Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 14(3) (2022) 2366- 2390

IJCI
International Journal of Curriculum and Instruction

A review of test use: The test anxiety inventory

Betül Alatlı [a] *

Balıkesir University, Faculty of Education, 10010, Balıkesir, Turkey

Abstract

This study was conducted to review the use of tests. For this purpose, 45 articles in which the Turkish form of the "Test Anxiety Inventory (TAI)," which is one of the tests frequently used in the field of education, was employed and that were published between 2000 and 2020 were examined in terms of factors that should be considered in the use of tests. A descriptive survey model was used in the study. Data were collected by using a "Control Form for the Use of Tests Adapted to the Target Culture", which was developed by the researcher. The findings indicated that in all studies, the feature intended to be measured with the inventory, the feature intended to be measured with the group, and the group that the test was applied were consistent. There were some incomplete or incorrect usages in terms of reporting the name of the TAI, the references related to the studies in which the inventory was developed and adapted, and the answers and scoring of the test. Only one study was found to present information about obtaining necessary permissions for the use of the inventory from the researchers who had adapted it. It was recommended that data about related properties should be reported and new analyses for validity and reliability should be conducted when necessary, even if the focus of test used in studies was not to determine psychometric properties. Accordingly, the validity and reliability data were mostly incomplete or not discussed at all in the articles examined within the scope of the study. The psychometric property that the researchers reported most and examined by using the data obtained from the related study was the "Cronbach's Alpha" coefficient. It was also noteworthy that there were inconsistencies between studies in reporting the psychometric properties of the TAI. Accordingly, in the case of test use, especially studies in which the test was developed in the related culture or studies in which the test was adapted to a specific culture should be reached as a primary source, and researchers should be supported by necessary education and sources in terms of factors that should be considered in the use of tests.

Keywords: Test use, Systematic Literature Review, Reliability, Validity, Test Anxiety Inventory

* Corresponding author Betül Alatlı. ORCID ID.: https://orcid.org/0000-0003-2424-5937
E-mail address: betulkarakocalatli@gmail.com

1. Introduction

Human beings need the knowledge to solve their problems. The most reliable way to reach knowledge, in other words, to solve factual problems, is the scientific method. Science, which is defined as the recognized body of knowledge, can be produced by scientific methods. The most common classification of science consists of natural sciences, social sciences, and mathematics. While mathematics is absolute, natural and social sciences are relative. The most important distinction between social and natural sciences lies in the difficulties in examining and controlling the content studied. Social science is the general name of the sciences that examine human relations and human behavior both socially and culturally. The novelty of social sciences and the difficulties that people may experience in examining their own species are among the most important limitations of this field of science (Karasar, 2009). Direct observation of human behavior has an important place in social science research. Direct observation of behaviors is a very time-consuming and expensive technique. Apart from this, it is impossible to make direct observations of some phenomena (Tezbaşaran, 2008). To measure the characteristics, such as intelligence, attitude, personality, belief, and perception, which cannot be observed directly and are frequently examined in social sciences, tests that contain items that stimulate the related feature are used. The general concept of "test" is used for measurement tools that are employed in education and psychology. These tests are frequently used by researchers because of their capacity to obtain valid and reliable measurement results, ease of application, and objectivity in scoring (Cronbach, 1951). Tests are considered very significant in large groups to achieve fast and scientifically valid, reliable data collection (Cronbach, 1960). Therefore, they are widely used in social sciences. With the increase in interaction in parallel with technological developments, the use of tests that have been developed in a certain culture and adapted to different cultures has been widely preferred in recent years (Deniz, 2007; Hambleton and Jong, 2003). In this study, the concept of "test" was used considering its general usage in the literature.

Tests have an important place in terms of their widespread use and advantages in social sciences, and it is equally important to develop them following the development and adaptation steps. Accordingly, many studies in the literature examine test development and adaptation studies in terms of related steps and many variables (Acar-Güvendir and Özer-Özkan, 2015; Çüm and Koç, 2013; Delice and Ergene, 2015; Deniz, 2007; Ergene, 2020; Hambleton, Meranda, and Spielberger, 2005; Karakoç and Dönmez, 2014; Murphy & Davidshofer, 2005; Şahin and Boztunç Öztürk, 2018; Yurdabakan and Çüm, 2017). Test development and adaptation studies should be carried out meticulously

to ensure that a test that has been developed or adapted and the scores obtained from this test are valid and reliable and that there are no problems with its use (Hambleton and Patsula, 1999; Şahin and Boztunç Öztürk, 2018) because test scores are used to make critical decisions about individuals. Similarly, the validity and reliability of tests have a very important place in the fairness and appropriateness of decisions made according to scores obtained from tests. Considering that there are 9016 test development and adaptation studies in the Turkish Measurement Tools Index (TOAD-2021 June) in Turkey, it can be seen that test use is highly prevalent and that tests are a very important field of study.

Adaptation studies are widely used for the comparison of individuals from different cultures in terms of personality, attitudes, success, and many psychological characteristics (Sireci and Berberoğlu 2000). For example, in addition to literacy tests in educational research, such as the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS), many scales are adapted to many cultures, and significant outputs can be obtained in terms of the education systems of countries. For example, 79 countries participated in the PISA 2018 application (MEB 2019), 64 countries participated in the TIMSS 2019 (MEB 2020), and 58 countries participated in the PIRLS 2016 (Mullis, Martin, Foy& Hooper, 2017); in other words, these scales were adapted to these countries. In addition, it is known that the "Wechsler Intelligence Scale for Children" is an intelligence test adapted to more than 50 cultures and that "Spielberger's Trait State Anxiety Inventory" is a personality inventory adapted to more than 50 cultures (Hambleton & Jong, 2003). Another personality test is the Big-Five Personality Traits Inventory. This inventory, which was developed by Benet-Martinez and John (1998), was adapted to 56 countries by Schmitt et al. (2007) within the scope of the “International Sexuality Description Project”. One more example of tests adapted to many cultures is the Test Anxiety Inventory (TAI). This test was developed following studies by Spielberger and a group of doctoral students at the University of South Florida between 1974 and 1979, and it was adapted to about 20 cultures. The test was developed to measure or assess test anxiety and was first published in 1980 together with a handbook. It was adapted to Turkish culture by Albayrak Kaymak (1985), and the handbook of the test was published by Öner (1990). It is a self-evaluation measurement tool. The English form of TAI was developed for university and high school students. In addition, 5th, 6th, 7th, and 8th-grade students were also included in the study group during its adaptation to Turkish (AlbayrakKaymak, 1985). This inventory is a measurement tool that can be administered to individuals aged 10 and older. The title of the inventory was changed to "Test Attitude Inventory" to prevent any negative situation that could be caused by the expression "anxiety" in students. TAI, which is a self-report psychometric scale, consists of 20 items, each of which has four choices [(1) almost never,

(2) sometimes, (3) often, and (4) almost always]. It has two factors, namely, “worry” and “emotionality”. The "emotionality" factor is based on the stimulation of the autonomic system, which constitutes the physiological and sensory aspect of test anxiety, while the "worry" factor is based on its cognitive aspect. The emotionality factor consists of ten items (items 1, 6, 7, 9, 10, 11, 13, 14, 15, 16, 18, and 19), and the worry factor consists of eight items (items 2, 3, 4, 5, 8, 12, 17, and 20). To examine the construct validity of the Test Anxiety Inventory, exploratory and confirmatory factor analyses were performed, respectively. The English form of TAI consists of two factors, too. As a result of the exploratory factor analysis conducted on data obtained from the Turkish version of TAI, the inventory was determined to have three factors: emotionality, nervousness, and worry. It is recommended that this three-factor structure should be preferred in clinical research or practical applications. The Turkish form of the inventory was tested for the emotionality and worry factors by using the confirmatory factor analysis. As a result, this two-factor structure was confirmed. The two-factor form is recommended for cross-cultural comparisons and studies examining the relationship between test anxiety and academic, cognitive, and emotional variables (Öner, 1990).

The scores obtained from the test include emotionality score (TAI-E), worry score (TAI-W), and the total score (TAI-T). Except for the first item, of the responses given to the rest of the items, “almost never (1)” shows high test anxiety, and “almost always (4)” indicates low test anxiety. The first item is reverse-scored and evaluated. The lowest and highest scores on the test range between 20 and 80 for the total score, 12 and 48 for the emotionality factor, and 8 and 32 for the worry factor. According to the scoring rules explained in the handbook of the inventory, if the count of invalid or incomplete answers on the entire inventory is more than two and it is more than one for each subtest, the form cannot be scored. However, for forms with one or two incomplete or invalid answers, a mean item value is calculated over the responded items, which is called "prorating". Accordingly, the mean value calculated is multiplied by 20, which is the total number of items, to obtain a total score. A similar process can be performed for subtest scores (Öner, 1990).

In the adaptation process of the test anxiety inventory, it was first translated into Turkish for translation and equivalence and then translated back into English by two American lecturers, who had a good command of the Turkish language. The back-translated and original English forms were compared and necessary corrections were made according to the differences between them. The English, Turkish, and two different mixed draft forms of the inventory were administered to 164 16-21-year-old female and male students who had a good command of English. Analysis of variance was used to compare the data obtained as a result of the administration of the four different forms. Accordingly, no significant differences were found between the groups that received

different forms. Thus, it was considered that the linguistic equivalence of the TAI was established (AlbayrakKaymak, 1985).

Cronbach's alpha coefficient was used to determine the reliability of the Test Anxiety Inventory as a measure of internal consistency. In the handbook of the inventory, reliability coefficients of the total test and subtests were presented for university, high school, and secondary school samples. The examination of Cronbach's alpha reliability coefficients that were calculated indicated that the highest value was 0.89 and it belonged to the university sample. The reliability of the emotionality subtest was found to be higher than that of the worry subtest. The lowest calculated Cronbach's alpha reliability coefficient was 0.73. In order to examine the item-level internal consistency, the item-total score correlation coefficients were calculated for the total test and subtests for all samples. When the calculated correlation coefficients were examined, it was determined that the lowest value was.46 for the overall test and .43 for the subtests. The test-retest method was used to examine reliability in terms of consistency and stability. Accordingly, the correlation coefficients calculated based on the administrations of the test at different intervals ranging from the same day to three weeks varied between .93 and .72. According to the reliability coefficients obtained, it can be interpreted that TAI is a reliable measurement tool.

In order to examine the criterion-based validity of the test anxiety inventory, the correlation coefficient between the total score and subtest scores on the inventory and the scores on the State-Trait Anxiety Inventory was calculated. The correlation coefficients between the TAI scores and the State-Trait Anxiety Inventory ranged from .39 to .70. This result was interpreted by the researchers who adapted the inventory as the quality of the test in meeting the theoretical expectations. However, the correlation coefficient between the scores on the Minnesota Counseling Inventory and its subtests and the scores on the TAI and its subtests, which ranged between r=.60 and .22, was found to show a positive and moderate relationship. In addition, the correlation coefficient between the TAI and the scores on the self-concept was found to range between r = -.31 and r = -.56 and evaluated as a negative and moderate relationship. In the adaptation study of the TAI, the relationship between students' grade point averages and average mathematics grades and the scores obtained from the TAI were examined, and the calculated correlation coefficients were found to be negative and range from an intermediate limit (r = -.43) to zero. Accordingly, the relationship between school success and the scores obtained from the TAI and worry subtest was determined to be significant, negative, and low. On the other hand, the relationship between the emotionality subtest of the TAI and school achievement was determined to be negative, low, and insignificant.

Very important and detailed information about the Test Anxiety Inventory was presented. In the process of test development or test use, the application and scoring of the test and reporting of the results are common processes that should be conducted by both developers and users. In addition, it is known that it is useful for test users to have some knowledge of test development principles so that they can better evaluate the test that they are planning to use. The test development process has certain standards, just like test use does (American Educational Research Association [AERA], the American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014). It is the test users' responsibility to use the test correctly. Individuals who will take responsibility for test use should have the necessary education and experience to handle it with the right technique and professional attitude, and they must be able to fulfill the qualifications specified in the handbook of the test. Test users should verify that the test is suitable for their intended use. They should evaluate written documents regarding validity and reliability in accordance with the intended use of the test. In some circumstances, it is recommended to obtain appropriate evidence for repeated use rather than existing validity and reliability data. Test users should ensure that if a significant change has been made in the language, scope, format, guidelines, or administration of the test, it should be validated for the relevant conditions, and if no additional validation has been done, reasonable justification should be provided. If the test has not been previously validated for its intended use, the user is responsible for providing information regarding validity. When the validity of the test is interpreted, general statements such as "validity of the test..." should be avoided since no test is valid for all situations or purposes. Similarly, statements like "the reliability of the test…." should also be avoided because reliability estimates are tested in many ways, each relying on different sources of error. If a test has subtest scores, reliability values should be reported for these scores as well. If a test is to be used for a purpose for which there is little or no evidence of validity, the user is responsible for documenting the rationale for the choice of the test and obtaining evidence of the reliability/precision of test scores and the validity of the test (Hovardaoğlu and Sezgin, 1998; the American Educational Research Association [AERA], the American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014).

It is important to remember that a test is neither reliable nor unreliable. Reliability is a property of scores on a test for a particular group of test-takers (Feldt and Brennan, 1989). Therefore, authors should provide coefficients of reliability for the scores of the analyzed data, even if the focus of their research is not psychometric. Researchers may be reluctant to change a previously developed measurement tool with weak psychometric properties after its introduction into the literature. In such cases, editors and reviewers should pay special attention to the psychometric properties of the measurement tools used and encourage revisions of such measurement tools (even if they

are not the developer of the scale) to prevent the accumulation of results based on relatively invalid and unreliable measurements (Wilkinson & APA Task Force, 1999). Before a scale that has been developed is used, the availability of the resources provided by the test developer, which summarize the objectives of the test, explain the procedures of implementation, describe the groups that take the test, and examine the validity and reliability of the interpretations of the scores obtained according to the available data, is very important for the test user. The user should carefully review and assess such resources provided by the test developer (AERA, APA, & NCME, 2014).

When studies on the use of scales are examined, it is seen that the reporting of reliability is a widespread practice. Wilson (1980) reviewed articles published in the American Educational Research Journal between 1969 and 1978. According to the results of the research, reliability coefficients were reported in 18% of them by referring to previous studies. The reliability coefficient in 37% of them was calculated and reported in line with the data obtained in the related study, whereas the reliability was not reported in 45%. Wilson (1980) recommended that researchers reported the reliability coefficients of their own data and that editors and reviewers rejected studies that did not report the psychometric properties of measurement tools (Green, 2011). Similarly, in their study on the use of scales in the articles published in the 1967, 1977, and 1987 volumes of the Journal of Counseling Psychology, Meier and Davis (1990) found that the psychometric properties of 95% of the scales in the 1967 volume, 85% of the scales in the 1977 volume, and 60% of the scales in the 1987 volume were reported. Vacha-Haase, Ness, Nilsson, and Reetz (1999) reviewed 839 articles published between 1990 and 1997 in three journals (Journal of Counseling Psychology, Psychology & Aging, and Professional Psychology: Research and Practice). Accordingly, it was determined that the reliability coefficient was reported in 36.4% of the articles based on the data obtained in the related research. Based on this, the editorial proposed changes in journal policies regarding the reporting of reliability coefficients. Vacha-Haase (1998) investigated 628 articles that had been published between 1984 and July 1997 and in which the Bern Gender Role Inventory had been used. According to the findings, 65.8% of the articles examined had not provided any data about the reliability of the related inventory. Reliability coefficients that had been obtained from the handbook or previous research were reported in 14.65% of the articles reviewed, and although no reliability coefficient was specified in 6.53% of the articles, it was reported that the reliability had been examined before. In 13.1% of the studies, reliability coefficients were calculated by using the data obtained in the related study, as recommended. Simmelink and Vacha-Haase (1999) selected 353 of 416 studies on the PsycINFO database, which had been published between 1958 and 1998 and contained the words "Rosenberg Self-Esteem Scale". Their findings indicated that reliability was not mentioned in 37.1% of the articles, the reliability obtained in previous studies was reported in 38.8%, and reliability coefficients

were reported based on the data obtained from the related study in 24.1%. Of the 816 articles that had been published between 1990 and 2000 and in which Spielberger's State-Trait Anxiety Inventory (STAI) was used, reliability was ignored in 73%, reliability coefficients previously reported in other sources were reported in 21%, and reliability coefficients calculated using the data obtained in the related study were reported only in 6% (Barnes, Harp, and Jung, 2002). When the relevant literature is examined, it is seen that studies on the reporting of reliability are widespread. In the current study, studies in which the Test Anxiety Inventory had been used were discussed in terms of the reporting of all psychometric properties.

Developed and/or adapted tests are frequently used in scientific research. Regarding the test decided to be employed in terms of the purpose of the research and the features to be measured, a detailed examination is required in terms of the features to be measured with the test, the group whose features are measured, and the psychometric properties of the test. It is expected that the compliance of the test with the research in which the test will be employed and the acceptability of the psychometric properties of the test are addressed in detail with the examinations done on the test. Only then the test, which is determined to be appropriate in line with these criteria, is considered appropriate for use in research. In addition, examinations of the psychometric properties of the test should be done within the scope of the related research, too. Considering the frequency of test use in social sciences and the decisions made in line with the scores obtained from tests, the points to be considered in the use of tests are considered very important. This study was conducted to discuss the process of test use through the Test Anxiety inventory to set an example for researchers. In this way, the relationship between test characteristics that needed to be reported and that was reported was clearly demonstrated. In the study, the articles in which the "Test Anxiety Inventory", which is frequently used in the literature and was adapted to Turkish culture, was used were examined in terms of factors to be considered during the use of a scale adapted to the related culture.

2. Method

A descriptive survey model was used in the study since it was conducted to reveal the tendency towards the use of an adapted scale. In descriptive studies, an existing phenomenon or characteristics of an individual or group are described as they are, and the existing situation is described quantitatively or qualitatively (Karasar, 2009).

2.1. Study Group

The study group consisted of 45 articles in which the Turkish adaptation of the Test Anxiety Inventory was used, which were published between 2000 and 2020, and

which had an accessible full text. First of all, the articles in which the inventory was used were searched by using the keyword string "Test Anxiety Inventory" and its Turkish translation. Afterward, 45 articles were randomly selected among studies published between 2000 and 2020 in which the Turkish adaptation of the Test Anxiety Inventory was employed. The distribution of the articles examined within the scope of the research by the year of publication, the index of the journal in which they were published, and the language of publication is given in Table 1.

Table 1. *Distribution of articles examined in the study by some variables*

Variable	f	%
Year		
2000-2005	3	6.67
2006-2010	6	13.33
2011-2015	11	24.44
2016-2020	25	55.56
The index of the journal		
Ulakbim	17	37.78
SSCI	6	13.33
Field index	2	4.44
Other	20	44.44
Language of the article		
Turkish	33	73.33
English	12	26.67
Total	45	100.00

As seen in Table 1, the majority of the articles (55.56%) examined in the study had been published between 2016 and 2020. The count of articles using the TAI indicated an increasing trend from 2000 to 2020. Of the total articles, 44.44% had been published in other indexes, 37.78% in Ulakbim, 13.33% in Social Sciences Citation Index (SSCI), and 4.44% in the Field index. Regarding the publication language of the articles, 73.33% had been published in Turkish and 26.67% in English. Since the Turkish form of the articles that had been published in both languages was examined, they were coded as Turkish publications.

2.2. Data Collection

The document review data collection technique was used to collect the study data. With the document review technique, data can be collected by examining existing records or documents (Karasar, 2009). For this purpose, the study data were collected by examining the full texts of the studies in which the Turkish form of the Test Anxiety Inventory was used for data collection. In the study, the "Control Form for the Use of Tests Adapted to the Target Culture", which was developed by the researcher, was employed. In the development of this control form and writing its items, the features expected from a valid and reliable test adapted to the target culture, as well as the important points to be considered in the use of a test in a study, were taken into account. After the items of the form were written, nine experts who completed their doctoral

education in the field of measurement and evaluation were consulted about the suitability of the items, suggestions for corrections, and items that should be added. The control form was finalized by making the necessary corrections in line with the expert opinions. The form is presented in Annex 1.

2.3. Data analysis

In this study, we analyzed 45 articles in which the Turkish version of the Test Anxiety Inventory was used to collect data. Data were analyzed by using categorical analysis and frequency analysis, which are among the content analysis methods. Content analysis is based on systematic coding within the framework of certain classifications and themes regarding qualitative or quantitative data (Cohen et al., 2011; Fraenkel et al., 2012). Through categorization, data is coded. Frequency analysis is used to present the scores of a variable as percentages, proportions, or counts according to the distribution characteristics. The data density of this variable expresses its significance (Büyüköztürk, 2007). Microsoft Office Excel 2010 software was used for content analysis. In this study, the items on the control form developed by the researcher were taken into consideration for content analysis. Frequency and percentage values were obtained for the items according to the coding made later.

For the reliability of the study, five of the 45 articles that were selected randomly were subjected to content analysis by a measurement and evaluation expert. The reliability coefficient formula developed by Miles and Huberman (1994) is as follows: "Confidence Coefficient = Count of Agreements / (Count of Agreements + Count of Disagreements)". The reliability estimation was performed by using this formula. The reliability coefficient calculated according to this formula was found to be 0.91. A reliability coefficient of 0.90 or greater is considered a high level of reliability (Cohen &Swerdlik, 2010). Accordingly, it can be said that the study had a high level of reliability.

3. Results

Within the scope of the research, 45 articles in which the Turkish adaptation of the "Test Anxiety Inventory" was used were examined. In the process of examining the articles, the findings and interpretations were discussed according to the "Control Form for the Use of Tests Adapted to the Target Culture", which was developed by the researcher. Table 2 shows the distribution of data for reporting the name of the test correctly.

Table 2. *Distribution of the articles in terms of reporting the name of the test*

Variable	f	%
Correctly reported	37	82.22
Partly correct	8	17.78
Total	45	100.00

As seen in Table 2, the name of the measurement tool had been correctly given as the "Test Anxiety Inventory" in 82.22% of the articles examined within the scope of the study. In 17.78% of the articles, the name of the measurement tool was partly correct. For example, the name was partly correct as “Test Anxiety Scale” in three articles (articles 37, 30, 28), as “Examination Anxiety Inventory” in article 33, as “Test Anxiety Scale” in article 16, as “Exam Anxiety Inventory” in articles 5 and 13, and as “Exam Attitude Inventory” in article 8. Since the expression "anxiety" was thought to evoke negative feelings in students, it was recommended to use the name of the inventory as "Test Attitude Inventory" during the application process (Öner, 1990). There was no explanation in the studies as to whether this was taken into account.

The distribution of the articles was examined in terms of consistency between features that were intended to be measured in the study in which the test was used and the features that the test measured. Accordingly, it was determined that the features that were intended to be measured in all of the 45 studies in which the TAI that was examined within the scope of the study was used were consistent with the test used. Table 3 presents the distribution of the articles in terms of correctly reporting the source study in which the TAI had been developed.

Table 3. *Distribution of the articles in terms of reporting the source study in which the test was developed*

Variable	f	%
Correct	30	66.67
Incorrect	8	17.78
Not specified	7	15.56
Total	45	100.00

As seen in Table 3, in 66.67% (f=30) of the 45 articles, in which the Turkish adaptation of the “Test Anxiety Inventory” was used, the reference for the study in which the scale had been developed was cited as “Spielberger (1980)”, which was the reference specified in the handbook of the inventory. However, 18.18% (f=8) of the articles were found to incorrectly cite the source study in which the measurement tool had been developed. For example, in studies 33 and 22, the author of the study in which the Test Anxiety Inventory had been developed was cited as Spielberger, but another source published in 1972 was cited. In articles 11 and 27, the study by "Schwarzer, Van Der Ploeg, and Spielberger (1987)" was cited as the development study of the inventory. In article 15, both the name of the author and the year of the study were cited incorrectly as "Spielberg (1962)". In article 12, "Spielberger and Vagg (1975)" was cited as the study in which the inventory had been developed. In article 10, the inventory was reported to have been developed by C.D. Spielberger et al., but the year was not specified. Moreover,

15.91% (f=7) of the articles reviewed had not included any information about the development study of the related inventory.

Table 4. Distribution of the articles in terms of reporting the study in which the test was adapted

Variable	f	%
Correct	36	80.00
Incorrect	5	11.11
Not specified	4	8.89
Total	45	100.00

In studies in which tests adapted to a different culture are used, the adaptation study should be examined in detail, and information about the source and quotations from the source should be handled correctly. The examination of the 45 articles in which an adapted measurement tool had been used indicated that 80.00% (f=36) of the articles reported the source of the "Handbook of the Test Anxiety Inventory," which is the source of the adaptation study, as Öner (1990) correctly (Table 4). The article "Albayrak Deniz, 1987," which included the adaptation study of the inventory and which was also included in the handbook of the inventory, and the thesis study "Albayrak Deniz, 1985" were also discussed in this review. However, in 11.11% (f=5) of the articles examined within the scope of the current study, the source of the adaptation study of the measurement tool that was used was included incorrectly. For example, there was a statement in article 37 as follows: "the Test Anxiety Scale adapted to Turkish by Öner." Also, the publication year related to the source was not cited. In article 26, the source was cited as "Öner and Albayrak Kaymak (1986)", but the publication year was given incorrectly. In article 40, the source for the adaptation study was given as "Öner (1989)". Similarly, in article 2, the publication year was given incorrectly as "Öner, (1980)."

After a suitable measurement tool is determined for the feature that is intended to be measured in a study, the permission of the authors who adapted the test - if it was adapted to the related culture - and the authors who developed it - if it was developed in the related culture - should be obtained. One of the important features of a study includes its reproducibility so that it can be robust in the face of all criticism (Karasar, 2009). The steps taken in the research process should be clearly stated. The articles analyzed within the scope of this research were also examined in terms of whether necessary permissions were obtained from the authors who had adapted the "Test Anxiety Inventory" were included. The findings obtained showed that only one of the 45 articles examined within the scope of the research included the permission statement for the use of the test. In article 11, the permission procedure was stated as "Permission for the use of the TAI was obtained from the YÖRET foundation, to which Necla Öner, who carried out the adaptation study of the scale, transferred the right to use the scale."

Another point to be considered in terms of test use in studies is that the group to which the test is adapted must be compatible with the study group or sample of the research. If the study group is not similar, validity and reliability studies should be repeated (AERA, APA, and NCME, 2014). Accordingly, it was determined that in all of the articles examined within the scope of the research, the group to which the test was adapted and the group to which the test was administered were compatible. It is thought that this may be related to the fact that the test anxiety inventory is a measurement tool suitable for a wide range of age groups. The TAI was developed for university and high school students, and in the process of adapting it to Turkish culture, 5th, 6th, 7th, and 8th-grade students were included in the study group. It was stated that the test could be used for all students aged 10 and older (Öner, 1990). Table 5 shows the findings of 45 articles examined within the scope of the research in terms of data about the answers given to the test.

Table 5. *Distribution of data about the answers given to the test*

Variable	f	%
Correct	10	22.22
Partly correct	20	44.44
Incorrect	1	2.22
Not specified	14	31.11
Total	45	100.00

The answers given to the items in the Test Anxiety Inventory consist of four choices: “almost never,” “sometimes,” “often,” and “almost always”. The respondent is expected to choose one of the appropriate answers. While 31.11% (f=14) of the 45 articles in which the related test was used did not include any information about answering the test, 22.22% of them were found to provide accurate information about the answers and answering the inventory (Table 5). However, information about answering the inventory was partly correct in 44.44% (f=20) of 45 articles that used the Turkish adaptation of the test anxiety inventory. For example, in article 37, it was described as follows: "a Likert-type scale that is scored between 1 and 4 and that individuals can apply on their own", and in article 32, it was described as in the following statement: "the TAI is a four-point Likert-type scale". In article 12, it was described as follows: “It is a psychometric scale using the self-report method. The inventory has a 4-point Likert-type scale and comprises 20 questions,” but no information was provided about what the options were. In article 15, while information about responses to the items was presented as follows: "The Test Anxiety Inventory consists of 20 items, each of which is a Likert-type with 4 points (1 = almost never; 4 = almost always)", the expression "a Likert type with 4 points" was not clearly stated. In article 17, it was described as follows: "The test, which is a 4 point

Likert-type scale", which did not provide enough information about responses to the items. In article 29, the expression "TAI is a four-point Likert-type inventory" was used. In article 24, it was described as a "Likert-type measurement tool". The rate of articles that provided inaccurate information about answers and answering the items was 20% (f=9). In article 44, it was stated that the test was a five-point Likert type (Table 5). The articles in which the related inventory was used were examined regarding also how the answers given to the test anxiety inventory were scored. The distribution of information regarding the scoring of the responses given to the items of the TAI in the reviewed articles is given in Table 6.

Table 6.*Distribution of the information regarding the scoring of the answers given to the items of the test*

Variable	f	%
Correct	6	13.33
Partly correct	25	55.56
Incorrect	1	2.22
Not specified	13	28.89
Total	45	100.00

When Table 6 was examined, it was found that only 13.33% of the studied (f=6) provided accurate information regarding the item and total scores of the TAI. However, 55.56% (f=25) of the articles examined within the scope of the study provided partly correct information regarding the scoring of the inventory. For example, in article 16, it was described as follows: “Participants responded on a 4-point scale ranging from 1 = Never, 2= Sometimes, 3= Often, to 4= Always,” and in article 12, it was stated as “the inventory has 4-point Likert type and comprises 20 questions.” In article 22, it was stated as “the weights of the answers given to the questions varied between 1 and 4 points. The lowest score to be obtained from the overall test is 20, and the highest score is 80," which was incomplete. In only one article, the TAI was described incorrectly as in the following statement: “The TAI is a 5-point Likert-type scale.”

None of the 45 articles reviewed within the scope of this research provided any information about the linguistic equivalence of the TAI, which is an important step in the adaptation process. According to the construct validity of the test anxiety inventory, the factor structure revealed in the development of the test and confirmed in the adaptation study included two factors, namely "worry" and "emotionality". The “emotionality” factor was based on the stimulation of the autonomic system, which constitutes the physiological and sensory aspect of test anxiety, while the “worry” factor was based on the cognitive aspect of it. The emotionality factor consisted of 12 items (items 1, 6, 7, 9, 10, 11, 13, 14, 15, 16, 18, and 19), and the "worry" factor consisted of eight items, (items 2, 3, 4, 5, 8, 12, 17, and 20). Table 7 shows the distribution of the information on the

construct validity of the inventory that was provided in the articles in which the test anxiety inventory had been used.

Table 7.*Distribution of the information on the reporting of the construct validity of the test*

Variable	f	%
Correct	6	13.33
Partly correct	31	68.89
Not specified	8	17.78
Total	45	100.00

As seen in Table 7, only 13.33% (f=6) of the studies provided complete and accurate information about the construct validity of the inventory. On the other hand, the majority of the studies in which the related test had been used reported incomplete information about the construct validity of the inventory (68.89%, f=31). For example, in article 8, this was reported as follows: “The scale has a total of 20 items and consists of worry (8 items) and emotionality (12 items) sub-dimensions, which represent two different dimensions.” The statement in article 12, "There are two sub-dimensions called worry and emotionality”, only provided the names of the sub-dimensions but did not include any information about the items under these sub-dimensions and how these dimensions were defined. Also, it was found that 17.78% (f=8) of the articles did not provide any information about the construct validity of the measurement tool. Validity is defined as the degree to which the test or the scores obtained from the test serve its purpose (Cronbach, 1990; Tekin, 2012). For this reason, it is expected that the structure of the test is confirmed with the data obtained in the related research to ensure that the decisions made according to the scores obtained from the tests are appropriate. The confirmatory factor analysis was not used in any of the articles, in which the TAI was used, to test whether the structure of the inventory was confirmed by the data obtained. Table 8 shows the distribution of articles that reported information about the criterion-based validity of the test anxiety inventory.

Table 8.*Distribution of information about criterion-based validity of the test*

Variable	f	%
Partly	13	28.89
No information	32	71.11
Total	45	100.00

During the examination of the criterion-based validity of the test anxiety inventory, its relationship with scores on the State-Trait Anxiety Inventory and the Minnesota Counseling Inventory, self-concept, GPAs, and grade averages from

mathematics was examined. As seen in Table 8, information about the criterion-based validity of the inventory was partly included in 28.89% (f=13) of the articles examined within the scope of this research. Regarding criterion-based validity, it was determined that information about achievement scores was not mentioned generally. For example, information about criterion-based validity that was provided in article 44 did not include achievement score relationships as in the following statement: "The TAI had positive correlations with trait anxiety (0.48), prior to testing state anxiety (.51), anxiety scale of the MMPI (0.27-0.46) and with issues on the student problem checklist (.27-.60). It had moderate-negative correlations with self-concept (-0.31-0.56)." Table 9 presents the distribution of studies in terms of reporting information about the internal consistency of the TAI.

Table 9. *Distribution of information about the internal consistency reliability of the test*

Variable	f	%
Partly	28	62.22
Not specified	17	37.78
Total	45	100.00

In the adaptation study, data about Cronbach's Alpha reliability coefficient, which was calculated to examine the reliability of the Test Anxiety Inventory, were presented for both the subtests and the total test. Some of the information regarding the internal consistency reliability of the inventory was included in 62.22% (f=28) of the 45 articles examined within the scope of this research (Table 9). The reliability coefficients for the subtests were not included. For example, in article 19, it was stated that "the internal consistency of the inventory varied between .89 and .73." It was found that the information given in the articles was inconsistent in terms of internal consistency reliability and that some of them were confused with other reliability coefficients. For example, the statement in article 36, "KR-20 value was between 0.70 and 0.90", actually reported values belonging to the test-retest reliability of the scale. In article 11, while the reliability coefficients for the total scale were given, they were not evaluated for subtests. The statement in article 2, "It was determined that the internal consistency coefficients calculated to determine the internal consistency of the scale items on the TAI ranged from .93 to .94", provided quite limited information. These coefficients were the values that had been obtained during the development of the inventory. The reliability coefficient reported in article 38 was not clearly specified whether it represented internal consistency or test-retest, as in the following statement: "A reliability study about TAI has been conducted and reliability coefficient has been found as, r = .95." On the other hand, it was seen that the information given about the internal consistency reliability of the Test Anxiety Inventory differed considerably in the articles, and no examples were

found in this sense. In some articles, it was seen that "Cronbach's alpha coefficient," which is used to examine internal consistency reliability, and the "Item-Total test correlation coefficient," which is calculated during the item analysis process, had been confused. For example, the information regarding the reliability of the scale was partly included in article 26 as in the following statement: "It was found that the internal consistency coefficients did not fall below .46 for the overall test and .43 for sub-dimensions." Reliability is defined as the degree of the error-free status of the test and the scores obtained from it (Cronbach, 1990; Tekin, 2012). Accordingly, it is recommended to conduct a separate examination for the reliability of the data obtained while using a test. Table 10 presents findings of whether an analysis of the internal consistency reliability of the TAI was conducted in the articles reviewed within the scope of the research.

Table 10. *Distribution of the studies in terms of whether an internal consistency reliability analysis was conducted*

Variable	f	%
Yes	12	26.67
No	33	73.33
Total	45	100.00

As seen in Table 10, 26.67% (f=12) of the articles that were reviewed within the scope of this study had examined the internal reliability of the Test Anxiety Inventory. For example, article 32 reported the findings of the examination as follows: "Cronbach's alpha values were .86 for the emotionality sub-dimension, .79 for the worry sub-dimension, and .90 for the total score". In article 22, the coefficient obtained within the scope of the study was reported as "Cronbach's alpha coefficient for the overall test was found to be .87 according to the scores obtained from the study group in this study." However, the majority of the articles (73.33%, f=33) had conducted no additional examination of the internal consistency reliability of the TAI (Table 10). The distribution regarding the reporting of test-retest reliability of the test is given in Table 11.

Table 11. *Distribution of information provided about the test-retest reliability of the test*

	f	%
Correct	14	31.11
Partly correct	3	6.67
Not discussed at all	28	62.22
Total	45	100.00

It was determined that the correlation coefficients calculated based on the administrations of the Test Anxiety Inventory at times ranging from the same day to

three-week intervals varied between .90 and .70 (Öner, 1990). Accordingly, the examination of the 45 articles, in which the related inventory had been used, within the scope of this study indicated that 31.11% (f=14) of the articles provided complete data about the test-retest reliability of the inventory (Table 11). However, 6.67% of the articles reported incomplete data. For example, it was stated in article 14 that the "test-retest reliability conducted at a two-week interval was .81." In article 15, the same issue was reported as follows: the "test-retest relationship for the Test Anxiety Inventory was .80." These are incomplete data about the test-retest reliability of the inventory. According to the findings in Table 10, it was seen that the test-retest reliability of the inventory had not been mentioned at all in 62.22% (f=28) of the articles in which the TAI was used. No additional examination was conducted for the test-retest reliability of the TAI in any of the 45 articles reviewed within the scope of the current study.

4. Discussion and Conclusion

In this study, a total of 45 articles in which the Turkish adaptation of the 'Test Anxiety Inventory' had been used as a data collection tool were examined in line with important points to be considered in the use of a test adapted to another culture. The findings obtained in the study were discussed in line with the control form developed by the researcher following expert opinions. The results of the study were also included under this heading

In the study, first of all, the articles were examined regarding the correct use of the name of the test. Accordingly, it was determined that most of the studies reported the name of the "Test Anxiety Inventory" correctly. However, it was also revealed that the name of the test was reported incorrectly in some articles due to researchers' carelessness about the distinction between a scale and an inventory or translation errors. It was found that the agreement between the features measured by the inventory and the features intended to be measured was achieved in all studies. Compliance with the test is very important in terms of the appropriateness of the decisions to be made regarding the features to be measured.

In some articles examined within the scope of the research, although the study "Spielberger (1980)," in which the Test Anxiety Inventory was developed, was reported correctly, it was seen that some of them reported it incorrectly or did not report it at all. This result showed that studies in which the related measurement tool was used, not the handbook or the study in which the measurement tool had been developed, were cited as a reference. Similarly, the reference for the adaptation study of the Test Anxiety Inventory was either cited incorrectly or not cited at all in one-fifth of the articles examined. The adaptation study should be the most basic source to be referenced when a measurement tool adapted to another culture is employed in a study. This showed that

researchers cited studies in which the measurement tool was used, not the adaptation study.

Permission for the measurement tools used in studies should be taken from the authors who developed it - if it was developed in the related culture - or from the authors who adapted the scale - if it was adapted to the related culture, and this must be specified in the research report. However, only one of the 45 articles examined within the scope of the research was found to provide information about permissions. This situation leads to considerable doubts about whether the researchers obtained permission for the use of the measurement tool.

The Test Anxiety Inventory is a measurement tool that can be applied to all individuals aged 10 and over. It was seen that the inventory was used in appropriate age groups in all of the articles examined within the scope of the research. This may be related to the usability of the TAI in a wide age range. Information about the responses to the Test Anxiety Inventory was given correctly only in one-fifth of the studies examined. It was found that the majority of studies provided partly correct or no information about the inventory. For example, some of them stated that the inventory had a five-point Likert-type structure, which is incorrect because it is a four-point Likert-type scale. This also raises doubts about whether the answers to the form were used correctly. A fairly extensive explanation is found in the handbook for scoring the Test Anxiety Inventory. Researchers obtain research findings, results, and comments in line with the scores obtained from the tests they use. For this reason, how the scores obtained from the measurement tool are calculated, in other words, how the items are scored, is one of the important issues. In approximately one-third of the articles examined within the scope of the research, no information was given regarding the scoring of the inventory. For example, the first item of the TAI needs to be reverse scored. The points to be considered in scoring should be addressed in related studies.

In this study, we investigated whether researchers in the studies reviewed had conducted a validity analysis of the interpretations made according to the scores obtained from the scales. In addition, the 45 articles in which the TAI was used were also examined in terms of whether the studies provided information about the validity obtained from the handbook of the measurement tool or related studies. Accordingly, approximately one-fifth of the articles in which the Test Anxiety Inventory was used had not included any information about the construct validity of the inventory. On the other hand, the majority of studies were found to partly include data about construct validity. None of the studies had provided an analysis of whether the structure of the test was confirmed by using data obtained in the related study. There are very important findings in the handbook of the TAI in terms of criterion-based validity. However, in the majority of studies in which the TAI was used, no information was included regarding the criterion-based validity of the inventory. It is recommended that information about the

psychometric properties of tests used in studies should be provided accurately and completely. Test users should evaluate the sources of information about the objectives of a test that is developed, its administration process, the group taking the test, and its validity and reliability. When a test is translated and adapted from one language to another, test users are responsible for describing methods that have been used to determine the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for its intended use (AERA, APA, NCME, 2014). Especially, since it is known that validity and reliability are features based on scores obtained from tests, it is recommended that psychometric analyses should be performed separately in studies in which the test is used. It is known that there is a very important consensus in the literature that validity is not about tests, but interpretations made according to the measurements obtained from tests (AERA, APA, & NCME, 1999, 2014; Cizek, 2016; Cronbach, 1971; Kane, 2006; Messick, 1989; Newton, 2012; Reynolds, Livingston, and Wilson, 2006). Maciver, Anderson, Costa, and Evers (2014) revealed that the concept of "user validation", which is based on the precision, effectiveness, and validity of interpretations based on measurements obtained from a test, is much more compatible with the contemporary definition of validity.

The examination of articles in terms of reliability indicated that there was inconsistency between the coefficients given for reliability, especially in terms of internal consistency. In one-third of the articles examined within the scope of this research, it was found that no information was given about the reliability of the TAI in terms of internal consistency found in the handbook or any other research. Based on the data obtained from the studies in which the inventory was used, it was determined that the internal consistency reliability was examined in only about a quarter of the studies. Regarding the reliability of the test anxiety inventory in terms of consistency and stability by using the test-retest method, no information was mentioned in the majority of the articles examined within the scope of this research. Also, there were inconsistencies between especially reliability coefficients that were reported. Similarly, test users' knowledge of the reliability of tests was mostly incomplete or was not referred to at all. In particular, according to the findings obtained from related studies, it was found that the lack of an additional reliability analysis was rarely discussed (Barnes, Harp, and Jung, 2002; Meier and Davis, 1990; Simmelink and Vacha-Haase, 1999; Vacha-Haase, 1998; Vacha-Haase, Ness, Nilsson, &Reetz, 1999; Wilson, 1980).Similar to validity, reliability is also defined as a feature of the scores obtained from measurement tools (AERA, APA, & NCME, 1999; Bademci, 2007; Brookhart &Nitko, 2008; Gronlund& Linn, 1990; Feldt& Brennan, 1989; Linn & Miller, 2005; McMillan, 2007; Nilsson, Schmidt & Meek, 2002; Reynolds, Livingston, &Willson, 2009; Thompson, 2003). For this reason, researchers are advised to examine the reliability coefficients related to the scores obtained from the measurement tools (Wilkinson & APA Task Force, 1999).

According to the results obtained from the research, it was determined that information about the name of the test that researchers used, its purpose, the resources related to the development or adaptation study, answers, scoring, and psychometric properties of the test was mostly incomplete, incorrect, or not mentioned at all. Also, some inconsistencies were determined in terms of coefficients between some studies. Accordingly, the handbook of the tests used in studies, the development study if it has been developed in the related culture, or the adaptation study if it has been adapted to the related culture, should be used as the primary source. Researchers developing or adapting tests should ensure that test users can access the handbook of the test and full texts of development or adaptation studies easily. Even if the focus of the research is not to determine the psychometric properties, the researcher should provide evidence for the reliability of the interpretations made based on the scores obtained from the related test (Wilkinson & APA Task Force, 1999). In the studies examined in the current study, it was determined that only one study reported information about obtaining permission for the use of the test. Necessary permission for the use of tests must be obtained from the researchers who developed or adapted them. Researchers can be educated on the importance of test use as well as test development or adaptation steps. This study can be repeated with larger data for different tests.

References

AERA, APA ve NCME (2014). *Standards for educational and psychological testing.* Washington: American Educational Research Association.

Albayrak, Kaymak D. (1985). The development of the Turkish form of the Spielberger test anxiety inventory: A study of transliteral equivalence and reliability. [Unpublished master thesis], Bogazici University.

Albayrak, Kaymak D. (1987) Sınav kaygısı envanterinin Türkçe formunun oluşturulması ve güvenirliği. *Psikoloji Dergisi, 6*(21):55-62.

Bademci, V. (2007). Ölçme ve Araştırma Yöntembiliminde Paradigma Değişikliği: Testler Güvenilir Değildir. Ankara: Yenyap Yayınları.

Bademci, V. (2019). Geçerlik: Nedir? Ne değildir? *Eğitim ve Toplum Araştırmaları Dergisi, 6*(2), 373-385. **https://dergipark.org.tr/en/download/article-file/904540**

Barnes, L. L. B. , Harp, D. & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*(4), 603-618. **https://doi.org/10.1177/0013164402062004005**

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Ankara: ÖSYM Yayınları

Brookhart, S. M. & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, New Jersey: Pearson/Prentice Hall.

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı; istatistik, araştırma deseni, SPSS uygulamaları ve yorum*. Ankara: Pegem Yayıncılık.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212-225. **https://doi.org/10.1080/0969594X.2015.1063479**

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). London: Routledge.

Cohen, R. J. & Swerdlik, M. E. (2010). Psychological testing and assessment. New York: McGraw-Hill Higher Education.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.) (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1990). *Essentials of psychological testing* (2nd ed.). New York:Harper Collins.

Deniz, Z. (2007). Psikolojik ölçme aracı uyarlama. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *40*(1), 1-16. **https://dspace.ankara.edu.tr/xmlui/bitstream/handle/20.500.12575/46794/1100.pdf?sequence=1**

Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: Mc Graw HIll.

Gronland, N. E. & Linn, R. L. (1990). *Measurement and evaluation in teaching*. (Sixth Edition). New York: Macmillan.

Hambleton, R. K. & De Jong, J.H.A.L. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-134. **https://doi.org/10.1191/0265532203lt247xx**

Hambleton, R.K. ve Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1*(1), 1-30. **http://www.jattjournal.com/index.php/atp/article/view/48345**

Hovardaoğlu, S. ve Sezgin, N. (Çevirenler) (1998). *Eğitimde ve psikolojide ölçme standartları*. Ankara: Türk Psikologlar Derneği ve ÖSYM Yayınları.

IEA (2021).Progress in International Reading Literacy Study 2021, 20 Years of Trends in Reading Achievement. **https://www.iea.nl/studies/iea/pirls/2021**

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education & Praeger.

Karasar, N. (2009). *Bilimsel araştırma yöntemi.* Ankara: Nobel Yayınları.

Linn, R. L. & Miller, M. D. (2005). *Measurement and assessment in teaching.* (Ninth Edition). Upper Saddle River, New Jersey: Merrill.

Maciver, R., Anderson, N., Costa, A.C., & Evers, A. (2014). *Validity of Interpretation: A user validity perspective beyond the test score. International Journal of Selection and Assessment, 22(2), 149–164.* **https://doi.org/10.1111/ijsa.12065**

McMillan, J. H. (2007). *Classroom assessment. Principles and practice for effective ınstruction.* (Fourth Edition). Boston: Allyn and Bacon.

MEB (2019). PISA 2018 Türkiye Ön Raporu. **http://pisa.meb.gov.tr/wp-content/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf**

MEB (2020). TIMSS 2019 Türkiye Ön Raporu. **http://www.meb.gov.tr/meb_iys_dosyalar/2020_12/10173505_No15_TIMSS_2019_Turkiye_On_Raporu_Guncel.pdf**

Meier, S. T., & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37,* 113-115. **https://doi.org/10.1037/0022-0167.37.1.113**

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed.) (pp. 13- 103). New York: American Council on Education and Macmillan Publishing Company.

Miles, M. & Huberman, M. A. (1994). *An expanded source book qualititive data analysis.* Landon: Sage Publications.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Boston College, TIMSS & PIRLS International Study Center **http://timssandpirls.bc.edu/pirls2016/international-results/**

Newton, P. E. (2012). Clarifying the consensus definition of validity: Commentary. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 1–29. **https://doi.org/10.1080/15366367.2012.669666**

Nilsson, J. E., Schmidt, C. K. & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement, 62*, 647-658.

Öner, N. (1990). *Sınav kaygısı ölçeği el kitabı*. İstanbul: Yöret Yayınevi.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education*. Boston: Pearson.

Schmitt, D. P., & International Sexuality Description Project. (2003). Universal sex differences in the desire for sexual variety: Tests from 52 nations, 6 continents, and 13 islands. *Journal of Personality and Social Psychology, 85*(1), 85–104. **https://doi.org/10.1037/0022-3514.85.1.85**

Simmelink, S., & Vacha-Haase, T. (1999). Reliability generalization with the Rosenberg Self-Esteem Instrument. Paper presented at the annual meeting of the Rocky Mountain Psychological Association, Fort Collins, CO.

Sireci, S.G. ve Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated adapted items. *Applied Measurement in Education*, *13*(3), 229-248. **https://doi.org/10.1207/S15324818AME1303_1**

Spielberger, C.D. (1980*). Test anxiety inventory: Preliminary professional manual*. Palo Alto, CA: Consulting Psychologist Press.

Şahin, M.G., ve Boztunç Öztürk, N. (2018). Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması. *Kastamonu Eğitim Dergisi*, *26*(1), 191-199. **https://doi.org/10.24106/kefdergi.375863_**

Tekin, H. (2012). Eğitimde Ölçme ve Değerlendirme. Ankara: Yargı yayınları.

Tezbasaran, A. (2008). *Likert tipi ölçek hazırlama kılavuzu*. Mersin: Türk Psikologlar Dernegi.

Thompson, B. (Ed.) (2003). *Score reliability. Contemporary thinking on reliability issues.* Thousand Oaks, California: Sage.

Türkiye Ölçme Araçları Dizini –TOAD (2021). https://toad.halileksi.net/. Erişim tarihi: 10.06.2021

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20. **https://doi.org/10.1177/0013164498058001002**

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *The Journal of Experimental Education, 67*(4), 335- 341. **https://doi.org/10.1080/00220979909598487**

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. **https://doi.org/10.1037/0003-066X.54.8.594**

Willson, V. L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher, 9,* 5-10. **https://doi.org/10.2307/1175221**