

Article

PROVAL: A Comprehensive Protocol for Developing Validation Studies

Isabel Benítez^{1,2}, Andrés González-Gómez^{1,2} and José Luis Padilla^{1,2}

¹ University of Granada (Spain)

² Mind, Brain and behavior Research Center (CIMCYC) Granada (Spain)

ARTICLE INFO

Received: 23/12/2025

Accepted: 05/05/2026

Keywords:

Validity evidence

PROVAL

Validation protocol

Sources of validity evidence

ABSTRACT

Introduction: Validity is a fundamental aspect of psychological assessment because it examines the extent to which interpretations of the scores produced by an instrument support its intended purpose. While validation studies are frequently conducted, they often emphasize specific phases or validation methods, such as pilot testing or expert judgments, while neglecting current additional sources of validity evidence that could improve an instrument's psychometric properties. This limitation is partly due to a lack of detailed easy-to-use guidelines, which make validation processes inaccessible to non-expert researchers and practitioners. Additionally, the traditional focus on quantitative methods often overlooks the potential benefits of complementary approaches to validation that will help people design more comprehensive validation studies. **Method:** To address these challenges, the present study introduces PROVAL, a nine-step protocol designed to guide researchers in conducting validation studies. **Results:** PROVAL promotes a reflective approach to help researchers and practitioners identify and prioritize the most important sources of validity evidence for their specific validation requirements. **Conclusions:** An illustrative application of PROVAL is presented, demonstrating its usefulness in optimizing resources and ensuring instruments effectively fulfill their intended purposes.

PROVAL: un Protocolo Integral para el Desarrollo de Estudios de Validación

RESUMEN

Introducción: La validez es un aspecto fundamental en la evaluación psicológica, ya que examina en qué medida las interpretaciones de las puntuaciones obtenidas mediante un instrumento respaldan su uso previsto. Aunque los estudios de validación son frecuentes, habitualmente reflejan las mismas fases y métodos (estudios piloto o evaluación de personas expertas), descuidando otras fuentes de evidencia que podrían mejorar las propiedades psicométricas de los instrumentos. Esta limitación se debe, en parte, a la falta de directrices, lo que a su vez dificulta que profesionales en áreas aplicadas y en investigación puedan desarrollar estudios de validación. Además, el enfoque tradicional centrado en métodos cuantitativos suele ignorar los beneficios potenciales de enfoques complementarios que contribuyan al diseño de estudios de validación más integrales. **Método:** Para abordar estos desafíos, el presente estudio introduce PROVAL, un protocolo de nueve pasos diseñado para guiar a los investigadores en la realización de estudios de validación. **Resultados:** PROVAL promueve un enfoque reflexivo que ayuda a identificar y priorizar las fuentes de evidencia de validez más relevantes según las necesidades específicas. **Conclusiones:** Se presenta una aplicación ilustrativa de PROVAL, que demuestra su utilidad para optimizar recursos y garantizar que los instrumentos cumplan eficazmente con sus propósitos previstos.

Palabras clave:

Evidencias de validez

PROVAL

Protocolo de validación

Fuentes de evidencia de validez

Validity is one of the most extensively discussed topics in psychometrics. Many authors have proposed definitions, theoretical proposals, and validation methods over the years (Crocker & Algina, 1986; Cronbach & Meehl, 1955; Lissitz & Samuels, 2007; Messick, 1989; Sireci, 2009; Zumbo, 2009). However, several long-standing issues and unresolved problems remain salient today. Validity is still mainly regarded as a property of an instrument, although Messick (1989) conceptualized it as an integrated and ongoing process that requires multiple perspectives to support the intended use of test and questionnaire scores. Likewise, most of validation studies follow the same routine process and structure, despite Kane (1992) introducing the argument-based approach to put the focus on the specific arguments supporting the creation of a test. This author also proposed practical guidelines for researchers and practitioners (Kane, 2006, 2013a), but these are not habitually considered. Despite the key role of the construct, the focus of validation studies should be the intended purpose and the interpretations of test scores, as Sireci (1998, 2009, 2013) emphasizes. However, the intended purpose is not even identified in most validation studies. Finally, researchers approach validity as an isolated issue, but it should incorporate cognitive and contextual variables of individuals responding to test items because, as Zumbo and colleagues (Chen & Zumbo, 2017; Zumbo & Hubley, 2017) noted, the “ecology” (i.e., personal, family, and societal factors), influences how people respond. The inability of psychometricians to eliminate these traditional and limiting ideas and practices has led to validity being treated as a formality, causing it to lose its primary purpose: ensuring that an instrument provides adequate measures of a construct that are suitable for the intended purpose. Addressing this challenge involves getting back to the concept and reviewing the main foundations of validity.

Starting with the concept, there appears to be a broad current consensus about the definition provided in the latest edition of the *Standards for educational and psychological testing* (hereafter, *Standards*) (American Educational Research Association [AERA] et al., 2014), where validity is defined as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA et al., 2014, p. 11). The two most recent editions of the *Standards* (AERA et al., 1999, 2014), conceptualize validity as a unified process centered on the construct, thereby burying the traditional division in types of validity. This unified perspective frames validity as a continuous process in which multiple sources of “data”, such as psychometrics and expert judgments, serve as “evidence” supporting (or failing to support) the interpretations derived from test scores. Specifically, the *Standards* describe five potential sources of evidence: test content, response processes, internal structure, relationships with other variables, and testing consequences. Depending on the validity argument formulated to guide the validation study and the specific purposes of the assessment tool, evidence will be collected from the most relevant and appropriate sources; however, there is still no clear agreement on which sources should be prioritized.

Among the five sources, evidence based on test content examines the degree to which a test’s content adequately represents the intended construct (for specific approaches, see Sireci & Faulkner-Bond, 2014). Evidence based on response processes evaluates the alignment between the construct being measured and participants’ actual response processes, aiming to identify factors that may elicit responses unrelated to the construct (see, for example, Engelhardt

& Goldhammer, 2019; Neuert & Lenzner, 2016; Padilla & Benítez, 2014; Rushton et al., 2015). Evidence based on internal structure focuses on whether the relationships among items and dimensions of the test are consistent with the theoretical interpretation of the scores (for a review of relevant procedures, see Rios & Wells, 2014). Evidence based on relationships with other variables assesses the extent to which associations between test scores and external measures align with the theoretical expectations derived from the construct (for an illustration, see Oren et al., 2014). Finally, evidence based on testing consequences concerns the evaluation of the intended and unintended outcomes of testing and assessment processes for all stakeholders starting by the examinees themselves. Several methodological approaches have been proposed to address this aspect (Ilescu & Greiff, 2021; Lane, 2014), particularly to support the formulation of interpretation and use argument (IUA) that “explicitly links the inferences from tasks to decisions and uses, and then to the resulting consequences” (Lane & Marion, 2025, p. 260). In fact, these authors expand the definition of validity presented in the *Standards* to explicitly incorporate the role of the consequences, referring to validity as “the extent to which theory and evidence support or refute proposed and actual or enacted test score interpretations, uses, and consequences...” (p. 192).

Despite numerous advances, guidance on how to gather and prioritize different sources of validity evidence remains relatively limited. To address this gap, several practical frameworks have been proposed. Both Kane and Sireci, in their previously cited works, developed models and general recommendations to guide validation by outlining steps and recommendations for identifying threats to validity and designing studies to examine them. More recently, Sireci and Benítez (2023) introduced a practical guide for conducting validation studies, detailing procedures for addressing the five sources of validity evidence and proposing a structured sequence of steps to support the validation process. This guide represents a valuable resource for approaching validation from a comprehensive perspective, offering detailed information on procedures for collecting different sources of validity evidence. However, both this guide and other existing frameworks often remain inaccessible to researchers and practitioners who are not experts in psychometrics. This limitation highlights the need to disentangle the validity process, make it more user-friendly, and present it in ways that facilitate rigorous yet practical application. Specifically, there is a need to support researchers in making decisions throughout the design and development of validity studies.

With this goal in mind, several studies have sought to illustrate how proposed frameworks can be applied in specific contexts. For example, Cook et al. (2015) developed a practical introduction to Kane’s framework to make it “accessible and applicable to a wide range of assessment tools and activities” (p. 561). Other authors have focused on outlining the essential elements of validation processes to establish a basic level of confidence in what an instrument measures. Yasir (2016), for instance, summarizes the processes of adaptation and validation, providing a detailed description of the tasks and steps involved in developing assessment instruments suitable for evaluating public health variables in cross-cultural studies. Although Yasir’s framework relies on definitions of validity that differ from those in the *Standards*, it nonetheless seeks to clarify the relationship between the concept of validity and the contributions of specific statistical techniques. Additionally, other tools illustrate

the necessary steps for creating or analyzing the psychometric properties of assessment instruments in specialized domains. For example, Tsang et al. (2017) outline the procedures for developing, adapting, and validating instruments in the field of pain medicine. Their guide offers step-by-step recommendations for producing sound measurement instruments and includes a glossary to make technical terminology more accessible. However, these proposals overlook some of the more established aspects of validity, such as identifying the intended purpose or the integration of multiple sources of validity evidence.

These specific applications offer valuable guidance for researchers seeking to develop or employ psychometrically sound instruments. However, many of these guides present “automatic” or overly standardized procedures that discourage reflection on the unique requirements of each instrument and the assessment context. Consequently, to promote a deeper and more comprehensive understanding of the constructs being measured it is essential to design validation processes that move beyond the mere fulfillment of formal requirements. This aim underscores the responsibility of psychometricians to provide resources that facilitate the rigorous development of validation practices in both research and applied contexts. As emphasized in the *Standards*, “validation should be understood as the process of constructing and evaluating arguments for and against the intended interpretations of scores and their relevance for the proposed use” (AERA et al., 2014, p. 11). The statement highlights the importance of engaging in thoughtful, context-sensitive validation rather than following routine and predetermined steps. It also calls for the continuous analysis of psychological and educational phenomena, recognizing that different constructs cannot be measured through identical procedures but instead require tailored approaches. This perspective shifts the focus toward the careful selection of methods and the critical understanding of tools that enable a more profound exploration of the phenomenon under study.

To address previous limitations, we need methodological approaches that prioritize the intended purpose of the instrument and help articulating it, foster reflection on its requirements, assessment context, and allow for the flexible integration of multiple sources of information. Therefore, it is necessary to move beyond traditional approaches that restrict the validation process to a limited set of questions and instead promote alternatives that consider the full range of possibilities for addressing potential threats. In this regard, pragmatism, one of the main foundations of Mixed Methods Research (MMR), becomes central. MMR has transformed the way research designs are conceptualized, as it enables researchers to approach phenomena from an open perspective in which specific procedures are not predetermined but are selected based on their ability to address the research question (Creswell & Plano-Clark, 2011; Tashakkori & Teddlie, 1998). Grounded in pragmatism epistemology foundations, MMR emphasizes the centrality of the research objective, using it as the key element guiding the selection of design, techniques, and procedures within a study (Johnson & Christensen, 2019). The contributions of MMR, along with empirical evidence supporting its usefulness for addressing complex phenomena, suggest that this approach may also offer valuable insights in the context of validity. A closer examination of the principles underlying both “modern” validity theory and MMR reveals a clear conceptual alignment between them. From the earliest discussions of validity, scholars

have emphasized its dynamic nature, framing validation as an evolving and context-dependent process. Messick (1989) described validation as a hypothesis-testing endeavor that involves the use of multiple procedures to gather information, the consistency of which is then evaluated to determine the extent to which the intended interpretations can be supported. He also underscored the influence of context and group characteristics on test scores, as well as the limitations of purely local validation efforts, issues later expanded upon in frameworks such as the ecological model (Chen & Zumbo, 2017). Moreover, the integration of data from multiple sources is a recurrent theme across these perspectives, as the accumulation of validity evidence inherently involves combining information to strengthen the overall validity argument.

MMR provides two others important contributions to the discussion of validity. On one hand, the incorporation of qualitative procedures to offer relevant information about the same phenomena examined through quantitative approaches, but from a different perspective (Dellinger & Leech, 2007). This idea helps recover aspects of validity that have been largely overlooked, particularly those sources of evidence that are, by their nature, difficult to access through quantitative methods like response processes and intended and unintended testing consequences. MMR can help place these sources on an equal footing with other types of evidence traditionally included in validation studies. The reviews by Cizek et al. (2008) and Zumbo (2025) highlight how empirical studies on validity often fail to incorporate certain essential elements. According to their findings, only a small minority of studies up to the date of publication adopt a unified view of the concept of validity, with most treating validity as a property of the test itself. Consequently, the conceptualization of validity in terms of evidence is often missing. Among the studies that do adopt this perspective, references to evidence based on response processes and testing consequences are rare, reflecting, to some extent, the uncertainty researchers face when addressing these sources of information. Adopting approaches that incorporate qualitative data and sources of evidence not typically collected does not imply that these elements should be always mandatory. In fact, validation studies based on more traditional sources of validity evidence can, in many cases, provide sufficient support. However, in other situations, the characteristics of the target construct, the assessment instruments or context itself may introduce threats related to response processes and the consequences of evaluation. Kane (2013a) emphasized that “our inability to specify all potential consequences in precise quantitative terms does not eliminate the need to evaluate the consequences” (p. 61). Moreover, fairness in measurement has gained prominence with the rise of comparative studies across linguistic, cultural and sociodemographic groups that can involve decision-making processes capable of significantly affecting the lives of those being assessed differently. In fact, fairness cannot be separated from the concept of validity (Lane & Marion, 2025). Therefore, evaluating consequences can, in many instances, contribute to ensuring equity among individuals and groups.

On the other hand, MMR can also contribute to validation by helping to “summarize” and “report” validity arguments, an unresolved issue in the latest *Standards*, as Lane and Marion (2025) point out, and one they describe as a “challenging activity” when professionals and practitioners must integrate and synthesize diverse types of validity evidence and supporting arguments. MMR emphasizes the development of a “narrative” that synthesizes and summarizes both

quantitative and qualitative findings when evaluating intervention programs, while also being adaptable to diverse audiences, including stakeholders, practitioners, academics, and policymakers.

MMR is not conceived as an alternative to traditional validation studies but rather as an approach that enriches the information obtained through more commonly used quantitative methods. The relevance of MMR within the validation process becomes evident when considering some of its defining features, such as its nature as a hypothesis-testing process, its dynamic character, and the centrality of the study's purpose (Johnson & Christensen, 2019). Furthermore, from a modern perspective on validity, there is a clear call to incorporate multiple sources of evidence. In this context, MMR can be viewed as a response to this call, as it facilitates the integration of information from often overlooked sources of evidence, such as response processes or the consequences of test use.

Some counterarguments to these ideas have also emerged in literature. For example, Russell (2022) contended that it is not possible to synthesize or integrate evidence, that test developers are not responsible for the potential misuse of scores, and that certain arguments, such as those related to consequences, cannot be addressed until some time has passed after the evaluation. However, integration becomes meaningful when it involves gathering different perspectives on the same phenomenon from the "parts" involved in the assessment to inform decision-making; while we cannot be accountable for all possible uses, guidelines can be established to help prevent misuse; and although consequences cannot be evaluated *a priori*, they can indeed be anticipated.

The previous reflections on measurement highlight several challenges that we, as psychometricians, must address. As argued, the limited practical applicability of proposals for developing validity studies, combined with the inherent dynamism of psychological and educational constructs in evolving contexts, which pose ongoing challenges for validation, has led researchers and practitioners to rely on tools that offer simplified validation processes, make use of available data, or involve the design of readily feasible empirical studies. However, as emphasized in the *Standards*, validation should aim to achieve a comprehensive understanding of the phenomenon under study, enabling the confirmation of the appropriateness of the interpretations made based on the test scores.

Three key considerations should guide the development of an approach that facilitates both the evaluation of how test-takers' responses support the intended interpretations and uses of scores, and the understanding of the arguments that justify these interpretations and uses. The first concerns the conceptualization of validity in terms of sources of evidence that can inform all aspects of the phenomenon under study, thereby ensuring that potential threats to the proposed interpretations are identified. The second involves adopting a methodological framework capable of addressing a phenomenon as complex as validity while providing guidance for integrating data from multiple sources of evidence and incorporating contextual factors that may influence measurement. The third relates to selecting procedures and techniques that best address the validation projects, including qualitative methods that allow for the examination of aspects of the phenomenon not accessible through quantitative approaches, such as the causes or origins of statistically identified factors. These methods should also facilitate the collection of less frequently explored types of evidence, such as those based on response processes and testing consequences.

All of this underscores the need for a validation protocol that, while methodologically rich and rigorous, provides the necessary flexibility to accommodate validation processes for instruments with varying characteristics, assessing constructs of different natures in diverse contexts. This must be done in such a way that, as described in the *Standards*, it is possible to generate "a coherent argument regarding the degree to which evidence and theory support the interpretation of test scores for specific uses" (AERA et al., 2014, p. 21). Such a protocol should also encourage potential users to reflect on validity arguments, which should be the cornerstone of the process beyond the specific tools required.

For these reasons, the present study aims to propose PROVAL, a mixed-methods protocol for obtaining validity evidence, designed to guide the formulation of validation studies that support the intended uses and interpretations of instruments assessing psychological and educational variables across various contexts. PROVAL guides the design and development of validation studies by posing specific questions that facilitate the application of descriptions from previous frameworks to concrete situations. In other words, PROVAL not only outlines the steps for validation but also indicates how to construct validity arguments by articulating the intended purpose and tracing the full path to identifying potential threats. The proposed protocol is intended as a starting point for incorporating innovations related to validity theory, guidance on conducting validation studies, or new procedures for gathering validity evidence. This work also aims to illustrate the use of PROVAL by showing a specific application.

Description of PROVAL

PROVAL is a proposal to address validity studies based on both the theoretical and practical approaches available in literature and the demands of professionals and researchers. PROVAL was developed following the previously described guidelines and tools. Specifically, it draws on the phases proposed by Sireci (2013) and Sireci and Benítez (2023), as well as on the guidelines for formulating validity arguments proposed by Kane (1992). Within the argument-based approach to validity, Kane (1992, 2006, 2013a, 2013b) distinguishes between the interpretation/use argument (IUA), which specifies the intended interpretations and uses of test scores, and the validity argument, which evaluates the plausibility of these interpretations based on empirical evidence. Although this distinction has been highly influential in contemporary validity theory, for the purposes of the PROVAL framework, we adopt a more integrated and operational use of the term "validity argument," consistent with the interpretation proposed by Sireci (2013), in which both the specification of intended interpretations and the evaluation of supporting evidence are considered within a unified argumentative structure. This choice was motivated by the applied nature of PROVAL, which aims to translate validity theory into a stepwise protocol that can be readily implemented by researchers and practitioners. Importantly, this does not imply a departure from Kane's framework, but rather a pragmatic adaptation of terminology to facilitate the operationalization of argument-based validation within a structured methodological guide. In line with the definition of the sources of evidence outlined in the *Standards* and the recommendations for the most appropriate procedures to gather each type of evidence, a series of steps are established to guide and structure the design and development of a validation study that

focuses on the demands arising from the intended meaning of the scores and the use of the target instrument.

PROVAL is a protocol composed of nine steps, each representing a single activity or task related to the validation process. This means that traditional questions, such as “articulating the intended purposes,” can be divided into multiple parts to ensure that users engage in a thorough reflection on all relevant elements, rather than providing general statements that omit important details. Each step includes guidelines aimed at helping researchers reflect on the elements that, depending on the content of the inference and the kind of instrument, may be relevant for obtaining validity evidence. To this end, each step is divided into three sections: description, reflections, and results. The first section provides a brief description of what the step entails. The second section outlines the questions that researchers and practitioners must ask themselves. The answers to these questions lead to the results in the third section. PROVAL is designed to be used for the validation of inferences from measurement instruments that assess psychological and educational variables. The application of PROVAL will culminate in providing evidence of the test’s utility in achieving the intended objective and offering arguments to support the intended purpose and scores’ interpretations. Both the Spanish and the English version of PROVAL are available in appendixes 1 and 2 in the following link: <https://osf.io/zcxjq/overview>. Table 1 summarizes the steps of the PROVAL protocol, a framework designed to guide the design and development of validation studies focused on the specific requirements of measurement instruments.

Although PROVAL is an evidence-based proposal, one of its main contributions is its accessibility to users and its responsiveness to practitioners’ needs. Consequently, the protocol will be regularly updated as additional evidence is collected, incorporating clarifications and improvements to continue facilitating the validation process. Final versions will be made available in the “Current version of PROVAL” folder within the OSF project.

Illustrating the Application of PROVAL with the Spanish Version of the Inventory of Attitudes Toward Sexist/Nonsexist Language (IASNL-G-S)

To make easier the understanding of each step proposed in PROVAL, the following presents its application to the Spanish version of the *Inventory of Attitudes Toward Sexist/Nonsexist Language* (IASNL-G-S). The instrument represents a typical research tool in psychology, exhibiting many of the common characteristics researchers encounter when selecting a scale. Additionally, it presents several challenges related to cultural considerations and the potential consequences for the assessed participants. The IASNL-G is an instrument developed by Parks and Roberton (1998) to evaluate attitudes toward the use of sexist/nonsexist language. The instrument consists of 21 items scored on a 5-point Likert-type scale, where lower scores indicate a negative attitude, and higher scores reflect positive attitudes. The items are divided into three sections: 1) The first section asks participants to rate their agreement with 12 statements; 2) the second section requires participants to indicate the level of sexism of specific words and expressions in four items; and 3) the final section includes five items that ask participants about their habitual use of certain expressions.

Three researchers collaborated on the Spanish adaptation of the IASNL-G (IASNL-G-S), employing a committee approach (Harkness & Schoua-Glusberg, 1998). The committee approach to translation design follows International Test Commission Guideline TD-2 (5), which advocates for a double translation and reconciliation procedure (International Test Commission, 2017). Two researchers independently translated the original English version into Spanish, while a third researcher acted as the adjudicator. The adjudicator reviewed and compared both translations, resolving discrepancies to create a final integrated version that incorporated contributions from both translators. The adjudicator’s role involved two main tasks. First, when both translators produced identical or comparable translations, the adjudicator accepted the consensus version. In cases where one translation was clearer, this version was selected. Second, for elements where discrepancies arose, the adjudicator marked specific items for discussion. A joint meeting was held to address these discrepancies, where each researcher presented their rationale, and final decisions were made on the best translation choices. All three researchers were familiar with the content of the scale and

Table 1
Summary of the PROVAL Validation Protocol

Step	Title	Description	Main Result
1	Intended purpose and scores’ interpretations	Defines the test purpose, context of use, targeted population, and the decisions to be made based on scores.	Formulation of the intended purpose and score interpretations.
2	Formulation of validity arguments and assumptions	Identifies arguments based on the five sources of validity evidence from the <i>Standards</i> (AERA et al., 2014).	Formulation of validity assumptions and hypotheses to be investigated.
3	Prioritization of validation studies	Organizes the validity assumptions based on their relevance according to the test’s intended use (e.g., diagnosis, selection).	A prioritized list of assumptions to guide the study.
4	Nature of the evidence needed	Determines who can provide the information (participants) and whether data will be quantitative or qualitative.	Identification of necessary instruments, participants, and specific data requirements.
5	Selection of collection and analysis procedures	Identifies appropriate methods for data collection and the specific analytical techniques to be used.	Detailed description of the study’s phases and methodology.
6	Design of the validation study	Combines previous steps to establish the order, priority, and nature of the design (qualitative, quantitative, or mixed).	Final configuration and illustration of the validation study design.
7	Development of validation studies	Executes data collection according to the design, addressing recruitment and ethical considerations.	Preparation of all materials and successful data collection.
8	Data analysis and interpretation of results	Performs analyses and relates findings back to the initial validity assumptions.	Evidence supporting or refuting the validity assumptions.
9	Integration of evidence and conclusions	Synthesizes all findings to make a final decision on the relevance of the intended interpretations.	Final conclusions on the usefulness and validity of the test.

had experience with methodological aspects of translation and adaptation processes.

The application of PROVAL framework in the present study shows the first seven steps, which cover everything from outlining the intended use of the instrument to describing the complete design of the validation study and main aspects related to the development of the validation studies. In each step, responses to the reflections and the results derived from those reflections are provided. Appendixes 3 and 4 show the application of PROVAL to the IASNL-G-S in both Spanish and English (<https://osf.io/zcxjq/overview>). The illustrated steps cover the main contributions of PROVAL, although the documents will be updated once the validation studies are completed and the remaining steps are finalized. This will provide a complete example of a comprehensive validation study, including the interpretation and integration of results. Both the protocol and the illustration are conceived as dynamic tools that will be updated to incorporate innovations related to the concept of validity and derived from evidence obtained through the application of the tool in different contexts.

The first step presents information related to the intended purpose and characteristics of both the target population and the instrument. Since this is an adaptation, the information was derived from the original version, which included details about score interpretations. When developing new measures, the scores' interpretations should be described in that step. In the example, the results from step 1 include the description of the scale, the scores interpretations and the target population.

Step 2 evaluates whether available information in the scientific literature about the intended instrument can be interpreted as validity evidence. An affirmative response would apply if prior studies provided validity evidence. In that case, previous results will be summarized to show the current knowledge and identify the outstanding validation questions. In this case, some studies were found where different versions of IASNL-G were applied and provided data about the instrument's functioning. For instance, one of the validation studies concluded that the construct is unidimensional and include both beliefs about nonsexist language and recognition of sexist language, whereas the section measuring "use or willingness to use nonsexist language" seems to represent an independent construct or a linguistic artifact (Parks & Robertson, 2000). Therefore, that information is summarized as part of validity evidence based on internal structure, while other sources of evidence show information as "There are not studies investigating the relations between the scores from IASNL-G and other instruments measuring theoretically related constructs. However, the study presenting the original version of IASNL-G indicates that the scores could be related to those from the Gender-Specific Language Scale (GSLs)". In the results of step 2, we proposed validity assumptions according to the previous evidence, as well as additional assumptions reflecting both our goals and other claims requiring support from validity evidence. For instance, for validity evidence based on relations to other variables, the validity assumption was: "The total score on the IASNL-S will be positively related to the total score on the Gender Equality Attitudes Scale (Lorenzo Moledo et al., 2023)." That was the process for the five sources of validity evidence.

In step 3, validity assumptions were prioritized based on their relevance. The prioritization helps to organize the process to ensure

that the most relevant validity assumptions are addressed. Although collecting the five sources of validity evidence will provide a wider perspective, sometimes it won't be possible and, in these cases, the order will guide how to proceed. For completing the step, PROVAL suggests using the indications on PETEYC, a tool based on previous classifications of test's uses and evidence collected through experts' evaluations. That tool is part of a research project and is publicly available (<https://osf.io/2x6mr/overview>). In the example, all the assumptions were included as the goal of the illustration is to present the most ideal situation. According to the reflections, assumptions were ordered by grouping them under the five sources of validity, and were prioritized as follows: test content, internal structure, relations to other variables, response processes, and testing consequences. That indicates that the main validation study should be focused on collecting validity evidence based on test content followed by those based on internal structure and relations with other variables.

Step 4 identified who can provide the intended data. In the illustration, we would need the participation of two groups of people: experts and participants from the target population. These two groups of informants were linked to specific sources of validity evidence, as questions to reflect require thinking on the intended evidence. For instance, in reflection about the test content, the answers show that: experts can confirm whether the items in the IASNL-G-S reflect the construct "attitudes towards sexist/non-sexist language" and identify causes when that does not occur. Therefore, experts will give quantitative and qualitative information.

Building on this, the reflections in step 5 helped with structuring the validation studies in different phases, aligning them with the groups providing the information and the procedures required to collect the intended data. Questions are designed to describe the phases around the groups of participants and the type of information needed. In the example, three different phases will allow extracting the intended information. First, a group of experts will participate in an expert's judgement task. Secondly, a representative sample from the target population will respond to the IASNL-G-S and other instruments in a pilot study. Finally, a smaller group from the target population will participate in cognitive interviews.

Taking into account these phases and their relevance, step 6 describes the study design as follows [(QUAN+qual) → QUAN → (quan+QUAL)], by using the standard nomenclature for mixed-methods designs (Johnson & Onwuegbuzie, 2004). The representation indicates that the design will have three phases, with two of them collecting both quantitative and qualitative data. The phases will be sequential, as each one will be based on results from the previous. Quantitative data will be the main source in the two initial phases, whereas qualitative data will have a most relevant role in the last one. Finally, some specifications were added in step 7 to facilitate the preparation of materials and the planning of tasks.

After step 7, the subsequent steps involve conducting the phases, analyzing data from each phase, and integrating the results to formulate comprehensive conclusions about the "support" for the intended use of the instrument. These conclusions will guide decisions on necessary modifications and lead to the creation of the final version of the instrument. The illustration will be updated to incorporate all the details related to the development of the study.

Discussion

The present study aimed to propose PROVAL, a mixed-methods protocol for obtaining validity evidence designed to guide the development of validation studies focused on supporting the intended purposes and scores' interpretations of instruments assessing psychological and educational variables in various contexts. The proposal, based on the definition of validity and sources of validity evidence from the *Standards* (AERA et al., 1999, 2014), provides a framework that integrates both the validation guide through validity arguments, as indicated by Kane (1992), and the general phases outlined by Sireci (2013). The result is a series of steps that help researchers reflect on the validity evidence required by the validity argument by responding to specific questions.

To facilitate the understanding of the process, a specific case illustrated the application of PROVAL. Specifically, PROVAL was used to design the validation studies for the Spanish version of the IASNL-G (IASNL-G-S), an instrument originally created to assess attitudes toward sexist/nonsexist language (Parks & Robertson, 2000). The result of the application was a mixed-methods sequential design validation study, where three different phases address the validity questions raised by the validity claims that support the intended interpretation and test use. The first phase focuses on validity evidence based on test content, incorporating expert judgment to assess the overlap between the items and the intended construct. The second phase proposes a pilot study to collect responses from participants within the target population, aiming to analyze the psychometric properties of the items and the scale, as well as to gather validity evidence based on the internal structure and relationships with other variables.

These two phases represent the most common tasks performed in validation studies and would likely be included even if PROVAL was not applied. However, the needs related to these phases become clearer when responding to the reflections, which will also guide the design of the experts' tasks, the identification of appropriate instruments for the pilot study, and the recognition of potential benefits that additional procedures, such as qualitative strategies, could offer.

In addition, the application of PROVAL provided two more relevant pieces of data: the response processes involved in answering the items and the consequences perceived by responders. In the third phase, interviews emerged as an appropriate procedure, as validity evidence based on response processes and testing consequences is needed to support the intended use of the instrument. Furthermore, the data to be collected during the application of this procedure is specified. Thus, PROVAL helps extend the validation process by incorporating additional sources of evidence that could provide a better understanding of the instrument's weaknesses.

The final design proposed by PROVAL would change for different purposes and scores' interpretations, but also if other gaps were detected in relation to the scale's functioning. That points to the flexibility of PROVAL to guide processes with diverse characteristics.

PROVAL helps fill some of the gaps in designing validation studies. First, it encourages approaching the validation process as inherently connected to the intended instrument and prompts reflection on the main limitations that should be addressed during the validation study, rejecting automatic processes where the same steps are conducted. Second, it emphasizes the role of validity evidence based on response processes and testing consequences, traditionally

“ignored” (Cizek et al., 2008), by making their contribution explicit and providing guidance on how to incorporate procedures to collect such evidence. Third, it highlights elements that could be better understood through qualitative procedures, proposing alternative designs such as those based on MMR. Finally, it presents the validation process as a comprehensive task that should not be automatically designed but instead enriched by a deep study of the target instrument, while remaining accessible to non-experts, as it consists of a step-by-step, ordered process.

Regarding the benefits of the proposal, the approach to validation of PROVAL promotes a comprehensive and reflective understanding of validity that goes beyond the automatic processes sometimes proposed. The reflections it encourages are linked to methodological issues that often generate uncertainty and “resistance” among professionals from other fields, who also need to deal with the measurement of psychological variables. Therefore, this project addresses a fundamental need: making validation procedures accessible to professionals and researchers, avoiding reliance on automatic and routine processes limited to the most familiar ones, and promoting reflective actions based on the identification of validity demands and the selection of the most suitable methodology for each situation. This practical and theoretical proposal also has an aspect related to the transfer of scientific knowledge to society. The dissemination of the tool will be facilitated, along with support for researchers and professionals who wish to apply it. Therefore, PROVAL promotes a closer connection between the concept of validity and users of measurement instruments. In addition, PROVAL is intended to serve as a permanent tool, incorporating new findings from both theoretical and empirical studies. As such, it is designed as a dynamic protocol that is open to users' experiences and responsive to innovations in the literature.

PROVAL complements other tools currently available in the field, such as the Test Review Model published by the European Federation of Psychologists' Associations (EFPA, 2025) or the COSMIN Taxonomy of Measurement Properties (Mokkink et al., 2010). The EFPA model identifies elements to consider when assessing test quality, including instrument description, the quality of test materials, reliability, interpretation norms, and validity, while primarily focusing on the utility of validity evidence provided by test authors and the three “classic” sources of validity evidence: content, internal structure, and relations with other variables. COSMIN, in turn, is widely recognized as one of the most comprehensive tools for evaluating the measurement properties of health-related instruments, particularly for evaluative purposes. Although PROVAL does not aim to assess measurement properties directly, it is designed to assist professionals and practitioners in conducting validation studies guided by the intended use of the measures across diverse contexts and fields, including education, social sciences, and health. Importantly, PROVAL includes the collection of validity evidence related to response processes and testing consequences, incorporating fairness within the latter, and adopts a broader perspective than EFPA or COSMIN, which focus primarily on specific sources of evidence or accessibility for particular populations. Therefore, PROVAL, conceived as a dynamic tool, could also be useful to address validity issues identified in both the EFPA and COSMIN frameworks.

Despite the contributions of this study, no evidence has been gathered yet regarding the utility of PROVAL beyond the specific

examples provided. Therefore, future studies will focus on gathering validity evidence to support the intended purpose of the tool. Specifically, we will conduct empirical studies focused on identifying potential limitations of PROVAL and implementing improvements that facilitate its application across different fields. We will also create a space for users of PROVAL to collect information about their experiences with the tool, and proposal to implement adjustments to improve its usability. This will promote the creation of a collaborative and dynamic product that will be periodically reviewed to incorporate any innovations and improvements detected by researchers.

Author Contributions

Isabel Benítez: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing. **Andrés González-Gómez:** Formal Analysis, Supervision, Validation, Visualization, Writing – Review & Editing. **José Luis Padilla:** Supervision, Validation, Writing – Review & Editing.

Funding

Grant funded by Consejería de Universidad, Investigación e Innovación and by ERDF Andalusia Program 2021-2027 (C-SEJ-245-UGR23); and by MICIU/AEI/ 10.13039/501100011033 and ERDF/EU (PID2023-147219NB-I00). This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, and the decision to submit the manuscript for publication

Declaration of Interest

The authors declare that there are no conflicts of interest.

Data Availability Statement

Materials are available at the following link: <https://osf.io/2x6mr/overview>

Acknowledgments

The Mind, Brain and Behavior Research Center receives funding from grants CEX2023-001312- M by MICIU/AEI/10.13039/501100011033 and UCE-PP2023-11 by the University of Granada.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Chen, M. Y., & Zumbo, B. D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 53-68). Springer. https://doi.org/10.1007/978-3-319-56129-5_4
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397-412. <https://doi.org/10.1177/0013164407310130>
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560-575. <https://doi.org/10.1111/medu.12678>
- Creswell, J. W., & Plano-Clark, V. L. (2011). *Designing and conducting mixed methods research*. Sage Publications.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/h0040957>
- Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research*, 1(4), 309-332. <https://doi.org/10.1177/1558689807306147>
- European Federation of Psychologists' Associations AISBL (EFPA). (2025). *Model for the review, description and evaluation of psychological and educational tests (Test review model)*. <https://www.efpa.eu/>
- Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology*, 10, Article 1131. <https://doi.org/10.3389/fpsyg.2019.01131>
- Harkness, J., & Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. Harkness (Ed.), *Cross-cultural survey equivalence* (pp. 87-126). Zentrum für Umfragen, Methoden und Analysen -ZUMA-. <https://www.ssoar.info/ssoar/handle/document/49733>
- Iliescu, D., & Greiff, S. (2021). On consequential validity. *European Journal of Psychological Assessment*, 37(3), 163-166. <https://doi.org/10.1027/1015-5759/a000664>
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests (Second edition)*. [www.InTestCom.org]
- Johnson, R. B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixed approaches*. Sage publications.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 33(7), 14-26. <https://doi.org/10.3102/0013189X033007014>
- 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. Breenan (Ed.), *Educational Measurement* (pp. 17-64). American Council on Education and Praeger.
- Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://psycnet.apa.org/doi/10.1111/jedm.12000> <https://doi.org/10.1111/jedm.12000>
- Kane, M. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, 50, 1115-1122. <https://doi.org/10.2307/23353803>
- Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, 26(1), 127-135. <https://doi.org/10.7334/psicothema2013.258>
- Lane, S., & Marion, S. F. (2025). Validity and validation. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed.,

- pp. 191–275). Oxford University Press. <https://doi.org/10.1093/oso/9780197654965.003.0004>
- Lissitz, R. W., & Samuels, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437–448. <https://doi.org/10.3102/0013189X07311286>
- Lorenzo Moledo, M., Quiroga-Carrillo, A., & Ferraces Otero, M. J. (2023). Estudio psicométrico de una escala sobre actitudes hacia la igualdad de género con mujeres reclusas. *Revista Española de Investigación Criminológica*, 20(2), Article e657. <https://doi.org/10.46381/reic.v20i2.657>
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (pp. 13–103). American Council on Education and Macmillan.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter LM, & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Neuert, C. E., & Lenzner, T. (2016). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology*, 19(5), 501–519. <https://doi.org/10.1080/13645579.2015.1049448>
- Oren, C., Kennet-Cohen, T., & Turval, E. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema*, 26(1), 117–126. <https://doi.org/10.7334/psicothema2013.257>
- Padilla, J.L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>
- Parks, J. B., & Robertson, M. A. (1998). Influence of age, gender, and context on attitudes toward sexist/nonsexist language: Is sport a special case? *Sex Roles*, 38(5), 477–494. <https://doi.org/10.1023/A:1018766023667>
- Parks, J. B., & Robertson, M. A. (2000). Development and validation of an instrument to measure attitudes toward sexist/nonsexist language. *Sex Roles*, 42, 415–438. <https://doi.org/10.1023/A:1007002422225>
- Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108–116. <https://doi.org/10.7334/psicothema2013.260>
- Rushton, P. W., Routhier, F., Miller, W. C., Auger, C., & Lavoie, M.-P. (2015). French-Canadian translation of the WheelCon-M (WheelCon-M-F) and evaluation of its validity evidence using telephone administration. *Disability and Rehabilitation*, 37(9), 812–819. <https://doi.org/10.3109/09638288.2014.941019>
- Russell, M. (2022). Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement: Issues and Practice*, 41(2), 25–35. <https://doi.org/10.1111/emip.12453>
- Sireci, S., & 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., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100–107. <https://doi.org/10.7334/psicothema2013.256>
- Sireci, S. G. (1998). The construct of content validity. *Social indicators research*, 45(1), 83–117.
- Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Information Age Publishing Inc.
- Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50(1), 99–104. <https://doi.org/10.2307/23353800>
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Sage.
- Tsang, S., Royse, C. F., & Terkawi, A. S. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia*, 11(1), 80–89. https://doi.org/10.4103/sja.SJA_203_17
- Yasir, A. S. (2016). Cross cultural adaptation & psychometric validation of instruments: Step-wise description. *International Journal of Psychiatry*, 1(1), 1–4. <https://doi.org/10.33140/IJP/01/01/00001>
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.) *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Information Age Publishing.
- Zumbo, B. D. (2025). Looking through the lens of DIF: Embracing the many ways of being human. *The Score*, 47(4), 2–5.
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research* (Vol. 69). Springer International Publishing. <https://doi.org/10.1007/978-3-319-56129-5>