IJETS@R
(ISSN: 2587-0238)

Eminoğlu-Özmercan, E. (2023). Content Analysis of Postgraduate Theses on Differential Item Functioning in Türkiye, *International Journal of Education Technology and Scientific Researches, 8(24), 2365-2376.*

DOI: http://dx.doi.org/10.35826/ijetsar.681

Article Type (Makale Türü): Research Article

CONTENT ANALYSIS OF POSTGRADUATE THESES ON DIFFERENTIAL ITEM FUNCTIONING IN TÜRKİYE

Esra EMİNOĞLU ÖZMERCAN
Instructor Dr., İstanbul University, İstanbul, Türkiye, esemcan@gmail.com
ORCID: 0000-0003-4105-9837

Received: 21.05.2023 *Accepted: 06.09.2023* *Published: 01.10.2023*

ABSTRACT

The purpose of the current study is to determine the general trends in postgraduate theses written on Differential Item Functioning (DIF) in the field of Educational Measurement and Evaluation in Türkiye from 2012 to 2022. To this end, the theses were examined in terms of the type of theses, publication year, which data set was used, according to which variable DIF was examined, which school subjects were used, what the examined school subjects were, which country's data were used, which DIF detection methods were used and the type of data used. Within the study, a total of 66 graduate theses written in Türkiye from 2012 to 2022 were analyzed. As a result of the analysis, it was determined that the theses related to DIF were mostly written in 2019, with master's and doctoral theses peaking in 2019 and 2016 respectively, that PISA data were frequently used, that the studies predominantly focused on the detection of DIF in relation to the gender variable, that mathematics data were commonly used, and after Türkiye, the United States data were used the most, that the Mantel-Haenszel (MH) method was predominantly used for the detection of DIF and that the studies were conducted using two-category data.

Keywords: Measurement and evaluation, Differential item functioning, research trends, content analysis.

INTRODUCTION

Measurements are performed to determine to what extent objects and individuals have certain qualities and characteristics. Therefore, two characteristics related to the measuring tool must be provided to allow the information obtained from different measurement tools to be used in practical situations for studies with different purposes. Validity and reliability are the two most fundamental attributes sought in measurement tools. Reliability refers to the degree to which measurement results are free from random errors (Turgut & Baykul, 2010). Validity is the process of obtaining evidence showing that the test provides accurate results for the evaluation decisions to be made (Messick, 1995). As for validity, expressed as the degree to which the measurement tool serves its purpose (Tekin, 1993; Turgut, 1983; Turgut & Baykul, 2010), all test items prepared are expected to effectively discriminate between individuals (Nunnally & Bernstein, 1994). Validity is not related to measurement results but to inferences made from measurement results. It is noted that when validity studies are not conducted, the inferences drawn from measurement results will be meaningless (Zumbo, 1999).

The validity of a test is influenced by systematic errors. Systematic error is a type of error that varies in quantity and direction from one measurement to another and has an identifiable source (Turgut, 1983). Bias, which refers to systematic error in measurement results against a particular group (Camilli & Shepard, 1994; Osterlind, 1983; Reynolds & Suzuki, 2003), affects all measurement results in the same direction, altering measurement results in favor of or against a specific group (Osterlind, 1983). In this case, bias has a negative impact on validity, and validity is affected by systematic errors. Since any item in the test providing an advantage to one group adversely affects validity, item bias is one of the issues that should be considered in test validity.

Zumbo (1999) defines item bias as the likelihood of one group giving a correct response to an item more than the other group due to the item's characteristic features or test conditions that are not suitable for the test's purpose. DIF is required, but not sufficient, for item bias. Angoff (1993) defines item bias as a type of invalidity that affects one group more than the other group.

When two groups of the same ability level take a test, but the probability of one group answering an item correctly differs from the other group due to characteristics of the test items or test conditions, students' test scores can be biased. In other words, even when individuals have equivalent levels of the measured trait (such as academic achievement or ability), their test scores may be biased due to variables such as culture, gender, school type, socioeconomic status, ethnic background, and others, which are unrelated to the intended construct being measured (Shepard et al., 1981; Zumbo, 1999). Therefore, it is important for psychological measurements or test scores to be unbiased and to measure the same construct for each individual. Therefore, conducting DIF analyses is important in the test development process and for ensuring validity (Walker, 2011). Determination of item bias begins with the determination of whether there is DIF in the item. For an item to be biased, it is necessary for the item to exhibit DIF. If individuals from different groups taking the test, matched for the ability being measured by an item, have different probabilities of success on that item, DIF occurs (Clauser & Mazor, 1998). However, for an item to be considered biased, exhibiting DIF alone is not sufficient. If DIF is detected in

an item, further investigation is needed to determine whether item bias is present or not (Zumbo, 1999). After potential causes of DIF have been identified by experts, a decision is made regarding whether the item is biased or not (Camilli & Shepard, 1994). In comparisons made based on test results, determining whether test items exhibit DIF in relation to the relevant variable is crucial for making more valid comparisons and ensuring less biased measurements. Additionally, this will provide test developers and practitioners with more favorable conditions.

When the literature is reviewed, it is seen that research articles on item bias and DIF are quite common both nationally and internationally. In addition, there are some studies on the trends in postgraduate theses in the field of Measurement and Evaluation in Education. However, no study conducted in Türkiye on differential item functioning (DIF) and item bias in the field of Measurement and Evaluation has been found. In this regard, Berrío et al. (2020) analyzed studies that used simulation data under various conditions to investigate DIF. Analyzing recent studies in the field is important for identifying the trends in topics, approaches, methods, etc., evaluating the development of the field, and providing recommendations and predictions for the future.

It is argued that descriptive content analyses categorizing studies conducted in a specific time frame in a specific field have contributed very little or even not at all to the field. However, studies conducted at intervals of 5 to 10 years can reveal gaps or excessive burdens in the research field (Dinçer, 2018). It is thought that identifying research trends in the subjects of item bias and differential item functioning in the field of measurement and evaluation, which is related to the feedback element of the open and societal system of education and all areas of education, will be beneficial to researchers in this field. This systematic review aims to determine the current state of the measurement and evaluation field in relation to DIF.

The current study aims to examine master's and doctoral thesis studies on Differential Item Functioning published in the Higher Education Council (YÖK) National Thesis Centre (YÖKTEZ) database between 2012 and 2022 from various aspects. The reasons for choosing this thesis from 2012 to 2022 are to determine current trends and to observe changes over the past 10 years. To this end, answers to the following questions were sought:

1. What is the distribution of the theses by type?
2. What is the distribution of the theses by publication year?
3. What data are used for DIF in the theses?
4. What variables are used for DIF in the theses?
5. Which school subjects/fields are examined for DIF in the theses??
6. Which countries are examined in DIF studies?
7. What are the DIF detection methods used in DIF studies?
8. What is the distribution of data types used in DIF studies?

METHOD

Research Model

This study is a systematic review study. Systematic review studies are conducted using three methods: meta-analysis, meta-synthesis, and descriptive content analysis. A systematic review is a structured and comprehensive synthesis of a large number of studies conducted with similar methods by experts in the field to determine the best available research evidence (Karaçam, 2013). Systematic reviews include the processes of defining selection criteria, conducting a search for relevant studies, critical evaluation, data analysis, and synthesis (Dybå & Dingsøyr, 2008). This study employs a descriptive content analysis to examine the theses published in the YÖKTEZ database between 2012 and 2022 according to the specified criteria and to reveal the trend in terms of the examined criteria (Çalık & Sözbilir, 2014).

Population and Sample

The population of this study consists of the theses written on DIF and found in the Higher Education Council (YÖK) National Thesis Centre. The sample of this study consists of 66 graduate theses written on DIF and found in the Higher Education Council (YÖK) National Thesis Centre between 2012 and 2022.

Data Collection and Analysis

In line with the aim of examining the postgraduate theses written on Differential Item Functioning (DIF) in Türkiye, a search was conducted in the Higher Education Council (YÖK) National Thesis Centre database under the "Advanced Search" section in the "Search" tab for each of the following keywords in English and Turkish: "Differential item functioning", "madde yanlı VEYA madde yanlılığı", "madde işlev farklı", "madde işlev farklılığı", "değişen madde fonksiyonu", "item bias" and "test bias". In the "Search Field" section, "All" was selected. The year range was filtered as 2012-2022, and "permitted" theses were searched. As a result of the search, a total of 218 theses were found, including recurrent theses. After removing the recurrent theses, 119 theses were obtained. Out of these theses, 12 were not related to the field of Education and Instruction, 3 were prepared outside Türkiye (abroad) and 18 were not related to DIF, so they were not included in the analysis. Thus, a total of 66 postgraduate theses were included in the analysis. The 66 theses included in the study were tabulated using percentages and frequencies in accordance with the research questions.

A specific analysis form tailored to the study was created by the researcher for examining the theses. The developed form includes areas for descriptive information about the theses (thesis number, year, etc.), thesis title, subject area, methods and techniques used for data analysis, software packages used for data analysis, and so on. After the information about each thesis was added to the analysis form, the data were transferred to a Microsoft Office Excel file. The data were then organized, grouped, and presented in a numerical format.

In the analysis of the data, categorical and frequency analysis, which are types of content analysis, were used. Content analysis is "the objective and systematic classification of the meaning and/or grammar of the message contained in verbal, written and other materials, the conversion of it into numbers and making inferences about it" (Tavşancıl & Aslan, 2001). The current study followed the stages of content analysis, including data preparation, defining the unit of analysis, developing coding schemes and categories, testing the coding scheme, conducting all coding, evaluating coding consistency, defining categories or themes, and reporting the results (Tavşancıl & Aslan, 2001; Yıldırım & Şimşek, 2011).

Before starting the analysis, five theses were randomly selected from the graduate theses, and they were independently coded by a different researcher to assess the reliability of the coding. To ensure consistency among the researchers, five theses were independently examined by both researchers. The obtained data were used to estimate reliability in terms of consistency, following the approach suggested by Miles and Huberman (1994).

$$\text{Reliability} = \frac{Number\ of\ Agreements}{Number\ of\ Agreements + Number\ of\ Disagreements} \times 100$$

The reliability coefficient between the researchers was calculated to be 0.95. This shows that the consistency between the researchers is high (Tavşancıl & Aslan, 2001). In addition, the Krippendorf Alpha coefficient was analyzed. Krippendorf Alpha coefficient is used to determine the agreement between two or more raters. The Krippendorf alpha coefficient was found to be 0.94 and this indicates that the inter-rater agreement is high. The data were analyzed using Microsoft Office Excel and IBM SPSS Statistics 28.0 software programs.

FINDINGS

Findings are presented in the order specified by the research questions.

Data on the distribution of the theses by publication year are presented in Table 1.

Table 1. Distribution of the Theses by Publication Year

Publication Year	f	%
2012	4	6.1%
2013	5	7.6%
2014	4	6.1%
2015	7	10.6%
2016	7	10.6%
2017	4	6.1%
2018	3	4.5%
2019	11	16.7%
2020	6	9.1%
2021	8	12.1%
2022	7	10.6%
Total	66	100%

When the distribution of theses on DIF by publication year is examined, it is seen that the highest number of theses was written in 2019 (f=11) and the smallest number of theses was written in 2018 (f=3).

Table 2. Distribution of Different Types of Theses by Publication Year

Thesis Type	Master's		Doctoral	
Publication Year	f	%	f	%
2012	2	5.1%	2	7.4%
2013	2	5.1%	3	11.1%
2014	2	5.1%	2	7.4%
2015	3	7.7%	4	14.8%
2016	2	5.1%	5	18.5%
2017	1	2.6%	3	11.1%
2018	2	5.1%	1	3.7%
2019	10	25.6%	1	3.7%
2020	3	7.7%	3	11.1%
2021	6	15.4%	2	7.4%
2022	6	15.4%	1	3.7%
Total	39	100.0%	27	100.0%

When the distribution of different types of theses written on DIF by publication year is examined, it is seen the highest number of master's theses was written in 2019 (f=10), followed by 2021 (f=6) and 2022 (f=6) and the smallest number of master's theses were written in 2017 (f=1). The highest number of doctoral theses was written in 2016 (f=5), followed by 2015 (f=4) while the smallest number of doctoral theses was written in 2018 (f=1), 2019 (f=1), and 2022 (f=1).

Table 3. Distribution of the Data Used in the Theses

Data Used	f	%
ABİDE	3	4.4%
Open High School (ALS) Exams	1	1.5%
Public Personnel Selection Exam (KPSS)	1	1.5%
University Placement Exam (LYS)	1	1.5%
Central Exams	1	1.5%
High School Entrance Exam	11	16.2%
Tests developed by the teacher	4	5.9%
Free Boarding and Scholarship Exam (PYBS)	1	1.5%
PISA	24	35.3%
Simulation	11	16.2%
Verbal Reasoning Aptitude Test	1	1.5%
TIMSS	5	7.4%
Attitude Scales	3	4.4%
Intelligence Test	1	1.5%

When the distribution of data used in the theses is examined, it is seen that in studies on DIF, PISA data were used the most (f=24), followed by data from the High School Entrance Exams (f=11) and simulation (f=11).

Table 4. Distribution of the Variables Used in the Theses

Variable	f	%
Mother's education level	1	1.0%
Father's education level	1	1.0%
Gender	35	35.4%
Geographical region	1	1.0%
Experimental weighted scoring	1	1.0%

Language	7	7.1%
Four-variable	1	1.0%
State of having a computer/tablet at home	1	1.0%
Two-variable	7	7.1%
Statistical regional units	1	1.0%
Living in an urban or rural area	1	1.0%
Type of booklet	2	2.0%
Having or not having a chronic disease	1	1.0%
Culture	9	9.1%
Item format	1	1.0%
Item subject area	1	1.0%
Ordering of items by difficulty level	1	1.0%
Graduated department	1	1.0%
School type	6	6.1%
School district	1	1.0%
Adequacy of the student's pocket money	1	1.0%
Disability status	2	2.0%
Amount of the student's weekly pocket money	1	1.0%
State of having had a serious illness	1	1.0%
Grade level	1	1.0%
Socio-economic level	2	2.0%
One-variable	2	2.0%
Expert judgement	1	1.0%
Three-variable	3	3.0%
Country	4	4.0%
Place of residence	1	1.0%

When Table 4 is examined, it is seen that the gender variable (f = 25) is the variable used most when determining whether there is DIF in the items, followed by the culture variable (f=9). Since the presence of Differential Item Functioning (DIF) in items was investigated in relation to more than one variable in some of the analyzed theses, the total number of variables used exceeds the total number of theses.

Table 5. Distribution of the School Subjects/Fields Used in the Theses

Course/Subject Area	**f**	**%**
Religion and Ethics	1	1.3%
Science	19	25.0%
General Aptitude Exam	3	3.9%
English	4	5.3%
Mathematics	21	27.6%
Reading Skills	5	6.6%
Social Studies	2	2.6%
History of Republican Reforms and Kemalism	2	2.3%
Attitude/Intelligence	12	15.8%
Turkish	7	9.2%

In Table 5, it is seen which school subjects/fields were used in determining the presence of DIF in items. As seen here, mathematics is the school subject most used in the investigation of the presence of DIF in items (f=21), followed by science (f=19). Moreover, the fields of attitude and intelligence (self-efficacy scales, attitude questionnaires, and intelligence scales, etc.) are among the popular fields used to investigate whether there is DIF in items (f=12).

Table 6. Distribution of the Countries Analysed in the Theses

Country	f	%	Country	f	%
America	11	11.7%	Kazakhstan	1	1.1%
Germany	1	1.1%	Korea	1	1.1%
Albania	1	1.1%	Costa Rica	1	1.1%
Australia	4	4.3%	Mexico	1	1.1%
UK	3	3.2%	Portugal	1	1.1%
Finland	2	2.1%	Singapore	1	1.1%
France	3	3.2%	Shanghai-China	1	1.1%
England	3	3.2%	Chile	1	1.1%
Ireland	1	1.1%	Tobago	1	1.1%
Sweden	1	1.1%	Trinidad	1	1.1%
Japan	1	1.1%	Türkiye	51	54.3%
Canada	1	1.1%	New Zeeland	1	1.1%

Table 6 shows which countries' data were used in the postgraduate theses related to DIF. As seen in Table 6, the highest number of theses used the data from Türkiye (f=51), followed by data from America (f=11), Australia (f=4), the UK (f=3), England (f=3) and France (f=3).

Table 7. Distribution of DIF Determination Methods Used in Theses

Used Method	f	%
Field indexes	1	0,6%
ANOVA	1	0,6%
b parameter difference	2	1,3%
BILOG MG DIF	1	0,6%
Cox β	1	0,6%
Multilevel mixture item response theory	1	0,6%
DINA-DMF	1	0,6%
Generalized Progressive Linear Modelling (GPLM)	2	1,3%
Generalized Mantel Haenszel (GMH)	1	0,6%
Mixed Logistic Regression Method	1	0,6%
Mixed Rasch Model	1	0,6%
Partial Point Model	1	0,6%
Likelihood Ratio-LR	1	0,6%
Liu- Agresti Statistics	1	0,6%
Logistic Regression (LR)	27	17,2%
Logistic Regression Likelihood Ratio Method	1	0,6%
Lord's $\chi 2$ Test Method	4	2,5%
LORDIF	2	1,3%
Item Response Theory - Likelihood Ratio Analysis	11	7,0%
Mantel Test	5	3,2%
Mantel-Haenszel (MH)	36	22,9%
Multiple Indicators Multiple Causes	4	2,5%
Ordinal Logistic Regression (OLR)	3	1,9%
Latent Class Analysis	2	1,3%
poly-SIBTEST	6	3,8%
Raju's area measurement methods	4	2,5%
Rasch Tree Method	2	1,3%
Rasch Model	6	3,8%
Ordered Logistic Regression	2	1,3%
Simultaneous Item Bias Test (SIBTEST)	24	15,3%
Standardization-DMF (ST-DMF)	2	1,3%

When Table 7 is examined, it is seen that DIF detection methods were used according to both Classical Test Theory and Item Response Theories. From among these methods, Mantel-Haenszel (MH) (f=36) one of the

classical test theory methods, is the method most frequently used method in the theses, followed by Logistic Regression (f=27) and Simultaneous Item Bias Test (f=24).

Table 8. Distribution of the Types of Data Used in the Theses

Type of Data	f	%
Multi-category data	9	13.6%
Two-category data	52	78.8%
Two-category and Multi-category data	5	7.6%
Total	66	100%

According to Table 8, two-category data (f = 52) is the most frequently used data type in the theses on DIF, followed by multi-category data (f=9) and both two-category and multi-category data (f=5).

CONCLUSION and DISCUSSION

A total of 66 postgraduate theses were found to have been written on DIF between 2012 and 2022. These theses were prepared within the universities that give postgraduate education in the field of Measurement and Evaluation in Education. The highest number of theses on DIF was written in 2019. In the bibliometric analysis of the years in which more articles on DIF were published, the highest number of articles was in 2019 and 2021. The research finding is similar to this finding (Eminoğlu-Özmercan, 2023).

When the type of postgraduate theses and the year they were published were examined, it was determined that the most postgraduate theses were written in 2019 and the least in 2017. It was also determined that doctoral theses were written most in 2016 and least in 2018, 2019, and 2022. Accordingly, in this study, it can be said that the tendency to study DIF is shifting towards master's theses from doctoral theses and that the tendency to study DIF in doctoral theses is gradually decreasing.

It was determined that in the theses examined, larger-scale tests that were easier to access were used as data sets. In addition, some studies using simulation data also used real data. Since both simulation and real data were used in two of the theses examined, the total number of data used is more than the total number of theses. It can be said that PISA data were preferred most because it is a large-scale test, it provides access to a lot of data about the participants and it offers the opportunity to work with more than one form. Gender was found to be the variable most frequently examined in the postgraduate theses. It was also seen that many different variables were used while conducting a DIF analysis. In studies, while it can be determined whether there is DIF in items based on only one variable, different variables can be used besides this one variable. In some of the theses, two, three, or four variables were used to examine their interaction.

In the postgraduate theses examined, mathematics and science were found to be the school subjects most preferred. It can be said that this is due to the fact that the main theme of each application is different in the applications of large-scale tests such as PISA and TIMSS, which are carried out at regular intervals.

Since the theses examined were written in Türkiye, it can be seen that they were mostly conducted with many different data from Turkey. The second most used data were from America. The reason why countries such as the USA, Australia, and the UK were used more than once is because international large-scale tests such as PISA and TIMSS, administered in different years, were evaluated. Almost all the other countries were used once. The reason for this may vary depending on the purpose of the thesis, or it may be that a country being compared was not wanted to be compared again because it had been compared before. Since in some studies, more than one country was examined, the total number of the countries used in the theses is higher than the total number of the theses.

When the methods used to determine whether items exhibit DIF were examined, Mantel-Haenszel (MH) was found to be the most frequently used method. This finding is similar to the finding that the Mantel-Haenszel (MH) method is prominent and widely used in detecting DIF (Diaz et al., 2021; Gomez-Benito & Navas-Ara, 2000; Guilera et al., 2013). At the same time, Wainer and Sireci (2005) stated that the Mantel-Haenszel (MH) method is one of the most frequently used methods in DIF detection. After the Mantel-Haenszel method, Logistic Regression and Simultaneous Item Bias Test methods are the most commonly used methods. Logistic Regression is one of the most effective and recommended methods among various DIF detection methods (Camilli and Shepard, 1994). Also, Berrio et al. (2020) stated that Mantel-Haenszel and Logistic Regression methods are the most widely used methods using simulated data under various conditions. In addition, the reason for using these methods may be that they are frequently used as a DIF detection method and can be used with two-category data.

In the theses on DIF, two-category data was most frequently used. It can be said that the reason for this may be that the data are in the form of 1-0 and the partially scored items are mostly converted into 1-0 data and used.

SUGGESTIONS

Within the current study, postgraduate theses on Differential Item Functioning (DIF) were examined. The theses used were from the theses open to access at the YÖK National Thesis Centre. Studies carried out between 2012 and 2022 but not accessible through the YÖK National Thesis Centre could not be included in the analysis process of the study. On the other hand, it is clear that research on DIF is not limited to postgraduate theses. It is possible to conduct a broader study by including conference proceedings and articles published in domestic and international journals.

ETHICAL TEXT

In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. The responsibility belongs to the author for any violations that may arise regarding the article. This study that does not require an ethics committee.

Author(s) Contribution Rate: The contribution of the author is 100% in this study.

REFERENCES

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. P.W. Holland ve H. Wainer (Ed.). *Differential item functioning* (pp. 1-23). Hillsdale, NJ:Lawrence Erlbaum Associates.

Berrío, Á. I., Gomez-Benito, J., & Arias-Patiño, E. M. (2020). Developments and trends in research on methods of detecting differential item functioning. *Educational Research Review*, 31, 100340. https://doi.org/10.1016/j.edurev.2020.100340

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 differentially functioning test items. *Educational Measurement Issues and Practice*. 17: 31 – 44.

Çalık, M. & Sözbilir, M. (2014). İçerik analizinin parametreleri. *Eğitim ve Bilim,* 39(174), 33-38. https://doi:10.15390/EB.2014.3412

Diaz, E., Brooks, G., & Johanson, G. (2021). Detecting differential item functioning: Item Response Theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education,* 8(2), 376-393. https://doi.org/10.21449/ijate.730141

Dinçer, S. (2018). Content analysis in scientific research: Meta-analysis, meta-synthesis, and descriptive content analysis. *Bartın Üniversitesi Eğitim Fakültesi Dergisi,* 7(1), 176-190. https://doi: 10.14686/buefad.363159

Dybå, T. & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology,* 50(9-10), 833-859.

Eminoğlu-Özmercan, E. (2023). The bibliometric analysis of the differential item functioning using VOSviewer. [Manuscript submitted for publication]...

Gomez-Benito, J., & Navas-Ara, M. J. (2000). A comparison of χ2, RFA and IRT based procedures in the detection of DIF. *Quality and Quantity,* 34(1), 17-31. https://doi.org/10.1023/A:1004703709442

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods,* 18(4), 553-571. https://psycnet.apa.org/doi/10.1037/a0034306

Karaçam, Z. (2013). Sistematik derleme metodolojisi: Sistematik derleme hazırlamak için bir rehber. *Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi*, 6(1), 26-33.

Messick, S. (1995). Validity of psychological assessment. validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologica Association*. 50(9). 741-749.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Osterlind, S. (1983). *Test item bias*, Sage Publications.

Reynolds, C. R. & Suzuki, L. A. (2003). Bias in psychological assessment an empirical review and recommendations. I. R. Weiner (Ed.). *Handbook of Psychology* (s. 82-113). John Wiley & Sons.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6(4), 317–375. https://doi.org/10.2307/1164616

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

Tekin, H. (1993). *Eğitimde ölçme ve değerlendirme.* (8. Baskı). Yargı Yayınları.

Turgut, M. F. (1983). *Eğitimde ölçme ve değerlendirme metotları*. Gül Yayınevi.

Turgut, M. F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme* (8.baskı). Pegem Akademi.

Wainer, H., & Sireci, S. G. (2005). *Encyclopedia of social measurement*. ScienceDirect.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment,* 29(4), 364-376. https://doi.org/10.1177/0734282911406666

Yıldırım, A. & Şimşek, H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri* (8. bakskı). Seçkin Yayıncılık.

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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.