The Routledge Handbook of Second Language Acquisition and Language Testing

Paula Winke, Tineke Brunfaut

Validating Assessments for Research Purposes

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781351034784-4
Andrea Révész, Tineke Brunfaut
Published online on: 28 Dec 2020

Accessed on: 08 Apr 2021
https://www.routledgehandbooks.com/doi/10.4324/9781351034784-4

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Background

When SLA researchers design or adapt an assessment for research purposes, a key concern is to validate the interpretations of the assessment. As elaborated upon by Chapelle (Chapter 2, this volume), validation in this context refers to the process of obtaining evidence for the claims made about research instruments in view of their intended use and in view of their appropriateness for the research population. It involves collecting information which demonstrates that accurate interpretations are made on the basis of the scores achieved on the assessment. If research instruments are flawed due to issues with validity, any conclusions reached about SLA theory and L2 instruction will lack credibility.

There have been repeated calls for attributing greater importance to validation work in the field of SLA (e.g., Chapelle, 1994, 1999; Douglas, 2001; Norris & Ortega, 2003, 2012; Shohamy, 2000). As early as the 1990s, Chapelle (1994, 1999) emphasised the need to validate SLA instruments for their proposed research purposes, and illustrated and discussed how measurement theory could be applied to the validation of assessments within SLA. An important contribution in SLA-related validation work was the publication of Norris and Ortega’s (2003) measurement framework, which aimed to provide a detailed guide to validity considerations and practices in order to address concerns in validity evaluation in SLA. Another valuable resource is Purpura, Brown, and Schoonen’s (2015) article, which demonstrated how Kane’s (2006, 2013) influential validity framework can be utilised to justify interpretation of scores obtained through the use of SLA elicitation devices.

Likely due to these publications, there is an increasing recognition of the importance of validating research instruments among SLA researchers. However, validation practices often still remain insufficiently aligned with published guidelines for good practice in developing, using and evaluating research instruments (Purpura et al., 2015; Schmitt, Nation, & Kremmel, 2019). With the ultimate aim of improving current validation practices in SLA, this chapter introduces some key frameworks that can be utilised in SLA research, provides illustrative examples of how an argument can be built for validating SLA assessments, and gives some recommendations to guide future research.
**Key Concepts**

**Construct**: Entity that is deduced from theory or observation but cannot be directly measured.

**Construct underrepresentation**: An assessment that fails to measure key aspects of the construct and is thus too narrow in what it assesses. The scores only partly represent the intended construct, and inferences about learners’ ability based on these scores may be incorrect.

**Construct-irrelevant variance**: An assessment that is systematically influenced by factors unrelated to the construct. The scores include systematic measurement error due to factors other than the intended construct, and score interpretations may be incorrect.

**Grounds or datum**: Data that is available.

**Claim**: Statement about intended interpretation based on data.

**Qualifier**: Specification of the extent to which the claim applies.

**Warrant**: Justification for the claim, based on the grounds/datum.

**Backing**: Evidence that corroborates the warrant.

**Rebuttal**: Statement that describes a situation where the claim would not be sustained; counterclaim.

**Interpretation/use argument**: Network of inferences and assumptions leading from observed data to interpretations.

**Validity argument**: A structure that provides the evidence needed in support of and against a proposed interpretation.

---

**Key Issues**

**Measurement and Validation Frameworks**

In SLA validation research, a principal aim is to seek construct-related evidence for the research instruments. When SLA researchers design empirical research studies, they often develop assessments that intend to measure constructs such as knowledge, development, memory, or aptitude, which come from theory or are deduced from observations but cannot be measured directly. Thus, as part of validation, researchers need to gauge how appropriately and precisely the interpretations based on the assessment match a construct’s theoretical definition, and to what extent that construct definition accounts for learners’ results on the assessment. In particular, two main threats need to be considered: **construct underrepresentation** and **construct-irrelevant variance** (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Construct underrepresentation refers to a situation where an assessment does not succeed in capturing key aspects of the construct because the assessment does not adequately sample all aspects of performance required to make inferences about the construct. For example, a grammaticality judgement test (GJT) would suffer from construct underrepresentation if it only included items targeting a small number of grammatical features but claimed to test L2 learners’ grammatical competence as a whole. In this case, construct underrepresentation could be resolved by designing a more comprehensive GJT entailing a broader spectrum of target grammatical features, or by modifying the claim about what the test measures to include only those aspects of grammar actually tested. Construct-irrelevant variance occurs when assessments are systematically influenced by factors that are not relevant to the construct assessors intend to measure. In other words, the assessment entails extraneous variables. For example, if in a GJT some items were misspelled, this could cause construct-irrelevant variance since some learners might classify these as ungrammatical when, in fact, they showed orthographical and/or lexical issues (which are not the intended construct of the assessment).
To help SLA researchers avoid construct underrepresentation and construct-irrelevant variance in their assessments, Norris and Ortega (2003) provided a comprehensive guide for developing and using SLA research tools. Their six-stage measurement framework explains how to proceed through each stage to achieve warranted interpretations about the construct under investigation. The first three stages are concerned with the conceptualisation of the measurement process. In particular, the first step involves defining the construct, that is, what exactly researchers would like to investigate in light of the interpretations or conclusion they would like to reach. Such construct definitions should signal the theoretical assumptions on which they are based and need to be specific so that they can be linked to observable behaviours. For example, a researcher might decide to examine the role of phonological short-term memory (PSTM) in L2 listening, where the construct of PSTM could be derived from Baddeley’s (1986) multidimensional model of working memory. The second step entails identifying the behaviours to be observed with a view to supplying sufficient evidence in support of the intended construct interpretations and to avoid construct underrepresentation. The process of associating behaviours with constructs is ideally informed by the collective theoretical knowledge and empirical work in the field. For example, in the case of PSTM, a well-established behaviour associated with this particular construct is the ability to repeat sound sequences immediately after they have been heard. Then, in the third step, the researcher specifies the tasks or situations that will be used to elicit or observe the target behaviours. Key to this stage is a careful consideration of whether the tasks/situations can indeed generate the envisioned type of evidence. For example, in the study of PSTM and listening, the researcher might decide to employ non-word or digit span tests, in which participants need to reproduce series of non-words or numbers increasing in length, to assess PSTM.

In the next three stages of Norris and Ortega’s (2003) measurement framework, the outcomes of the conceptualisation process are proceduralised. The behaviours are elicited through careful operationalisation and administration of the research instruments, that is, the actual elicitation devices are developed and then employed to collect the data. To ensure construct representation in the research instruments, researchers need to make sure that the tasks fully capture the behaviours they plan to elicit. For example, if the aim is to measure whether learners’ ability to use an L2 construction has improved as a result of an instructional treatment, it is advisable to obtain evidence through a range of tasks creating various contexts for the use of the target feature. There are also several potential sources of error at this stage (e.g., unclear instructions, noisy computer lab), which can be pre-empted through careful piloting of instruments and data collection procedures, to ensure consistent and comparable results (see, for example, Bachman & Palmer, 2010; Mackey & Gass, 2016 for overviews). Once behaviours have been elicited, the observations are coded and/or scored with regard to aspects that are relevant to the constructs examined. Important to achieving validity at this point is to develop a coding and/or scoring scheme that is well aligned with the construct definition. For example, if the goal of the study is to capture syntactic complexity, researchers will need to employ a variety of measures to capture the multidimensional nature of this construct rather than a single index (e.g., Bulté & Housen, 2012). Coding and scoring consistency is a further priority, which can be facilitated through careful coder and rater selection and training, and checking coder and scorer reliability (see Révész, 2012). It is also crucial at this stage to provide statistical evidence that the instruments function as intended, for instance, in the form of internal consistency reliability estimates (Bachman, 2004; Green, 2013). These reliability measures themselves also need to be chosen carefully, since unsuitable ones are sometimes used or their assumptions not checked (see Chapelle, Chapter 2, this volume). Finally, the scores obtained on the assessments need to be subjected to proper statistical analyses to interpret scores in relation to the research questions (Norris et al., 2015; Plonsky, 2013). If the statistical analyses are not appropriately selected and rigorously conducted, interpretations from the instruments may be skewed, misleading, or simply incorrect.
In sum, Norris and Ortega’s (2003) framework provides guidance for designing and implementing research instruments that are appropriate to reach meaningful interpretations about theoretical constructs. After all, an important purpose of using the instruments is to empirically test SLA theories. A limitation of Norris and Ortega’s (2003) framework, however, is its lack of emphasis on validating the use of the instruments. While this is a general weakness of the framework, it constitutes a drawback especially in the context of SLA studies that investigate the effectiveness of L2 instructional treatments. When applied to such situations, the framework does not direct researchers to substantiate whether the interpretations reached are warranted for the relevant learner populations and instructional contexts (Drackert, 2015a). Purpura et al. (2015) and Drackert (2015a) convincingly argue that Kane’s (2006, 2013) argument-based validation framework can successfully address this limitation as it provides a framework for evaluating both intended interpretations and uses of assessments.

Kane’s (2006, 2013) argument-based approach draws on Toulmin’s (1958) six-component model of argument, explained in Table 3.1. Kane (2006, 2013) called for building an argument along the following two stages:

a. Drawing up an interpretation/use argument (IUA) about the proposed interpretations and uses of the assessment, i.e., the plan. This means specifying the series of claims made in the

<table>
<thead>
<tr>
<th>Component</th>
<th>Definition</th>
<th>SLA example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data/ Grounds</td>
<td>Facts or evidence that can be utilised to verify or refute the argument made.</td>
<td>Second language data elicited through a research instrument.</td>
</tr>
<tr>
<td>Claim</td>
<td>Statement which entails conclusions and intended interpretations that are deduced based on the data.</td>
<td>If a group of L2 learners achieves higher scores on a test assessing their knowledge of a target L2 feature after being exposed to one type of instructional treatment than another group of learners receiving a different type of treatment, the claim could be that the former treatment was more effective.</td>
</tr>
<tr>
<td>Qualifier</td>
<td>Specification of the conditions under which the claim applies.</td>
<td>The advantage observed for one of two instructional treatments applies only if there were no initial differences in the two groups’ knowledge of the target construction.</td>
</tr>
<tr>
<td>Warrant or inference</td>
<td>Linking of claims and data, including a set of assumptions.</td>
<td>Inference related to the reliability of coding procedures and the internal consistency of the research instruments: If the coding process and instruments prove reliable, we could conclude that the instrument produces consistent results, thus providing evidence for the claim.</td>
</tr>
<tr>
<td>Backing</td>
<td>Corrobortations of the assumptions entailed in the warrants, involving the provision of theoretical and/or empirical evidence in support of them.</td>
<td>Backing for scoring and instrument reliability could be provided by calculating coder and instrument reliability statistics.</td>
</tr>
<tr>
<td>Rebuttal</td>
<td>Counter-argument that describes situations where the argument would not be sustained.</td>
<td>The better performing group in an experimental study might receive additional instruction in their regular language class, which could account for their higher gains.</td>
</tr>
</tbody>
</table>
interpretation/use argument, including inferences about the domain description, evaluation or scoring, generalisation, explanation, extrapolation, and utilisation (see Chapelle, Chapter 2, this volume for a description of these stages). If the validation process yields evidence against any of the claims, the research instrument under scrutiny will be put to doubt.

b. Providing a validity argument which evaluates the appropriateness of the claims outlined in the interpretation/use argument (IUA) through accumulating theoretical and empirical evidence against or in support of the intended interpretations and uses, i.e., the actual argument.

Purpura et al. (2015) proposed that “the appropriateness and adequacy of evidence needed in a validity argument be determined by the purpose of the study, the research claims we would like to make, and the consequences of using the scores for the intended research purposes” (p. 49), arguing that it may not be practical or possible to gather and analyse data to test all claims. They emphasise, however, that it is important for researchers to be aware of the claims implicated in their research instruments and to take into account whether they have or have not provided evidence for the relevant claims when interpreting their research findings.

Issues in Validation in SLA

Against this background, we now turn to three issues in SLA validation work: how to proceed with validating theory-derived research instruments, how to validate research instruments for various uses, and how to plan validation as an integral phase of SLA research studies.

Validating Theory-Derived Research Instruments. Here, we describe how SLA researchers could approach the validation of research instruments that are constructed on the basis of SLA theory. We clarify this with examples from the context of research on implicit and explicit learning, since probably the most comprehensive and influential validation work in SLA has concerned the development of measures of implicit and explicit knowledge to inform the longstanding theoretical controversy regarding the relationship between these types of knowledge.

Ellis and colleagues (Ellis, 2005; Ellis et al., 2009) initiated work into the validation of interpretations based on measures of implicit and explicit knowledge. Although not explicitly following Norris and Ortega’s (2003) measurement framework, the steps the researchers took exemplify the stages outlined in that model. Drawing on existing theoretical and empirical work, Ellis et al. derived a constellation of observable behaviours assumed to be associated with implicit and explicit knowledge. Through this process, implicit knowledge came to be defined as intuitive, procedural, systematic, automatic, non-verbalisable, and developmentally-constrained, and explicit knowledge as conscious, declarative, actively controlled, verbalisable, and learnable at any age. Next, they designed a battery of assessments to elicit the observable behaviours associated with the construct definitions of implicit and explicit knowledge. An elicited imitation task, an oral narrative task, and a timed GJT were constructed to tap implicit knowledge, and an untimed GJT and a metalinguistic knowledge test were developed to gauge explicit knowledge. Then, a series of studies was conducted to test whether the assessments were indeed successful in eliciting behaviours linked to the constructs of implicit and explicit knowledge (e.g., more versus less systematic behaviour), i.e., whether the assessments’ interpretations would be valid. In addition, statistical analyses were carried out to test the hypothesised links between the participants’ performance on the assessments and their relationships to other factors assumed to be related to the theoretical constructs (e.g., age of starting to learn the L2). Based on the results of the project, the researchers concluded that the instruments could be used as valid measures of implicit and explicit knowledge. As Norris and Ortega (2012) noted, Ellis and colleagues’ project serves as a useful example for future validation work, in several respects. First, the researchers defined the constructs—in this case, implicit and explicit knowledge—synthesising previous theoretical and empirical work, thereby arriving
at construct definitions that reflected researchers’ shared knowledge in the domain. Also, they assessed hypothesised links between the tests that were aligned with theoretical assumptions and expectations shared within the research community.

The project, however, has also been subjected to criticisms, some of which are relevant from the perspective of validation. In particular, several concerns have been raised regarding the appropriateness of the factor analyses conducted. Isemonger (2007) argued that confirmatory rather than exploratory factor analysis should have been employed in Ellis (2005) given that a priori hypotheses were made regarding the nature of the constructs assessed by the instruments. Although Ellis and Loewen (2007) addressed this limitation in a follow-up study, further issues have been highlighted by Vafaee et al. (2017), who pointed out that rival models were not tested in the factor analysis. From the perspective of Kane’s argument-based approach, it could be argued that Ellis and colleagues’ work lacked consideration and evaluation of some key rebuttals. To address this shortcoming and to test new, theoretically derived, comprehension-based measures of implicit knowledge, Vafaee et al. (2017) performed a new study to validate the interpretations drawn from measures of implicit and explicit knowledge. The results challenged the findings of previous research, indicating that while self-paced reading and word-monitoring tasks appear to measure implicit knowledge, timed GJTs may not constitute tests of implicit knowledge. This line of research provides an excellent example of how ongoing validation work, building on previous attempts to operationalise key constructs in the field, can lead to the development of more and more refined instruments that allow for more appropriate interpretations about the constructs of interest.

**Validating Research Instruments for Various Uses.** In SLA research, instruments that have been used with a certain learner population for one purpose in one study are often utilised in new studies for a different purpose and with different groups of participants. This practice, in and of itself, is not questionable. However, the problem is that researchers often borrow instruments without validating them for the populations and uses in their own research; that is, they fail to make sure that the instruments enable reaching warranted interpretations about the learners and contexts they investigate.

Drawing on Drackert (2015a), we illustrate how researchers can validate SLA research instruments for different uses and how Kane’s (2006, 2013) argument-based approach to validity can be utilised in validating SLA assessments. Drackert (2015a) set out to validate a Russian elicited imitation (EI) test as a proficiency assessment for two uses: (1) controlling for participants’ initial oral proficiency at the beginning of a study and (2) screening participants into a study on the basis of their overall listening and speaking proficiency. In validating the EI test for the first use, Drackert took the following steps. The project started with specifying the intended test context and use, followed by the development of the test and the construction of the interpretation/use argument. As shown in Figure 3.1, the first inference in the interpretation/use argument concerned the elicitation of test performance, which for example comprised assumptions about test administration procedures, test delivery quality, and comprehensibility of instructions. This inference is not specified in Kane’s (2006) original model, but was added by Drackert given that data elicitation is a crucial stage in SLA research. The second, scoring/evaluation inference entailed assumptions about the monitoring of the choice of test items, the development of the scoring rubric, and the consistency of scoring. In the third, generalisation inference, the assumptions were related to the reliability of the test, the appropriateness of the sample size, and the consistency of the test across educational settings and different first language backgrounds. The assumptions under the fourth inference, extrapolation, were derived from previous theoretical and empirical work, and outlined expected statistical relationships of the EI test to relevant background variables and self-assessments of L2 skills. The last inference, utilisation, specified the uses for which the test had been validated.

Once the interpretation/use argument had been developed, Drackert posed a number of evaluation questions, listed in Figure 3.2. Then, the process of building the validity argument began...
 Validating Assessments for Research

CONCLUSION: The test score reflects a certain degree of Russian oracy. It can be used as a covariate for group-comparison SLA studies, in other words studies that need to control for the initial levels of participants' Russian oral skills. The employment of the EIT for the intended use will enable benchmarking, replication, and generalizations across SLA studies in the Russian language.

1. Correlations show expected relationship with Russian learning. Students with longer Russian learning history and study abroad experience perform better on the EIT.
2. Correlations between the learners' self-assessment and the scores on the EIT are higher for oral skills (listening and speaking) than for written skills (reading and writing).

1. The observations made in testing are representative of the universe of observations.
2. The sample is large enough to reduce reasonably sampling error.
3. Test reliability is demonstrated in both settings.
4. The test demonstrates the ability to elicit a wide range of scores in both settings.
5. The items are free of bias and do not advantage either of the groups.

1. Statistical characteristics of stimuli are investigated and the combination of best-functioning items is chosen to comprise the final test.
2. Rubrics are developed, trialed, and revised.
3. The scoring is applied consistently and fairly: Raters display high levels of inter-rater reliability.

Test procedure and instrumentation are standardized and trialed. The EIT does not lead to unexpected behavior. Participants understand the instructions and are able to perform on the test as intended.

1. Oracy skills are critical for oral communication (listening and speaking).
2. EIT can assess oracy skills.

**Figure 3.1** Interpretation/use argument 1: EIT scores as a covariate (Drackert, 2015a, p. 112).

through addressing the evaluation questions. Drackert administered the EI test and a background questionnaire to 97 Russian-FL learners in the USA and Germany. The results were subjected to a range of statistical analyses, including descriptive statistics and differential item functioning, item response theory, correlational and interrater reliability analyses, to answer the evaluation questions. The findings confirmed that the EI test can reliably be used for one purpose, to control for initial levels of Russian oral proficiency in SLA studies.

To investigate whether the test can also be used for selecting participants for SLA studies based on their listening and speaking proficiency, a second validation study ensued. This
study followed the same overall design, starting with the construction of the interpretation/use argument, followed by deducing and investigating evaluation questions based on the interpretation/use argument. Sixty-seven learners of Russian in the USA took the EI test, a Russian speaking test, a C-test, and a listening comprehension test. Results from an array of descriptive and inferential statistical analyses showed that the test can distinguish between four levels of Russian-FL speaking proficiency and predict a threshold for listening ability at the intermediate level.

To sum up, Drackert’s work shows that Kane’s argument-based approach to validity can act as a framework for validating SLA research instruments for different uses. For additional worked examples of the argument-based approach in L2 instrument development and validation, we refer the reader to Voss (2012) and Kremmel (2017). 


testify into Research Designs. Finally, we demonstrate how researchers can integrate validation of the instruments used in SLA research into the design of smaller-scale SLA studies through an example. For the past two decades a lot of instructed SLA research has explored the effects of task complexity (i.e., inherent cognitive demands of tasks) on L2 outcomes measures such as complexity, accuracy and fluency in order to test cognitive models of task-based language learning (Robinson, 2001; Skehan, 2009). Until recently, a methodological weakness in existing research was that no evidence was provided in support of task complexity manipulations. Researchers typically designed a complex and simple version of a task (e.g., one with more and one with less reasoning demands), but failed to test whether the task version designed to be more cognitively demanding was indeed more cognitively complex. In response to calls to address this shortcoming (Norris & Ortega, 2003; Révész, 2014), more and more studies now include independent measures of cognitive load or mental effort into their design (e.g., Baralt, 2013; Révész et al., 2014; Zalbidea, 2017), such as subjective self-ratings, subjective time estimations, dual-task methodology and expert judgments. Some researchers have also trialled various ways to assess task-generated cognitive load (e.g., Révész et al., 2016; Sasayama, 2016) to inform future task complexity validation practices. Figure 3.3 provides an illustration of a possible argument structure for studies with independent measures of task complexity, hypothesising that the task version created to be more complex is indeed more cognitively demanding.

1. Which combination of items differentiates most accurately and reliably between learners of different levels of oracy skills in Russian? […]
2. Is the scoring rubric used appropriately, in particular: Are all points of the rubric utilized? Do raters display high inter-rater reliability? […]
3. Does the test function independently of the educational setting?
   3.1. Does the test function reliably in two educational settings? […]
   3.2. Does the test demonstrate the ability to elicit a wide range of scores in both settings? […]
   3.3. Are the items free of bias for both learner groups? […]
4. Are there any substantial correlations between the length of studying Russian, study/residence abroad, frequency of use of Russian, age of first exposure, and the scores on the EI? […]
5. Are the correlations between the learners’ self-assessment and the scores on the EI higher for oral skills (listening and speaking) than for written skills (reading and writing)? […]

*Figure 3.2 Questions posed to evaluate the interpretation/use argument (from Drackert, 2015a, p. 114).*
Validating Assessments for Research

Recommendations for Practice

With the ultimate goal of improving validation practices within the field, we would like to end this paper with some practical recommendations that may guide SLA researchers when constructing and validating research instruments. Several of the suggestions made below mirror the comments put forward by Chapelle (Chapter 2, this volume) and other researchers in the context of broader methodological discussions in the field (Ercikan & Pellegrino, 2017; Marsden & Plonsky, 2018; Norris & Ortega, 2003, 2012; Norris et al., 2015; Plonsky & Gass, 2011; Purpura et al., 2015).

- When designing an empirical study, consider each step outlined in Norris and Ortega’s (2003) measurement framework (construct definition, behaviour identification, task specification, behaviour elicitation, observation scoring, data analysis). Draw up an interpretation/use argument and formulate evaluation questions for each stage and a plan for the validation argument (Kane, 2006, 2013). Make sure to provide evidence in support of the claims in the interpretation/use argument while considering possible ways to provide rebuttals against the evidence provided. An example of how this can be done was provided in the previous section. In short, evidence for the validity of each stage of the measurement process should be given rather than assumed.

![Figure 3.3 Argument structure (based on Purpura et al., 2015).]
• In particular, keep the target construct in mind at all times and carefully evaluate whether your instrument operationalises it in a sufficiently comprehensive manner (to avoid construct underrepresentation) and does not require or involve the use of other knowledge, skills or abilities (to avoid construct-irrelevant variance).

• Construct underrepresentation, or even misrepresentation, is something to look out for when measuring constructs. Checking for representation problems could be done through, for example, validation work that investigates learners’ cognitive-response processes (Nicols & Huff, 2017). That is, researchers can collect data on how learners mentally process and work through items during assessments through, for example, think-alouds, retrospective interviews, eye-tracking, or key-stroke-log methods, as explained by Nicols and Huff, and use that data to verify the mental processes assumed to underlie scores.

• When it is not possible to collect and analyse data to address claims put forward in the validity argument, highlight this and take it into account when interpreting research findings.

• Do not assume that tests which have been validated for one purpose with one population of learners in one context will also be valid for a different purpose, a different population of learners, and/or a different setting. Instruments need to be validated for their intended purposes, populations, and contexts.

• Follow the guidelines for reporting quantitative methods and results outlined in Norris et al. (2015). This will enable fellow researchers to replicate previous research as well as evaluate and critique the validity of it.

• Share your instruments and data in open-access repositories such as IRIS—a digital repository of instruments for research into second language learning and teaching (https://www.iris-database.org/iris/app/home/index), accompanied by the original publication/dissertation with information on the purpose, population and context in which the instruments were used and data generated. This will not only allow other researchers to better assess the validity of research instruments and designs used in previous research but also facilitate ongoing validation endeavours aimed at the refinement of SLA elicitation devices (see, for example, continuous validation work on the development of implicit and explicit knowledge measures).

Recommended Readings


This monograph provides a step-by-step illustration of how Kane's argument-based approach to validation can be applied to developing and validating SLA assessments.


This chapter discusses key stages in the SLA measurement process with a view to achieving validity and reliability.


This article shows how to construct a validity argument for an SLA research instrument, and also highlights typical pitfalls in SLA research with respect to validity.
Validating Assessments for Research

References


