BMIJ
ISSN: 2148-2586

Review Article

BUSINESS & MANAGEMENT STUDIES:
AN INTERNATIONAL JOURNAL
Vol.:8 Issue:3 Year:2020, 2694-2726

Citation: Sürücü, L. & Maslakçı, A., Validity And Reliability In Quantitative Research, BMIJ, (2020), 8(3): 2694-2726, doi: http://dx.doi.org/10.15295/bmij.v8i3.1540

VALIDITY AND RELIABILITY IN QUANTITATIVE RESEARCH

Lütfi SÜRÜCÜ [1]

Ahmet MASLAKÇI [2]

Received Date (Başvuru Tarihi): 13/06/2020

Accepted Date (Kabul Tarihi): 19/07/2020

Published Date (Yayın Tarihi): 25/09/2020

In the article, the first author is in the role of the Corresponding Author.

ABSTRACT

The Validity and Reliability of the scales used in research are essential factors that enable the research to yield beneficial results. For this reason, it is useful to understand how the Reliability and Validity of the scales are measured correctly by researchers. The primary purpose of this study is to provide information on how the researchers test the Validity and Reliability of the scales used in their empirical studies and to provide resources for future research. For this purpose, the concepts of Validity and Reliability are introduced, and detailed explanations have been provided regarding the main methods used in the evaluation of Validity and Reliability with examples taken from the literature. It is considered that this study, which is a review, will attract the attention of researchers.

Keywords:

Scale,

Validity,

Reliability,

Quantitative Research

JEL Codes:

C12, *C15, C42*

NİCEL ARAŞTIRMADA GEÇERLİLİK VE GÜVENİLİRLİK

ÖZ

Araştırmada kullanılan ölçeklerin geçerliliği ve güvenilirliği araştırmanın sağlıklı sonuçlar vermesini sağlayan önemli faktörlerdir. Bu nedenle, güvenilirliğinin ve geçerliliğinin araştırmacılar tarafından nasıl doğru ölçüldüğünü anlamak faydalıdır. Bu çalışmanın temel amacı, araştırmacıların ampirik araştırmalarında kullandığı ölçeklerin geçerliliği ve güvenirliliğini nasıl test edecekleri konusunda bilgi sunmak ve ileride yapılacak olan araştırmalara yönelik kaynak sağlamaktır. Bu maksatla çalışmada, geçerlik ve güvenirlik kavramları tanıtılmış ve literatürden alınan örneklerle geçerlik ve güvenirlik değerlendirilmesinde kullanılan ana yöntemlere ilişkin ayrıntılı açıklamalar yapılmıştır. Derleme olarak yapılan bu çalışmanın araştırmacıların ilgisini çekeceği değerlendirilmektedir.

Anahtar Kelimeler:

Ölçek,

Geçerlik,

Güvenirlik,

Nicel Araştırma

JEL Kodları:

C12, *C15, C42*

[1] Phd, European Leadership University, lutfi.surucu@elu.edu.tr, https://orcid.org/0000-0002-6286-4184

[2] Asst. Prof., Cyprus Science University, ahmetmaslakci@csu.edu.tr, https://orcid.org/0000-0001-6820-4673

Bu makale, araştırma ve yayın etiğine uygun hazırlanmış ve iThenticate intihal taramasından geçirilmiştir.

1. INTRODUCTION

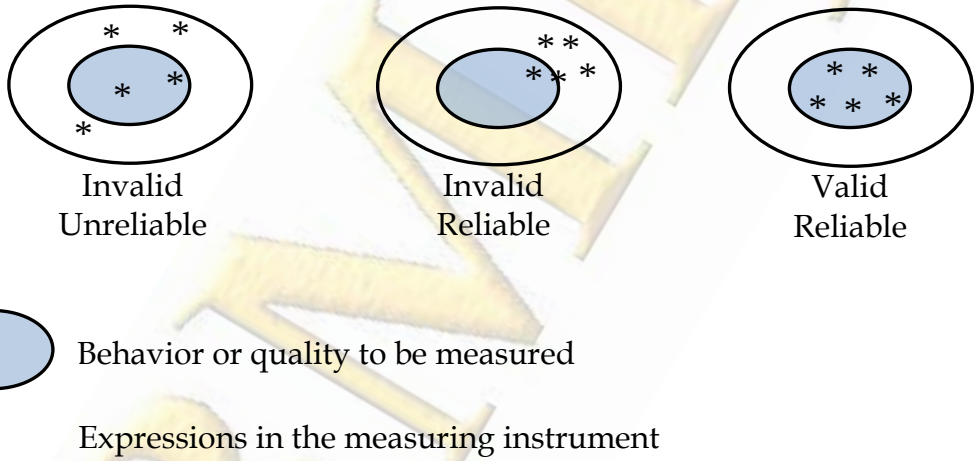
Researchers develop a conceptual model that includes certain variables for the problems they identify in their study or for the topics they want to investigate. Then, they collect and analyze the data obtained through a questionnaire to test the conceptual model they have developed. Since most empirical studies are included in this paradigm, in order to achieve beneficial results in the study, the instrument used to measure the survey must have certain qualities. The first of these qualities is the validity of the scale. Validity is concerned with whether the measuring instrument measures the behaviour or quality that it is intended to measure and is a measure of how well the measuring instrument performs its function (Anastasi and Urbina, 1997). Another feature that the scale should have is Reliability. Reliability is an indicator of the stability of the measured values obtained in repeated measurements under the same circumstances using the same measuring instrument. Reliability is not only a feature of the measuring instrument, but it is also a feature of the results of the measuring instrument. According to the current literature, it is clear that the Reliability and Validity of the measuring instrument are two indispensable features. A study conducted using a measuring instrument that does not possess one or both of these features will not yield beneficial results. For this reason, the measuring instrument used in the study must have both Validity and Reliability.

This study aims to introduce researchers to the main methods used in the literature to provide information about the Validity and Reliability of scales and to evaluate the Validity and Reliability of the scales used in the researches. The study provides a general overview to raise awareness among researchers instead of developing a critical view of studies on Validity and Reliability.

The study consists of five parts. The first part of the research is the introduction part, where general information is presented to the readers. In the second part, information about Validity and Reliability, in the third part, increasing Reliability and in the fourth part, threats to Validity and Reliability are included. The last part is the conclusion and discussion part.

2. VALIDITY AND RELIABILITY

Although the concepts of Validity and Reliability are closely related, they express different properties of the measuring instrument. Generally, a measuring instrument may be reliable without being valid, but if a measuring instrument is valid, it is also likely to be reliable. However, Reliability alone is not sufficient to ensure validity. Even if a test is reliable, it may not accurately reflect the desired behaviour or quality (Figure 1). For this reason, researchers must test both the Validity and Reliability of the measuring instrument they intend to use. The measuring instrument must satisfy these two conditions. Otherwise, it will not be healthy for researchers to interpret the research findings.



2.1. Validity

Validity refers to whether the measuring instrument measures the behaviour or quality it is intended to measure and is a measure of how well the measuring instrument performs its function (Anastasi and Urbina, 1997). Validity is determined by the meaningful and appropriate interpretation of the data obtained from the measuring instrument as a result of the analyses. Whiston (2012) defined validity as obtaining data that is appropriate for the intended use of the measuring instruments. In this case, validity tests, which determine whether the expressions in the scale make suitable measurements according to the purpose of the research, come to the fore. Testing the validity of the measuring instrument is more difficult but more

important than assessing its Reliability. In order for the research to yield beneficial results, the measuring instrument must measure what it claims. The use of a validated measuring instrument ensures that the findings obtained as a result of the analyses are valid.

In order to determine the validity of the measuring instrument, different types of validity have been suggested in the literature (Oluwatayo, 2012). These can be listed as follows: Predictive Validity, Concurrent Validity, Content Validity, Criterion-Related Invalidity, Internal Validity, External Validity, Construct Validity, Face Validity, Systemic Validity, Theoretical Validity, Jury Validity, Consequential Validity, Cultural Validity, Interpretive Validity, Descriptive Validity, Evaluative Validity, Statistical Conclusion Validity, and Translation Validity. Although it is possible to expand this list further, two types of validity are generally accepted to have particular importance in the literature, namely content validity and construct validity. Although the types of validity named above are generally used for different purposes, some are valid forms of substitution. For example; Predictive Validity is obtained by calculating the correlation between the estimated score obtained from a scale and the criterion known to measure the properties desired to be measured. Concurrent validity, on the other hand, is a form of validity used as a measure of the convergence of the results of an alternative instrument used to measure the same structure or as a measure of the sameness. When the definitions of both validities are examined, the closeness of these two types of validity attracts attention. From this perspective, researchers must decide for themselves which validity to be tested in line with their needs and purposes. However, if the researcher does not develop a new scale and uses one that has been previously created and tested for Validity and Reliability in the local language of the country in which the original research was conducted, it is sufficient to test the content and construct validity.

2.1.1. Content Validity

Bollen (1989) defined content validity as a qualitative form of validity that evaluates whether the expressions contained in the measuring instrument represent

the phenomenon intended to be measured. In line with this definition, it can be said that a content validity of a measuring instrument is a validity study that reveals the extent to which each item in the measuring instrument serves the purpose. Content validity used in scale development or adaptation of the developed scale for the relevant culture and language provides the determination of the most appropriate expressions to improve the quality of the expressions in the measuring instrument and to serve the purpose of the scale. Thus, it is ensured that there is a useful scale with content capability that serves the purpose of the prepared measuring instrument to measure any behaviour or quality. In studies in the field of social sciences, in particular, the content area of many concepts used is unclear. Therefore, there is no consensus in the literature regarding the definitions and content of most concepts. A researcher who performs a content validity study must develop a theoretical definition of the relevant concept and determine the content (dimensions) of that concept.

In the literature, several methods have been proposed for determining content validity. Among them, taking expert opinions and statistical methods are the two most frequently applied methods.

In the first method, evaluation by more than one referee is also a method of obtaining expert opinions. This method is a process that transforms qualitative studies based on expert opinions into quantitative statistical studies (Yeşilyurt and Çapraz, 2018). In this method, the researcher consults the experts to evaluate each expression in the developed measuring instrument in terms of the content of the scale or terms of appropriateness and evaluates each expression in line with the opinions of the experts (Rubio et al., 2003). In obtaining objective results in the calculations to be made for determining the content validity, the quality and number of experts have significant importance (Ayre and Scally, 2014). Qualified experts are crucial for the results to be consistent and unbiased. Therefore, care should be taken when choosing experts and academicians or practitioners with extensive knowledge should be preferred for the measuring instrument that is intended to be developed.

Evaluation of content validity according to expert opinion is a form of a statistical analysis based on the content validity of whether the items in the measuring instrument should be on the scale or not, and it is calculated according to the below formula (Lawshe, 1975):

$$CVR = \frac{N_e}{\frac{N}{2}} - 1 \quad \text{or} \quad CVR = \frac{N_e - \frac{N}{2}}{\frac{N}{2}}$$

Expressions in the formula denote:

CVR = Content Validity Ratio

N= Total number of experts evaluating items in the measuring instrument.

N_e = Number of experts evaluating the relevant item as appropriate.

According to Lawshe (1975), each statement in the pool of items created is presented to experts to obtain their opinions. Experts score these statements as "Appropriate", "Appropriate But Should Be Corrected" and "Subtracted". If half of the experts express their opinion on the statement in the measuring instrument as 'Appropriate', CVR= 0, if more than half of them state "Appropriate", CVR> 0, and if less than half of the experts state "Appropriate" then CVR<0. If the CVR is 0 (zero) or negative, that expression must be subtracted from the measuring instrument (Yeşilyurt and Çapraz, 2018).

The second method is to test the content validity using statistical methods. Among the different statistical methods, the most frequently used is factor analysis. Factor analysis emerged in the early 1900s through the study of Charles Spearman and is used in many fields such as social sciences, medicine, economics, and geography. Factor analysis uses mathematical procedures to simplify interrelated measurements to explore patterns in a set of variables. It summarizes the data to easily interpret and understand the relationships and patterns of observed variables in the measuring instrument.

In factor analysis, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are the two most widely used factor analysis techniques. The

researcher's experience shapes EFA, and in general, EFA is intuitive. EFA is also a type of exploratory analysis applied by researchers to learn about the number or nature of expressions on the scale. This analysis allows the researcher to identify the main dimensions of a relatively sizeable hidden structure represented by several elements (Henson and Roberts, 2006; Swisher et al., 2004). The primary purpose is to examine a large number of expressions related to the structure of the scale, to identify fewer expressions that explain the structure of the scale, and to increase the explanatory power of the scale structure. It is usually performed to reduce the number of variables observed in scale development studies and to determine what factors it creates. In CFA, the primary purpose is to test the accuracy of the previously validated scale or model.

A common approach in empirical research is to perform CFA in order to test the accuracy of the scale and the model in studies conducted using pre-tested scales. In the CFA performed, if the threshold values are not provided, or the structure of the measuring instrument is not verified, then EFA should be conducted. With the EFA, the relationship pattern between the expressions and factors in the measuring instrument is explored, and the necessary corrections are made, thus allowing CFA to be performed. These procedures are essential for research to provide healthier results.

2.1.2. Construct validity

Construct validity is concerned with the degree to which the instrument measures the concept, behaviour, idea or quality- that is, a theoretical construct- that it purports to measure. In other words, it is the ability to distinguish between participants with and without the behaviour or quality to be measured. For example, when measuring instrument developed by Allen and Meyer (1990) is used to measure organizational commitment in employees, if employees with high organizational commitment have high scores and employees with low organizational commitment have low scores, this implies that the measuring instrument has construct validity. In summary, the fact that the measuring instrument has construct validity means that it proves the construct to be measured; in other words, it can

reveal the construct. Construct validity is widely used in research and is based on the logical relationships between variables. However, a constructed fit of the study is not sufficient when examining validity. Testing the convergent and discriminant validities after the construct validity test is of great importance for the research to yield beneficial results. Many methods have been developed in the literature to test construct validity. Among these methods, in the method developed by Cronbach and Meehl (1955), inter-construct relations come to the fore rather than item-construct relations. Campbell and Fiske developed the multitrait-multimethod matrix in 1959, which is relatively easy to calculate and is based on testing convergent/discriminant validities. Although the developed matrix was relatively easy for researchers, it still contained some difficulties. Therefore, Fornell and Larcker (1981) proposed a technique for measuring convergent and discriminant validity based on the average explained variance (AVE) value obtained from each factor as a method of determining construct validity. This proposed technique has been generally accepted in the literature.

Convergent Validity

Convergent validity states that the expressions related to the variables are related to each other and the factors they create, and this means that the measuring instrument designed to measure particular construct measures this intended construct correctly. Convergent validity expresses that expressions should be related to each other and to the factor that is purported to measure the same concept. Convergent validity shows the degree of the relationship between the observed variables that measure the latent variable (Hair et al., 1998).

In order to provide convergent Validity, AVE values must be less than the composite Reliability (CR), and each AVE value must be greater than 0.5. AVE is obtained by dividing the sum of squares of the covariance loadings of the expressions related to the factor by the number of expressions. CR, which is formulated by Werts et al. (1978) refers to the level of Reliability of the relationship between observed variables and latent variables of a measurement instrument. CR is used to measure the general Reliability of heterogeneous but similar expressions, and

the value obtained is essential for determining the Reliability of the scale. The formulas of AVE, Square root value of AVE, and CR are shown below.

$$AVE = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{k} \qquad \text{Square root value of } AVE = \left(\frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{k}\right)^2$$

$$CR = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} - \lambda_i^2}$$

Where:

k = number of observed variables

λ_i = represents the factor loading of item *i* onto a single common factor.

Another method for determining convergent validity is based on the square root value of AVE being smaller than the CR, Cronbach's alpha (α), and AVE. In this case, it is stated that the scale provides convergent validity. The scale is considered to be reliable when the CR value is more significant than 0,7, as with the Cronbach's alpha value. In determining the convergent validity, a separate evaluation should be made for each factor construct.

It would be useful to give an example in order to provide a better understanding of the subject. For example, the "Standardized Regression Weights" values obtained from the confirmatory factor analysis of the innovative work behaviours scale developed by Janssen (2000) with the help of the AMOS program are shown in the table below.

Table 1. Assessment of the Measurement Model

Items	Standardized loadings	Cα	AVE	CR
Item 1	0,710			
Item 2	0,717			
Item 3	0,828			
Item 4	0,788			
Item 5	0,765	0,895	0,512	0,903
Item 6	0,643			
Item 7	0,640			
Item 8	0,605			
Item 9	0,711			

The factor loads of the items in the scale are presented in the table above as a result of the DFA conducted to determine the convergent validity of the Innovative Work Behaviors scale developed by Janssen (2000). When the table is examined, it is seen that the AVE value is greater than the lower threshold value of 0.5 (AVE = 0.512), and the CR value (CR = 0.903) is greater than the AVE value. Findings show that the scale has convergent validity. Also, the fact that the CR value is greater than the lower threshold of 0.7 indicates that the scale is reliable. Although many programs (e.g. LISREL, R, EQS 6.1, and Excel) are used for calculating CR and AVE values, AVE and CR values can be calculated with simple programming in Excel. The programming that can be done in Excel is shown below.

	A	B	C	D
1				
2	Item Number	Factor Loading λ	λ^2	$1-\lambda^2$
3	1	.710	=B3*B3	=1-C3
4	2	.717	=B4*B4	=1-C4
5	3	.828	=B5*B5	=1-C5
6	4	.788	=B6*B6	=1-C6
7	5	.765	=B7*B7	=1-C7
8	6	.643	=B8*B8	=1-C8
9	7	.640	=B9*B9	=1-C9
10	8	.605	=B10*B10	=1-C10
11	9	.711	=B11*B11	=1-C11
12				
13	COUNT	=COUNT(B3:B11)	=COUNT(C3:C11)	=COUNT(D3:D11)
14	SUM	=SUM(B3:B11)	=SUM(C3:C11)	=SUM(D3:D11)
15	SQUARE	=B14*B14		
16				
17	AVE	=C14/C13		
18	CR	=B15/(B15+D14)		

In the table above, data should be entered into the "Factor Loading" column by the researchers. AVE and CR values will be calculated automatically after the required formulas are used in the related fields in Excel.

Discriminant Validity

Discriminant validity is utilized to ensure whether the observed variables used in the measurement model measure the latent variable. Discriminant validity refers to the situation where expressions in the scale refer to one specific factor and are less related to the other factors; in other words, one item is related to one factor. In order to determine the discriminant validity, the Maximum Squared Variance (MSV) and the Average Shared Square Variance (ASV) values must be calculated. The MSV is the square of the highest correlation coefficient between factors. ASV value is obtained by dividing the sum of the squares of the variance shared by other factors by the number of shared variances. In other words, the arithmetic mean of the sum of the squares of the correlation coefficients between factors gives ASV value. Formulas for calculating ASV and MSV values are presented below.

$$ASV = \frac{\sum r^2}{d} \qquad MSV = [\text{Maksimum } (r)]^2$$

Where:

r= The number of Factor Correlation

d= The number of factors

For a better understanding of the subject, we can consider the determination of the discriminant validity of the innovative work behaviour scale developed in 3 dimensions by Janssen (2000). Firstly, a correlation analysis is performed with the help of the SPSS program. The results of the correlation analysis performed as an example are given below.

Table 2. Correlation Results

Variables	**1**	**2**	**3**
Idea Generation	1		
Idea Promotion	.293	1	
Idea Realization	.521	.402	1

The highest correlation in the table is between idea realization and idea generation (r = .521). Since MSV is the square of the highest correlation coefficient between factors, the MSV value is = .271 (.521 * .521= .271). The ASV value is the arithmetic average of the sum of the squares of the correlation coefficients. So the ASV value is = .173 $\left(\frac{.293^2 + .521^2 + .402^2}{3} = .173\right)$

Three conditions must be met in order for discriminant validity to be established. The first of these conditions is that the MSV value must be less than the AVE value (MSV <AVE). The second condition is that the ASV value must be less than the MSV value (ASV <MSV). The last condition is that the square root value of AVE should be greater than the correlation between factors. If all these conditions are met, it can be said that the measuring instrument provides discriminant validity.

The fact that convergent and discriminant validities are provided indicates that the factors in the scale can be evaluated together and that a total score can be obtained from the scale. Additionally, the scale's discriminant validity indicates that the factors have different characteristics in terms of their scope. In this context, it can be argued that each factor can be evaluated independently; in other words, total points can be obtained from each factor.

2.1.3. Face validity

Face validity is a subjective decision based on the researcher's feelings, thoughts, and intuition about the functioning of the measuring instrument. Researchers such as Kaplan and Saccuzzo (2017), and Whiston (2012) claimed that face validity could not be considered as an indicator of validity. Researchers believe that the results of face validity are not supported by statistical data and can be perceived to show validity even though the measuring instrument does not measure the structure it is intended to measure. Also, since face validity is a subjective decision, a measuring instrument that provides face validity according to individual researchers may not be considered persuasive for other researchers. Therefore, face validity is often seen as a weak form of structural validity.

In line with the current literature, concluding that face validity is insignificant is not an appropriate approach. It is a validity factor that should be taken into consideration when reporting validity procedures in social sciences research.

Assessment of face validity is performed by expert staff or by academic staff on the structure the measuring instrument is trying to determine. Within the evaluation criteria, different criteria are included such as (a) the purpose of each statement is appropriate for the measuring instrument, (b) the statements in the scale are clearly understood by the participants, (c) the readability of all statements in the measuring instrument, (d) the attractiveness of the questionnaire created, (e) the difficulty of each item appropriate for the level of the participants.

It is not sufficient for researchers to report that the face validity of the measuring instrument has been determined to be satisfactory by experts. In order to

make their research healthier and more meaningful, it would be beneficial for researchers to state experts' comments on the specified criteria in their research.

2.2. Reliability

Reliability refers to the stability of the measuring instrument used and its consistency over time. In other words, Reliability is the ability to measure instruments to give similar results when applied at different times. Of course, it is unlikely that the same results will be given every time due to differences at the time the measuring instrument is applied, as well as changes in the population and the sample. However, a strong positive correlation between the results of the measuring instrument is an indication of Reliability. The Reliability of the measuring instrument is an essential consideration for the results of the study to be healthy. Therefore, researchers should ensure that measuring instrument used is reliable.

Different methods are used to determine the Reliability of the scales used in empirical research. Among these, the most frequently applied methods are test-retest reliability, alternative forms, and internal consistency tests. Internal consistency tests can be applied in three different ways (split-half, item-total correlations, and alpha reliability coefficient).

In scale development studies, researchers can test the Reliability of the scales they develop by doing one or more of the test-retest Reliability, alternative forms, and internal consistency tests. On the other hand, the researchers who used the scale previously developed and whose Reliability was tested; they just need to do one of the internal consistency tests. The most preferred internal consistency tests are the alpha reliability coefficient.

2.2.1. Test-Retest Reliability

Test-retest Reliability refers to the consistency of the results obtained when the measuring instrument is applied to the same sample group at different times. The questionnaire prepared to test the Reliability of the measuring instrument with the test-retest method is firstly applied to a sample group. Then, the same questionnaire is applied again to the same sample group after a specific time. A high correlation between comparable survey data obtained at different times is an indication that the

measuring instrument has test-retest Reliability. The vital aspect is the period between the survey applications. The consensus is that the period should be long enough for the participants to remember their responses to the statements in the questionnaire, but short enough that the behaviour or quality of the scale will not change. For example, a researcher who wants to measure organizational commitment chooses to test the Reliability of the scale with the test-retest method: After applying the first survey application to an organization's employees, it would not be appropriate to apply the second survey a year later. This is because, during that year, the employees' organizational commitment may have decreased or increased due to the emergence of different factors within the organization. Generally, intervals of 1-2 weeks or 10-15 days will be sufficient for the test-retest method.

Researchers aiming to determine Reliability using the test-retest method generally predict the Reliability by using the Pearson correlation coefficient or comparing the data using the t-test (Oluwatayo, 2012). It should be remembered that the population from which the sample data comes must have a normal distribution in order to perform the T-test. Although there are different opinions in the literature regarding the interpretation of the obtained data, the general opinion is that a correlation value of 0.80 and above indicates that the measuring instrument provides test-retest Reliability (Whiston, 2012).

The implementation of this method involves some difficulties and limitations. Firstly, it can be challenging to reach the same sample group at different times for various reasons. For example, this could be caused by participants in the sample group taking annual leave at the time of the second test, leaving the job or their lack of willingness to participate in the research again. Consequently, failure to reach the same participants may threaten the test-retest Reliability of the scale. Another problem can occur when the period between the first and second test is too short. When the time between the two tests is short, the participants may remember the answer to the questions in the first test and answer from memory without thinking in the second test (Drost, 2011).

Another problem is that the participants may have been interested in the research and could give different answers in the second test by learning or researching the subject. Both potential scenarios could negatively affect the Reliability. On the other hand, if the period between the two tests is too long, this represents another problem. For example, suppose that a researcher may prefer the test-retest method to Determine the Reliability of a "job satisfaction" scale which is developed by Judge et al. (1999). If there is a long-time-interval between the two tests, the participants' job satisfaction levels may change as a result of certain events. In this case, the participants will give different answers to the measuring instruments. If the same sample group gives different answers to the expressions in the measuring instrument, it will negatively affect the Reliability of the measuring instrument. Therefore, the researchers should consider that when the time between the two tests is too long, the participants may be exposed to events that change their views, attitudes, and feelings about the behaviour being examined, or that situational factors may change. As a result, researchers who want to apply the test-retest method should ensure that the time difference between the two tests is sufficiently long that the participants do not remember their responses to the previous test, but also short enough so that the participants are not exposed to events that could cause changes in the behaviours desired to be measured. There is no generally accepted view in the literature regarding the time difference between the two tests to determine test-retest Reliability. However, for the test-retest, it is believed that if the tests are performed with an interval of 1-2 weeks or 10-15 days, this can yield beneficial results.

2.2.2. Alternative Forms or Parallel Forms

Another method used to test Reliability is the alternative forms method. Before applying this method, two different measuring instruments need to be developed that measure the same behaviour or quality. In addition to measuring instruments with the same content area, they must have the same number of items as well as similar features. In this method, the first measuring instrument is applied to the participants, followed by the second alternative form; the results are then evaluated to estimate the reliability coefficient. The focus is on the similarity of the

results obtained from both measuring instruments and the extent to which they match. The data obtained from both sample groups are compared using Pearson statistics or t-test statistics. The fact that there is no significant difference between the means of two groups or the correlation between the alternative forms is high (0.80 and above) indicates that there is no measurement error and it has equivalence reliability (Bowling, 2014). Unlike the test-retest method, in this method, both developed measuring instrument scan be applied simultaneously on two homogeneous groups.

When the above literature is examined, the alternative forms technique seems to be similar to the test-retest method, apart from the fact that two different measuring instruments are used to measure the same behaviour or quality (Drost, 2011). From this perspective, it can be said that some of the limitations of the test-retest method are also valid for the alternative forms technique.

2.2.3. Methods of Internal Consistency

Internal consistency is related to the Reliability of expressions contained in the measuring instrument. The measuring instrument measures the consistency of the items within it and questions how well the measuring instrument measures a particular behaviour or quality. The internal consistency of the measuring instrument depends on the correlation of each item that constitutes the measuring instrument.

Many different methods have been used for determining Reliability based on internal consistency in previous studies. Among these, the most preferred methods are Split-half, item-total correlations, and alpha coefficient (Kuder-Richardson-20 & 21 and Cronbach's alpha). However, among the existing methods, the most preferred and widely used method is to determine the internal consistency according to the Cronbach's alpha value. A comprehensive summary of Cronbach's alpha, as well as all other methods, is provided below to provide researchers with an understanding of the relevant methods.

2.2.3.1. Split-Half Method

Nunnally (1978) recommended that the Split-Half Method should be used when measuring the variability of behaviour for short periods when alternative

forms are not available. Unlike the test-retest and alternative form methods, the Split-half method is usually applied in the same period. In this method, the measuring instrument is firstly applied to a sample group. Before the analysis is conducted, the items included in the measuring instruments are divided into two halves in terms of content and degree of difficulty. Splitting the measuring instrument into two parts is usually achieved by assigning all the odd-numbered items in the measuring instrument to one group and all the even-numbered items to another group. The results of the two tests obtained by dividing the measuring instrument are expected to match. Reliability can be calculated using the Spearman-Brown formula:

$$\text{Reliability} = \frac{2r}{1+r}$$

Where *r* refers to the correlation between the two tests.

The value obtained as a result of the Spearman-Brown formula ranges from 0 to 1. The obtained value is 0.70 and above indicates that the scale is reliable. The literature states that this value can be tolerated up to 0.60 for descriptive and explanatory research. However, the size of the reliability coefficient is closely related to the number of items in the scale. Beneficial results are not obtained on scales consisting of a small number of items (8 items and below). According to Kline (1993); this method should not be used for scales with less than 10 items.

The split-half method has several advantages over the test-retest and alternative forms methods. Firstly, as in the test-retest and alternative form methods, the measuring instrument is applied simultaneously to a single sample group instead of being applied to the same sample group at different times. Therefore, the participants cannot answer from memory without thinking about their responses. Besides, as in other methods, there are no difficulties in terms of reaching the same participants. In this context, accessing the data in the split-half method is more comfortable and less costly (Bollen, 1989).

It should be noted that apart from the existing benefits, the split-half method also has certain disadvantages. In this method, Reliability is measured by the correlation of the items between the two halves. Therefore, the correlation (i.e.

Reliability) will vary slightly depending on how the items are divided (Drost, 2011). This method should be preferred when all questions measure the same structure, and there are too many items in the measuring instrument for this structure to be measured. This means that in cases where the measuring instrument measures more than one structure, it would not be appropriate to use the split-half method. For example, the measuring instrument developed by Luthans et al. (2007) that determines the lively psychological capital of the participants consists of a 4-dimensional scale comprised of 24 items (self-efficacy, hope, resiliency, and optimism). Therefore, it is not appropriate to use the split-half method to determine the Reliability of the positive psychological capital scale.

2.2.3.2. Item-Total Correlations

Correlation can be used to determine the Reliability of the scales as well as to measure the relationship of the measuring instruments with each other. The Item-Total Correlations is used in this context to refer to how much the score of each item in measuring instrument is related to the total score of all items in the measuring instrument. It is also stated that the item-total correlations for the items in the measuring instrument will be between 0,30 and 0,80, and the items with these values are considered suitable. However, Norman and Streiner (2003) stated that in order to meet the scaling assumptions, it is sufficient for one item to correlate with the total score of all items in the measuring instrument by more than 0,20. In general, if the correlation is less than 0.3, the items do not represent the conceptual structure, but if it is above 0.80, the items are interpreted as representing only a particular aspect or a specific area of the conceptual structure. Therefore, the total correlation values between 0.3 and 0.8 mean that the items are sufficiently homogenous and contain the original variance.

There should be between 100 and 200 respondents for item-total correlation analysis. The stated sample sizes are valid not only for the main study but also for the pilot study. Below is an example of Item-Total Correlation Matrix.

Table 3. Item-Total Correlation Matrix

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1,000	0,322	0,104	0,364	0,410
Item 2	0,322	1,000	0,201	0,383	0,360
Item 3	0,104	0,201	1,000	0,250	0,241
Item 4	0,364	0,383	0,250	1,000	0,451
Item 5	0,410	0,360	0,241	0,451	1,000

When the Item-Total Correlation Matrix is examined, it is seen that the correlation of the 3rd Items is lower than the lower threshold value of 0,3 and does not represent the conceptual structure. Removing the third item will increase the Reliability of the scale. However, removing the item from the scale should not be the first alternative. When such items are removed from the scale, there is a possibility that the diversity in the scale decreases and so the items are collected in a very narrow area. Therefore, instead of removing the item, the effect of the removed item on the alpha value should be examined first. If the alpha value of the item has little increase effect, it will be more appropriate not to remove this item.

Finally, when researchers decided to test the Reliability of a measuring instrument using the item-total correlations method, it is useful to examine the correlation after removing the contribution of the item they have identified to the total correlation. If the measuring instruments have dichotomous responses (e.g., Yes / No, Agree / Disagree, True / False, etc.), researchers are generally recommended to use the Point biserial correlation.

2.2.3.3. Alpha Coefficient

Cronbach's Alpha Coefficient

The most popular method used in research to test internal consistency is the determination of the alpha coefficient. In the literature, different calculations have been developed for the alpha coefficient. Despite this diversity in the literature, the Cronbach's alpha coefficient, which was developed by Cronbach (1951) and is named after the researcher who developed the coefficient, is generally accepted in the

literature. As the Cronbach's alpha coefficient, the value of which is between 0 and 1, approaches +1, it is stated that internal consistency is high.

Formula;

$$\alpha = \left[\frac{N}{N-1}\right]\left[\frac{S_x^2 - \sum S_i^2}{S_x^2}\right]$$

N = Number of items in the measuring instrument

S_i^2 = variance of each item

S_x^2 = sum of variance points of each item in the measuring instrument

Although the Cronbach's alpha is interpreted in different ways in the literature, the generally accepted approach is presented in the table below.

Table 4. The Classification of Cronbach's Alpha Coefficient

Cronbach's Alpha Coefficient	**Interpretation of Cronbach's Alpha Coefficient**
≥ 0,9	The internal consistency of the scale is high,
0,7 ≤ α < 0,9	The scale has internal consistency,
0,6 ≤ α < 0,7	The internal consistency of the scale is acceptable,
0,5 ≤ α < 0,6	The internal consistency of the scale is weak,
α ≤ 0,5	The scale has no internal consistency.

A specific issue needs to be considered here. Firstly, researchers believe that a measuring instrument is very reliable when the Cronbach's alpha value, which shows the internal consistency of the scale, is measured as 0,95 and above. This should not be viewed as a correct approach. This high value indicates that some expressions found in the measuring instruments are the same and do not have any distinctive features. In other words, this indicates there are more expressions in the measuring instrument than necessary and that this behaviour or quality can be measured with fewer expressions. When an article is submitted for publication in an

SSCI / SCI-indexed journal, referees are likely to criticize this issue. Likewise, a Cronbach's alpha value between 0,6 and 0,7 may not be sufficient for journals with an SSCI / SCI index. For this reason, a Cronbach's alpha value of 0,7 and above is an indicator of the internal consistency of the scale.

Kuder-Richardson-20 and 21 (KR-20 and 21)

Although not widely used, the method developed by Kuder and Richardson (1937) can be used to determine the internal consistency of the measuring instrument. In this method, the homogeneity of the items is evaluated. The most frequently applied homogeneity index is KR-20. This method is based on the ratio of correct and incorrect answers to the answers given to each item in the measuring instrument. KR-20 is valid for tests whose items are divided into two (True / False) (Oluwatayo, 2012). By calculating the percentage of correct answers given to each item in the test, the internal consistency can be estimated by the Kuder-Richardson 20 and 21 formulas. The coefficient obtained shows the internal consistency of the measuring instrument.

If there is scoring with different weights among the items in the measuring instrument, this method cannot be used. If the difficulty levels of the items in the measuring instrument are not different from each other, the KR-21 method can be used. In KR-21, all items in the test are assumed to be of equal difficulty and the calculation is much simpler. In order to reduce the time required and to use easily accessible data, the KR-21 method is frequently preferred by researchers.

Resultantly, since Kuder-Richardson 20 and 21 are based on the logic of the correctness and inaccuracy of the answers, they can be applied in tests that score points for correct answers, and no points for incorrect or empty answers. The formula developed by Kuder and Richardson (1937) to determine the homogeneity of the items is shown below.

The formula of KR-20:

$$r_{20} = \frac{k}{k-1}\left[\frac{S_x^2 - \sum pq}{S_x^2}\right]$$

In the formula;

r_{20} = Reliability of the measuring instrument

k = Number of items in the measuring instrument

S_x^2 = Variance value of the whole measuring instrument

p = Number of correct answers to each item in the measuring instrument

q = The number of incorrect answers given to each item in the measuring instrument.

As stated in the formula, the reliability result is obtained by calculating each item (pq) and collecting data for each item ($\sum pq$). The formula of K-21, where all the items in the test are of equal difficulty and which is more straightforward in terms of calculation, is:

$$r_{21} = \frac{k}{k-1}\left[1 - \frac{M(K-M)}{KS^2}\right]$$

In the formula;

r_{21} = Reliability of the measuring instrument

k = Number of items in the measuring instrument

S^2 = Variance of scores

M = Mean of the scores.

When both formulas are analyzed, it is seen that the Reliability of the measuring instrument will take less time to calculate in the KR-21 method due to the fact there are fewer procedures and the information used can be obtained easily.

3. INCREASING RELIABILITY

As a result of the analysis of the scale used in the research, the primary reason for a reliability coefficient being low is that the researcher chose to use the wrong scale. In order to avoid this problem, researchers should prefer to use scales whose Validity and Reliability have already been tested. The literature emphasizes that the

scales developed and applied to vary according to the population and culture in which the research was conducted. For this reason, it is essential to adapt the scales developed in different cultures to the language of the research and to test their Validity and Reliability. A mistake frequently made by researchers is that they translate the scale developed in a different language into their native language and apply it in the research. This often causes problems such as the low Reliability of the scale and the inability to achieved beneficial results. However, the population on which the research is conducted is also essential. Researchers should choose to use scales that have been previously tested have been determined to be reliable and valid for the population they plan to investigate. For example, a scale named "Attitude towards evidence" developed by Ruzafa - Martínez et al. (2011) for nurses may not produce the same results when applied to a different population (soldiers, hotel employees, students, etc.).

When it is determined that the measuring instrument used does not have sufficient Reliability as a result of the analyses, several methods can be applied.

1. Despite all efforts, the best way of increasing the Cronbach's alpha coefficient is to increase the sample size. Increasing the number of samples will likely increase the Cronbach's alpha coefficient, which indicates the internal consistency of the scale. However, it is not true that increasing the number of samples will increase the Cronbach's alpha value to a great extent.

2. One of the methods used in scale development studies to increase the Reliability of the measuring instrument is to increase the number of expressions in the measuring instrument. As the number of expressions increases, the Cronbach's alpha value of the scale will increase. This is related to the formula used to calculate the Cronbach's alpha value. For this reason, researchers now state the Cronbach's alpha value as well as the "Composite Reliability" value in their research.

3. If the Cronbach's alpha value is too low (0,40 and below), there are two possibilities. Firstly, this means that the vast majority of the participants in the sample group answered the statements in the measuring instrument without reading or gave random answers. In this case, the researcher should examine the data of each

survey and exclude the survey data that is filled in a certain systematic way or is considered to have been filled randomly from the scope of the research. In this case, the best method is to evaluate the data in terms of outliers, which should be excluded from the research. In cases where the researcher does not apply the questionnaires to the participants themselves (where a human resources manager or surveyors give them), this problem is frequently encountered.

Another possibility is that the researcher used a formative scale. In the literature, there are two types of scales: formative and reflective. The main difference between the two scales is based on the relationship between cause (formative) and effect (reflective). It is essential to determine the direction of the relationship to obtain accurate and logical results in the research. If the direction of this relationship is from the latent variable to the observed variables, the scale is reflective, and from the observed variables to the latent variables, the scale is formative. In the formative scale, items expressions determine the latent variable and show the reasons, not the effects, of the latent variable (Aksay and Ünal, 2016). Therefore, there may be a negative or zero correlation between expressions informative scales. In this case, the Cronbach's alpha value is too low, indicates that the measuring tool works very well. Despite these differences, even researchers discussing the objectionable aspects of formative scales draw attention to the fact that it is incorrect to analyze formative scales as if they were reflective scales. As a result, researchers should know whether the scale used in their research is formative or reflective and use a scale appropriate to the research model and theory.

4. One of the main reasons why the Cronbach's alpha value is meagre is due to the inability of the researchers to fully recognize the scale used. Among the items in the scale used, reverse codes can be found. If the researcher is not aware or does not notice this, the Cronbach's alpha value could be low.

Reverse coded items are when all statements in the scale are positive, some statements are negative, or all items are negative, while complimentary items are. The issue of whether the items are positive or negative should be addressed in terms of "the meaning of the items". For example, some of the examples of positive

statements on the organizational commitment scale developed by Allen and Meyer (1990) are "I enjoy discussing my organization with people outside it.", "I feel as if this organization's problems are my own." and "I would be thrilled to spend the rest of my career with this organization." As an example of negative statements are "I do not feel like 'part of the family' at my organization.", "I do not feel a strong sense of belonging to my organization." and "It would not be too costly for me to leave my organization now." All negative items on this scale are "Reverse Coded".

Reverse coded items could be used in scales prepared in the Likert type. One of the main reasons for using Reverse Coded Items is the idea that designing some propositions with negative expressions while determining the items will make the answer more reliable. When there are reverse-coded items in the scale, participants will not be able to establish a specific logic during the answering period and will carefully read all items to understand whether all items are positive or negative. In this case, the answers given to the items will be more accurate, and the validity of the scale will increase even more (Carifio and Perla, 2007). With this perspective, while many researchers develop scales, they prefer to use reverse coded items in scales. If the researchers know which items are reverse-coded in their study, they can reverse-scored them while transferring the data of the indicated items to the SPSS environment (i.e., 1 = 5, 2 = 4, 3 = 3. 4 = 2, 5 = 1). If the researcher does not know which of the items are reversed coded in the scale, it is useful to examine the correlations between the items of the scale. In this way, it can be understood which items are reverse coded items in the research. As an example, let us assume that the researcher aims to measure work stress and the Cronbach's alpha coefficient of the scale is determined as 0.220 as a result of the analysis performed. In order to make the control to increase the Cronbach's alpha coefficient, it is necessary to select the following module of SPSS program (Analyze / Scale / Reliability Analysis) and the related items in the scale should be assigned to the "Items" box in the next window that opens. In the next transaction, from the "Statistics" tab, "Scale", "Scale if item deleted" and "Correlations" should be marked. The output file after these operations is as follows.

Table 5. Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
JobStress1	12,8292	4,656	,326	,227	-,064[a]
JobStress2	12,6975	4,097	,504	,329	-,276[a]
JobStress3	12,8648	4,303	,377	,352	-,145[a]
JobStress4	12,6762	4,384	,377	,483	-,136[a]
JobStress5	13,6441	10,623	-,617	,429	,737

In the "Item-Total Statistics" table, the "Cronbach's Alpha if Deleted" column shows the Cronbach's Alpha value that can be obtained by deleting each expression. When the "Cronbach's Alpha if Item Deleted" column in the table is carefully examined, it can be seen that by deleting the expression "Job Stress 5", the Cronbach's alpha value of the scale will increase from 0.220 to 0.849 (it is useful to recall that the Cronbach's alpha value was measured as 0.220 in our first analysis). However, it is not the right approach to extract items from a scale whose Validity and Reliability have been previously tested. It would be an even worse approach to extract items from a scale where the number of items is so limited. Scale development studies are long-term projects conducted by a team of experts, including linguists, sociologists, and practitioners. Items in the created scale have a meaning and a reason. For this reason, an excellent theoretical framework should be developed to remove the item or items from the scale, and this should be effectively explained to the reader.

The "Corrected item-total Correlation" should be interpreted before the item is removed from the scale. When this column is analyzed, it is seen that the "Job Stress 5" item has a negative correlation with other items in the scale (r = -0,617). This shows that the related expression is reverse-coded. For this reason, the relevant item should be reverse coded before conducting the analysis. In the SPSS program, the

Cronbach's alpha value will be increased by reverse coding the relevant item (Transform / Recode into Different Variables or Transform / Recode into Same Variables). The new Cronbach's alpha value for the same example is shown below:

Table 6. Cronbach's Alpha Values

First Cronbach's Alpha Value
(Incorrect Coded Reverse Item)

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,220	,252	5

New Cronbach's Alpha Value
(Correct Coded Reverse Item)

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,792	,791	5

4. VALIDITY AND RELIABILITY THREATS IN STUDIES

There are many threats to Validity and Reliability in empirical research. For this reason, researchers should be aware of these threats and take appropriate measures to reduce bias in their studies. By taking these measures, it will be ensured that the research gives beneficial results. Below are the main types of bias that threaten the Validity and Reliability of the research.

4.1. Conceptual and Theoretical Bias

Conceptual and theoretical bias is caused by the fact that the researchers do not create a hypothesis according to the literature and theoretical framework. Researchers should thoroughly analyze the literature while creating a hypothesis in their studies and should form their hypothesis on a specific theoretical basis. Otherwise, the findings will be conceptualized and interpreted in the wrong direction. This bias is usually caused by the fact that researchers do not have full control of the research procedure or act in a biased manner. A common mistake is to obtain the data given by the questionnaire first and then to create the hypothesis according to the analysis. In this case, while creating a hypothesis, theoretical bias is encountered.

4.2. Sample Bias

Sample bias is related to the fact that the sample group included in the research does not represent the population. This bias occurs in two situations. In the first situation, sample bias occurs when the sample group included in the research is not similar to the universe in terms of structure and features. For example, in a study on the psychological effects of smoking, the sample group should be composed of participants who smoke. Another example of sample bias is where the sample group does not sufficiently represent the population (the number of samples is low).

4.3. Bias Caused by the Expectation of the Hypothesis to be Confirmed

The hypothesis is an estimate of the possible outcomes of the research. Hypotheses should be consistent with the objectives of the study, be prepared in a way that can be tested and measured, and should include all the variables to be used in the analysis. Correctly structured hypotheses are indicative of which statistical analysis method to use and which of the variables will be dependent and independent. For this reason, the researcher should state his/her hypotheses in the research. However, the idea that the hypotheses that the researchers determined for the research should be confirmed causes bias in the study. For this reason, researchers should share all hypotheses that are meaningful or meaningless as a result of statistical tests. If the hypotheses tested at the end of the research are not supported (meaningless), the researcher should explain this obtained result on a theoretical basis. The fact that researchers change their hypotheses or exclude unsupported hypotheses from the scope of the research in line with the obtained data as a result of the analyses causes bias. However, the findings obtained as a result of the research are essential. Findings that exceed the expectations are even more critical. Factors such as the sample group that the research was conducted on, the culture of the sample, or the time of the research may be essential details in terms of obtaining these findings. The researcher's interpretation of the findings on this theoretical basis will be an essential contribution to the literature.

4.4. Transaction Bias

Specifying the sample refers to the sum of all errors, from data collection to analysis. The fact that the measuring instrument used in the study is valid or reliable, the expectation of verifying the hypotheses and the application of manipulation or inappropriate analysis techniques performed in the analyses are actual examples of transaction bias.

The bias mentioned in the above literature can significantly affect the results of the research. It also threatens the Validity and Reliability of the findings obtained as a result of the research. For this reason, to reduce the validity and reliability threats in their research, researchers should: (1) clearly define the research problem, (2) construct hypotheses on a theoretical basis, (3) reach a sufficient sample size to represent the population, (4) select the sample group objectively, (5) use a valid/reliable measuring instrument to collect data, (6) analyze the data with appropriate statistical techniques and finally, (7) researchers should not expect to validate the hypotheses during the analyses. Thus, researchers can prevent Type I and Type II errors when interpreting the results.

5. CONCLUSION AND DISCUSSION

In quantitative research, most of the predictor and outcome variables are abstract concepts known as theoretical structures. The use of a valid and reliable measuring instrument to measure such abstract concepts is an essential factor in determining the quality of the research. This study emphasizes that the Validity and Reliability of the scales used in quantitative research are essential in addition to the creation of literature on Validity and Reliability. Therefore, if researchers pay attention to Validity and Reliability throughout their research, it is thought that valid and reliable findings will be obtained. The validity of the measurement instrument to measure accurately without confusion with another feature is defined as "validity". Validity is the degree to serve by the intended use of the scale. Reliability is that the measurement instrument gives consistent results under the same circumstances. When the existing definitions are examined, it is clear that Validity and Reliability are two critical features that should be present in every measurement instrument.

Evaluating studies with scales that do not have any of these features will not be ethically correct. Studies that are not ethically correct become scientifically controversial.

Many methods and techniques can be used to test the Validity and Reliability of the scales used in quantitative research. In this study, literature was created to inform researchers about the commonly used methods and techniques that are generally accepted in the literature; thus, this study furthers the understanding of Validity and Reliability, which is of great importance for researchers conducting and/or evaluating empirical research. Finally, it is strongly recommended in their studies that researchers should include the construct validity (convergent validity and discriminant validity), which are among the validity types, and methods of internal consistency (Cronbach's alpha and CR), which are among the Reliability tests.

REFERENCES

Allen, N. J., & Meyer, J. P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology*, *63*(1), 1-18.

Aksay, B., & Ünal, A. Y. (2016). Yapısal Eşitlik Modellemesi Kapsamında Formatif Ve Reflektif Ölçüm. *Çağ Üniversitesi Sosyal Bilimler Dergisi*, *13*(2), 1-21.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.

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.

Bollen, K. A. (1989). The consequences of measurement error. *Structural Equations with Latent Variables*, John Wiley & Sons, Inc., 151-178. https://doi.org/10.1177/1038416217724516.

Bowling, A. (2014). *Research methods in health: Investigating health and health services*. McGraw-hill education (UK).

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106-116.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Drost, E. A. (2011). Validity and Reliability in social science research. *Education Research and Perspectives*, *38*(1), 105.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis 5th ed. *New Jersey, NJ: Printice-Hall*.

Henson R.K., & Roberts J. K. (2006) Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. Educational and Psychological Measurement. 66(3).

Janssen, O. (2000). Job demands, perceptions of effort-reward fairness and innovative work behaviour. *Journal of Occupational and Organizational Psychology*, *73*(3), 287-302.

Judge, T. A., Thoresen, C. J., Pucik, V., & Welbourne, T. M. (1999). Managerial coping with organizational change: A dispositional perspective. Journal of Applied Psychology, 84(1), 107-122.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues*. Nelson Education. USA.

Kline, P. (1993). Psychometric theory and method. *The handbook of Psychological Testing*, 5-170.

Kuder, G.F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

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

Luthans, F., Avolio, B. J., Avey, J. B., & Norman, S. M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology*, *60*(3), 541-572.

Norman, G. R., & Streiner, D. L. (2003). *PDQ statistics* (Vol. 1). PMPH-USA.

Nunnally, J. C. (1978). Psychometric Theory McGraw-Hill Book Company. *INC New York*.

Oluwatayo, J. A. (2012). Validity and reliability issues in educational research. *Journal of Educational and Social Research*, *2*(2), 391-400.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94-104.

Ruzafa-Martínez, M., López-Iborra, L., & Madrigal-Torres, M. (2011). Attitude towards Evidence-Based Nursing Questionnaire: development and psychometric testing in Spanish community nurses. *Journal of Evaluation in Clinical Practice*, *17*(4), 664-670.

Swisher L.L., Beckstead J.W., Bebeau M.J. (2004) Factor analysis as a tool for survey analysis using a professional role orientation inventory as an example. *Physical The*rapy. 84(9):784-99.

Werts, C. E., Rock, D. R., Linn, R. L., & Jöreskog, K. G. (1978). A general method of estimating the Reliability of a composite. *Educational and Psychological Measurement*, *38*(4), 933-938.

Whiston, S. C. (2012). *Principles and applications of assessment in counseling*. Cengage Learning. USA.

Yeşilyurt, S., & Çapraz, C. (2018). Ölçek geliştirme çalışmalarında kullanılan kapsam geçerliği için bir yol haritası. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, *20*(1), 251-264.