Bartin University Journal of Faculty of Education
https://doi.org/10.14686/buefad.1676585

BARÜ

Research Article

Development of Test and Item Analysis Competency Perception Measurement Tool: Validation and Reliability Study

Test ve Madde Analizi Yeterlilik Algısı Ölçme Aracının Geliştirilmesi: Geçerlik ve Güvenirlik Çalışması

Münevver Başman*, Fatih Kezer, Müge Uluman Mert*****

* Marmara University, Department of Educational Measurement and Evaluation, İstanbul, Türkiye. E-mail: munevver.kaya@marmara.edu.tr
** Kocaeli University, Department of Educational Measurement and Evaluation, Kocaeli, Türkiye. E-mail: fatih.kezer@kocaeli.edu.tr
*** Marmara University, Department of Educational Measurement and Evaluation, İstanbul, Türkiye. E-mail: mugeulumann@gmail.com

Citation: Başman, M., & Kezer, F., & Uluman Mert, M. (2026). Development of test and ıtem analysis competency perception measurement tool: Validation and reliability study. *Bartın University Journal of Faculty of Education, 2026*(2), 353-367. https://doi.org/10.14686/buefad.1676585

Article Information

Article History

Received:
15 April 2025
Accepted :
14 January 2026
Published:
12 March 2026

Keywords

Teacher competencies
Test and item analysis
Scale development

Abstract

Purpose: *The quality of the tests depends on the test and item analysis. Determining teachers' competence levels in this area will reveal which areas they need to be supported. This study aims to develop measurement tools (a scale and a checklist) that can measure teachers' perceptions of test and item analysis competence and to determine teachers' perceptions of competence in this area.* ***Method:*** *A total of 1091 teachers working in public schools, selected through convenience sampling, participated in this survey model study. In the scale development phase, 393 teachers' data were used for EFA, 306 teachers' data for CFA, and 92 teachers' data for test-retest analysis. The final form of the scale and the checklist developed by researchers about test and item analysis competency was applied to 300 teachers, and an attempt was made to determine teachers' perceptions of test and item analysis competence.* ***Findings:*** *It was observed that the scale comprised 22 items and a single-factor structure. It was found that the one-dimensional structure of the scale was confirmed by examining the values of the fit indexes with confirmatory factor analysis. When the reliability coefficients were examined, it was seen that the Cronbach's alpha value and test-retest analysis values were .968 and .975, respectively. Furthermore, the scale ensures measurement invariance across gender. A checklist consisting of 42 items was also developed.* ***Conclusion:*** *Based on the findings, the developed measurement tool is considered valid and reliable. It was concluded that teachers had deficiencies in test and item statistics and that they were particularly inadequate in subjects requiring advanced statistics.*

Makale Bilgisi

Makale Tarihçesi

Geliş Tarihi:
15 Nisan 2025
Kabul Tarihi:
14 Ocak 2026
Yayımlanma Tarihi:
12 Mart 2026

Anahtar Kelimeler

Öğretmen yeterliliği
Test ve madde analizi
Ölçek geliştirme

Öz

Amaç: *Testlerin kalitesi, test ve madde analizine bağlıdır. Öğretmenlerin bu alandaki yeterlilik düzeylerinin belirlenmesi, hangi alanlarda desteklenmeleri gerektiğini ortaya koyacaktır. Bu araştırmanın amacı, öğretmenlerin test ve madde analizi yeterliliklerine ilişkin algılarını ölçebilecek ölçme aracı (ölçek ve kontrol listesi) geliştirmek ve öğretmenlerin bu alandaki yeterlilik algılarını belirlemektir.* ***Yöntem:*** *Tarama modelindeki bu çalışmaya, kolayda örnekleme yoluyla seçilen, kamu okullarında çalışan toplam 1091 öğretmen katılmıştır. Ölçek geliştirme aşamasında 393 öğretmenin verisi AFA, 306 öğretmenin verisi DFA ve 92 öğretmenin verisi ise test-tekrar test analizi için kullanılmıştır. Araştırmacılar tarafından geliştirilen test ve madde analiz yeterliği ile ilgili ölçek ve kontrol listesinin son hali 300 öğretmene uygulanarak öğretmenlerin test ve madde analiz yeterliği algıları belirlenmeye çalışılmıştır.* ***Bulgular:*** *Ölçeğin 22 maddeden oluşan tek faktörlü bir yapıya sahip olduğu görülmüştür. Ölçeğin tek boyutlu yapısının doğrulayıcı faktör analizi ile uyum indeksi değerlerinin incelenmesiyle doğrulandığı bulunmuştur. Güvenirlik katsayıları incelendiğinde Cronbach Alfa değeri ve test-tekrar test analizi değerlerinin sırasıyla .968 ve .975 olduğu görülmüştür. Ayrıca, ölçek cinsiyete göre ölçme değişmezliğini sağlamaktadır. Bununla birlikte 42 maddelik kontrol listesi de geliştirilmiştir.* ***Sonuç:*** *Elde edilen bulgular doğrultusunda geliştirilen ölçme aracı geçerli ve güvenilir kabul edilmektedir. Öğretmenlerin test ve madde istatistiklerinde eksiklikleri olduğu ve özellikle ileri istatistik gerektiren konularda yetersiz oldukları sonucuna varılmıştır.*

Introduction

To monitor and evaluate students' development in the teaching and learning process, measurement and evaluation methods and techniques are used. It is important to create valid and reliable measurement tools in the measurement and evaluations conducted to determine the extent to which the specified gains have been achieved. Multiple-choice tests are frequently used in in-class, national and international measurements.

Multiple-choice tests, known for their effectiveness and economy, are widely preferred in educational evaluations in various content areas (Gierl et al., 2017). Multiple-choice tests stand out for their ability to make objective measurements, ease of application and scoring, ability to measure at most levels of learning in relation to the cognitive domain, ability to be applied to a large number of students simultaneously, ability to estimate item and test statistics, and ability to cover a wide range of content (Fellenz, 2004; Saadat et al., 2020; Simkin & Kuechler, 2005). On the other hand, creating multiple-choice tests is a complex and time-consuming task since the items must be tested in terms of standards and quality (Haberman et al., 2019; Tarrant et al., 2006). To test the items and create valid and reliable tests, item analyses must be performed. Item analyses guide in determining the characteristics of the items and the quality of the test (Gronlund, 1993; Singh et al., 2009).

A statistical method for choosing and removing test items according to their difficulty index, discrimination index, and distractor efficiency is item analysis (Sharma, 2019). It

assists us in identifying items that are excessively challenging or easy, items that have unjustifiable distractions, or objects that cannot differentiate between students who are and are not familiar with the subject (Lange & Mehrens, 1967). Sharma (2019) and Freeman (1962) state that each test item determines the test's quality. The quality of the developed multiple-choice tests can also be revealed through item analysis. Sim and Rasiah (2006) and Zubairi and Kassim (2006) add that item analysis gives teachers input on what has to be changed to make multiple-choice questions appropriate for the test. Item analysis is not only for the benefit of teachers but also for the benefit of students. In this way, valid and reliable information about students is obtained.

Teachers and instructors should create test items that are most appropriate to students' skills and abilities and that can discriminately measure what students have learned. There may be hints in poorly structured multiple-choice tests that let students estimate the right response without knowing the answer, and they may not be able to adequately distinguish between high- and low-achieving students. Using item analysis, well-structured items that are able to differentiate between pupils with high and low ability (Schuwirth & Van Der Vleuten, 2003) can be developed. Analysing the items will help to value high-quality items and create quality tests. However, many teachers prefer to use test items found in books or on the Internet, or they create similar items and use them without testing the quality of the items. This can lead to the creation of tests that are not valid and reliable. In fact, many teachers do not have sufficient knowledge about item analysis (Karaca, 2003; Pektaş, 2010; Tarrant et al., 2006). Understanding instructors' proficiency in item analysis is the greatest way to address this issue (Fook et al., 2013).

In the studies conducted, it is seen that there are scales for determining the measurement and evaluation competence perceptions of lecturers (Sever & Saban, 2015) and teacher candidates (Evin-Gencel & Özbaşı, 2013; Karaca, 2003; Karaduman & Yanpar-Yelken, 2020; Karaman & Şahin, 2014; Pektaş, 2010; Sabancı & Yazıcı, 2017). In these studies, the scales developed by Bütüner et al. (2010), Karaca (2003), Nartgün (2008), Pektaş (2010) and Sever and Saban (2015) were used. Although the developed scales include those that can measure the measurement and evaluation competence perceptions of lecturers and teacher candidates, it is seen that there is no measurement tool that can measure the measurement and evaluation competence perceptions of teachers and, more specifically, their test and item analysis competence perceptions. A measurement tool is also needed to determine teachers' competence in test and item analysis. The purpose of this study is to ascertain teachers' judgements of their own competence in test and item analysis, as well as to develop a measurement tool that can assess teachers' perceptions of that competence. The following queries were sought for this purpose:

- What is the validity of the Test and Item Analysis Competence Perception Scale?
- What is the reliability of the Test and Item Analysis Competence Perception Scale?
- What is the validity of the Test and Item Analysis Competence Perception Checklist?

- What is the reliability of the Test and Item Analysis Competence Perception Checklist?
- What are the teachers' perceptions of test and item analysis competence?

Methodolgy

Research Design

This study aimed to develop valid and reliable measurement tools (a scale and a checklist) to assess teachers' perceptions of their competence in test and item analysis and to determine their perceived competence levels in this area. To achieve these aims, the study employed a survey research design, which is appropriate for describing existing situations as they are and collecting quantitative data from a large group of participants (Karasar, 2023). The survey research design supports both objectives: it enables the systematic development and validation of the measurement tool and provides descriptive data reflecting teachers' perceived competence levels.

Study Group

The study group consists of teachers working in public schools, determined through convenience sampling. In the scale development study, data were collected from 428 teachers for EFA, 325 teachers for CFA, and 92 teachers for test-retest in Istanbul province in the 2023-2024 academic year due to its cosmopolitan structure and easy accessibility for researchers. The developed teachers' competence perception scale and checklist for test-item analysis were applied to 308 teachers working in public schools. After excluding missing data and outliers, the analyses were continued with the remaining 393 participants for EFA, 306 for CFA, and 92 for test-retest. The final version of the scale and checklist was administered to 300 participants in Istanbul in the 2023-2024 academic year, and descriptive statistics were obtained. Demographic information about the teachers is given in Table 1.

Table 1. *Distribution of Demographic Information of the Participants*

	EFA		CFA		Test-Retest		Competency Perception Scale	
Variables	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Gender								
Female	290	73.8	229	74.8	57	62	222	74.0
Male	103	26.2	77	25.2	35	38	78	26.0
School type								
Secondary School	211	53.7	153	50.0	211	53.7	190	63.4
Imam Hatip Secondary School	10	2.5	8	2.6	10	2.5	8	2.7
Anatolian High School	71	18.1	61	19.9	71	18.1	64	21.3
Science High School	6	1.5	3	1.0	6	1.5	2	.7
Imam Hatip High School	14	3.6	11	3.6	14	3.6	12	4.0
Vocational and Technical Anatolian High School	25	6.4	22	7.2	25	6.4	24	8.0
Seniority								
0-5 years	267	62.4	196	60.3	57	62	176	58.7

6-10 years	53	12.4	44	13.5	12	13	39	13.0
11-15 years	19	4.4	38	11.7	14	15.2	38	12.7
16-20 years	8	1.9	16	4.9	6	6.5	16	5.3
20 years and above	41	9.6	31	9.5	3	3.3	31	10.3

Note. EFA: Exploratory Factor Analysis, CFA: Confirmatory Factor Analysis.

When Table 1 is examined, it is seen that 209 (%73.8) of the data obtained for EFA were collected from female teachers and 103 (%26.2) from male teachers. For CFA, 229 (%74.8) of the data were collected from female teachers and 77 (%25.2) from male teachers. To obtain descriptive statistics using the developed competence perception scale and checklist, 222 (74%) of the data were collected from female teachers and 78 (26%) from male teachers. For all applications, data were collected from teachers who worked mostly in secondary schools. When the seniority years are considered, it is seen that there are more teachers with 0-5 years of seniority.

Measurement Tool Development Process

The eight steps suggested by DeVellis (2017) for the scale development process will be followed during the development of the scale: (1) clearly defining the construct to be measured, (2) creating the item pool, (3) determining the measurement format, (4) reviewing the initial item pool by experts, (5) considering the validity items, (6) applying the items to the scale development sample, (7) evaluating the items, (8) optimizing the scale length. In order to develop a measurement tool that measures teachers' perceptions of competence in testing and item analysis, first, the relevant literature was reviewed, and measurement tools in this field were examined. A literature review was conducted to better understand and clearly define the construct and to identify existing scales. The item pool was created by considering the literature. After creating items that could measure teachers' perceptions of competence in the test and item analysis, the measurement format was determined. No reverse-scored items have been written. In order to make the measurement tool ready for pilot application, four expert opinions were used, three from the Measurement and Evaluation field and one from the Turkish Language and Literature field. In line with expert opinions, 2 items were rewritten and 3 items were added. As a result of all the evaluations, some items in the scale were rearranged, and a five-point Likert-type measurement tool consisting of a total of 22 items with options from “Absolutely Inadequate” to “Absolutely Adequate” was obtained.

The trial form of the scale was shared with participants via Google Forms, and sent through both e-mail and phone messages. Before the scale was filled in, the study group was informed about the research and their consents regarding their voluntary participation were obtained. In addition to the scale developed to determine teachers' perceptions of competence, a checklist was also developed. First of all, the literature was examined, and the features that should be included in the item and test statistics were determined. The candidate items in the checklist were created with Yes-No options and presented to two experts in the field of Measurement and Evaluation for their opinions. After the necessary corrections were made, a 42-item checklist was created for testing and item statistics. Sample items of the checklist are as follows: I can calculate the

discrimination index of each item, and I can interpret the skewness coefficient of test scores based on the measurement results. The content validity ratio (CVR) was determined by a total of five experts, and all items were deemed appropriate by each expert (CVR=1), ensuring content validity (Ayre & Scally, 2014; Lawshe, 1975). To determine the reliability of the checklist, interrater reliability was calculated (dividing the number of agreements by the total number of agreements and disagreements) (Miles & Huberman, 1994) and its value was found as 100%. The scale and checklist were applied to 300 teachers, and data were obtained.

Data Analysis

Data were analysed with IBM SPSS Statistics (Version 26), Jamovi (Version. 2.4.14) and R Program (Version 4.3.2). SPSS program was used for data organization, item analysis, exploratory factor analysis and descriptive statistics, Jamovi program (semlj module) was used for confirmatory factor analysis, and R program (difR package) was used for measurement invariance. When the data obtained for exploratory factor analysis was examined, it was seen that there was no missing data. First, the univariate and multivariate extreme values of the data were examined. Z scores were examined to determine univariate outlier whether they were within the ±3 range. Mahalanobis distances were calculated for multivariate outliers, and the significance threshold was set at $p<0.001$ (Çokluk et al., 2023). Since univariate extreme values were detected in 13 of the data and multivariate extreme values in 22, they were removed from the data set. The item analysis and exploratory factor analysis were continued with the remaining 393 participants.

When analysing the data, item analysis was first performed. The Pearson product-moment correlation coefficient was calculated for the item analysis. As a result of the item analysis, it was seen that the item-total correlation values were .30 and above for all items. According to the statement, items with a correlation of .30 or above are good at discriminating between people, things between .20 and .30 can be included in the scale if it is thought necessary, and items with a correlation of less than .20 shouldn't be included (Büyüköztürk, 2019). Accordingly, as a result of the item analysis, all items were found suitable for use in the scale.

Secondly, it was examined whether the data were suitable for factor analysis. The results of the Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity tests were analysed to determine whether the data were appropriate for factor analysis. The findings of the analysis are given in Table 2.

Table 2. *KMO and Bartlett Sphericity Test Values*

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.937
Bartlett's Test of Sphericity	Approx. Chi-Square	10065.930
	df	231
	p	.000

When Table 2 is examined, it is seen that the KMO value is .937 and the result of the Barlett test is significant (p<.05). Field (2000) stated that the result of the KMO test should be .50 as the lower limit. Tavşancıl (2010) expressed the KMO value as perfect for .90 and above. From this perspective, it is seen that the value obtained as a result of the KMO test in the study is in a good possible value range, and the data have a suitable structure for factor analysis.

The principal axis factor analysis method and the promax rotation method were used for EFA since there is a relationship between the factors (Tabachnick & Fiddell, 2007). Cronbach's alpha reliability coefficient was used in the reliability analysis of the scale.

In order to determine whether the developed scale supports the construct validity, CFA was performed on a similar group, and data were collected from 325 teachers. No univariate extreme value was observed in the data. There are 19 data points with multivariate extreme values, and they were removed from the data set. CFA was performed using the maximum likelihood estimation method with the remaining 306 data sets, as the data provided multivariate normality. As a result of the analysis, fit indices were examined. In addition, the measurement invariance of the data according to gender was also examined. For measurement invariance, four nested hierarchical models were tested, namely configural invariance, metric invariance, scalar invariance and strict invariance. In order to ensure measurement invariance, both fit indices should be examined and |ΔCFI | ≤.01, |ΔGamma hat|≤.001 and |ΔMcDonald's NCI|≤.02 values should be between (Cheung & Rensvold, 2002).

In addition, whether the data obtained from the measurement tool were reliable, the Cronbach alpha coefficient obtained from the data collected for EFA, data were also collected from 95 teachers for test-retest at a two-week interval and analyzed. There was no missing data. After excluding the excluded outliers, the analysis was conducted with 92 data points. Since the data distribution showed a normal distribution, the Pearson Correlation coefficient was calculated. 308 teachers were reached with the developed measurement tool and checklist, and descriptive analyses were conducted with the 300 data obtained from both the competence perception scale and the checklist.

Results

The findings regarding construct validity, Cronbach's alpha reliability coefficient and test-retest calculations for validity and reliability analyses while developing the measurement tool and the findings obtained from the measurement tool and checklist developed to determine teachers' perceptions of competence are given below.

Exploratory Factor Analysis Findings

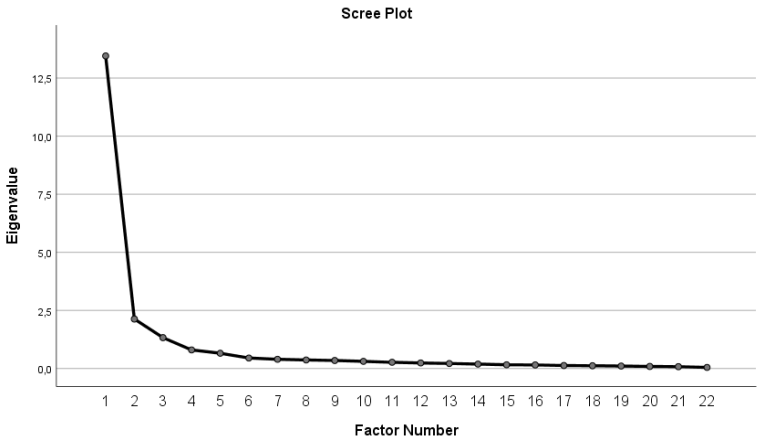
In line with the data obtained for exploratory factor analysis, eigenvalues, parallel analysis results and scree plot were examined to determine the construct of the scale. The findings are given in Table 3 and Figure 1.

Table 3. *EFA and Parallel Analysis Eigenvalues*

Factor	EFA Eigenvalues	PA Eigenvalues	Result
1	13.451	1.453	Accepted
2	2.129	1.371	Accepted
3	1.327	1.318	Accepted
4	.798	1.266	Rejected

Note. EFA: Exploratory Factor Analysis, PA: Parallel Analysis.

Figure 1. *Scree-Plot Graph*



When the eigenvalues and scree-plot graph were examined, it was concluded that the first factor explained 59.958% of the variance, and considering that the second (8.635%) and third (4.681%) factors made much less contribution to explaining the variance, and expert opinions were also taken, it was concluded that the scale had a single factor. When limited to a single factor, it was seen that it explained 59.349% of the variance. Factor loadings, item-total correlation and item-remainder correlation of the single-factor structure of the scale are given in Table 4.

Table 4. Factor Loadings of Items According to Exploratory Factor Analysis Findings

Items		Factor 1	Item-Total Correlation	Corrected Item-Total Correlation
1	I can decide on the quality of the item by examining the item statistics.	.741	.739**	.716
2		.779	.771**	.751
3		.753	.744**	.722
4		.718	.715**	.687
5		.750	.755**	.728
6	I can make inferences about the success of each student	.749	.748**	.726
7		.763	.760**	.736
8		.807	.804**	.784
9		.802	.805**	.785
10		.806	.808**	.789
11		.612	.655**	.612
12		.719	.756**	.723
13		.787	.815**	.791
14		.754	.779**	.751
15		.748	.774**	.745
16		.845	.855**	.838
17		.829	.840**	.821
18	I can interpret measures of central	.810	.825**	.803

	dispersion of test scores (range, standard deviation, variance, etc.).			
19		.841	.848**	.830
20		.825	.836**	.816
21		.824	.838**	.818
22		.636	.675**	.634

Note. **p<.01

According to Table 4, the factor loadings of the items are between .612 and .845, the item total correlations are significant between .655 and .855, and the corrected item total correlations range from .612 to .838.

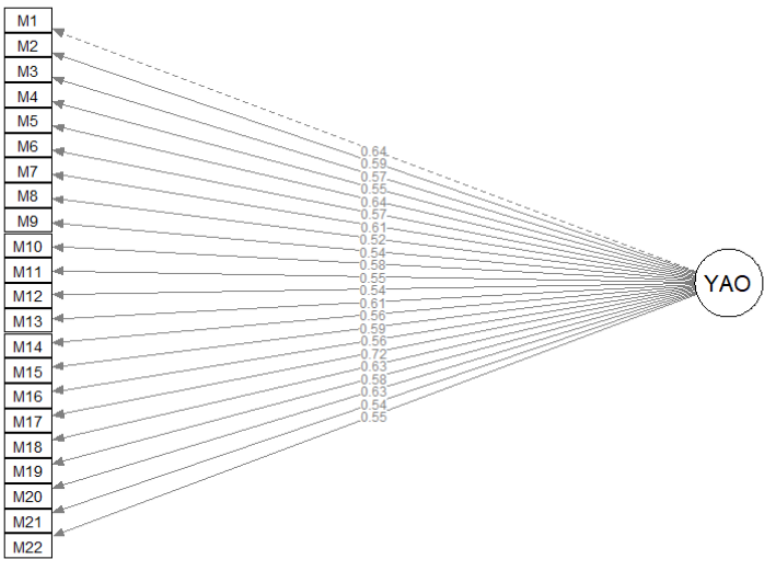
Confirmatory Factor Analysis and Measurement Invariance Findings

CFA was conducted to evaluate the accuracy of the structure obtained from the exploratory factor analysis. As a result of the CFA, 22 items were gathered under a single factor; the fit index values are shown in Table 5, and the factor loadings are shown in Figure 2.

Table 5. Fit Index Values

Fit Indexes	Values obtained from the scale	Good Fit Indicator Criteria	Acceptable Fit Indicator Criteria
χ^2/sd	1.995	≤3	3<X^2/sd ≤5
RMSEA	.057	≤.05	.05<RMSEA≤.08
GFI	.889	≥.95	.90≤GFI<.95
CFI	.909	≥.95	.90≤CFI<.95
RMR	.042	≤.05	.05<RMR≤.08
SRMR	.047	≤.05	.05<SRMR≤.08
TLI	.899	≥.95	.90≤TLI<.95

Figure 2. *Confirmatory factor analysis findings*



As a result of the confirmatory factor analysis, it is seen that the 22-item single-factor structure is confirmed. The goodness-of-fit statistics obtained were found as χ2/sd=1.995, RMSEA=.057, CFI=.909, GFI=.959, SRMR=.047, TLI=.899. These values are considered acceptable values (Çokluk et al., 2023; Tabachnick & Fidell, 2013). EFA and

CFA were examined to demonstrate the validity of the measurement tool. In addition, whether the measurement tool showed gender bias or not was examined to determine whether there was measurement invariance according to gender.

The measurement invariance of the scale by gender was examined by testing four nested hierarchical models, namely configural invariance, metric invariance, scalar invariance and strict invariance. The multi-group CFA findings regarding the single-factor measurement invariance of the scale by gender are given in Table 6.

Table 6. Measurement invariance fit indices by gender

Model	χ2 (df)	RMSEA	SRMR	CFI	TLI	RNI
Configural	694.670 (418)	.066	.058	.882	.870	.882
Metric	716.847 (439)	.064	.067	.882	.876	.882
Scalar	741.471 (460)	.063	.068	.880	.880	.880
Strict	759.332 (482)	.061	.069	.882	.887	.882
Model	Δχ2 (Δdf)	ΔCFI	ΔGamma hat	ΔMcDonald's NCI		
Configural	-	-	-	-		
Metric	22.177 (21)	.000	.000	.000		
Scalar	24.624 (21)	-.002	.000	.000		
Strict	17.861 (22)	.002	.000	.000		

When Table 6 is examined, it is seen that the fit indices are close to acceptable values and the difference values meet the conditions in the findings obtained from the measurement invariance analysis of the scale according to gender (Cheung & Rensvold, 2002). The fact that the obtained Δχ2 values were not significant (p>.05), and the ΔCFI, ΔGamma Hat and ΔMcDonald's NCI values were appropriate, can be stated as indicators that configural, metric, scalar and strict invariances were met according to gender. Considering the results of the fit indexes, it is accepted that the scale provided full measurement invariance by gender subgroups. As a result, the comparisons made regarding gender in the model will be meaningful within the framework of these findings.

Reliability Findings

To determine the reliability of the measurement tool, Cronbach's alpha coefficient and test-retest were used as internal consistency reliability coefficients. The findings are presented in Table 7.

Table 7. *Reliability coefficients of the scale*

Scale	Cronbach's Alpha Value	Test-Retest Value
Test and Item Analysis Competency Perception	.968	.975

As seen in Table 7, the Cronbach's alpha coefficient for the Test and Item Analysis Competency Perception Scale was calculated as .968. In addition, a test-retest was applied with a two-week interval, and the correlation between them was examined. When the Pearson product-moment correlation coefficient was examined for test-retest, it was found to be .975 (p<.05). In psychological tests, reliability coefficients of .70 and above are sufficient for reliability (Büyüköztürk, 2019). When the values obtained are examined, the scale can be described as a reliable scale.

As a result of the EFA, CFA and reliability analyses conducted to reveal the validity and reliability of the measurement tool, it can be concluded that the scale can measure teachers' perceptions of test and item analysis competence with sufficient validity and reliability. In addition, the fact that the measurement tool provided measurement invariance according to gender also supports the validity of the tool.

Determination of Teachers' Perceptions of Test and Item Analysis Competence

Descriptive statistics obtained from the developed scale and the frequency of responses given for each item in the checklist were used to determine the level of teachers' perceptions of test and item analysis competence. Since the final version of the test and item analysis competency perception scale has a total of 22 items, the expected lowest score is 22.00, the highest score is 110.00, and the range is 88.00. The highest score obtained by teachers regarding the level of having these determined competencies is 110.00, and the lowest score is 22.00. The range was found to be 88.00. It is seen that the scale covers the expected width. The scale mean was calculated as 74.59, the median as 78.00, and the standard deviation as 18.66. The skewness coefficient was found as -.636, and the kurtosis coefficient was .160. These values show that the distribution of scale scores is close to a normal distribution.

Considering the scale results, the three competencies that teachers think they have the most among the determined test and item analysis competencies are that they can make inferences about the success of each student (65.7%)and the group (63.7%) by examining the test statistics, and that they can identify students' learning gaps by examining test and item statistics (62.7%). The three competencies that they consider themselves the least competent in are being able to create a histogram graph of test scores (25.3%), being able to use at least one program that can calculate test and item statistics (36%), and being able to create a frequency table of test scores (38.3%).

In addition, the number of those who said 'No' in calculating and interpreting the statistics of the items in the test is higher than the number of those who said 'Yes'. Although the number of those who said 'Yes' is higher than those who said 'No' only in the item 'Calculating the percentage of options marked', it is seen that there are still more No respondents in interpreting. It is seen that teachers have the most difficulty in interpreting the standard deviation and variance values in item statistics.

In test statistics, there are more Yes than No respondents in calculating simpler statistics such as the Arithmetic mean, Median, Mode, and Range, while there are more No than Yes respondents in calculating other test statistics. Considered in general, it is seen that teachers have the most difficulty in calculating and interpreting the skewness and kurtosis coefficients of test scores and the KR-20 and split-half reliability coefficients based on measurement results.

Discussion

In this study, a valid and reliable measurement tool was developed to measure teachers' perceptions of competence in test and item analysis and teachers' perceptions of

competence in this area were tried to be determined. For the validity and reliability studies of the measurement tool, content validity, construct validity, measurement invariance and reliability coefficients were examined, and the findings obtained showed that the developed scale was valid and reliable. The developed scale is a 22-item five-point Likert-type and one-dimensional. In addition to the developed scale, a 42-item checklist was created so that teachers' perceptions of competence in test and item analysis could be determined in more detail.

When the responses given to the scale items were examined, the three scale items that teachers thought they were least competent in were creating a histogram graph, using a program that can perform test and item analysis and creating a frequency table for test scores. In the checklist used to obtain more detailed information in addition to the developed scale, this situation was in the direction of deficiencies in calculating and interpreting item and test statistics. There are more people who think they are not competent enough to calculate and interpret item statistics than those who think they are competent enough. In test statistics, it is seen that teachers can do simpler statistical operations more easily, but have difficulty with advanced statistics. This situation can be stated that in the scale developed by Karaca (2003) to measure teacher candidates' perceptions of their measurement and evaluation competencies, the competencies that the candidates think they have the least are knowing and being able to perform statistical operations appropriate to the types of measurement tools. In the study conducted by Pektaş (2010), the measurement and evaluation general competency perception scale for teacher candidates developed by Nartgün (2008) was used, and it was found that it was moderately sufficient in statistical analysis and reporting. The subjects included in statistical analysis and reporting are item difficulty index and item discrimination power calculation and interpretation, normal distribution, skewness, kurtosis, etc. Calculation and interpretation, determination of the correlation technique appropriate for the structure of the data, making and interpreting the calculation, and calculating and interpreting statistics such as t-test, F-test, etc. It was stated that teacher candidates especially have many deficiencies in statistical knowledge.

In general, while teacher candidates can see themselves as moderately competent in the statements given in the given scale, they specifically stated that they are less competent in the items given in the checklist. Based on the information obtained with the checklist, it can be stated that teachers are partially competent. Based on the scores obtained from the test and item analysis competency perception scale and the checklist, it can be stated that teachers are moderately competent in test and item statistics. This situation is consistent with the studies of Yeşilyurt (2012), Erdoğdu and Kurt (2012), Sabancı and Yazıcı (2017) and Karaduman and Yanpar Yelken (2020). The scale developed by Nartgün (2008) was used in the mentioned studies, and in all four studies, it was found that teacher candidates were moderately competent in statistical analysis and reporting. In the study conducted by Çakan (2004) with teachers working in primary and secondary schools, it was concluded that teachers found themselves inadequate or deficient in measurement and evaluation.

Conclusion and Recommendations

In summary, this study has primarily presented a valid and reliable measurement tool that can measure teachers' perceptions of test and item analysis competence and provide more detailed information in statistical analyses. With the developed scale and checklist, an attempt has been made to determine teachers' perceptions of test and item analysis competence. When the obtained data is evaluated, it is seen that teachers are partially and moderately competent. This situation reveals that teachers have deficiencies and that it is an area that needs to be developed. Therefore, it can be suggested that in-service training, such as measurement literacy, applied statistics and statistical software training can be provided to teachers in this area and that programs that perform test and item statistics should be introduced. In addition, since it was not determined whether there are significant differences in teachers' perceptions of competence in different groups in this study, it can be suggested that variables such as gender and different departments be taken into consideration and that there are differences.

The study group consists of teachers working in public schools, and in this respect, the study is limited only to teachers working in public schools; teachers working in other institutions could not be taken into account. Teachers' perceptions of their test and item analysis competencies are limited to their responses to the items on the developed scale and checklist.

Ethical Approval: Ethical approval and written permission were obtained from the Marmara University Institute of Educational Sciences Research and Publication Ethics Committee (approval granted at the meeting dated 31 August 2022, application no. 06-42).

Author Contributions

Münevver Başman: Conceptualisation, Theoretical Background, Methodology, Data Collection, Data Analysis, Data Visualisation, Writing, Review and Editing

Fatih Kezer: Methodology, Data Collection, Data Analysis, Data Visualisation, Writing, Review and Editing

Müge Uluman Mert: Conceptualisation, Theoretical Background, Data Collection, Review and Editing

Conflict of Interest: The authors declare no conflict of interest.

Plagiarism Checks: The manuscript was screened for plagiarism using iThenticate. The overall similarity index was found to be 12%, and no evidence of plagiarism was detected.

Grant Support: This research was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) (Project No: 122K062).

Artificial Intelligence Usage Statement: Portions of this manuscript were assisted by the use of ChatGPT (OpenAI, San Francisco, CA, USA), an artificial intelligence language model,

for purposes such as improving clarity and grammar. The authors reviewed and edited the content generated by the model and took full responsibility for the final version of the manuscript.

REFERENCES

Ayre, C., & Scally A. J. (2014). Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development, 47* (1), 79–86. https://doi.org/10.1177/0748175613513808.

Çakan, M. (2004). Comparison of elementary and secondary school teachers in terms of their assessment practices and perceptions toward their qualification levels. *Ankara University Journal of Faculty of Educational Science, 37*(2), 99-114. https://doi.org/10.1501/Egifak_0000000101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2023). *Multivariate statistics for the social sciences: Applications of SPSS and LISREL* (7th ed.). Pegem.

Erdoğdu, M. Y. & Kurt, F. (2012). The analysis of teacher's competency perceptions on measurement and assessment according to the some variables. *Electronic Journal of Education Sciences, 1*(2), 23–36.

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment & Evaluation in Higher Education, 29*(6), 703-719. https://doi.org/10.1080/0260293042000227245

Fook, C. Y., Yunus, F. W., & Sidhu, G. K. (2013, December). *Teachers' knowledge on item analysis and item analysis software*. In 2013 IEEE Conference on e-Learning, e-Management and e-Services (pp. 30-33). IEEE.

Freeman, F. (1962). *Theory and practice of psychological testing.* Oxford & Ibh publishing.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*, 1082–1116. https://doi.org/10.3102/0034654317726529

Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Allyn and Bacon.

Haberman, S. J., Liu, Y., & Lee, Y. H. (2019). Distractor analysis for multiple-choice tests: An empirical study with international language assessment data. *ETS Research Report Series, 2019*(1), 1-16. https://doi.org/10.1002/ets2.12275

IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

Karaca, E. (2004). Development of a Likert type competence perception scale towards measurement and evaluation competencies of teacher candidates. *Dumlupınar University Journal of Social Sciences, 9*, 1-19.

Karaduman, B., & Yelken, T. Y. (2020). Identification of prospective teachers' evaluation preferences and general proficiency perceptions in measurement and evaluation. *Journal of Çukurova University Institute Social Sciences, 29*(1), 339-353. https://doi.org/10.35379/cusosbil.651030

Karasar, N. (2023). *Scientific research method: Concepts, principles, techniques*. Nobel.

Lange, A., Lehmann, I.J. & Mehrens, W.A. (1967). Using item analysis to improve tests. *Journal of Educational Measurement, 4*(2), 65-68.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology, 28*(4), 563-575.

Miles, M. B. (1994). Qualitative data analysis: An expanded sourcebook. *Thousand Oaks*.

Nartgün, Z. (2008). Measurement and evaluation common competency perception scale for prospective teachers: a validity and reliability study. *Bolu Abant İzzet Baysal University Journal of Faculty of Education, 8*(2): 85-94.

Pektaş, S. (2010). The analysis of teacher candidates' competency perceptions on measurement and assessment. [Master's Thesis, Bolu Abant İzzet Baysal University]. National Thesis Center of the Council of Higher Education.

R Core Team (2019). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Saadat, S., Noori, M., Alipour-Anbarani, M., Mousavi Bazaz, N., Babakhanian, M., Montazeri KHadem, A., & Azadi, H. (2021). Students' challenge in answer-changing on multiple-choice exams; doubting the answer or not? a systematic review. *Medical Education Bulletin, 2*(1), 137-144. https://doi.org/10.22034/meb.2021.293511.1008

Sabancı, O., & Yazıcı, K. (2017). Examining pre-service teachers' efficacy perceptions towards measurement and evaluation. *Trakya Journal of Education, 7*(7), 128-153.

Schuwirth, L. W., & Van Der Vleuten, C. P. (2003). Written assessment. *Bmj, 326*(7390), 643-645. https://doi.org/10.1136/bmj.326.7390.643

Sharma, L. R. (2019). Item analysis: An evaluation of multiple choice questions based on research methodology in the internal examination. *International Journal of New Technology and Research, 5*(12). 01-08.

Sim S.M., & Rasiah R. I. (2006). Relationship between item difficulty and discrimination indices in true/falsetype multiple choice questions of a para-clinical multidisciplinary paper. *Annals Academy of Medicine Singapore, 35*, 67-71.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. *Decision Sciences Journal of Innovative Education, 3*(1), 73-98.

Singh T, Gupta P, & Singh D. (2009). *Test and item analysis. Principles of Medical Education* (3rd ed.). Jaypee brothers N Delhi.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed). Pearson.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med. Educ. 9*(40). https://doi.org/10.1186/1472-6920-9-40

The jamovi project (2024). jamovi (Version 2.4.14) [Computer Software].

Zubairi A.M. & Kassim N.L. (2006). Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysain Journal of English Language Teaching Research, 2*, 1-20.