009) and Yen (1993) state that if the local independence assumption cannot be met, it may affect the parameter estimates based on individuals and the reliability and validity results of the results obtained from the measurement tool. Kılıç *et al*. (2022) state that only 68% of the studies examined in their study controlled for unidimensionality and 30% controlled for local independence.

Since nine of the theses were instruments with two response categories, the two-category Rasch model was used. Partial Credit Model was preferred for measurement tools with multiple response categories. However, no model comparison was made in any of the studies. There are advantages of using the Partial Credit Model. Krishnan & Idris (2018) ention this point in their study entitled Using the Partial Credit Model to Improve the Quality of an Instrument. However, despite these advantages, a model comparison will provide more detailed information. This finding is similar to Kılıç *et al*. (2022), who explain that model comparison was conducted in only one study. When the model-data fit was analyzed, it was determined that only one thesis did not provide information on item model fit. Although there are different approaches to test item-model fit in studies, Infit and Outfit values are generally interpreted. Bond & Fox (2015) state that fit statistics always take positive values and when the fit statistic values are 1.00, they indicate excellent model-data fit. Furthermore, they express that the fit statistic criterion may change according to the characteristics and purpose of the measurement tool used. However, although the theses examined were in different fields, it was determined that the range of 0.50 and 1.50 was used. Again, unlike the findings of this study, Kılıç *et al*. (2022) state that item fit was not tested in the majority of the studies. At the same time, within the scope of this research, only one thesis provided information about the estimation method. As stated by Hambleton & Swaminathan (1985), Marginal Maximum Likelihood is the most commonly used estimation method, but Joint Maximum Likelihood, Conditional Maximum Likelihood and Bayesian Estimation method are also among the estimation methods used. It is among the results obtained that there is a lack of information about these estimation methods in the theses. In this study, it was revealed that item parameters were given in 20 theses. Sixteen of the theses had multiple response categories, but only five studies gave threshold values and checked whether the step transitions were regular. Point Biserial values of the items were also included in two studies. In parallel with the findings of this study, Kılıç *et al*. (2022) also state that item parameters were given in 79% of the studies.

It was determined that none of the theses examined in this study included all the item information functions, only sample items. In one study, expected and observed item characteristic curves were included and it was determined that the expected and observed probabilities were compatible. Apart from this, it is also among the results that comments were made on the Person-Item map in seven theses. Linacre (2008) stated that these maps, also called Wright Maps, are informative in showing the distribution of item difficulties and individuals' responses. Again, unlike the findings of this study, Kılıç *et al*. (2022) stated that almost half of the studies included item information functions and test information functions.

Nearly all of the theses examined presented results on reliability, but the Person separation index, which should be given within the scope of the Rasch model, was not included in six studies. While two of these studies did not provide any information on reliability, four of them provided reliability estimates based on the CTT. Walker *et al*. (2012) argue that .70 should be taken as a criterion for the reliability index obtained from the Rasch model as in internal consistency coefficients. The criterion of .70 was also taken into consideration in the studies. In addition, in four theses, information was also provided with the individual dissociation index used to separate individuals and the item dissociation index used to verify the hierarchy of items. When the Changing Item Function results were analyzed, 10 studies examined whether the items showed DIF, but in most of these studies, which DIF determination method was used was not explained and no information was given about the DIF size in the studies. It was also found that expert opinion on item bias was not taken. Kılıç *et al*. (2022) also explain that in the articles they examined within the scope of

IRT, Marginal Reliability value was given in almost half of the studies, and the item with DIF was removed from the measurement tool only in one study.

Although this research has some findings, it also has some limitations. In this study, only theses in the National Thesis Center in Turkey were analyzed. Although there are some studies in which measurement tools are scrutinized within the scope of CTT, there are a limited number of studies in which measurement tools are examined within the scope of IRT. Since there is no study that only evaluates within the scope of Rasch model, it is thought that this study will be informative for researchers who will develop measurement tools using Rasch model. For this reason, it is recommended to evaluate the articles in which only the measurement tools related to the Rasch model are examined. In line with the results obtained, it is unraveled that there is no common systematic in terms of developing or adapting measurement tools within the scope of Rasch model. Therefore, it is suggested that more studies explaining this systematic in detail should be conducted.

References

Acar Güvendir, M., & Özer Özkan, Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi, 52*, 23-33. https://doi.org/10.17755/esosder.54872

Ağır, H. (2019). *Fiziksel ve sosyal katılım öznel indeksi (subjective index of physical and social outcome) anketinin Türkçe adaptasyon, geçerlik ve güvenirlik çalışması* [Turkish adaptation, validity and reliability of the subjective index of physical and social outcome]. [Uzmanlık tezi, Kırıkkale Üniversitesi]. Ulusal Tez Merkezi

Akşehirli Seyfeli, M.Y. (2023). *Objektif yapılandırılmış sınav aracının klasik test kuramı, genellenebilirlik kuramı ve madde tepki kuramı ile değerlendirilmesi* [Evaluation of objective structured examination tool with classical testing institution, generalizability theory and item response theory] [Yüksek lisans tezi, Erciyes Üniversitesi] Ulusal Tez Merkezi

Akşehirli, Ö. (2022). *Gebelerde doğum şekli hakkında bilgi düzeylerinin belirlenmesine yönelik test geliştirilmesi* [Development of a test for determining the level of knowledge about the delivery method in pregnant women] [Doktora tezi, Ankara Üniversitesi] Ulusal Tez Merkezi

Al-Deges, W. (2019). *Pelvik taban sağlığı bilgi testi geliştirme, geçerlik ve güvenirliği development* [Validity and reliability of pelvic floor health knowledge quiz] [Yüksek Lisans Tezi, Ankara Yıldırım Beyazıt Üniversitesi] Ulusal Tez Merkezi

Alınca, G. (2018). *Grup çalışmasına yönelik tutum ölçeğinin geçerlik ve güvenirliğinin incelenmesi* [Validity and reliability of the attitude scale for group work] [Yüksek lisans tezi, Ege Üniversitesi] Ulusal Tez Merkezi

Appendix 1. List of Reviewed Articles

Azrilah, A.A., Mohd, S.M., & Azami, Z. (2017). *Asas model pengukuran rasch: pembentukan skala dan struktur pengukuran* (1st ed). Penerbit Universiti Kebangsaan Malaysia.

Barış Pekmezci, F., & Ayan C. (2020). Confusion of scale development: Investigation of self-efficacy scales. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 48*, 130-151. https://doi.org/10.9779/pauefd.529986

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: fundamental measurement in the human sciences (3rd ed.)*. Mahwah, NJ: L. Erlbaum.

Bond, T.G., & Fox, C.M. (2015). *Applying the rasch model: fundamental measurement in the human sciences (3rd ed.)*. Mahwah, NJ: L. Erlbaum.

Boone, W.J. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE-Life Sciences Education*, *15*(4), 1-7.

Bowen, G.A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27-40.

Brentari, E., & Golia, S. (2008). Measuring job satisfaction in the social services sector with the Rasch model. *Journal of Applied Measurement*, *9*(1), 45-56. Retrieved from http://www.unibs.it/sites/default/files/ricerca/allegati/10061.pdf

Brinthaupt, T.M., & Kang, M. (2014). Many-faceted rasch calibration: An example using the self-talk scale. *Assessment*, *21*(2), 241-249.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* SAGE Publications.

Clauser, B.E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

Courville, T.G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics* (Unpublished doctoral dissertation). Texas A&M University, Texas. http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/1064/etdtamu-2004B-EPSY-Courville-2.pdf?sequence=1&isAllowed=y

Çüm, S., & Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi, *Eğitim Bilimleri ve Uygulama*, *12*(24), 115-135. Retrieved from https://www.idealonline.com.tr/IdealOnline/pdfViewer/index.xhtml?uId=5928&ioM=Paper&preview=true&isViewer=true#pagemode=bookmarks

Çiçekçi, H.C. (2019). *Tıp eğitiminde klinik öncesi eğitim dönemi ve klinik eğitim dönemi testlerinin psikometrik özelliklerinin incelenmesi: Ege Üniversitesi Tıp Fakültesi Örneği* [The psychometric properties of the preclinical and clinical phase testing: A sample of the Ege University] [Yüksek lisans tezi, Ege Üniversitesi] Ulusal Tez Merkezi

de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

Delice, A., & Ergene, Ö. (2015). Ölçek geliştirme ve uyarlama çalışmalarının incelenmesi: Matematik eğitimi makaleleri örneği. *Karaelmas Eğitim Bilimleri Dergisi*, *3*(1), 60-75. Retrieved from https://dergipark.org.tr/tr/pub/kebd/issue/67216/1049114

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

Doğan, E. M. (2009). *Türkiye'deki psikolojik çalışmalarda kullanılan testlerin psikometrik özelliklerinin incelenmesi: Kültürel açıdan test uyarlama çalışmalar*. [Yayımlanmamış yüksek lisans tezi], Muğla Üniversitesi Sosyal Bilimler Fakültesi, Muğla.

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice Hall.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.) *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Council of Europe/Language Policy Division.

Elhan, A.H., & Atakurt, Y. (2005). Ölçeklerin değerlendirilmesinde niçin Rasch analizi kullanılmalıdır? *Ankara Üniversitesi Tıp Fakültesi Mecmuası*, *58*(1), 47-50.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists. Quality of Life Research*. Lawrence Erlbaum Associates.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Erkuş, A. (2007). Ölçek geliştirme ve uyarlama çalışmalarında karşılaşılan sorunlar. *Türk Psikoloji Bülteni,* *13*(40), 17-25. https://bursa.psikolog.org.tr/tr/yayinlar/dergiler/1031828/tpb134004.pdf

Erol, R. & Eskici, M. (2022). Analysis of developed/adapted scales for distance education. *Journal of Educational Technology & Online Learning*, *5*(4), 936-951.

Esen, Y. (2013). *Development of a test for assessing teachers mathematical content knowledge for teaching geometric measurement at elementary grade level* [Middle East Technical University] Ulusal Tez Merkezi

Fidan, Ç. (2021). Türkiye'de geliştirilen dindarlık ölçekleri (1989-2015): dindarlık ölçme çalışmaları üzerine değerlendirmeler. *Türk Din Psikolojisi Dergisi*, (4), 101-118.

Finch, W.H., & French F.B. (2007). Detection of crossing differential item functioning: a comprasion of four methods. *Educational and Psychological Measurement*, *67*(4), 565-582.

Garrett, P. (2009). *A Monte Carlo study ınvestigating missing data, differential item functioning and effect size* (Doctoral Dissertation). Georgia State University.

Gül Ş., & Sözbilir, M. (2015). Fen ve matematik eğitimi alanında gerçekleştirilen ölçek geliştirme araştırmalarına yönelik tematik içerik analizi. *Eğitim ve Bilim, 40*(178), 85-102. http://dx.doi.org/10.15390/EB.2015.4070

Güler, G., & Ayan, C. (2020). Review of attitude scales developed in Turkey between 2002-2018 regarding the scale development process. *Journal of Faculty of Educational Sciences, 53*(3), 839-864. https://doi.org/10.30964/auebfd.658488

Hagquist, C., Bruce, M., & Gustavsson, J.P. (2009). Using the Rasch model in nursing research: an introduction and illustrative example. *International journal of nursing studies*, *46*(3), 380-393.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and appliccations*. Springer Science and Business Media, LLC.

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

Hinkin, T. R. (1995). A review of scale development practices in the study in organizations. *Journal of Management*, *21*(5), 967-988. https://doi.org/10.1177/014920639502100509

Hubley, A.M., & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, *123*(3), 207-215.

Irmak, D.E. (2021). *Dikkat eksikliği hiperaktivite bozukluğu (dehb) olan çocuklarda "çocuk aktivite öz değerlendirme ölçeği'nin (cosa)" Türkçe uyarlamasının geçerlilik ve güvenilirliğinin incelenmesi* [Evaluation of the validity and reliability of the Turkish adaptation of the "child activity self-assessment scale (cosa)" in children with attention deficit and hyperactivity disorder] [Yüksek Lisans Tezi, Hacettepe Üniversitesi] Ulusal Tez Merkezi

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14,* 329–349.

Keskin, E. (2023). *İnme sonrası mobiliteyi değerlendiren Abıloco anketinin Türkçe versiyonu, geçerlik ve güvenilirliğinin araştırılması* [Investigation of Turkish version, validity and reliability of Abıloco questionnaire evaluating mobility after stroke] [Doktora Tezi, Hacettepe Üniversitesi] Ulusal Tez Merkezi

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

Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, *2*(4), 335-357.

Koç, G. (2022). *İş ve sosyal uyum ölçeğinin çocuk ve ebeveyn formunun Türkçe geçerlilik ve güvenilirlik çalışması* [Tıpta Uzmanlık Tezi, Sağlık Bilimleri Üniversitesi] Ulusal Tez Merkezi

Koşar, C. (2015). *Hasta aktiflik düzeyi ölçüm aracının (patient activation measure) Türkçe 'ye uyarlanması: geçerlik ve güvenirlik çalışması* [Adaptation of patient activation measure into Turkish: reliability and validity] [Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi] Ulusal Tez Merkezi

Krishnan, S., & Idris, N. (2018). Using partial credit model to improve the quality of an instrument. *International Journal of Evaluation and Research in Education*, 7, 4, 313-316.

Kulak, E. (2020). *Hasta aktiflik düzeyinde klinisyen desteği ölçüm aracının (clinician support for patient activation measure) Türkçe'ye uyarlanması: geçerlik ve güvenirlik çalışması* [Adaptation of clinician support for patient activation measure into Turkish:

Reliability and validity study] [Uzmanlık Tezi, Marmara Üniversitesi] Ulusal Tez Merkezi

Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch measurement transactions, 16*(2), 878.

Linacre, J.M. (2015). *A user's guide to Winsteps® Rasch-model computer programs*. Beaverton, Oregon.

Lord, F.M. (1980). *Aplications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Maindal, H.T., Sokolowski, I., & Vedsted, P. (2009). Translation, adaptation and validation of the American short form Patient Activation Measure (PAM13) in a Danish version. *BMC public health, 9*, 1-9.

Mallinson, T. (2007). Why measurement matters for measuring patient vision outcomes. *Optometry and Vision Science*, *84*(8), 675-682.

Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, *10*(1), 17-29.

Maxwell, J.A. (1996) *Qualitative research design: An interpretive approach*, Thousand Oaks, CA: Sage.

Miles, M., & Huberman, M. A. (1994). *An expanded sourcebook qualitative data analysis.* Sage Publications.

Moral, E. (2020). *Besleme uygulamaları ve yapısı anketi Türkçe geçerlilik güvenilirlik çalışması* [Feeding practices and structure questionnaire Turkish validity and reliability study] [Yüksek lisans tezi, Marmara Üniversitesi] Ulusal Tez Merkezi

Morizot, J., Ainsworth, A.T., & Reise, S.P. (2007). Toward modern psychometrics. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personalty psychology*. The Guilford Press.

Murphy, K.R., & Davidshofer, C.O. (2005). *Psychological testing: principles and applications*. New Jersey: Pearson Education International.

O'Leary, Z. (2017). *The essential guide to doing your research project.* SAGE Publications Inc.

Özalp Ateş, F.S. (2015). *Ölçeklerde yapı geçerliliğinin değerlendirilmesinde faktör analizi ve Rasch analizi yaklaşımları* [Factor analysis and Rasch analysis in the evaluation of construct validity of scales] [Yüksek lisans tezi, Ankara Üniversitesi] Ulusal Tez Merkezi

Özdemir Deniz, P. (2021). *Çocuk ve ergenlerde dısabkıds astım modülünün geçerlilik ve güvenilirliği, astım yönetiminde video eğitiminin yaşam kalitesi üzerine etkisi* [Validity and reliability of the disabkids asthma module in children and adolescents, the effect of video education on quality of life in asthma management] [Uzmanlık Tezi, Aydın Adnan Menderes Üniversitesi] Ulusal Tez Merkezi

Özdeş, N. (2018). *Lise öğrencilerinin fiziksel uygunluk bilgi düzeylerinin incelenmesi* [Physical fitness knowledge levels (PFKL) of high school students] [Yüksek lisans tezi, Çanakkale Onsekiz Mart Üniversitesi] Ulusal Tez Merkezi

Öztuna, D. (2008). *Kas-iskelet sistemi sorunlarının özürlülük değerlendiriminde bilgisayar uyarlamalı test yönteminin uygulanması* [An application of computerized adaptive testing in the evaluation of disability in musculoskeletal disorders] [Doktora Tezi, Ankara Üniversitesi] Ulusal Tez Merkezi

Pallant, J.F., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18.

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

Riazi, A., Aspden, T., & Jones, F. (2014). Stroke Self-efficacy Questionnaire: a Rasch-refined measure of confidence post stroke. *Journal of rehabilitation medicine*, *46*(5), 406-412.

Røe, C., Damsgård, E., Fors, T., & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic musculoskeletal pain. *BMC Musculoskelet Disord, 15*(1), 95. https://dx.doi.org/10.1186/1471-2474-15-95, PubMed 24646065

Roever, C. (2005). That’s not fair! Fairness, bias, and differential item functioning in language testing. *SLS Brownbag*, 9(15), 1-14.

Sakınmaz, E. (2019). *Çocuklar için istismar bilgi ölçeğinin Türkçe’ye uyarlanması, geçerlik ve güvenirliği* [Adaptation, validity and reliability of exclusive information survey for children] [Yüksek lisans tezi, Akdeniz Üniversitesi] Ulusal Tez Merkezi

Slavec, A., & Drnovsek, M. (2012). A perspective on scale development in entrepreneurship research. *Economic and Business Review, 14*(1), 39-62. http://ojs.ebrjournal.net/ojs/index.php/ebr/article/view/69/pdf

Soycan, M., & Babacan, E. (2019). Müziksel işitme, okuma ve yazma ile ilgili geliştirilmiş ölçme araçlarının incelenmesi: içerik analizi çalışması. *Elektronik Sosyal Bilimler Dergisi*, *18*(69), 343-353.

Sumintono, B. (2017) *Rasch Model Measurement as Tools in Assessment for Learning.* In: International Conference on Educational Innovation (ICEI 2017), 14 October 2017, Wyndham Hotel, Surabaya, Indonesia.

Şahin, A. (2022). *İnhalasyon uygulamalarına yönelik ebeveyn bilgi ölçeğinin geliştirilmesi* [Development of parental knowledge scale for inhalation practices] [Doktora Tezi, Atatürk Üniversitesi] Ulusal Tez Merkezi

Şengül Avşar, A., & Barış Pekmezci, F. (2023). Examination of motivation scales: Is the purpose academic promotion or the need to measure psychological constructs? *Psycho-Educational Research Reviews, 11*(3), 774-791. doi: 10.52963/PERR_Biruni_V11.N3.19

Şenol, A. (2018). *Klinik hemşirelik uygulamalarına yönelik öz düzenlemeli öğrenme ölçeğinin geçerlik güvenirliğinin incelenmesi* [The investigation of the validity and reliability of the self-regulated learning scale in clinical nursing practice] [Yüksek lisans tezi, Ege Üniversitesi] Ulusal Tez Merkezi

Tavşancıl, E., & Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri.* Epsilon Yayınları.

Tavşancıl, E., Güler, G., & Ayan, C. (2014). *2002-2012 yılları arasında Türkiye'de geliştirilen bazı tutum ölçeği geliştirme çalışmalarının ölçek geliştirme sureci acısından incelenmesi*. IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi (Uluslararası Katılımlı) 9-13 Haziran, Hacettepe Üniversitesi, Ankara.

Teke, C. (2017). *Pozitif mental sağlık ölçeğinin Türkçe geçerlilik ve güvenirliği* [Reliability and validity of the positive mental health questionnaire in a sample of Spanish University students] [Yüksek lisans tezi, İzmir Kâtip Çelebi Üniversitesi] Ulusal Tez Merkezi

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?. *Arthritis Care & Research*, *57*(8), 1358-1362.

Tetik, E. (2022). *Afazide aile yaşamı etki ölçeğinin Türkçe 'ye uyarlanması, geçerlik ve güvenirlik çalışması* [Adaptatıon of the family aphasia measure of life impact scale into Turkish: validity and reliability study] [Yüksek lisans tezi, Anadolu Üniversitesi] Ulusal Tez Merkezi

Thorndike, R.L. (1982). *Aplied psychometrics*. Houghton Mifflin Company, Boston

Tosun, C., & Taşkesenligil, Y. (2015). The instruments used in science education in Turkey: A descriptive content analysis. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi*

Dergisi, *15*(2), 364- 383. Retrieved from https://dergipark.org.tr/tr/download/article-file/17460

Tunay, Z. Ö. (2013). *Çocuklarda kullanılan 25-maddelik Cardiff görsel yeti ölçeğinin Türkçe sürümünün geçerlilik ve güvenilirliği* [The reliability and validity of Turkish version of 25-item cardiff visual ability questionnaire for children] [Yüksek lisans tezi, Ankara Üniversitesi] Ulusal Tez Merkezi

Tüzüngüç, B. (2019). *Ortaöğretim öğrencilerinin sosyobilimsel muhakeme yeteneklerinin araştırılması* [Investigation of socio-scientific reasoning skills of high school students] [Yüksek lisans tezi, Marmara Üniversitesi] Ulusal Tez Merkezi

Walker, C.M., Beretvas, S.N., & Ackerman, T.A. (2001). An examination of conditioning variables used in computer adaptive testing for DIF. *Applied Measurement in Education, 14,* 3-16.

Walker, E.R., Engelhard, G., & Thompson, N.J. (2012). Using Rasch measurement theory to assess three depression scales among adults with epilepsy. *Seizure*, *21*(6), 437-443. http://dx.doi.org/10.1016/j.seizure.2012.04.009

Wei, S., Liu, X., & Jia, Y., (2014). Using rasch measurement to validate the instrument of students' understanding of models in science (SUMS). *International Journal of Science and Mathematics,12*, 1067.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Wright, B.D., & Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* *8*(3), 370-371. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

Wright, B.D., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

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.

Yılmaz, Ö. (2017). *Yeterlik kazanma ölçeğinin geçerlik ve güvenirliğinin incelenmesi* [İnvestigation of validity and reliability of ascent to competence scale] [Yüksek lisans tezi, Ege Üniversitesi] Ulusal Tez Merkezi

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.