International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 163–172
https://doi.org/10.21449/ijate.1368458
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Implications of current validity frameworks for classroom assessment

Ezgi Mor [1,*], **Rabia Karatoprak Erşen** [1]

[1]Kastamonu University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Türkiye

ARTICLE HISTORY

Received: Sep. 29, 2023
Accepted: Dec. 14, 2023

Keywords:
Validity,
Classroom assessment,
Educational assessment,
Argument-based approach,
Educational standards.

Abstract: The argument-based approach is the current framework for validity and validation. One of the criticisms is that understanding and applying this approach to practice are complicated and require abstract thinking. Teachers or school administrators in teaching and learning need support in their validation practice. Due to the abstract structure of validity, the test users and instructors who are not familiar with psychometrics may face problems in gathering validity evidence. Especially in classroom assessment, teachers may deal with understanding the complex methods of validation. In line with this need, the purpose of this study is to help instructors validate their assessment practices by providing a pathway to guide them through their validation processes and to make the validation process more obvious in classroom assessment. For this purpose, a checklist including the validity indicators for classroom assessment is developed. In this development process, Sireci's (2020) 4-step validation which is based on AERA et al. (2014) Standards and Bonner's (2013) study as a framework were followed. The validity indicators were composed by simplifying the AERA's standards and the ones which are relevant to classroom assessment were selected. In addition to the standards, the aforementioned studies were investigated and the validity indicators that may be applicable in classroom assessment were determined.

1. INTRODUCTION

In social sciences, the researchers appeal tests in order to gather information about the people for such a wide range of purposes. Educational and psychological tests are widely used by researchers, employers, and psychologists to make many crucial decisions which are diagnosis, treatment, certification, and evaluation. The consequences of these decisions can be high-stakes in individuals' lives such as enrollment in undergraduate programs, or being licensed to practice their jobs. Hence it is a well-known fact that the tests are valued universally, however, the actual value of the tests is determined by the accuracy level of these decisions. This argument was supported by Sireci and Benitez (2023), who stated that the real value of the tests depends on the quality of the test scores and the provided validity evidence related to the recommended usages of the tests.

In educational and psychological assessment, there is more than one problematic issue that should be handled in a detailed way and one of these issues is the validity of the scores. Validity

*CONTACT: Ezgi Mor ✉ ezgimor@gmail.com Kastamonu University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

is one of the concepts frequently considered by almost everyone in education, psychology, or social sciences who has collected data and made inferences based on it. Although it may seem like a rather abstract and technical subject that only those in the field of psychometry can comprehend, the definition of validity has actually undergone radical changes throughout its history in order to make it more unified, observable, and operative.

The concept of validity has been discussed since the early 1900s and is stated as the most vital psychometric quality of test scores (Sireci, 2020). Even though it is explained as the degree to which the test measures the quality it aims to measure, there is not a clear and straightforward definition of validity upon which most of the scholars in the field of educational and psychological measurement agree (e.g., Cizek, 2012; Newton & Shaw, 2014, 2016; Markus, 2016). Validity and validation are defined differently in the primary sources of educational and psychological measurement such as Educational Measurement (Brennan, 2006) and Standards for Educational and Psychological Testing (American Educational Research Association [AERA] et al., 2014). Table 1 presents these definitions.

Table 1. *Definitions of validity and validation.*

	Validity	Validation
Kane (2006, p. 17 in *Educational Measurement*)	the extent to which the evidence supports or refutes the proposed interpretations and uses	evaluating the plausibility of proposed interpretations and uses
Standards for Educational and Psychological Testing (AERA et al., 2014, p. 11)	Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests	accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations

The definitions given in Table 1 are made according to the argument-based approach. Even though it dates back to Cronbach (1988), Kane (1992, 2006, 2013) made the argument-based approach more known and accessible (Sireci, 2020). These definitions refer to not only interpretations but also uses of test scores. According to the argument-based approach, if the argument makes sense and is complete, its inferences are plausible, and the challenges about inferences and assumptions are cleared, then the interpretations/uses (IU) can be considered plausible, in other way, valid.

Kane's argument-based approach first appeared in Kane (1992). Even though it has been around for over 30 years, it has not been widely adopted by professionals in practical settings. Authors such as Newton (2013), Newton and Shaw (2014), and Sireci (2013) criticize the argument-based approach such that understanding and applying it to practice are complicated and require abstract thinking. Furthermore, Moss (2013) and (2016) state that it does not address the assessment needs of teachers or school administrators in teaching and learning and they need support in their validation practice. According to Kane (2013), users are responsible for validation in most cases. However, Moss points out that the information from the test may not have sufficient quality as evidence. Instead, the capacity of how local users use the test data determines the quality of data use. Therefore, validation should be a collaborative practice of test developers and test users.

The validity issues have been accepted as a concern of psychometrics for a long time. Due to the abstract definition and structure of the validity, it may be problematic for instructors who are not interested in psychometry and statistics. The ones who are not familiar with psychological testing or psychometry may be confused while studying the definitions and requirements of validity in AERA et al. (2014) Standards. For this reason, there is a need to develop more concrete ways to analyze the validity of scores, especially in the classroom assessment. As a response to this need, in this study, researchers aim to describe and discuss

the latest validity definitions and develop a checklist including the validity requirements that may be applicable in educational settings. Hence, the purpose of this study is to help the instructors validate their assessment practices by providing a pathway to guide them through their validation processes. This paper starts with a summary of the conceptual evolution of validity and validation. It continues with a part investigating the implications of the validation process in educational settings, especially in classroom assessments based on the research of Bonner (2013) and Sireci (2020). Upon all of the theoretical discussions and analyses, a checklist proposed by the authors was presented followed by the conclusion.

1.1. Validity and Validation

Theoretical discussions about the validity concept may be traced back over a century with Thorndike's (1904) thoughts. His thoughts were accepted as prime for standardized testing in the United States and many European countries (Sireci, 2009). Upon Thorndike's studies, the other prominent development in the concept of validity was observed in the 1940s and 1950s. It was the first time that the researchers reached a consensus about the validity in the Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal (APA,1952). In these technical recommendations report for psychological and educational testing, validity was re-conceptualized in different ways and these ways were named as the faces of validity: these are a) content, b) construct, c) predictive, and d) concurrent validity. This report may be accepted as the primary version of professional guidelines on test development, use, and evaluation. After this report, Cronbach and Meehl published a paper in order to discuss the construct validity in 1955 and this research was considered as one of the most influential studies covering the validity issue in a detailed way. Cronbach and Meehl (1955) proposed construct validity as a framework for interpretations about traits that are defined in terms of performance or tasks or behaviors shown by the individuals who have the trait interpreted in terms of lasting characteristics of individuals. The framework was not easy to apply in practical settings but it was influential in setting the term construct instead of trait.

Conceptualizing validity as a unified approach has been tried for a long time. In the 1970s, application-specific practical measurement settings used validity types separately whenever they were appropriate for the interpretations or uses (Kane & Bridgeman, 2021). Messick (1989) provided a unified validity framework centered on construct validity. It was much more comprehensive in defining construct validity than Cronbach and Meehl (1955). However, it was still difficult to apply.

Messick (1980) considers validity as an evaluation and Cronbach (1988) suggests applying the logic of evaluation argument as a framework for validation. Cronbach (1988) connects evaluation with an argumentative approach such that the argument should connect "concepts, evidence, social and personal consequences, and values" (p. 4). Argument-based approach (Cronbach, 1988) provides a framework for validation when there is no formal theory of construct. This approach helps eliminate ambiguity and open-endedness from the validation process by specifying a validity argument. The validity argument provides an evaluation of the proposed score interpretation and use through investigating any evidence related to attempted claims.

Kane (2013) proposes a way to make the reasoning behind claimed score interpretations more explicit and clearer so that the evaluation of that reasoning becomes more manageable. He suggests developing an Interpretation/Use Argument (IUA), where both interpretation and use of scores have equal importance. This is in contrast to his previous (2006) work where the emphasis was only on score interpretations. Kane defines IUA as "a network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions" (p. 8). According to this definition, interpretations consist of claims about a unit of analysis, and score uses include decisions about this unit of analysis.

Both interpretation and use have a major role in test development such that they define the purpose of the test. Even though it may seem that interpretations and uses are distinct, they may not be in practice.

The argument-based approach to validity has two steps (Kane, 2013). The first step of validation is to develop the IUA, which helps to understand the required evidence for validation and set the criteria for adequacy of validation. After the IUA has been developed, the IUA is evaluated using a validity argument. A validity argument is the evaluation of the evidence needed to evaluate the inferences and assumptions of the IUA. If the IUA is judged to be clear, coherent, and complete, and its inferences and assumptions are reasonable, then the claimed interpretation or use is valid.

Kane and Bridgeman (2021) stated that there exists an incompatibility between unified and application-specific frameworks since the early conceptualizations of validity. Standards for Educational and Psychological Testing published in 1974 and 1999 could not provide a solution for this incompatibility. 1985 standards necessitate evidence specific to the IU of interest. However, it did not contain much explicit guidance about combining different kinds of evidence. Although 2014 standards are in line with the argument-based approaches, the chapter on validity is written in terms of five kinds of evidence: "evidence based on test content, on response processes, on internal structure, and relations to other variables, as well as evidence for validity and consequences of testing".

One of the criticisms of the argument-based approach is that it does not address validation of local uses (Moss, 2013, 2016) such that the IUA framework mostly focuses on testing programs and intended uses and does not address how actual IU in local contexts can be validated. For instance, the need to validate the consequences of decisions about improving teaching and supporting learning is an example of the local context. In that case, teachers or school administrators are local users, and the actual IU depends on the purpose of these local users. The purpose might be to enable local users to incorporate the information gathered from the test into instructional practices and use the test results to make decisions about classroom activities, which the current validity theory does not support in a simple way.

2. THE IMPLICATION OF THE VALIDATION PROCESS IN THE EDUCATIONAL SETTINGS

Validity has been a primary concept in educational assessment and in line with validity, the test scores and their usage in educational settings are accepted as essential in the whole education process. Education assessment activities are designed and administered to gather information about students' learning processes in order to detect learning deficiencies and determine the students' achievement levels. Teachers are all convinced about these issues, such as deciding the purpose of tests, and the importance of educational assessments, especially in formative however, they have some problems with validity issues. Many teachers expressed concerns about accountability testing with respect to fairness, accessibility, representativeness, and alignment (Welch, 2021). In this point of view, it is clear that most of the teachers need some support in gathering validity proofs for their tests.

As aforementioned, validity has gained so many meanings throughout the history of psychometry and the debates have continued about the additional meanings and implications of it. Welch (2021) stated that there is a gap in understanding validity issues between teachers and measurement experts and in order to bridge the gap, reframing the messages around validity to help teachers understand the theoretical debates in more observable ways. Alignment of the curriculum, relevance, utility of information, comparability, replicability, stability under different modes, and content representativeness in adaptive tests are all areas that are equally important as alignment. One approach may be to relate additional sources of validity to

elements in the peer-review process of the teacher-made test and the scores obtained from it. In response to this, in this part of the study, the researchers aimed to conceptualize the validity in more concrete ways and paraphrased the already mentioned issues of validity in more observable ways, especially in classroom assessment (CA). While doing this, the resources stated in the first part of the study were used, and especially the AERA standards were benefited mainly. In addition to the APA standards on validity, Bonner's (2013) work on validity in CA was investigated in detail.

2.1. Validity in Classroom Assessment

Bonner (2013) asserted that CA differs from other educational assessments in a radical way in terms of purposes, hence validity may be a secondary purpose for CA. The researcher also stated that in CA, validity or appropriateness of inferences about test scores should be the real concern and it is recommended that teachers and researchers may use the validity analyzing methods to judge the propriety of the inferences.

Bonner (2013) proposed five critical principles that may be used in CA and if these criteria are taken into account, the researcher claims that the sensitivity to individual learners and learning outcomes may be reflected in the assessment process. Also, these principles are equally relevant to validity claims of the researchers and both types of data; qualitative and quantitative. These criteria are listed below:

1. Assessment should be aligned with instruction: It is stated that the curricular standards are not enough for achievement tests in CA. The tests should be aligned based on the tasks used in instruction. Nitko (1989) also supported this idea long before Bonnes (2013) study by defining the appropriate uses of tests that are linked with or integrated with instructional materials and procedures. Bonnes (2013) improves this claim by stating that if the CA is aligned with the instruction poorly, CA may have negative impacts on students' attitudes, motivation, and classroom environment. It is suggested to analyze test content represented on a test by comparing the instruction time or emphases on lesson plans.

2. Bias should be minimal at all phases of the assessment process: This criterion is so crucial, especially for the multicultural classroom environment. Students are open to many diverse factors' effects on the testing process. Some items may be in favor of fluent readers in paper-and-pencil tests, glib writers in essay formats, and personality attributes and performance assessments. Also, the teachers may be affected by biases when scoring the items. In order to minimize the influences of bias in CA, which also increases the validity of CA, tests and tasks can be analyzed by subject-matter experts, a group of teachers, or reviewed and debriefed assessments with a small group of students. Methods to reduce scoring bias, use of rubrics, co-scoring, and multiple-raters for samples of student work may be preferred.

3. Assessment processes should elicit relevant substantive processes: Thinking processes and task-relevant behaviors that are consistent with cognitive perspectives on assessment should be included in CA. Using cognitive processes in the tests may provide better diagnostic information about students' learning levels. Also, these cognitive processes should be included not only in tests but also in scoring phases by using rubrics.

4. Effects of assessment-based interpretations should be evaluated: The results and decisions based on test scores should be justified by strong logical arguments or evidence. Both cognitive and affective consequences of the tests should be analyzed. Especially for formative assessments, teachers should attempt to provide opportunities for students to be reassessed if the results of tests are ineffective or inappropriate.

5. Validation should include evidence from multiple stakeholders: Teachers should know and accept that the validity of their assessment-based decisions, but these decisions may be questioned by the other stakeholders. However it is a fact that there is no requirement of the

getting the approval of all the stakeholders' about the CA decisions. Kane (2006) emphasized the importance of the other stakeholders, who are not in the development processes of the tests, including the consequences of the tests, without this inclusion, the assumption that our assessment-based decisions are all valid. Teachers, who are in the development process of CA, are primarily responsible for evaluating their assessment processes and the assessment-based consequences. Hence the stakeholders may be colleagues, mentors, or professional test developers. As a principle, responsibility for assessment validation should be dependent on the judgment of a single individual.

These five criteria emphasize the importance of validity in CA and teachers are able to apply most of the stated procedures in order to validate the test scores. The other research that focuses on the validation process in a more applicable way is Sireci's (2020) work. In the following part of the study, the research is presented and Sireci's (2020) stepwise perspectives on the validation process are analyzed.

2.2. Sireci's Validation Steps

In the previous part, the five criteria proposed by Bonner (2013) were explained in detail, and it is a fact that these criteria do not differ radically from the AERA et al. (2014) Standards. Actually, most of the validity studies are based on these standards, and one of the most prominent and current studies investigating the validity in line with the AERA et al. (2014) Standards is Sireci's (2020) work. In this research, the researcher investigated the history of the validity concept and updated his previous Sireci (2013) study for the validation process. In Sireci (2013), the researcher proposed a three-step validation process based on AREA et al. (2014) Standards. These steps involved 1) clear articulation of testing purposes, 2) consideration of potential test misuse, and 3) crossing test purposes and potential misuses with the Standards' five sources of validity evidence. In the updated study, Sireci (2020) added one more step and it is 4) prioritizing the validity of studies to be conducted. In this part, these steps were explained concisely and the validity investigation ways that may be adapted in CA were emphasized especially.

Step 1. Articulating the Purposes of the Test: The process of validation includes gathering and analyzing evidence in order to defend the purpose of test usage. In line with the AERA et al. (2014) Standards of validation, Sireci (2020) also emphasized that the validation process begins with the explicit statement of the proposed interpretations of the test scores and of course, this purpose should be supported by a rationale. The important issue is that the intended purposes should be defined in an explicit and concise way and most of the time, the purposes are composed in a general, unclear, and complex way.

Step 2. Identifying Potential Negative Consequences of Test Use: As Messick (1989) stated, it is not enough to determine the intended test usage. It is also crucial to define the potential negative effects of the testing programs. Sireci (2020) suggested criticizing testing programs' adverse effects at the public level. For the large-scale assessment test, it may be stated that it has the potential to influence the curriculum negatively. These potential negative effects should be investigated at test level.

Step 3: Crossing test purposes and potential misuses with the Standards' five sources of validity evidence: In this step, the sources of validity evidence defined in the standards were included. These sources are test content, response processes, internal structure, relations with other variables, and testing consequences. The sources are explained in detail in the Standards, and Sireci (2020) exemplified their usages of them in the validation process with the Massachusetts Adult Proficiency Tests (MAPT) by using the technical manual of this test. Upon analyzing the questions; the ones that may be related with the CA were found and given below by adapting the CA settings:

1. Does the test actually measure students' achievement/ability/ skill /knowledge in the related course?
2. Does it measure these knowledge and skills as they are defined in the curriculum framework?
3. Are the test scores useful for evaluating students' progress toward meeting educational goals?
4. Are the test results useful for evaluating the related program/curriculum of the course?
5. What are the effects of the test on instruction in the education process?

The questions stated above do not stem from only the explicit testing purposes of test use but, some of them especially the last one emanates from implied test purposes, too. However, in the CA, nearly all of the stated questions should be investigated by the teachers who developed a test.

Step 4: Prioritizing the Validity Studies to be conducted: It is a well-known fact that all validity evidence suggested by the standards and/or the related research, are not possible to be gathered, hence some prioritization is needed in order to use time and resources efficiently. This prioritization should be applied based on the purpose of the test and sufficient validity evidence should be gathered as parallel with the test's purpose.

This four-step approach serves as the investigation of validity in an argument-based approach and within this approach, Sireci (2020) emphasized that the limitation of this approach is that it requires responsible test developers and evaluators to clearly articulate testing purposes and the intended information. The other drawback of this approach is stated that applying this approach requires prioritization and it may be problematic to select the type of validity evidence to be gathered. Of course, gathering all types of validity evidence and answering all the research questions for validation is not feasible and that's why prioritizing research questions is needed (Sireci, 2020).

Despite the hottest debates on the approaches and definitions of validity, it is clear that there are still several open-ended and questionable points of the validity investigations for the researchers. Actually, the validity issues may be analyzed in a more direct and easier way for educational settings because the tests used in classroom assessment are developed mainly for determining learning levels and monitoring students' progress. Hence the purposes of teacher-made tests are more obvious and the validity evidence may be gathered easily. In order to make the validation process more trackable and objective, the validity indicators that may be efficient in CA were determined and prepared as items that are open to be questioned by teachers or instructors who developed the tests. These indicators are presented below:

2.3. Validity Indicators in CA

In this part, the determined validity indicators are given. While composing this checklist, the researchers studied collaboratively and the draft of the checklist was analyzed by two different measurement specialists, who had doctorate degrees in measurement and assessment. Based on the experts' views, the indicators were prepared as a form of the checklist format which is composed of 17 items with three grading categories, satisfied, not-satisfied and not applicable. This checklist is presented in Table 2 below.

The validity indicators were prepared to cover the whole validation process. Hence, the checklist was composed by adopting an inclusive approach in which the whole validation process was considered. If the items are investigated in detail, the validation process can be observed. The checklist starts with the definition of the main purpose of the test, which is the first step of the development process of any test. The second and third items are aligned with the first one, it is stated that the possible usages of the test should be described in a detailed way. This is specifically important when the current validity frameworks (e.g., Kane, 2013; Sireci, 2013) include both score interpretations and score uses in the validation process. The third item is closely related to the first indicator in which teachers/ instructors are expected to

relate all the test items by considering the main purpose of the test. In the fourth item, characteristics of the test takers are emphasized and the test-takers should be defined in a detailed way. The fifth and sixth items are related to the scores' meaning and these items emphasize the usage of the scores. These items are so essential that the total score of the test is expected to reflect the level of the measured trait, which depends on the items' scoring. The next two items, seventh and eighth, are related to organizing the administration process of the test. Then in the next items, the content of the items and the relationships among the items are also considered and the determination of item formats is included in the validation process. The item formats should be selected as parallel with curriculum and teaching activities. The scoring criteria and weighting of the items are also included in the validation process. Lastly, the reliability evidence was emphasized in the context of the validation process. In brief, with these indicators, we exerted to cover all the validation steps which were determined according to the primary sources such as AERA et al. (2014) standards, Kane (2006, 2013), Bonnes (2013) and Sireci (2020). These indicators are suggested to be essential for the tests used in CA made by teachers/instructors

Table 2. *Validity indicators checklist.*

Validity Indicators	Satisfied	Not Satisfied	Not Applicable
1. The main purpose of the test is defined.			
2. The proposed test uses are stated in a detailed way.			
3. The test is designed in order to measure students' features; such as achievement/ability/skill/knowledge.			
4. The group of students for which the test is intended is specified.			
5. Test scores are composed to provide useful information for evaluating students' progress toward meeting educational goals.			
6. Test scores are composed to provide useful information for evaluating the related program of the course.			
7. Test administration procedures are determined before the test administration.			
8. The procedures followed in generating test content are justified.			
9. Both the item formats and the content of the items are aligned with the curriculum.			
10. In addition to the included content domains, the areas of the content domain that are not included are indicated.			
11.The test scoring procedures are described in detail.			
12. If it is claimed that the test is unidimensional, such a claim is justified with statistical analysis.			
13.The relationships among the items are investigated using item scores.			
14. Reliability evidence for each reported score is provided.			
15. If a test provides more than one score, the distinctiveness of the separate scores is justified.			
16. If a test provides a composite test score, the basis and how the test scores are combined are justified.			
17. If a differential weighting is proposed by test developers/teachers, the rationale behind the scoring is specified.			

3. CONCLUSION

The validity chapter written by Kane in the current edition of *Educational Measurement* (Brennan, 2006) and *Standards for Educational and Psychological Testing* (AERA et al, 2014) adopt the argument-based approach for validity and validation. According to Kane (2013), it is flexible in accommodating various applications such as achievement testing or experimental designs where causal inferences are made. As long as the claims made according to test scores are plausible and representative of the test scores, and justified empirically, then IUA (i.e., interpretation/use argument) is valid. However, Sireci (2013) argues that the development of interpretive argument, especially scoring, generalization, and extrapolation inferences can be complex and overwhelming, which might discourage practitioners from using the IUA framework. Another criticism is the lack of support for the professionals working in teaching and learning (Moss, 2013; 2016). Sireci (2020) in which 4-step validation using AERA et al. (2014) Standards as a framework for validation practices is proposed can be a practical guidance towards these criticisms.

In the second part of this study, upon evaluating and analyzing the primary studies in this field, the implications of the validation process in educational settings, especially in CA were determined. By investigating the argument-based approach (Kane, 2006; 2013) and Sireci's (2020)'s ideas and suggestions on the validation process, we proposed a checklist in which the essential validation indicators are included. While preparing these items, the clarity and simplicity of the statements were essentially paid attention. Due to the complexity of the Standards, teachers/ instructors may face some problems in understanding and applying these standards in their tests and test scores. Hence, we aimed to develop a short, brief instrument by prioritizing the CA applications and needs. Hence, we aim that the checklist may be used by a wide range of researchers who may be unfamiliar with psychometric issues in depth. With this checklist, teachers or instructors are able to evaluate their test scores by using this checklist, and in order to obtain more valid scores from the tests, they may evaluate their test items, testing conditions, scoring, and the process of test development in terms of these indicators. These indicators are stated as a three-point grading format in which the teachers/instructors may select the appropriate option for their tests and test scores. All items are designed as applicable for all types of tests that may be administered in CA.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Ezgi Mor: Investigation, Resources, and Writing-original draft. **Rabia Karatoprak Erşen:** Methodology, and Writing-original draft.

Orcid

Ezgi Mor https://orcid.org/0000-0003-0250-327X
Rabia Karatoprak Erşen https://orcid.org/0000-0001-8617-1908

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bonner, S.M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J.H. McMillan (Ed.), *SAGE handbook of research on classroom assessment*, (pp. 87-106). SAGE.

Cizek, G.J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31. https://doi.org/10.1037/a0026975

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum Associates, Inc.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 174-203.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

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

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M., & Bridgeman, B. (2021). The evolution of the concept of validity. In B.E. Clauser & M.B. Bunch (Eds.), *The history of educational measurement* (pp. 181-205). Routledge.

Markus, K.A. (2016). Alternative vocabularies in the test validity literature. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 252-267. https://doi.org/10.1080/0969594X.2015.1060191

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moss, P.A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*(1), 91-98. https://doi.org/10.1111/jedm.12003

Moss, P. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236-251. https://doi.org/10.1080/0969594X.2015.1072085

Newton, P.E. (2013). Two kinds of argument?. *Journal of Educational Measurement*, *50*(1), 105–109. https://doi.org/10.1111/jedm.12004

Newton, P., & Shaw, S. (2014). The deconstruction of validity: 2000–2012. In *Validity in educational and psychological assessment* (pp. 135-182). Sage.

Newton, P., & Shaw, S. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*(2), 178-197. https://doi.org/10.1080/0969594X.2015.1037241

Sireci, S.G., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema 35*(3) 217-226. https://doi.org/10.7334/psicothema2022.477

Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*(1), 99-104. https://doi.org/10.1111/jedm.12005

Sireci, S.G. (2020). De-"constructing" test validation. *Chinese/English Journal of Educational Measurement and Evaluation*, *1*(1), Article 3. https://doi.org/10.59863/CKHH8837

Welch, C.J. (2021). Rethinking measurement 101: Lessons learned from teachers. *Educational Measurement: Issues and Practice*, *40*(4), 13-17. https://doi.org/10.1111/emip.12479