SAPIENZA GRUPO EDITORIAL
Publisher: Sapienza Grupo Editorial
R. Santa Cruz, 2187, Vila Mariana
São Paulo, Brazil
editor@sapienzaeditorial.com

Sapienza
International Journal of Interdisciplinary Studies
ISSN: 2675-9780

OPEN ACCESS
BY NC ND
Signatory of DORA

# Validity in evaluation: where is the argument-based approach heading?

Validade na avaliação: para onde está indo a abordagem baseada em argumentos?

Validez de la evaluación: ¿hacia dónde se dirige el enfoque basado en argumentos?

**Karla Karina Ruiz Mendoza**
https://orcid.org/0000-0001-8978-8364
Professor and Researcher at Universidad Autónoma de Baja California UABC, Mexico
ruiz.karla32@uabc.edu.mx (correspondence)

**Luis Horacio Pedroza Zúñiga**
https://orcid.org/0000-0002-5256-2967
Professor and Researcher at Universidad Autónoma de Baja California UABC, Mexico

**Alma Yadhira López García**
https://orcid.org/0000-0002-7474-5799
Professor and Researcher at Universidad Autónoma de Baja California UABC, Mexico

**ARTICLE HISTORY**

**Received:** 28-03-2024
**Revised Version:** 26-05-2024
**Accepted:** 13-06-2024
**Published:** 01-07-2024
**Copyright:** © 2024 by the authors
**License:** CC BY-NC-ND 4.0
**Manuscript type:** Article

**ARTICLE INFORMATION**

**Science-Metrix Classification (Domain):**
Economic & Social Sciences
**Main topic:**
Validity and Evaluation
**Main practical implications:**
Explores how ethical and cultural considerations can improve the accuracy and fairness of psychometric testing in diverse educational and professional contexts.
**Originality/value:**
The article provides a new perspective on how to integrate ethical aspects in test validation, enriching both theory and methodology in educational evaluation.

## ABSTRACT

The evolution of the concept of validity is examined in the context of the integration of Generative Artificial Intelligence and ethical stances, and with it, informed decision-making. The methodology used includes the history of concepts as laid out by Koselleck, analyzing how the concept of validity is a fundamental concept. The method used is a literature review, analyzing historical and contemporary perspectives and arguments from influential authors such as Messick and Kane. This conceptual journey leads us to recognize that validity is not a monolithic entity, but a complex fabric of multiple theoretical and practical threads, ranging from the internal logic of evaluations to the repercussions of their application in society. Furthermore, validity is recognized as a complex construct that cannot be simplified to a single aspect or characteristic of a test or evaluation, differentiating between validity and validation. The five historical periods distinguished in the literature that reflect paradigmatic changes in the understanding of validity were: gestational, crystallization, fragmentation, reunification, deconstruction, culminating with the period of diffusion. The most relevant conclusion is that validity is not static but dynamic, evolving with context and application. It also emphasizes the need for continuous validation adapted to emerging challenges, such as Generative Artificial Intelligence (GenAI), with the goal of ensuring that evaluations are accurate and fair amid a growing trend on ideas of quantum computing.

**Keywords:** psychometrics; validity; educational assessment; generative artificial intelligence.

## RESUMO

Examina-se a evolução do conceito de validade num contexto de integração da Inteligência Artificial Gerativa e das posturas éticas, e com isso, a tomada de decisões informadas. A metodologia utilizada compreende a história dos conceitos proposta por Koselleck, analisando como o conceito de validade é um conceito fundamental. O método utilizado é a revisão da literatura, analisando perspectivas históricas e contemporâneas e argumentos de autores influentes como Messick e Kane. Este percurso conceitual leva-nos a reconhecer que a validade não é uma entidade monolítica, mas um tecido complexo de múltiplos fios teóricos e práticos, que vão desde a lógica interna das avaliações até às repercussões da sua aplicação na sociedade. Além disso, reconhece-se a validade como um constructo complexo que não pode ser simplificado a um único aspecto ou característica de um teste ou avaliação, diferenciando entre validade e validação. Os cinco períodos históricos que se distinguiram na literatura que refletem mudanças paradigmáticas na compreensão da validade foram: gestacional, cristalização, fragmentação, reunificação, desconstrução, culminando com o período de difusão. A conclusão mais relevante é que a validade não é estática, mas dinâmica, evoluindo com o contexto e a aplicação. Onde a Abordagem Baseada em Argumentos de Kane é teorizada mas não levada à prática, marcando uma preocupação pelos alcances do entendimento do conceito. Enfatiza-se, também, a necessidade de validação contínua e adaptada aos desafios emergentes, como a Inteligência Artificial Gerativa (IAGen), com o objetivo de garantir que as avaliações sejam precisas e equitativas diante de uma tendência crescente sobre as ideias de computação quântica.

**Palavras-chave**: psicometria; validade; avaliação educativa; inteligência artificial gerativa.

## RESUMEN

Se examina la evolución del concepto de validez en un contexto de la integración de la Inteligencia Artificial Generativa y las posturas éticas, y con ello, la toma de decisiones informadas. La metodología utilizada comprende a la historia de los conceptos dispuesta por Koselleck, analizando cómo el concepto de validez es un concepto fundamental. El método utilizado es la revisión de la literatura, analizando perspectivas históricas y contemporáneas y argumentos de autores influyentes como Messick y Kane. Este recorrido conceptual nos lleva a reconocer que la validez no es una entidad monolítica, sino un tejido complejo de múltiples hilos teóricos y prácticos, que van desde la lógica interna de las evaluaciones hasta las repercusiones de su aplicación en la sociedad. Además, se reconoce la validez como un constructo complejo que no puede ser simplificado a un solo aspecto o característica de una prueba o evaluación, diferenciando entre validez y validación. Los cinco periodos históricos que se distinguieron en la literatura que reflejan cambios paradigmáticos en la comprensión de la validez fueron: gestacional, cristalización, fragmentación, re-unificación, deconstrucción, culminando con el periodo de difusión. La conclusión más relevante es que la validez no es estática sino dinámica, evolucionando con el contexto y la aplicación. Donde el Enfoque Basado en Argumentos de Kane es teorizado pero no llevado a la práctica marcando una preocupación por los alcances del entendimiento del concepto. Se enfatiza, también, la necesidad de validación continua y adaptada a los desafíos emergentes, como la Inteligencia Artificial Generativa (IAGen), con el objetivo de garantizar que las evaluaciones sean precisas y equitativas ante una tendencia creciente sobre las ideas de la computación cuántica.

**Palabras clave**: psicometría; validez; evaluación educativa; inteligencia artificial generativa

## INTRODUCTION

Validity is a current concept in a context where Generative Artificial Intelligence (GAI) has emerged, where both lead to an ethical discussion about its conception and use (Chomsky, Roberts & Watumull, 2023; Hornberger, Bewersdorff, & Nerdel, 2023). Validity has served as a means of quality control in test measurement (Thorndike, 1997), and its importance has been highlighted in numerous investigations that attempt to delve into the interpetation and use of test results by framing a differentiation between validity and validation (Kane, 2013; Koretz, 2008; Markus & Borsboom, 2013; AERA, APA & NCME, 2014; Newton & Shaw, 2014; Chapelle, 2021). However, the multifaceted nature of validity has presented challenges in its practical application, especially in psychology and education, as it is still referred to as construct validity rather than evidence based on test content (Lavery et al., 2020).

In the educational context, the validity of tests is crucial to accurately and ethically assess learning and performance within a school community or, alternatively, in a specific context (Gafni, 2016). In this area, some contemporary authors emphasize the need for valid tests with a rigorous validation process for informed and fair decision-making, which has been key in the understanding and interpretation of the concept (Kane, 2006a, 2013a; Chapelle, 2012; Zumbo & Chan, 2014; Lopez & Willms, 2019).

Therefore, this article proposes a review of the concept of validity, exploring the paradigmatic and theoretical changes over time, based on the perspectives of different authors such as Markus and Borsboom (2013) and Newton and Shaw (2014), as well as the perspectives proposed by Messick (1989) and Kane (2006a, 2013a), until reviewing the current perspectives from the Systematic Literature Review (SLR) of Lavery et al. (2020), as well as the most recent proposal by Carol Chapelle (2021), ending with a reflection on the implications of the IAG for the validation process currently known.

## METHODS

In order to broaden the paradigm of this analysis, Reinhart Koselleck (2000) who argues that conceptual history is vital to understand the evolution of ideas and concepts over time was taken up. Applying this approach, this study examines how definitions and understandings of validity have changed over time. Newton and Shaw (2014) and Garcia Medina et al. (2017) suggest that, rather than a linear evolution, validity has undergone significant transformations depending on historical context and social dynamics. To delve into the historical evolution of the concept of validity, a citation network analysis was applied based on the methodology proposed by Garfield (1979).

**Literature selection and data sources**: A citation network analysis was performed to identify the most influential works in the field of validity, based on citations to the works of authors such as Markus and Borsboom (2013). The literature review was conducted in academic databases such as Springer, ERIC, Elsevier, Web of Science (WoS) and Scopus. This review was complemented by the application of the SLR (Systematic Literature Review) of Lavery et al. (2020) to capture the latest conceptions of validity and its practical application, especially within the framework of Michael Kane's (2006) Argument-Based Approach.

**Categorization and content analysis:** Using Bardin's (2011) content analysis technique, analytical categories were established to reflect conceptual changes in the perception of validity. The categories identified include (1) gestational, (2) crystallization, (3) fragmentation, (4) reunification, (5) deconstruction, and (6) diffusion. Each category corresponds to a historical period with distinctive characteristics in the conceptualization of validity. This categorization allowed for a systematic coding and analysis of the collected content.

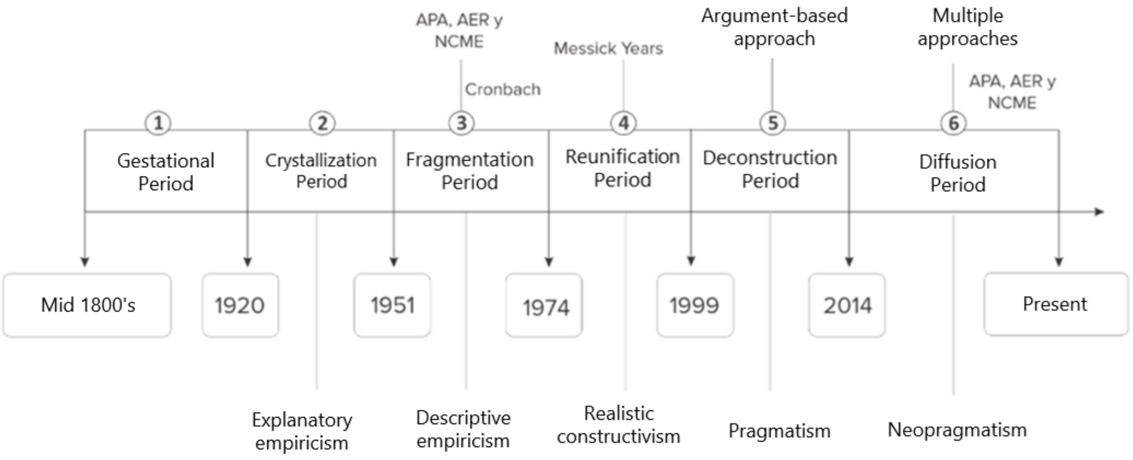
**Timeline as Analytical Tool:** A timeline was designed to trace the conceptual changes in validity throughout the identified periods. This analytical tool facilitates the detailed tracking of the theoretical and practical evolution of the concept, allowing for an in-depth understanding of how interpretations of validity have been influenced by sociocultural conditions and theoretical developments in the field of educational and psychometric assessment.

## VALIDITY CONCEPT TIMELINE

For the formation of an overview of the concept, the proposals of Markus and Borsboom (2013), Newton & Shaw (2014) and García-Medina (2017) were consulted, the first is the most cited on the subject, while the second addresses the concept of validity from a philosophical perspective, and finally García-Medina takes up Markus and Borsboom for his Spanish version. Figure 1 clearly expresses how each of the periods have been divided according to the proposals of Markus and Borsboom (2013), Newton & Shaw (2014) and García-Medina (2017). The periodization has focused on the development of the concept itself rather than classifications by type of validity or specific authors (Shaw & Crisp, 2011; Sireci, 2007). The diffusion period is a proposed name for this last stage, which is conceived as the one that is ongoing.

Each of the periods represents the culmination of the updating of the Standards for Educational and Psychological Testing. The Standards (so named hereafter) was initially published in 1966 by AERA, APA, and NCME and has undergone substantial updates in the years 1974, 1985, 1999, and 2014 (APA, AERA, NCME, 2014). Likewise, the publication and development of *Educational Measurement*, whose editions from 1951 to 2006, has been influential in the consolidation of various concepts proposed in the Standards, where the North American vision abounds with authors such as Cronbach (1951), Messick (1989), and Kane (2006). Each of these periods is presented below in Figure 1.

**Figure 1.** Timeline of the validity concept



***Note***. Own elaboration based on the proposal of Newton and Shaw (2014) and Markus and Borsboom (2013).

## GESTATIONAL PERIOD (MID 1800S - 1920S)

The gestational period of validity, between the mid-19th century and 1920, was characterized by a significant advance in statistical methodology that influenced the structuring of psychometric tests. These advances were aligned with the scientific and technological progress of the time, such as the discovery of X-rays by Röntgen (1895), the development of quantum theory by Planck (1900), and the publication of the theory of special relativity by Einstein (1905) (Watson, 2002). During this time, scientific work became more analytical and theoretical, with an emphasis on mathematical formulation rather than direct experimentation (Watson, 2002; Hobsbawm, 1998), which gave way to theorizing in other fields.

In the field of psychology and education, which were emerging as independent disciplines, a pronounced interest in the measurement of intelligence and other cognitive abilities was generated (Newton & Shaw, 2014). Intelligence tests, also known as *tests* for measuring intelligence quotient (IQ), were developed, as were the Binet-Simon tests, which assessed verbal, numerical, and spatial competencies (Watson, 2002; Newton & Shaw, 2014). In addition, innovators such as Francis Galton and James McKeen Cattell explored the connections between mental abilities and physical attributes, laying the foundations for psychometric measurement and construct validation (Newton & Shaw, 2014).

Therefore, Pearson was instrumental in this development by inventing the correlation coefficient in 1896, providing a tool to assess the relationship between test scores and mental abilities (Markus & Borsboom, 2013). Research from this era revealed significant correlations between brain size and certain cognitive abilities, which prompted the creation of tests to measure memory and reaction time, among other mental abilities (Garcia Medina et al., 2017).

Thus, this fervor to measure and understand human cognition was evidenced in the application of tests during World War I for the selection of recruits in the United States, marking a period of evaluation and reflection on the validity of the psychometric instruments used (Newton & Shaw, 2014; García Medina et al., 2017).

## CRYSTALLIZATION PERIOD (1921 - 1951)

During the Crystallization Period (1921-1951), the application of tests to judge students' abilities provoked intense scrutiny and led to deep reflection on the intentions and use of assessments (Newton & Shaw, 2014). In 1921, the North American National Association of Directors of Educational Research highlighted the need to reach consensus on the emerging measurement movement, ushering in an era where validity was established as an essential concept (Newton & Shaw, 2014). This period saw the development of two main approaches to validity: one focused on logical content analysis and the other based on empirical evidence of correlation (Lissitz, 2009). Often, validity was primarily supported by empirical correlations, although in situations where sample size was insufficient to obtain robust evidence, logical analysis between predictor and criterion and construct validity was called upon (Newton & Shaw, 2014).

Then, the crystallization period gave way to multiple definitions from the perspective of descriptive empiricism, focused on observation and data collection for the understanding of measured phenomena (see Table 1). This approach to empiricism, however, faced the challenge of how to interpret the observed results without falling into measurement errors (García-Medina et al., 2017), therefore, as shown in Table 1, the first definitions were more related to the conceptualization of the gestational period, where the aim was to empirically correlate the findings.

**Table 1** *Definitions of validity during crystallization period*

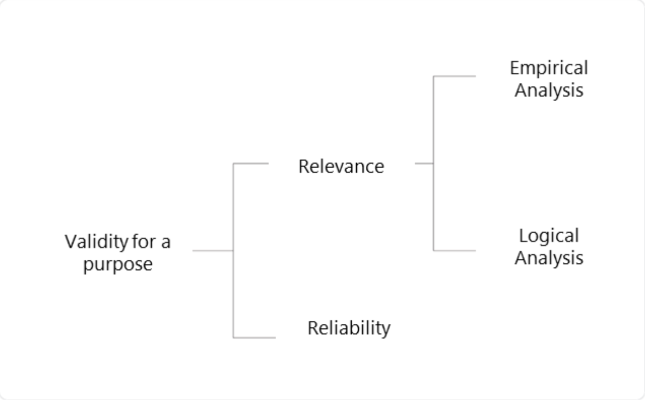
| Author | Year | Definition |
|---|---|---|
| Garrett | 1937 | "(...) the validity of a test is the fidelity with which it measures what it purports to measure (...)" (quoted in Lissitz, 2009, p.23). |
| Bingham | 1946 | "(...) the correlation of scores on a test with some other objective measure of what the test is intended to measure (...)" (quoted in Lissitz, 2009, p.23). |
| Guilford | 1946 | "(...) a test is valid for whatever it correlates with (...)" (quoted in Messick, 1989, p.18). |

***Note**: Own elaboration based on Newton and Shaw (2014).*

In the 1940s, as in Bingham and Guilford's definition (see Table 1), a more detailed analysis of content and correlations was introduced (Lissitz, 2009). Towards the end of this stage, Cronbach (1955) differentiated logical validity from empirical validity, the former associated with content analysis and the latter with empirical evidence linked to the correlation of test scores (Newton & Shaw, 2014); commonly known as content and criterion validity.

In the first edition of Educational Measurement in 1951, Cureton attempted to unify content and criterion validity under a common theory to demonstrate the representativeness of the measured domain and how well a test fulfilled its intended purpose (Cureton, 1951). Validity, according to Cureton, is composed of two aspects (see Figure 2): the relevance of the *test* with respect to its purpose and its reliability (cited in Chapelle, 2021). The discussion and reflection on these issues paved the way for a shift in thinking towards explanatory empiricism, where empirical evidence is paramount to explain and understand the object of research (Markus & Borsboom, 2013). However, reliability would become a different concept from validity, since a test could be reliable but not necessarily valid.

**Figure 2** *Diagram on the components of validity according to Cureton (1951)*



***Note**: based on Chapelle (2021).*

## FRAGMENTATION PERIOD (1952-1974)

During the period of fragmentation, spanning from 1952 to 1974, the concept of validity underwent significant diversification. The drive to standardize and delineate clear parameters led to the seminal publication "Technical Recommendations for Psychological Tests and Diagnostic Techniques" by the *American Psychological Association* (APA) in 1954, a draft of which dates back to 1952 (Newton & Shaw, 2014). Subsequently, the *American Educational Research Association* (AERA) and the *National Council on Measurement Used in Education* (NCMUE) developed "Technical Recommendations for Achievement Tests," published by the *National Education Association* (NEA) around 1955 (APA, AERA, NCME, 2014).

Reflecting on the limitations of tests that did not measure specific knowledge or skills, content-based tests or predictive abilities began to be validated, moving beyond mere score analysis (García-Medina et al., 2017). The emergence of construct validity became a mainstay during this period, addressing questions such as why individuals proficient in one task often excelled in others (Markus & Borsboom, 2013), thereby facilitating the assessment of relationships between various variables, paving the way for more accurate and reliable assessments based on hypothesized constructs (Markus & Borsboom, 2013).

Lee Cronbach emerged as a central figure, advocating that tests should be evaluated not only for their content, but also for their ability to measure abstract constructs (Lissitz, 2009; Newton & Shaw, 2014). Together with Paul Meehl, Cronbach proposed nomological network analysis to establish the validity of a test (Cronbach & Meehl, 1955). And on the other hand, proposals arose such as that of Jane Lovinger, who suggested that all forms of validity could be considered subcategories of construct validity (cited by Markus & Borsboom, 2013).

From 1954 to 1974, three editions of the Standards were published (APA, AERA & NCME, 2014), introducing content validity, criterion validity, and construct validity. However, their initial presentation implied mutual exclusivity, later clarified to suggest that these types of validity were interrelated. Each type of validity was associated with a specific category of test: content validity with achievement tests, criterion validity with aptitude tests, and construct validity with personality tests. These associations and their respective descriptions are detailed in Table 2.

**Table 2** *Approaches to Validity During the Fragmentation Period*

| Validity Approach | Description |
|---|---|
| Content Validity | Evaluates whether a test adequately measures the specific content or domain it is intended to measure. |
| Predictive Validity | Evaluates whether a test can accurately predict a future outcome or behavior. |
| Concurrent Validity | Evaluates whether a test can distinguish between presently known groups. |
| Construct Validity | Evaluates whether a test measures an abstract or theoretical concept, such as a psychological construct. |

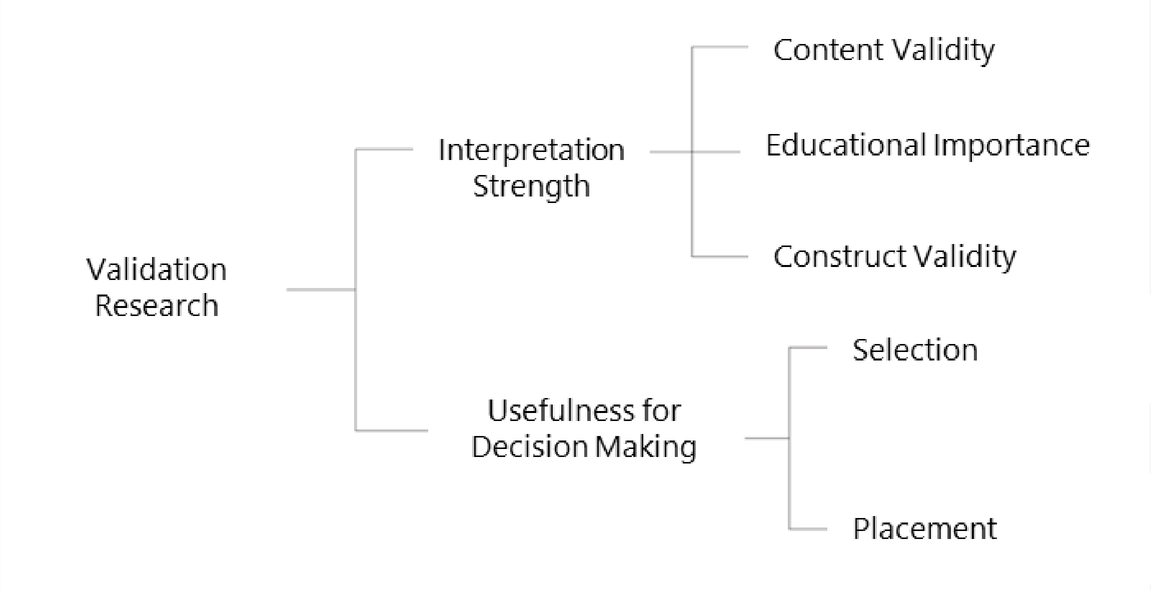
***Note:*** Own elaboration.

The year 1966 marked the debut of the first edition of the Standards for Educational and Psychological Testing (APA, AERA, NCME, 2014). This publication crystallized the fragmentation of validity into three types: content, criterion-related, and construct (Newton & Shaw, 2014).

By 1971, the second edition of *Educational Measurement* was launched, where Cronbach reflected on the misunderstandings between the concepts of validity of a test and its validation, arguing in favor of the validation of the interpretations of the test data, not of the test itself (Cronbach, 1971; Chapelle, 2021), i.e., one can actually validate the interpretation given to it at the precise moment, which could favor the understanding of why we can make the same test with different interpretations with respect to its time, or when to determine to omit a test to put another one or to update it. Thus, the discussion moved towards a broader validation approach, recognizing the need for diverse validation methods tailored to different types of tests.

Figure 3 illustrates the validation process from Cronbach's perspective, highlighting the fundamental concepts of content validity, construct validity and their educational importance, as well as their usefulness in decision-making processes.

**Figure 3**. Diagram of the types of validation research defined by Cronbach in 1971



***Note**:* based on Chapelle (2021, p.7).

By the mid-1970s, the discussion of validity and validation had evolved significantly, with validity defined as the desired quality or property of a test or measurement, i.e., the degree to which a test measures what it is supposed to measure. Validation was recognized as the process of inquiry to determine the legitimacy of a test or measurement (AERA, APA & NCME, 1974). This period recognized the complexity of validity and the need for a customized approach to validating different types of tests.

## REUNIFICATION PERIOD (1975-1999)

During the period known as the "Messick Years" (1975-1999) (Newton & Shaw, 2014; García Medina et al., 2017), there was a significant consolidation in the conceptualization of construct validity. Messick highlighted the multidimensionality of psychological constructs, such as anxiety or intelligence, arguing that these exist as multifaceted attributes of the individual (Markus & Borsboom, 2013).

This realist constructivist approach (Markus & Borsboom, 2013) enabled the unification of different forms of validity in educational measurement, providing a basis for a more integrated understanding of validity (Messick, 1989; Newton & Shaw, 2014).

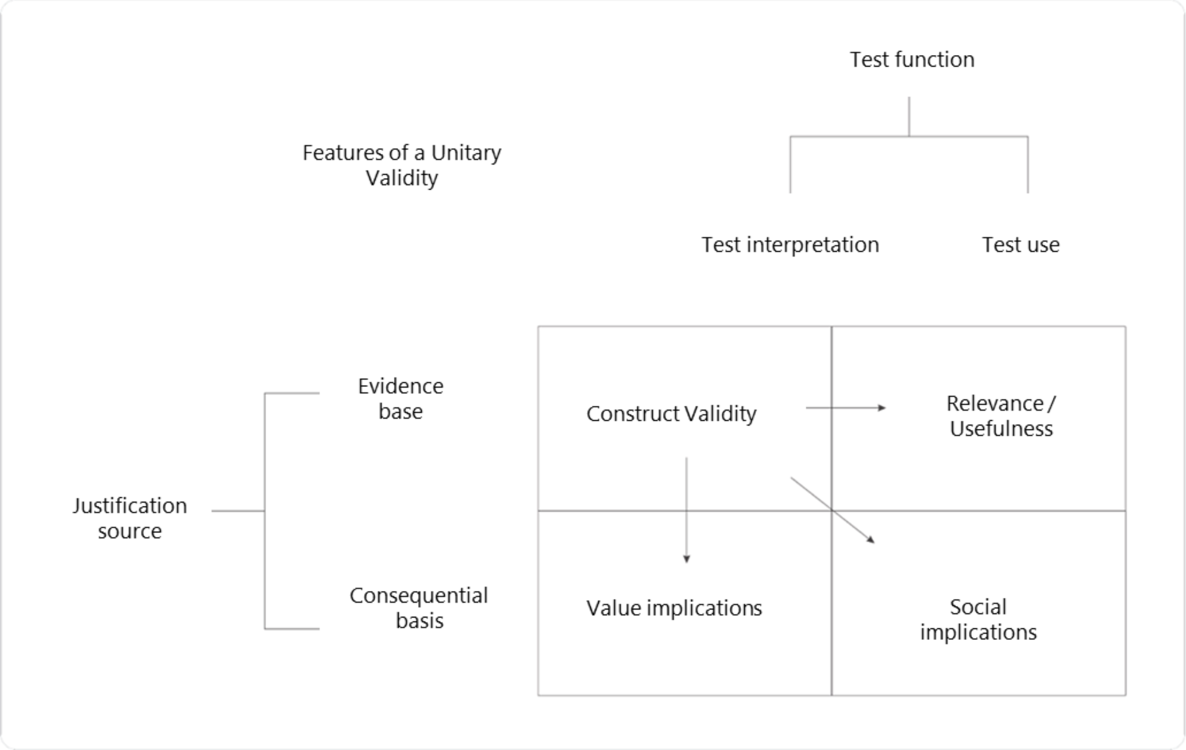
Messick (1989) proposed a definition of validity that included both the content of the *test* and the consequences of its use, emphasizing that validity depends on the specific use of the *test* and the context in which it is applied. This theoretical advance was important or transcendent for a more complete understanding of the concept of validity, stressing the need for a solid evidence base and considering the practical and social implications of *test* use (Chapelle, 2021). Messick (1989) also emphasized the importance of validation as an ongoing and systematic process that requires the accumulation of empirical and theoretical evidence to support validity claims. This process, guided by a clear conceptual framework, allows for the evaluation of all relevant facets of the measured construct (Markus & Borsboom, 2013).

The validation process, according to Messick (1989), is a methodology that goes beyond the simple accumulation of statistical evidence to affirm the validity of a *test*. It is a comprehensive approach involving multiple layers of analysis and review. Messick emphasized that validation is an ongoing task that does not end with the development of the *test*, but continues through its application and use in a variety of contexts. This process considers not only the technical aspects, such as reliability and construct validity, but also the ethical and social consequences of *test* use.

To illustrate this concept, Messick (1989) devised a framework that can be visualized in Figure 4. This figure, which details the facets of validity, can be conceptualized as follows: it implies that validation is a dynamic process that must consider multiple interrelated aspects of the test in question. It is not only about whether the test correctly measures a theoretical construct, but also about the ethical and practical consequences of the interpretation and use of the test results.

Thus, validation becomes a rigorous process that assesses the quality of tests in terms of their evidence base, construct validity, value implications, relevance/utility, and social consequences, all within the specific function and context of *test* use. This holistic approach to validation aims to ensure that tests are fair, accurate, and have a positive impact on both individuals and society. Thus, Messick provided the basis for the validity argument precisely by initiating discussion of the social impact and practical consequences of score-based decision making (Markus & Borsboom, 2013).

**Figure 4** Diagram of the facets of validity defined by Messick in 1989.



***Note***: based on Chapelle (2021).

## DECONSTRUCTION PERIOD (2000-2014)

The period of validity deconstruction (2000-2014) saw a significant transition in the theoretical approach to validity. This shift was characterized by a detailed and critical scrutiny of validity as a unified concept, and manifested itself in a plurality of perspectives that challenged traditional assumptions and promoted a more nuanced understanding of the construct (Newton & Shaw, 2014; García-Medina et al., 2017).

This period was marked by Messick's influence in the theoretical area of validity and the presence of Michael Kane, who proposed the Argument-Based Approach (ABA) (1992, 2006, 2009, 2013), highlighting the need to contextualize validation in terms of specific purposes and circumstances. Kane suggested that validity is not a static property of a *test*, but depends on the interpretation and use of *test* results in given contexts. Consequently, EBA emphasizes the utility and practical impact of tests, arguing that validity should be based on a sound causal theory rather than a mere nomological network (Kane, 2001, 2002, 2006, 2013).

However, there were also other international theoretical developments in this period. Table 3 presents these schools along with the relevant authors and their orientation toward the concept of validity. The academic debate on validity in psychometric assessment has been fertile ground for several schools of thought, each bringing unique dimensions to the concept. In the United States, the aforementioned conglomerate of scholars such as Messick (1989), Kane (1992, 2006, 2007, 2009, 2013), Cronbach (1971, 1988), Embretson (2001), and Linn (2004, 2005), among others, have delineated and delved deeper into construct validity and ABE, offering complex and adaptive structures that consider specific contexts and practical applications of tests advocating for a more contextualized assessment grounded in argumentative logic (Messick, 1989; Kane, 2006; Cronbach, 1971).

**Table 3** *Schools of the deconstruction period and their relation to validity*

| School | Relevant Authors | Orientation |
|---|---|---|
| United States | Messick (1989), Kane (1992, 2006, 2006, 2007, 2009, 2013), Cronbach (1971, 1988), Embretson (2001, 2007), Linn (2004, 2005), Schilling (2004), (Haertel & Lorié, 2004), (Haertel, 2013), (Sireci, 2007), (Chapelle, 2010; 2012). | On the one hand there was emphasis on construct validity (e.g., Messick, 1989; Embretson & Gorin, 2001) and on the other on the Argument-Based Approach; there were those who propose more structured approaches (e.g., Schilling, 2004) and one that depends on circumstances, situations, and contexts (Kane, 2007, 2012, 2013). |
| Canada | Zumbo (2004, 2007, 2014), Chan (2010, 2014). | They focused on construct theory as a conceptual framework and incorporated an interpretive perspective on the results as a core part of validity. |
| United Kingdom | Bachman (1990, 2005), Bachman and Palmer (2010), Shepard (1993, 1997). | They start from a pragmatic approach, i.e., their concern is focused on the usefulness, applicability and consequences of the tests; both in the individual and in the educational system(s) from validation based on argumentation. |
| Netherlands | De Jong Gierveld (1987, 1992, 2011), Sijtsma (2009) | They use item response models and theories with a comprehensive approach to validity, i.e., they consider multiple sources of evidence and perspectives for their assessment. |
| Australia | Mislevy, Steinberg and Almond (2003) | They focus on the degree to which tests can predict or relate to external criteria, i.e., criterion validity. They add ecological validity where the environment and context must be taken into account. |
| Germany | Borsboom (2004, 2009, 2013), Borsboom and Mellenberg (2004). | They are based on the psychometric approach, thus emphasizing the reliability and precision of the measurements. On the other hand, they have been interested in the area of questioning validity from a philosophical point of view. |

***Note**: Own elaboration*

On the other hand, in Canada, theorists such as Zumbo (2004, 2007, 2014) and Chan (2010, 2014) focused their attention on construct theory as the central axis, promoting an interpretation of the results that goes hand in hand with the intrinsic validity of the tests. From the United Kingdom, Bachman (1990, 2005) and Bachman and Palmer (2010), along with Shepard (1993, 1997), opted for a pragmatic approach, focusing on the utility, applicability, and consequences of assessments, both for individuals and for educational systems. In the Netherlands, De Jong Gierveld (1987, 1992, 2011) and Sijtsma (2009) applied item response theory models to offer a comprehensive approach to validity that encompasses multiple sources of evidence.

In Australia, the work of Mislevy, Steinberg and Almond (2003) focused on criterion validity, with special consideration of ecological validity and the relevance of context in the interpretation of test results. Finally, in Germany, Borsboom (2004, 2009, 2013) and Mellenbergh (2004) have contributed from a psychometric perspective, emphasizing reliability and accuracy of measures, and have explored the questioning of validity through a philosophical lens. These diverse approaches reflect the richness and complexity of the field of validity, highlighting the importance of a holistic and contextual understanding of the term.

Thus, validity began to be considered not as an inherent property of the *test*, but as a characteristic of the relationship between the *test* and the construct measured, adopting an ontological perspective (Borsboom et al., 2004, 2009) that had already been suggested by Cronbach (1971) decades earlier. Finally, the APA, AERA, and NCME (2014) provided a more robust and refined definition of validity in the Standards for Educational and Psychological Testing, culminating a period rich in debate and conceptual expansion that will undoubtedly continue to evolve and adapt to new contexts and understandings.

## VALIDITY AND THE STANDARDS: A 2014 UPDATE

The 2014 update of the Standards for Educational and Psychological Testing marked an important consolidation in the understanding of validity, emphasizing its nature as a unitary concept and the cumulative process of validation. At that year's NCME Annual Meeting, diverse views on how to use and understand validity converged in a series of papers that reflected the spectrum of opinions in the field, from ultraconservative to liberal views (Newton & Baird, 2016).

It should be noted that the research by Newton and Shaw (2014) did not consider the latest edition of the Standards due to its simultaneous publication with their work. For their part, Garcia-Medina et al. (2017) included the definition of validity of the Standards without establishing a defined cut-off in the period, suggesting that the implementation of these changes had not been studied.

The 2014 Standards define validity as the degree to which evidence and theory support interpretations of test scores for proposed uses. This approach highlights the importance of a validation process that involves the accumulation of relevant evidence to provide a sound scientific basis for proposed interpretations of test scores (APA, AERA, & NCME, 2014). This perspective rejects the idea of a typology of validity and instead recognizes different types of validation evidence that should be considered.

> "(...) degree to which evidence and theory support interpretations of a test's scores for proposed test uses. (...) The validation process involves accumulating relevant evidence to provide a sound scientific basis for proposed score interpretations." (p.11)

Table 4 summarizes the types of validation evidence according to the 2014 Standards, detailing how each type contributes to support for the validity of a test. This unitary perspective on validity moves away from the idea of classifying validity into distinct types and focuses on the cumulative evidence that reinforces the interpretation of scores.

**Table 4**. *On types of validation evidence according to the Standards (2014).*

| Type of Evidence | Description |
|---|---|
| Content-based evidence | The correspondence between the content of the test items and the construct they are intended to measure is evaluated. |
| Evidence based on response processes | The participants' response process is analyzed, including strategies used, response time and eye movements. |
| Evidence based on internal structure | The relationship between test items and construct components is examined to assess the internal alignment of the test. |
| Evidence based on relationships with other variables | The relationship between test scores and external variables is investigated as evidence of convergent, discriminant and generalization validity. |
| Evidence of validation and consequences of the tests. | The likely consequences of testing are analyzed, considering interpretations of scores, claims made about use, and unintended consequences. |

***Note***: *Own elaboration.*

Table 5 provides an overview of the evolution of the concept of validity throughout the editions of the Standards from 1966 to 2014. It shows a static view of assessment toward a continuous and dynamic validation process, emphasizing the accumulation of multiple pieces of evidence to strengthen the validity of test interpretations. These changes reflect greater flexibility and awareness of the context in which tests are applied and how their results are interpreted. The latest version of the Standards highlights a commitment to pragmatism, emphasizing that the validation process should be continuous and guided by the utility and impact of tests in specific contexts (Markus & Borsboom, 2013).

**Table 5** *Changes in the concept of validity according to AERA, APA and NCME*

| Year | Changes in the concept of validity | Category |
|---|---|---|
| 1966 | The aim is to measure the ability of the test to check whether it measures what it is supposed to measure, for which they established two types of validity: predictive validity and content validity. | *Types:* criterion related, construct related, content related |
| 1974 | The definition is broadened where questions on validity give rise to inferences that are appropriate and useful. They kept the two types of validity introduced in 1966 and added a third type: concurrent validity. | *Aspects:* criterion-related, construct-related, content-related |
| 1985 | It was emphasized that validity is a continuous and dynamic process that requires multiple sources and methods to assess valid evidence. The three types of validity introduced in 1974 were maintained and emphasized the joint use of multiple sources and methods to evaluate validity evidence. | *Categories:* criterion-related, construct-related, content-related, content-related |
| 1999 | They stated that empirical evaluation is fundamental to establishing valid evidence for the intended use. The Standards emphasized that all evidence should be continually evaluated to ensure its validity. | *Sources of evidence:* content, response processes, internal structure, relationships with other variables, consequences of the test. |
| 2014 | The importance of the continuous and cumulative process of validation rather than a single static assessment is emphasized. And it speaks of types of validity evidence, rather than distinct types of validity. | *Types of evidence based on:* content; response processes; internal structure; relationships with other variables; and, validation and consequences of the evidence. |

***Note***: Review of Camara (2006), the AERA, APA and NCME (1999, 2014) as well as Newton and Shaw (2014), and interview conducted with Sireci in 2017 by Nurl Doğan were taken up.

The 2014 revision and update of the Standards, therefore, marks a milestone in the conceptualization of validity, recognizing the complexity of the validation process and the need for an approach that incorporates a wide range of evidence relevant to the interpretation of test scores. Moreover, as Newton & Shaw (2014) have delineated, these updates seem to mark the junctures of academic and scientific debates on the subject, hence, one would have to wait for a new publication to observe what they conclude from a new period towards the present day.

## DIFFUSION PERIOD (2015 - TO PRESENT)

Although in the previous period there was a consensus on the concept and certain fundamental aspects of validity, such as the importance of constructing an argument with multiple types of evidence, there are still disagreements and controversies about how to define and assess validity in contexts and with different types of instruments; such as standardized tests, surveys or questionnaires, or to synthesize and better understand the validation processes (Lavery, et al., 2020). For this reason, it has been decided to call this diffusion period with a neo-pragmatic approach, since researchers have not yet mastered these conceptions and often continue to use the triad as the validity itself, instead of talking about evidence of validity from a validation process (Lavery, et al., 2020).

Neopragmatism (Santamaría, 2012) is characterized by looking to practice and experience to understand and solve problems, likewise, there are no definitive answers or absolute truths as it depends on the situation and context we face. It is this discussion that Chapelle (2021) maintains about how context can change the testing situation, especially when considering the different audiences and discourse communities that interact with the concepts of assessment, testing, and validity. Thus, both Kane (2016) and Shepard (2016) acknowledge that evaluators will end up interacting with groups that have their own languages, goals, and traditions.

The deconstruction stage allowed the elements and dimensions of validity to be broken down, recognizing its complexity and diversity (Markus & Borsboom, 2013). Although a conceptual definition was arrived at by the AERA, APA and NCME in 2014, the transition to neopragmatism seeks a clearer and more pragmatic understanding, integrating cultural, social and political factors. This shift is reflected in the work of Chapelle (2021), Kane (2016), and Shepard (2016), who emphasize the need to contextualize testing situations.

Although the AERA, APA, and NCME (2014) and Kane (2014) share the general framework of validity, they differ in the construction of arguments. Kane advocates strong, coherent arguments as the basis for validity, whereas the Standards focus on evidence supported by arguments, without specifying the nature of the arguments. This discrepancy has been addressed by Kane through his Argument-Based Approach (ABA), based on the logic of Toulmin's (1958) model, influencing other researchers such as Carrillo et al. (2020), Pedrosa et al. (2014) and Cook et al. (2015).

According to Lavery et al. (2020), the theoretical proposals of this period have lacked practical applications, highlighting the difficulty of transcending from theory to reality. However, current research indicates a gradual adoption of these theories in practice, particularly in university examinations, suggesting that the adoption of the neopragmatic approach, although complex, is beginning to influence practical validation methodology (see the corresponding SLR section of the present project).

Taking up the above, Chapelle (2012, 2015, 2021), who was a disciple of Kane, has made one of the most definite and practical proposals for those who are responsible for the evaluation of language *tests*, since, together with other colleagues, she has been in charge of the revision of the *Educational Testing Service's Test of English as a Foreign Language* (TOEFL.®), and at the same time has developed a robust design based on the EBA; which aims to clarify and help in defining a methodological design guide but not determinant. In his book *Argument-Based Validation in Testing and Assessment* (2021), he exposes the discussions that have existed on validity from four exponents that have guided the discussion through *Educational Measurement,* Cureton (1951), Cronbach (1971), Messick (1989) and Kane (2006).

In the field of assessment, Kane et al. (2020) sought to promote practical and exemplary validity assessments through the journal *Educational Assessment*. Papers published in this medium expand the understanding of validity, highlighting the importance of content- and criterion-based evidence, and consideration of the context and social consequences of assessments. However, as mentioned above, there is a great diversity of publications on validity and validation processes. On the one hand, there are those who take up the meanings of the past to bring them to the present and carry out another review, as in the case of Fabrigar, Wegener and Petty (2020), who choose to study the replication of studies based on validity, but they elaborate on the classic proposal of Cook and Campbell (1979) where four types of validity were proposed: statistical, internal, construct and external conclusion validity. At the same time, there are those who try to encompass social problems as is the case of alternative assessments designed to measure the progress and achievement of students with disabilities who cannot fully participate in traditional standardized assessments (Gotch & French, 2020).In addition, due to the literature consulted, one can perceive a growing interest by universities in the Middle East in validity through the argument-based approach (Esfandari et al., 2018), as well as mixed approaches (Jawhar et al., 2021; Fan, 2014) including corpus linguistic analysis (LaFlair & Staples, 2017) or systematic literature reviews (Cizek, Kosh & Toutkoushian, 2018; Lavery et al., 2020). It is likely that due to the clarifications described in the Standards, authors will gradually dare to develop mixed perspectives: "Qualitative studies are also relevant for supporting validity arguments (e.g., expert reviews, focus groups, cognitive labs)." (AERA, APA & NCME, 2014, p.73).

Among other perspectives, Pellegrino, DiBello, and Goldman (2016) have integrated the work of Kane and other authors, such as Mislevy, Riconscente, the Standards, and Wilson, into a framework for validating classroom assessments. Their approach analyzes cognitive, instructional, and inferential components of validity, and the importance of considering Kane's EBA in broader contexts has been highlighted. Wools, Eggen, and Béguin (2016) extended this approach by using multiple assessments to arrive at a single decision.

Likewise, other authors continue to take up Messick's ideas. Embretson (2016) has proposed on several occasions an Integrated Framework for Construct Validity, which focuses specifically on the development of tests and the consequences associated with their use. This theoretical framework offers a comprehensive perspective to address validity in assessment, considering key aspects related to test construction and impact. Likewise, Messick has been taken up as a better option for validating instruments or for academic tests (Chong et al., 2022), with the proposition of an argument being the meeting point for both proposals.

To conclude this timeline, in Table 6, a division of the periods was elaborated with their definition of validity, scope, as well as some representative authors and how they define or conceptualize this concept. This table was made from the analysis of the literature and the analysis elaborated by Lavery, et al. (2020), and Im, Shin and Cheng (2019) on contemporary positions towards the theoretical aspect. Against this background, it can be affirmed that the definition of validity could be clarified if the contextual, political and language aspects, among other characteristics, are taken into account, as developed by Chapelle (2021).

**Table 6** *Periods and their definitions of validity*

| Period | Definition of Validity | Scope | Representative authors |
|---|---|---|---|
| Gestational Period (mid 1800s-1920s) | Recognition of the importance of evaluating the ability of tests to measure what they are supposed to measure, without a clear and unified definition. | Just measure | Diversity of perspectives |
| Crystallization Period (1920-1951) | "The degree to which a test measures what it is supposed to measure," emphasizing the content of the test and its relationship to the construct being assessed. | Just measure | Louis Thurstone, Lee Cronbach, and Paul Meehl |
| Fragmentation Period (1952-1974) | Emergence of diverse perspectives and lack of consensus on the definition of validity. | Measure and decide | Lee Cronbach, and the AERA, APA, NCME |
| Re-unification Period (1975-1999) | Proposal of an expanded definition of validity that includes the content of the test and the consequences of its use. | Measure, make decisions and analyze their consequences | Samuel Messick |
| Deconstruction Period (2000-2014) | Questioning of the unitary concept of validity and emphasis on considering multiple facets of validity. | Measure, make decisions and analyze their consequences from a social and cultural perspective. | Michel Kane, Linn, Robert L. Brennan, David Carless, Mary James, and Paul Newton, among others. |
| Period of diffusion (2015-present) | Validity refers to the degree to which evidence and theory support interpretations of a test's scores for proposed uses of the tests. (AERA, APA & NCME, 2014, p.11) | Measure, make decisions and analyze their consequences from a social and cultural perspective. | AERA, APA, NCME |
| | - | Measure, make decisions and analyze their consequences | Samuel Messick |
| | Validity is a property of the interpretation and use of results. In addition, validity is discussed as part of *testing policies*, i.e., it is part of a set of rules and guidelines, such as quality standards, reliability, equity, transparency in results, management and feedback, to name a few. | Measure, make decisions and analyze their social and cultural consequences, accountability, among others. | Kane; Chapelle; Sireci; Embretson; and others. |

***Note***: Own elaboration.

### Thoughts on the future of validity and AI

Finally, from what is consulted in Lavery et al. (2020) it seems that there is little practice of ABE applied to validation processes, as if this understanding of how to be sufficiently rigorous to achieve the objectivity and truth of an instrument is not yet closed; and suddenly it is already coexisting with the boom of IAGen specifically with ChatGPT (a *Large Language Model*, LLM), being considered for psychometric and educational tests (Nasution, 2023), or to have a validation process of different tools such as rubrics. However, this type of approach is something that has already been envisioned with the introduction of *Learning Management Systems*, or Intelligent Tutoring Systems (ITS) as digital learning support tools that have the potential to create individualized and adaptive practice environments (Schmidt & Strasser, 2022).

Chapelle has followed these advances in language teaching and, together with Sauro, published *The Handbook of Technology and Second Language Teaching and Learning* (2017), which offers a review of computers and technologies as an expansion of existing pedagogical and assessment options, especially in the area of language. Likewise, there was an exploration of writing and how new technologies are shaping these practices. Although Chapelle has not published anything on this - he surely will - the introduction of ChatGPT in 2023 enables other new ways of thinking about validation processes, such as new types of evidence, and at the same time a new way of thinking about validity, as we move towards more agile processing of information, hence of results.

An example of the above is Nasution's (2023) study of a biology exam conducted by ChatGPT, which concluded that 20 of the 21 questions generated by ChatGPT were valid and the reliability, as measured by Cronbach's alpha coefficient, was 0.65, indicating acceptable reliability for the twenty valid questions. To generate the questions, the researchers asked ChatGPT AI to create multiple-choice questions with one correct and four incorrect answer choices on basic biology topics discussed in high school and college education. The questions generated by ChatGPT AI were subsequently evaluated by experts and presented to students in both English and Indonesian under strict supervision to ensure that the answers were based solely on their own knowledge (Nasution, 2023). The above, leads to rethinking the whole process, as we could see elaboration work disappearing and only supervisory work being carried out, especially if we are dealing with the allusions that IAGen has (Lingard, 2023).

Furthermore, Aloisi's (2023) analysis of the use of AI and machine learning in high-stakes assessments illustrates how these advances can compromise validity and, consequently, trust in educational assessment systems. These dilemmas highlight the need for a methodical approach to validation in the digital age, one that considers contextual variability and interaction between assessors and stakeholders, as underscored by Chapelle's practical approach to language test assessment and robust methodology based on Kane's principles (Chapelle, 2021; Kane, 2016).

However, there is still a long way to go with the integration of AI in the educational field, the study conducted on university students in Germany - being a developed country - by Hornberger, Bewersdorff, and Nerdel (2023), showed that those who have more skills and interest in this area are engineering students, followed by natural and exact sciences, medicine, and finally social sciences. On the other hand, Schmidt and Strasser (2022) address how an adaptive and intelligent learning environment could allow students to practice language skills in an individualized manner, tailored to their ability levels, interests, and motivation. In response, Schmidt and Strasser propose task-based language learning, a methodological approach that promotes the use of the target language as a communicative activity and focuses on the content and meaning of the message, seeking authenticity and defined linguistic outcomes.

In this sense, reflection on the inclusion of AI in educational testing takes us straight to the heart of the validity debate in the digital age. Chapelle and Kane have been instrumental in the evolution of validity theory with an EBA, which holds that the validity of a test must be established through a series of logical and empirical arguments that support its intended use (Chapelle, 2021; Kane, 2016).

The application of AI in education requires continuous and dynamic validation that accumulates multiple pieces of evidence to support the interpretation of test results, as reflected in the changes in the concept of validity in the editions of the AERA, APA, and NCME Standards from 1966 to 2014. The current neopragmatic approach stresses the importance of considering contextual variability and the interaction between raters and stakeholders. In this sense, Chapelle's work is relevant for its practical approach to language *test* evaluation, developing a robust design based on Kane's approach, which clarifies and helps to define a methodological design in test validation (Chapelle, 2021).

Therefore, ABE should be deepened in the validation processes of digital tools to ensure that psychometric assessments are accurate, fair and beneficial to society, considering recent aspects such as the use of AI for learning. This dynamic process of validation becomes a focus of interest that may lead to a reformulation of the concept of validity itself, recognizing that validity is not a static property, but evolves with context and application (Gallent-Torres et al., 2023).

Finally, the future of language learning projected by Schmidt and Strasser (2022) for 2040, who envision a synthesis of digital and traditional methods, reflecting an evidence-based environment with adaptive and multimedia resources, is increasingly plausible. In this emerging educational landscape, teachers proficient in data and technology management will be crucial for the effective integration of these tools into learning processes; and, in terms of high-impact testing, expertise in data science, models and algorithms will be increasingly required for better analysis

## CONCLUSIONS

The conceptual trajectory of validity in psychometric testing has been characterized by a constant effort to capture the essence of what it means to measure accurately, fairly, and relevantly (Newton & Shaw, 2014; Lissitz, 2009). This journey has been marked by seminal milestones highlighted in Standards and benchmark work within the field (AERA, APA & NCME, 2014).

The most recent stage, which we have termed the Diffusion period, emerges from a constructive critique of the previous consensus and recognizes a plurality of contexts and purposes of application. It is so named because of the dispersion of the concept of validity into multiple facets and domains of application, a recognition of its inherently situated and contextualized nature (Chapelle, 2021; Kane, 2016).

More recent proposals, such as those of Kane (2020) and Chapelle (2021), suggest an orientation towards a pragmatic and contextual approach to validation, stressing the need for rigorous and coherent argumentative reasoning. They urge a deeper reflection on the consequences and social impact of psychometric testing, an echo of the principles of fairness and equity that have become central to contemporary discussion.

### Limitations and future research

However, despite efforts to provide a comprehensive analysis of the conceptual evolution of validity, this study faces several limitations. First, reliance on the available literature in specific databases may have excluded relevant studies published in other media or languages, limiting the generalizability of the findings. Second, the content analysis, although systematic, is subject to the interpretation of the researcher, which could introduce bias in the categorization and analysis of the data. Third, the temporal delimitation of the review to 2020 may have omitted recent developments that could influence the current understanding of the concept of validity.

In response to these limitations and based on the results discussed, the following structured research points are proposed to address and expand the study of validity:

**Integration of ethical and social variables:** Future studies should explore how ethical and social considerations influence the interpretation and application of validity in different contexts. This could include studies that examine the impacts of testing on diverse populations, especially marginalized groups, to ensure that testing is fair and beneficial.

**International comparative analysis:** Comparative studies between different countries or regions are recommended to understand how cultural norms and educational policies influence the conceptualization and practice of validity. This would help identify best practices and common challenges globally.

**Development and testing of causal models:** Future research should formulate and test causal models that relate dependent variables such as student achievement or satisfaction with assessment to independent variables such as test characteristics (e.g., design, content) and application contexts (e.g., higher education, job assessment), as theorized by Chapelle (2021).

**Application in specific contexts:** The application of the findings in specific contexts, such as the assessment of competencies in the workplace or the measurement of learning outcomes in online education, is proposed. These studies could offer insights on how to adapt and validate assessment tools in dynamic and technologically advanced environments.

**Implications for educational practice and policy:** The results of this research could inform the development of policies and practices that promote the ethical and effective use of psychometric tests, ensuring that assessments are not only accurate but also equitable and culturally responsive.

In this sense, progress must continue to be made toward an even more holistic understanding of validity, one that fully recognizes its multidimensional nature and its fundamental role in informed and responsible decision making in education and psychology, envisioning the integration of digital tools empowered with Generative AI. Every statement in this passage reflects a commitment to constructive criticism and forward-looking vision, in line with the most rigorous scholarly tradition and conforming to APA 7th edition standards (American Psychological Association, 2020). Surely, in the future we will see the Standards intertwined with rigorous criteria with the use of Generative AI, such as ChatGTP, or integrating an international system for the elaboration of items and criteria; or perhaps there will be other models and languages that go beyond the above.

## REFERENCES

Acree, J., Hoeve, K.B., Weir, J.B. (2016). *Approaching the validation of accountability systems. Unpublished paper and presentation*. ERM 600: Validity and Validation, University of North Carolina at Greensboro.

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, 58, 98–110. https://doi.org/10.1111/ejed.12542

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). American Educational Research Association.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bardin, L. (2011). Análisis de contenido (3ª ed.). Ediciones Akal.

Borsboom, D. (2009). Educational Measurement (4th ed.). Structural Equation Modeling-a Multidisciplinary Journal, 16 (4), 702-711. https://doi.org/10.1080/10705510903206097Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Borsboom, D., Cramer, A., Kievit, R., Scholten, A., & Franic, S. (2009). The end of construct validity. In *The concept of validity* (pp. 135-170).

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brennan, R. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 36, 295–317.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 171–191.

Carrillo, B.; Sánchez, M., & Leenen, I. (2020). El concepto moderno de validez y su uso en educación médica. *Investigación en Educación Médica*, 98-106. https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, 29(1), 19-27. https://doi.org/10.1177/0265532211417211

Chapelle, C. (2021). *Argument-Based Validation in Testing and Assessment.* SAGE.

Chapelle, C., & Sauro, S. (2017). Introduction to the Handbook of Technology and Second Language Teaching and Learning. *The Handbook of Technology and Second Language Teaching and Learning*, 1–9. https://doi.org/10.1002/9781118914069

Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. Routledge.

Chapelle, C., Enright, M., & Jamieson, J. (2010). Does an Argument-Based Approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3-13.

Chomsky, N., Roberts I., & Watumull, J. (8 de marzo de 2023). Noam Chomsky: The False Promise of ChatGPT. The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2018). Gathering and Evaluating Validity Evidence: The Generalized Assessment Alignment Tool. *Journal of Educational Measurement*, 55(4), 477–512.

Cohen, L., Manion, L. & Morrison, K. (2007). *Research methods in education*. Routledge.

Cook, D., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*, 49, 560-575. doi: 10.1111/medu.12678

Cronbach, L. J. (1971). Test Validation. In R. Thorndike (Ed.), *Educational Measurement* (2nd ed., p. 443). American Council on Education.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), Educational measurement (pp. 621– 694). Washington, DC: American Council on Education.

De Jong Gierveld, J. (1987). Developing and testing a model of loneliness. Journal of *Personality and Social Psychology*, 53(1), 119-128. https://doi.org/10.1037/0022-3514.53.1.119

De Jong Gierveld, J., & van Tilburg, T. G. (1992). Triangulatie in operationalization method. In G. J. N. Bruinsma & M. A. Zwanenburg (Eds.), *Methodologie voor Bestuurskundigen: Stromingen en Methoden* (pp. 273-298).

De Jong Gierveld, J., & van Tilburg, T. G. (2011). *Manual of the Loneliness Scale 1999*. Vrije Universiteit, Department of Social Research Methodology.

Delgado-Rico, E.; Carretero-Dios, H., & Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology*, 12(3), 449-459. https://www.redalyc.org/pdf/337/33723713006.pdf

Derrida, J. (1997). Una filosofía deconstructiva. *Zona erógena*, 35.

Diana Arya, Anthony Clairmont, Daniel Katz & Andrew Maul (2020). Measuring Reading Strategy Use. *Educational Assessment*, 25:1, 5-30. https://doi.org/10.1080/10627197.2019.1702464

Embretson, S. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher*, 36, 449-455. https://doi.org/10.3102/0013189X07311600

Embretson, S. (2016). An Integrative Framework for Construct Validity. https://doi.org/10.1002/9781118956588.ch5.

Embretson, S., & Gorin, J. (2001). Improving Construct Validity With Cognitive Psychology Principles. *Journal of Educational Measurement*, 38(4), 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Evelyn S. Johnson, Angela Crawford, Laura A. Moylan & Yuzhu Zheng (2020). Validity of a Special Education Teacher Observation System, Educational Assessment, 25:1, 31-46, DOI: 10.1080/10627197.2019.1702461

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A Validity-Based Framework for Understanding Replication in Psychology. Personality and Social Psychology Review, doi:10.1177/1088868320931366

Fan, J. (2014). Chinese test takers' attitudes towards the Versant English Test: a mixed-methods approach. Language Testing in Asia, 4(1). doi:10.1186/s40468-014-0006-9

Ferrara, S. (2007). Our field needs a framework to guide development of validity research agendas and identification of validity research questions and threats to validity. Measurement: Interdisciplinary Research and Perspectives, 5(2–3), 156–164.

Gafni, N. (2016). Comments on implementing validity theory. *Assessment in Education: Principles, Policy & Practic*e. https://doi.org/10.1080/0969594X.2015.1111195

Gallent-Torres, C., Zapata-González, A., & Ortego-Hernando, J.L. (2023). El impacto de la inteligencia artificial generativa en educación superior: una mirada desde la ética y la integridad académica. *RELIEVE*, 29(2), art. M5. http://doi.org/10.30827/relieve.v29i2.29134

García-Medina, A.; Martínez-Rizo, F.; Cordero-Arroyo, G., & Caso-Niebla, J. (2017). Evolución del concepto de validez en la medición educativa. https://www.researchgate.net/publication/325346472_Evolucion_del_concepto_de_validez_en_la_medicion_educativa

Garfield, E. (1979). Citation Indexing—Its Theory and Application in Science, Technology, and Humanities. Wiley.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, 11(1-2), 1-18.

Hoeve, K.B. A validity framework for accountability: educational measurement and language testing. Lang Test Asia 12, 3 (2022). https://doi.org/10.1186/s40468-021-00153-2

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence*, 5, 100165. https://doi.org/10.1016/j.caeai.2023.100165

Jawhar, S., Al, M., Alhawsawi, S. & Alkushi, A. (2021). Validating English Language Entrance Test at a Saudi University for Health Sciences. Arab World English Journal (AWEJ), 12(2), 49-71. DOI: https://dx.doi.org/10.24093/awej/vol12no2.4

Jong-Gierveld, J. (1987). Developing and testing a model of loneliness. Journal of Personality and Social Psychology, 53(1), 119–128. https://doi.org/10.1037/0022-3514.53.1.119

Kane, M. (2002). Inferences about Variance Components and Reliability-Generalizability Coefficients in the Absence of Random Sampling. *Journal of Educational Measurement*, 39 (2), 165-181.

Kane, M. (2006a). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Kane, M. (2006b). Current Concerns in Validity Theory. *Journal of Educational Measurement*, 38(4), 319-342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2011). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17. doi:10.1177/0265532211417210

Kane, M. (2013a). 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. (2013b) The Argument-Based Approach to Validation, School Psychology Review, 42:4, 448-457.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement,* 38(4), 319-342. https://www.jstor.org/stable/1435453

Kerlinger, F. y Lee, H. (2001). Investigación del comportamiento: métodos de investigación en ciencias sociales. McGraw Hill.

Koretz, D. (2008). *Measuring up. What educational testing really tells us.* Harvard University Press.

Koselleck, R. (2000). *Los estratos del tiempo: estudios sobre la historia.* Paidós Ibérica.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. Language Testing, 34(4), 451–475. doi:10.1177/0265532217713951

Lavery, M., Bostic, J., Kruse, L., Krupa, E., & Carney, M. (2020). Argumentation Surrounding Argument-Based Validation: A Systematic Review of Validation Methodology in Peer-Reviewed Articles. Educational Measurement: Issues and Practice. doi:10.1111/emip.12378

Lindquist, E. F. (Ed.). (1951). Educational measurement. American Council on Education.

Lingard L. Writing with ChatGPT: An Illustration of its Capacity, Limitations & Implications for Academic Writers. *Perspectives on Medical Education*, 12(1): 261–270. DOI: https://doi.org/10.5334/pme.1072

Lissitz, R. (2009). *The concept of validity: Revisions, new directions, and applications*. Information Age Publishing.

Markus, K. & Borsboom, D. (2013). *Frontiers of Test Validity Theory. Measure, Causation and Meaning.* Routledge.

Messick S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* (18), 2, 5-11.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.

Nasution, N. E. A. (2023). Using artificial intelligence to create biology multiple choice questions for higher education. *Agricultural and Environmental Education*, 2(1), em002. https://doi.org/10.29333/agrenvedu/13071

Newton, P., & Shaw, S. (2014). *Validity in educational & psychological assessment*. SAGE.

Paul E. Newton & Jo-Anne Baird (2016) The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23:2, 173-177. https://doi.org/10.1080/0969594X.2016.1172871

Pedrosa, I., Suárez-Álvarez, J., & García-Cueto, E. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica*, 10(2), 3-18. https://dx.doi.org/10.5944/ap.10.2.11820

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educational Psychologist, 51(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Santamaría, F. (2012). De la analítica al (neo) pragmatismo. El giro de la filosofía anglosajona. *Revista Colombiana de Humanidades*, 80, 105-143. https://www.redalyc.org/pdf/5155/515551990007.pdf

Schilling, S. G. (2004). Conceptualizing the Validity Argument: An Alternative Approach. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 178–182.

Schmidt, T. , & Strasser, T.(2022). Artificial Intelligence in Foreign Language Learning and Teaching Anglistik, Volume 33, Issue 1 (2022), 165 – 184. DOI: https://doi.org/10.33675/ANGL/2022/1/14

Shepard, L. (2016) Evaluating test validity: reprise and progress. *Assessment in Education: Principles, Policy & Practice*, 23(2), 268-280. https://doi.org/10.1080/0969594X.2016.1141168

Sijtsma, Klaas. (2009). Correcting Fallacies in Validity, Reliability, and Classification. *International Journal of Testing*, 9, 167-194. https://doi.org/10.1080/15305050903106883.

Sireci, S. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50(1), 99–104.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, 36(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sireci, S. G. (2016). On the validity of useless tests. Assessment in Education: Principles, *Policy and Practice*, 23. https://doi.org/10.1080/0969594X.2015.1072084.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100–107.

Sireci, Stephen & Doğan, Nuri. (2017). Interview with Stephen G. Sireci on Validity. *Eğitimde ve Psikolojide Ölçme ve DEğerlendirme*, 8, 158-168.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Merrill Publishing Co/Prentice-Hall.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Watson, P. (2002). Introducción: la evolución de las leyes del pensamiento. En *Historia intelectual del siglo XX*, (pp.11-15). Crítica.

Zumbo, B. & Chan, E. (Ed.) (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. Springer Cham.

## Contribution of each author to the manuscript:

| | % of contribution of each author | | |
|---|---|---|---|
| **Task** | **A1** | **A2** | **A3** |
| A. theoretical and conceptual foundations and problematization: | 50% | 25% | 25% |
| B. data research and statistical analysis: | - | - | - |
| C. elaboration of figures and tables: | 70% | 15% | 15% |
| D. drafting, reviewing and writing of the text: | 70% | 15% | 15% |
| E. selection of bibliographical references | 30% | 30% | 40% |
| F. Other (please indicate) | - | - | - |

## Indication of conflict of interest:

There is no conflict of interest

## Source of funding

There is no source of funding

## Acknowledgments

There is no acknowledgments