Studies in Second Language Acquisition (2025), 1–25
doi:10.1017/S0272263125101058

RESEARCH ARTICLE

Understanding proficiency assessment practices in SLA research

Insights from researcher beliefs and practices

Hae In Park[1], Megan Solon[2] and Kwangmin Lee[3]

[1]Kyung Hee University, Seoul, Republic of Korea; [2]Indiana University, Bloomington, IN, USA and [3]Western Michigan University, Kalamazoo, WI, USA
Corresponding author: Hae In Park; Email: haeinpark@khu.ac.kr

(Received 28 September 2024; Revised 26 April 2025; Accepted 12 June 2025)

Abstract

While much discussion has focused on what researchers do and should do in second language proficiency assessment, less attention has been given to why persistent trends continue. This study investigated second language acquisition (SLA) researchers' beliefs, reported practices, and decision-making rationales regarding proficiency assessment. Using an online survey, we collected responses from 111 SLA researchers. Findings revealed that while researchers generally endorsed recommended methodological standards, practical constraints—such as time, accessibility, and ease of administration—frequently influenced their reported practices. A consistent belief–practice gap emerged across several key areas. Notably, reduced redundancy tests were rated favorably for both validity and practicality, reflecting a growing shift toward efficient, validated tools. These findings suggest that although methodological awareness is high, practical barriers continue to challenge the adoption of more rigorous proficiency assessment practices in SLA research.

Keywords: L2 proficiency; proficiency assessment; survey research; methodological rigor; researcher beliefs

Introduction

Central to second language acquisition (SLA) research is the construct of language proficiency, broadly defined as "knowledge, competence, or ability in the use of a language" (Bachman, 1990, p. 16), though multifaceted in nature involving various knowledge and skill components. SLA researchers assess and report second language (L2) learners' proficiency for a variety of research-related purposes, including "to justify the sampling of participants into a study or to assign participants to distinct groups" and to "aid readers of research when deciding the extent to which findings can be

generalized to other samples and populations" (Norris & Ortega, 2012, p. 580). Since understanding learner variability in L2 performance lies at the core of SLA research, assessing proficiency with valid and reliable measures is critical for securing high internal and external validity of research.

As laid out by Thomas (1994) and others, establishing the L2 proficiency of learners sampled in research is important for numerous reasons including being able to generalize a particular finding beyond a specific sample of learners to, for example, a level or stage of learning, being able to contribute to our understanding of developmental sequences whether using longitudinal or cross-sectional data, and being able to adequately control for potential differences between groups of learners sampled. Despite the fact that L2 proficiency occupies an important place in the study of L2 acquisition, research agendas have not always paid sufficient attention to L2 proficiency assessment. The construct of proficiency is often treated as a lurking variable (i.e., a variable that is not accounted for in a study but may affect the relationship between the variables under study; Norris, 2010) and is often taken for granted in research design (Hulstijn, 2012; Leclercq & Edmonds, 2014). According to previous systematic reviews (Park, Solon, Dehghan-Chaleshtori & Ghanbar, 2022; Thomas, 1994, 2006; Tremblay, 2011), some of the persisting issues with proficiency assessment practices include over-reliance on proxy measures of proficiency (e.g., institutional status) that are less accurate and less generalizable than objective measures and a lack of uniformity across studies in the methods of proficiency assessment, which can limit cross-study comparability and aggregation. As will be reviewed in greater detail in the Background section, the collective view of proficiency reporting practices provided by these surveys suggests that many of the noted weaknesses in proficiency assessment practices have remained relatively unchanged over 3 decades (1988–2019) and that great strides are needed for the field to establish more robust proficiency assessment standards.

Recently, the broader field of SLA has undergone what Byrnes (2013) termed a "methodological turn," marked by a heightened focus on rigor, transparency, and ethics. The field has seen a significant increase in awareness of methodological issues, including the need for more replication research (e.g., Porte, 2010), the promotion of stronger ethical research practices (e.g., Isbell & Kim, 2023; Larsson et al., 2023), critical reflection on methodological practices (e.g., Larson-Hall & Plonsky, 2015), and the application of more advanced analytical techniques (e.g., Plonsky, Egbert, & Laflair, 2015). Importantly, scholars now recognize that methodological rigor and ethical integrity are closely intertwined, asserting that even seemingly minor methodological missteps—such as employing an assessment tool without a well-founded validity argument or failing to be transparent about a data collection decision—can be framed as ethical issues (Plonsky et al., 2024). The heightened interest in research methods is evidenced by a spate of recent publications on this topic, including general methods books (e.g., Mackey & Gass, 2016), discipline-specific books focused on quantitative methods (e.g., Plonsky, 2015) and ethical issues (De Costa, Cinaglia, & Rabie-Ahmed, 2024), special journal issues dedicated to quantitative analysis (e.g., *Applied Linguistics* in 2016, *Language Learning* in 2015), and numerous journal articles and book chapters aimed at improving quantitative methods (e.g., Al-Hoorie & Vitta, 2019; Larson–Hall, 2017). Many of these contributions explicitly address proficiency assessment practices (e.g., Hulstijn, 2010, 2012; Leclercq & Edmonds, 2014; Norris & Ortega, 2003, 2012), highlighting issues such as the complexity of selecting appropriate proficiency assessment tools based on theoretical and psychometric considerations as well as the critical importance of rigorous construct definition and validation processes to ensure that measures genuinely reflect the intended language constructs. Considering this increasing awareness and emphasis on the importance of sound research

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

methods, the modest progress in proficiency assessment practices over the 3 decades is rather surprising. The persistent gap between recommended practices and actual practice underscores the need for an examination of potential factors that may hinder methodological improvements in proficiency assessment. This need deepens further when we consider that adopting more robust proficiency assessment standards is not merely a methodological necessity but also an ethical obligation.

To better understand the underlying reasons behind the persistent gap between recommended standards and actual practices in proficiency assessment, the present study examines SLA researchers' beliefs and reported practices regarding proficiency measurement. Specifically, we investigated researchers' perspectives on proficiency assessment, seeking to identify potential reasons why previous calls for methodological rigor have not resulted in substantial changes. Our exploration considers whether limited progress might be attributed to researchers' insufficient awareness of current proficiency assessment recommendations or perhaps practical constraints when selecting and using assessment tools. By providing insights into researchers' decision-making processes, we seek to understand the challenges they encounter in meeting the field's recommended methodological standards. To our knowledge, this study represents the first empirical effort to explore researchers' accounts of their proficiency assessment practices, offering valuable implications for fostering methodological improvements and ethical integrity in the field.

Background

Defining language proficiency

Despite the fact that defining a model of what constitutes L2 proficiency is an essential part of assessing language proficiency, it is by no means a trivial matter. Over the last 60 years, a number of language proficiency models have been proposed by scholars from different theoretical approaches (Norris & Ortega, 2003, 2012; Purpura, 2016). While earlier models of L2 proficiency (also known as "skills-and-elements" models; e.g., Carroll, 1968; Lado, 1961) primarily focused on linguistic knowledge, subsequent models expanded the concept of L2 proficiency to include variables related to social aspects of communication such as sociolinguistic competence and strategic competence (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980). These models have laid the groundwork for the dominant perspective in the field of assessment, which views language proficiency as multicomponential, comprising various interrelated domains such as grammatical knowledge, textual knowledge, and pragmatic knowledge.

More recently, Hulstijn (2015, 2019, 2024) has proposed a comprehensive framework of language proficiency that incorporates both the automaticity of processing and core linguistic knowledge. This model conceptualizes language proficiency in two dimensions: a cognitive dimension and a knowledge dimension. The cognitive dimension is similar to proficiency models and Bialystok (2001) and Cummins (1980) in that it makes a fundamental distinction between basic language cognition (BLC) and extended language cognition (ELC). BLC refers to the implicit, unconscious language cognition in the oral domains while ELC pertains to literacy skills, as taught in school (Hulstijn, 2024). The knowledge dimension differentiates between core and peripheral components of proficiency. The core components consist of essential linguistic knowledge (i.e., vocabulary, grammar, phonology, and pragmatics) as well as the speed at which this knowledge is processed. In contrast, the peripheral components encompass

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

less-linguistic or nonlinguistic knowledge, such as interactional abilities, strategic competence, and metalinguistic knowledge.

Despite widespread agreement that language proficiency is a highly complex and multicomponential construct, there remains little consensus on its precise definition. As this brief review illustrates, different models offer varying perspectives on the scope, components, and granularity of L2 proficiency. Given this variability, it is crucial for researchers to achieve conceptual clarity when defining proficiency in their work, to articulate their definitions explicitly, and to select assessment methods that best capture the specific facets of proficiency relevant to their research objectives.

Proficiency assessment practices in SLA

The importance of designing and implementing principled assessments of L2 proficiency has received considerable attention in SLA research over the past few decades. As part of an expanding line of research in the SLA-assessment interface (e.g., Bachman, 1988; Bachman & Cohen, 1998; Derrick, 2016; Winke & Brunfaut, 2021), a number of studies (Hulstijn, 2010; Leclercq & Edmonds, 2014; Norris & Ortega, 2012; Schoonen, 2011) have provided explicit guidance on how to define and measure L2 proficiency in research contexts. These conceptual discussions converge on several foundational principles: Researchers must define the construct of proficiency clearly; select assessment tools that match the study's goals and context; and consider trade-offs between validity, reliability, and practicality (Hulstijn, 2010; Norris & Ortega, 2012). As no single test can maximize all three, researchers are expected to justify their choices accordingly. As Schoonen (2011) notes, "doing so requires an understanding of choices we as researchers have in designing language assessments and of the consequences of choosing one of the options over another" (p. 701). In this regard, language assessment literacy—a set of knowledge and skills required to carry out assessment-related practices—may be viewed as an essential component of the research skillset for SLA researchers (Harding & Kremmel, 2021).

Complementing these theoretical contributions is a series of reviews of L2 proficiency assessment practices (Park, Solon, Dehghan-Chaleshtori, & Ghanbar, 2022; Thomas, 1994, 2006; Tremblay, 2011) examining the conventions for proficiency assessment practices in SLA research. Using different inclusion and exclusion criteria, these syntheses surveyed articles from major SLA journals to identify potential trends in proficiency reporting practices and to provide suggestions as to how researchers might implement more rigorous proficiency assessment standards in their research. The first two reviews, undertaken by Thomas (1994, 2006), surveyed studies within a 5-year time span (1988–1992 and 2000–2004, respectively) from four key journals (*Applied Linguistics, Language Learning, Second Language Research*, and *Studies in Second Language Acquisition*) and assigned them to one of the following four categories based on the approach to accounting for participants' L2 proficiency: (a) impressionistic judgment (the researcher's unsupported evaluation of learners' L2 proficiency); (b) institutional status (proficiency defined by their positions in a hierarchically organized social group); (c) in-house assessment (proficiency defined by locally developed and administered instruments); or (d) standardized test (proficiency assessed by standardized instruments). A heterogeneous "Other" category was also created for assessment techniques that did not fall into any of the four categories. In both studies, Thomas found that institutional status was, by far, the most commonly used means of assessing proficiency, followed by use of a standardized test. Thomas (2006) also found

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

that "relatively little aggregate change" occurred in the proportion of studies using each of these four broad techniques (p. 287).

Since the publication of Thomas (1994, 2006), two subsequent reviews have further documented the evolution of proficiency assessment practices over time. Tremblay's (2011) corpus surveyed articles from three SLA journals (*Second Language Research, Studies in Second Language Acquisition*, and *Journal of French Language Studies*) during a 9-year interval between 2000 and 2008, and Park et al. (2022) surveyed studies from five SLA journals (*Applied Linguistics, Language Learning, The Modern Language Journal, Second Language Research*, and *Studies in Second Language Acquisition*) published between 2012 and 2019. In these two reviews, categorization of proficiency assessment practices included a wider range of assessment techniques (nine techniques each in Tremblay, 2011 and Park et al., 2022 versus five techniques in Thomas 1994, 2006). Newly emerged categories included cloze tests or C-tests, oral tests, vocabulary tests, existing proficiency scores, and self-ratings, among others (note that the categories used by these two more recent reviews were not identical). Figure 1 presents the distribution of studies across assessment techniques by research synthesis.

Despite the expansion of assessment categories over time, the aggregate view of proficiency reporting practices portrayed by the four reviews suggests that there has been relatively little change in conventions for assessing L2 proficiency over the past 3 decades. The use of independent proficiency measures (i.e., with some form of test) has remained modest (ranging from 36% to 43%), while institutional status has remained the most frequent method of assessment across reviews (described as classroom level or years of instruction in Park et al., 2022, and Tremblay, 2011). This continued reliance on indirect or proxy measures of proficiency (i.e., those that do not employ an independent test as part of research procedures, such as institutional status, existing proficiency scores, self-ratings, or impressionistic judgments) is surprising given that such approaches to accounting for proficiency have been subject to repeated criticism for various shortcomings (e.g., Norris & Ortega, 2012; Thomas, 1994, 2006;

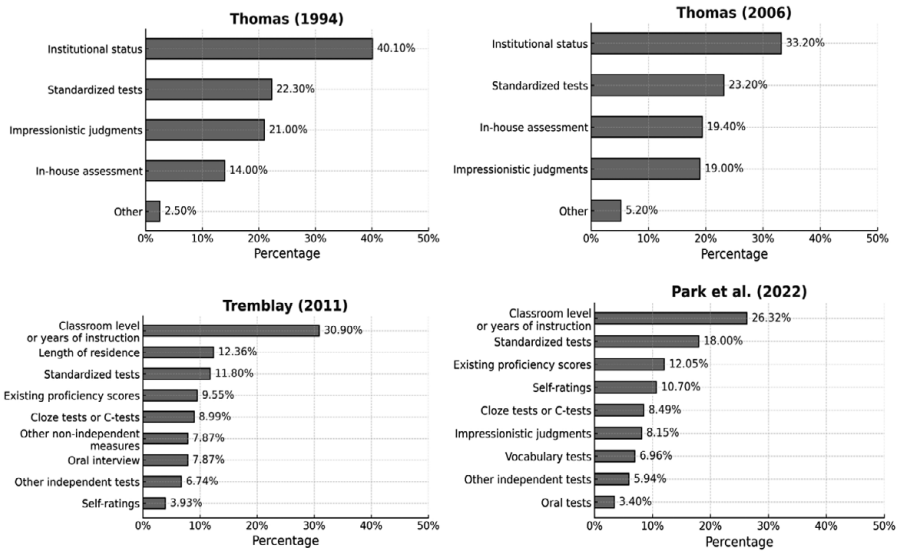

Figure 1. Distribution of studies across proficiency assessment techniques by research synthesis.

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Tremblay, 2011). First and foremost, indirect methods of assessment may lack precision in capturing learners' L2 proficiency at the time of testing if, for example, the collected information is outdated (e.g., participants' Test of English as a Foreign Language scores may be from several years ago) or unreliable (e.g., participants may under- or over-estimate their ability in the L2 in self-ratings). Furthermore, the use of indirect measures of proficiency may limit the generalizability of findings beyond the sample used or make intergroup comparability regarding proficiency across studies less ideal. For example, the reliance on institutional status as a proxy for L2 proficiency poses a serious threat to the validity of research that pools learners of the same status from multiple institutions, as institutions use different standards to assign L2 learners to different curricular levels and base promotion on different criteria. Furthermore, grouping research subjects by course level can result in large variability within each level and overlap between levels. Thus, groups determined solely by course level may not represent distinct and internally homogeneous levels of proficiency. Due to these noted shortcomings, the use of independent proficiency measures has been consistently recommended as a more rigorous practice to enhance the validity of proficiency reporting (Hulstijn, 2012; Norris & Ortega, 2012; Park et al., 2022). Nevertheless, institutionally defined proxies continue to be widely used as a means of proficiency assessment in SLA research.

In addition to these major trends that have persisted over time, Park et al. (2022) observed several other notable patterns within proficiency assessment practices of the past decade of SLA research. Specifically, they found that when proficiency was estimated with institutional status only, it typically did not enter into the research design. This suggests that studies that do include proficiency as an independent or grouping variable tend to employ measures other than institutional status or multiple proficiency measures including institutional status. In fact, approximately 25% of the studies reviewed by Park et al. utilized multiple proficiency assessment instruments. They also found that the use or not of an independent test of proficiency appeared to vary as a function of study characteristics and research design features. Rates of independent test use were relatively higher for studies investigating adult learners (as compared to child learners), for studies conducted in a laboratory setting (versus a classroom setting), and for studies using quantitative or qualitative methods (versus mixed methods). Additionally, studies investigating Dutch, German, Russian, and Spanish exhibited rates of use of an independent proficiency test at or above 50%, whereas studies targeting L2 French, for example, employed an independent test only 15% of the time.

By reviewing and reflecting on various issues pertaining to L2 proficiency assessment practices, previous studies have made a concerted effort to sensitize SLA researchers to the need for employing more rigorous and informative assessment techniques. Based on a review of widely cited conceptual papers (Hulstijn, 2012; Norris & Ortega, 2012; Schoonen, 2011) and research syntheses (Park et al, 2022; Thomas, 1994, 2006; Tremblay, 2011), we identified and summarized key field-recommended standards for assessing and reporting L2 proficiency, which are presented in Table 1.

Researcher cognition, beliefs, and self-reported practices

The "methodological turn" (Byrnes, 2013, p. 825) previously described for the field of SLA has led to increased reflection among researchers about methodological practices, training, and know-how. Integral to this movement is an investigation of researchers'

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Table 1. Field-recommended standards for L2 proficiency assessment in SLA research

Standard	Description	Key references
Define language proficiency explicitly	Researchers should clearly articulate how L2 proficiency is defined, taking into account the study's goals, research questions, and theoretical framing.	Norris & Ortega (2012); Park et al. (2022); Schoonen (2011)
Use independent proficiency measures	Proficiency should be assessed using direct, independently administered instruments rather than relying solely on proxies.	Hulstijn (2012); Tremblay (2011)
Report validity and reliability	Researchers should report the psychometric quality (e.g., validity, reliability) of the measures used, even when using pre-validated tools.	Norris & Ortega (2012); Park et al. (2022); Thomas (1994)
Routinely measure proficiency	Proficiency should be assessed even when it is not the primary focus of the study to support comparability and interpretation.	Park et al. (2022); Thomas (1994)
Avoid overreliance on proxy indicators	If using proxy measures (e.g., institutional status, years of instruction), researchers should justify their use or supplement with direct measures.	Park et al. (2022); Thomas (1994)
Use multiple measures when feasible	Employing more than one assessment tool enhances robustness and allows for a more comprehensive understanding of learner proficiency.	Park et al. (2022); Tremblay (2011)

beliefs, self-reported knowledge, and practices related to various aspects of research methodology and design. For example, a series of surveys (e.g., Gonulal, Loewen, & Plonsky, 2017; Loewen et al., 2014) explored SLA researchers' beliefs and self-reported knowledge, training, and practices related to statistical analysis, an area central to methodological rigor in SLA. Loewen et al. (2014) found that only 14% of SLA doctoral students and 30% of professors reported feeling that their statistical training was adequate. Loewen et al. (2020) further revealed that while most respondents understood basic statistical concepts, their performance on test items requiring more advanced statistical knowledge was lower. Interestingly, by comparing participants' self-rated confidence with their actual test scores, Loewen et al. demonstrated that perceived knowledge and actual knowledge (i.e., overall confidence ratings and overall statistical knowledge test scores) did not always align.

Such findings highlight the value of investigating researchers' perspectives on methodological issues more broadly. Exploring researchers' beliefs alongside their reported practices can shed light on potential gaps between what researchers think they should do and what they actually do, providing insights into persistent challenges in achieving the field's recommended standards of methodological rigor.

Motivation for and objective of the present study

Building on this line of inquiry, the present study focuses specifically on L2 proficiency assessment practices. While methodological discussions have emphasized what researchers ought to do when assessing proficiency, relatively little attention has been given to understanding why certain trends persist despite calls for change. For instance, do researchers' practices reflect their underlying beliefs about what constitutes sound proficiency assessment or about the value of specific assessment tools? Or are there discrepancies between researchers' beliefs about research practices and their actual

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

implementation? What are the reasons behind researchers' preferences for specific assessment tools? To date, no empirical study has systematically examined SLA researchers' beliefs, self-reported practices, and underlying rationales regarding proficiency assessment—insights that are critical for explaining why methodological gaps have remained over decades.

To address this gap, we conducted a mixed-methods survey study (following a convergent design; Sato, 2022) to explore SLA researchers' beliefs, reported practices, and preferences regarding different proficiency assessment techniques. The following research questions (RQs) guided this investigation:

RQ 1. To what extent do SLA researchers' beliefs about proficiency assessment practices align with methodological standards recommended by the field?

RQ 2. How do SLA researchers evaluate different L2 proficiency assessment techniques in terms of various test characteristics?

RQ 3. To what extent do SLA researchers' self-reported proficiency assessment and reporting practices align with methodological standards recommended by the field?

RQ 4. Which assessment techniques do SLA researchers report using to assess L2 proficiency in their research?

RQ 5. What are SLA researchers' main reported reasons for choosing particular assessment techniques?

Method

Participants

The target population for the current study was researchers who self-identified as engaging in research on some aspect(s) of SLA. This included novice scholars, defined as doctoral students with experience conducting research, regardless of their publication experience. The survey data were gathered between January and February 2022 via an online questionnaire distributed by the Qualtrics platform (http://www.qualtrics.com). Following previous survey-based studies (Isbell et al., 2022; Loewen et al., 2020; McManus, 2022), we used multiple techniques to obtain a large and broad sample of SLA researchers. First, names and email addresses from the 2016–2020 Second Language Research Forum conference programs were extracted. Using this method, 962 researchers with identifiable contacts received an invitation to the survey. Second, survey links were distributed through several listservs (e.g., *GU Linguist*) and Facebook groups (e.g., *Applied Linguistics Research Methods*, *Collaborative Efforts in Linguistics*) that are likely to be subscribed to by SLA researchers. We also shared the survey link with known SLA researchers around the world via email, encouraging them to pass along the invitation to eligible colleagues.

Due to the snowball sampling method used, an exact response rate was not available to report. However, 292 respondents started the survey by clicking on the survey link, and 111 of them completed the survey by providing a response for most of the questions. If respondents missed more than one sub-section of Part II of the survey (see the Instrument section for more information), their data were excluded from the study. The final sample resulted in 111 researcher respondents from 19 different countries: A majority of the respondents (72.07%) were located in North America (n = 80; Canada = 10, USA = 70), 13.51% in Europe (n = 15; England = 1, France = 1,

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Germany = 1, Greece = 1, Hungary = 1, Poland = 1, Portugal = 3, Spain = 2, Sweden = 2, Switzerland = 2), 9.91% in Asia (n = 11; China = 2, India = 1, Japan = 6, Korea = 2), 1.8% in South America (n = 2; Columbia = 1, Uruguay = 1), and 2.70% in Australia (n = 3). In terms of research experience, 9.91% of the respondents (n = 11) reported not having published any peer-reviewed research articles (five of these had an article either under review or in press); 24.32% (n = 27) as having published 1–5 articles; 19.81% (n = 22) as having published 6–10 articles; 24.32% (n = 27) as having published 11–20 articles; and 20.72% (n = 23) as having published 20 or more. The average number of years since receiving the Ph.D. for 109 respondents (two respondents gave no response) was 8.39 (standard deviation [*SD*] = 9.89), with the range of –3–42 (a negative value indicates the number of years left until the expected graduation year). In terms of research orientation, 6.31% of the respondents (n = 7) reported conducting qualitative research only, 37.83% (n = 42) conducting quantitative research only, and 54.96% (n = 61) conducting more one than one type of research (quantitative, qualitative, and/or mixed-methods research).

Instrument

Survey development involved two steps. First, an initial pool of items was developed to address the five RQs, based on previous syntheses of L2 proficiency assessment and reporting practices in SLA research (Park et al., 2022; Thomas, 1994, 2006; Tremblay, 2011). Next, the survey was piloted among two experienced SLA researchers not otherwise involved with the project to optimize content validity and minimize measurement bias. It was then modified based on their feedback concerning the survey's questions, response options, and organization. The final version of the survey comprised two sections and took approximately 20 min to complete. The survey is available in Appendix A of the online supplementary material as well as on the Open Science Framework (https://osf.io/fv8ug/?view_only=f34b7ed61f2e4b8887ebf87661c9604d).

Section I included a study eligibility check question (i.e., [yes/no] I consider myself a second language acquisition researcher) and nine items related to personal and professional demographics. If participants responded yes to the eligibility check question, they proceeded to the next items, and if no, they withdrew from the survey. The demographic questions elicited information about participants' current position, country of residence, year Ph.D. was received (or was expected to be received), Ph.D. granting institution, journals in which they have published in the last 5 years, main research areas, research orientation, target languages under study, and number of peer-reviewed articles published.

Section II included a concise definition of several key terms (e.g., second language proficiency,[1] independent test, indirect measure) as well as items relating to SLA researchers' beliefs and practices regarding L2 proficiency assessment and reporting. The items in Section II were mainly multi-item scales (e.g., agreement-disagreement

[1]In our survey, we had defined L2 proficiency according to Hulstijn's BLC, focusing specifically on oral proficiency and frequent lexical and grammatical structures. However, an anonymous reviewer pointed out that this definition does not fully capture the broader, multifaceted conceptualizations common in SLA research (e.g., Bachman's 1990 model of communicative language ability). We acknowledge the importance of this point. To explore whether this narrow definition influenced survey responses, we conducted a retrospective interview with five participants, who reported paying minimal attention to the provided definitions due to their existing familiarity with these concepts. While this suggests minimal impact on participants' survey responses, future research should consider employing comprehensive definitions that better reflect the diverse proficiency constructs relevant to various SLA contexts.

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

with statements) although some were open-ended questions. When developing multi-item scales, an equal number of pro and con statements was used following Dörnyei's (2010) recommendations for questionnaire design. All the items were grouped into two themes, which formed separate sub-sections within Section II: beliefs and practices.

The *beliefs* section elicited respondents' views on various aspects of L2 proficiency assessment and reporting practices. They were asked to (a) indicate their level of agreement with a set of statements (e.g., "Knowing learners' L2 proficiency is important for understanding most types of SLA research") on a 6-point scale ranging from "strongly agree" to "strongly disagree" and (b) evaluate a set of L2 proficiency assessment techniques (e.g., standardized tests) in terms of various test attributes (e.g., trustworthiness, accuracy, availability, validity,[2] etc.), using 7-point semantic differential scales (e.g., with trustworthy on one end and untrustworthy on the other). The *practices* section elicited participants' reflections on their own practices of L2 proficiency assessment and reporting. They were asked to indicate (a) how true a set of statements (e.g., "I test learners' L2 proficiency when I conduct an SLA study") is for them on a 6-point scale ("almost always true" to "almost never true") and (b) how frequently they employed various proficiency assessment tools in the past 5 years on a 6-point scale ("always" to "never"). Lastly, participants were asked to indicate their most and least preferred means of proficiency assessment and provide reasons for them. While we recognize that the appropriateness of a proficiency assessment method is inherently context-dependent (Norris & Ortega, 2012), this question was designed to capture participants' general impressions and preferences. These responses were not intended to represent judgments about universal applicability, but rather to provide insight into researchers' broad attitudes, experiences, and perceptions of commonly used assessment tools.

Analyses

Following the recommendations of Plonsky (2015), we adopted a "back-to-basics" approach to data analysis. This primarily involved calculating descriptive statistics for responses to each survey question, including mean (*M*), *SD*, 95% confidence interval (CI), median, and interquartile range (IQR). Given that analyses differed for different types of survey questions and their resulting data, detailed descriptions of our methods of analysis are provided in the Results section.

In this study, missing data were handled at the item level. As a result, the number of responses varied across different items related to belief and practice, among the 111 respondents. For each assessment technique, respondents with missing data for that technique were excluded from the analysis of that item. This method ensured that

[2]An anonymous reviewer raised an important point about the definition of validity, noting that it is not an inherent property of a test but rather refers to the appropriateness of the interpretations and uses of test scores for specific purposes, populations, and contexts (AERA, APA, & NCME, 2014). We fully agree with this definition and acknowledge that validity is not a fixed characteristic of any given assessment tool. In the present study, our goal was not to evaluate the actual validity of each technique but rather to capture researchers' general perceptions of validity—impressionistic judgments likely shaped by their disciplinary training and exposure to studies validating proficiency assessment techniques. These subjective impressions, while not tied to specific research contexts, provide valuable insight into broader methodological beliefs and the perceived credibility of commonly used assessment approaches in SLA.

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

the descriptive statistics, including *M*s, *SD*s, and CIs, were calculated based on the available responses for each technique. By maximizing the use of available data for each item, this approach ensured that the descriptive and inferential statistics were based on the most complete dataset possible for each specific item. The actual data spreadsheet used for analysis is available on the Open Science Framework: https://osf.io/fv8ug/?view_only=f34b7ed61f2e4b8887ebf87661c9604d.

Results

RQ 1: To what extent do SLA researchers' beliefs about proficiency assessment practices align with methodological standards recommended by the field?

Ten questionnaire items (items 1-1 to 1-10 in the survey) assessed whether participants' beliefs were aligned with methodological standards regarding proficiency assessment and reporting practices. Using a 1–6 Likert scale ("strongly disagree" to "strongly agree"), participants indicated their level of agreement with 10 statements that reflected either desirable (k = 5) or less desirable (k = 5) research practices. To evaluate item performance, we examined factor loadings, communalities, and uniqueness values for each item (see Appendix B in the online supplementary material). A majority of items demonstrated adequate factor loadings, suggesting meaningful contributions to the underlying construct. However, two items (items 2 and 7)—both negatively worded—exhibited the weakest psychometric properties (factor loadings below .10) and were removed from the final analysis to improve the scale's reliability and construct coherence.[3] In addition, item 6 was removed following reviewer feedback regarding its ambiguous wording, which could lead to different interpretations and compromise the validity of responses. After the removal of these items, the internal consistency of the belief scale (k = *7*), as measured by McDonald's omega, was .75, indicating acceptable reliability given the broad nature of belief constructs and relatively small number of items (Clark & Watson, 1995).

Descriptive statistics of responses for the remaining seven items (desirable, k = 5; less desirable, k =2) are presented in Table 2, and histograms for each statement are available in Appendix C in the online supplementary material. Respondents generally reported a high level of agreement (overall M of five items = 4.67) with statements reflecting the field's recommendations: Understanding learners' L2 proficiency is crucial for SLA research (item 1-1, M = 5.35); assessing participants' L2 proficiency should be part of the research design (item 1-3, M = 5.05); direct assessment with an independent test is preferable to indirect methods (item 1-5, M = 4.54); and checking and reporting the quality of the L2 proficiency measure used is essential (item 1-10, M = 5.06). In contrast, respondents indicated a relatively lower level of agreement with statements reflecting less desirable research practices, with average values ranging from

[3]An anonymous reviewer pointed out that negatively worded items may reduce response reliability. We acknowledge this concern, as it has been well documented in the literature (e.g., Dodeen, 2023) and note that our item-level analysis (Appendices B and F) showed that several negatively worded items exhibited lower factor loadings and communalities than their positively worded counterparts. Despite these limitations, we retained several negatively worded items in the belief and practice constructs because they reflect misconceptions or outdated proficiency assessments practices that persist in SLA research. Including such items was necessary to assess whether researchers continue to endorse these views. While these items require careful interpretation, we believe they add important conceptual value by capturing a fuller spectrum of beliefs and reported practices.

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Table 2. Beliefs about proficiency assessment and reporting practices

	Researcher belief	*n*	M	*SD*	95% CI	Median	IQR
Statements consistent with field recommendations							
1–1	Knowing learners' L2 proficiency is important for understanding most types of SLA research.	110	5.35	1.01	[5.16, 5.54]	6	1.00
1–3	It is important for SLA researchers to assess participants' proficiency in the L2 as part of their research design.	111	5.05	0.93	[4.88, 5.23]	5	1.00
1–5	It is better practice to assess L2 proficiency directly with an independent test (e.g., standardized tests, cloze test) than indirectly (e.g., through classroom or institutional level, learner self-ratings, researcher impressionistic judgments).	111	4.54	1.27	[4.30, 4.78]	5	2.00
1–9	Indirect measures are not methodologically sound methods for accounting for L2 proficiency.	111	3.37	1.34	[3.12, 3.62]	3	2.50
1–10	It is important for SLA researchers to check and report the quality (e.g., validity, reliability) of the L2 proficiency test that they use.	111	5.06	0.88	[4.90, 5.23]	5	1.00
Statements not consistent with field recommendations							
1–4	Knowing learners' L2 proficiency is not important if L2 proficiency is not a variable of interest (either dependent or independent) in the research design.	110	2.38	1.20	[2.15, 2.61]	2	1.75
1–8	Non-independent proficiency measures can be as good as independent proficiency tests in getting information about learners' L2 proficiency.	111	3.10	1.26	[2.86, 3.34]	3	2.00

Note: Item codes correspond to the numbering in the survey provided in Appendix A in the online supplementary material. Six-point Likert scale for all items: 1 = strongly disagree; 6 = strongly agree.

2.38 to 3.10 (overall M of two items = 3.93). Interestingly, respondents appeared more ambivalent about the soundness of indirect or non-independent measures, showing average response values closer to the middle of the scale (item 1-8, M = 3.10; item 1-9, M = 3.37).

RQ 2: How do SLA researchers evaluate different L2 proficiency assessment techniques in terms of various test characteristics?

Seven survey items (items 2-1 to 2-7 in the survey) were designed to gauge respondents' attitudes toward various methods of assessing L2 proficiency. Using a semantic differential scale, respondents rated seven different types of or approaches to L2 proficiency assessment according to nine characteristics. These characteristics fit into two broader dimensions: validity-related characteristics (e.g., accuracy, representativeness, trustworthiness,

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

validity) and practicality-related characteristics (e.g., accessibility, cost effectiveness, ease of administration, scoring ease, time efficiency). Across the seven assessment methods, item factor analyses showed consistently strong model fit, with all comparative fit index and Tucker–Lewis index values exceeding .95. Although root mean square error of approximation and standardized root mean squared residual were less than optimal in a few cases, most values remained within or near acceptable thresholds (≤ .08; Kline, 2023). To prevent superficial responding or a positive response set, the positioning of "positive" and "negative" poles was varied on the survey (e.g., untrustworthy – trustworthy, accurate – inaccurate; Aiken, 1996; Dörnyei & Taguchi, 2009). In presenting the results, we have reverse-scored any positive-negative items so that responses for each trait are summarized on a 1–7 scale whereby a higher number indicates a more positive evaluation for that particular characteristic. The mean ratings for each proficiency method according to each characteristic are graphically represented in Figure 2 (see Appendix D in the online supplementary material for detailed descriptive statistics). Dark gray areas represent validity-related characteristics, while light grey areas represent practicality-related characteristics.

Figure 2 reveals important differences in how researchers think about and evaluate different proficiency assessment methods. For example, standardized tests and oral interviews generally received relatively high ratings on validity-related characteristics,

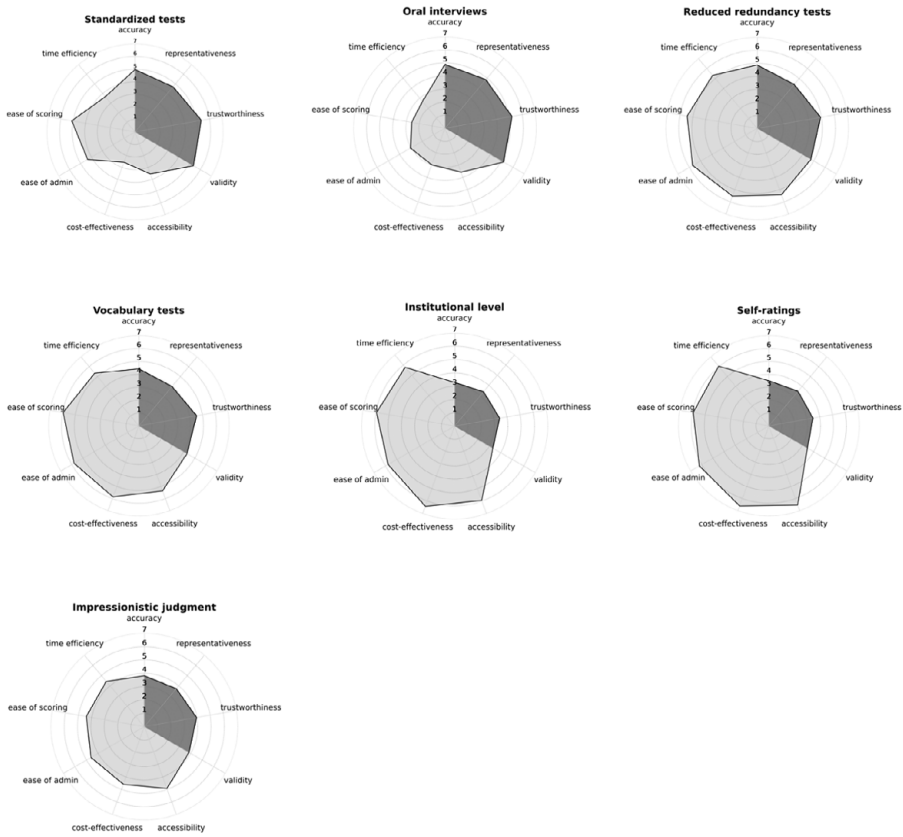

Figure 2. Semantic differential ratings for various characteristics of proficiency assessment methods.

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

such as accuracy, representativeness, trustworthiness, and validity (as represented by the bigger dark gray volume), while they were rated lower on several of the characteristics related to practicality, such as accessibility, cost effectiveness, ease of administration, and time efficiency (as represented by the smaller light gray volume). Classroom or institutional level, self-ratings, and impressionistic judgments, in contrast, were rated quite highly on practicality-related qualities but lower on validity-related characteristics. Interestingly, reduced redundancy tests were rated relatively highly on all characteristics, though practicality-related characteristics (overall M = 5.48) demonstrated descriptively higher ratings than validity-related characteristics (overall M = 4.74). Vocabulary tests showed similar patterns, though with more distinction between practicality- and validity-related characteristics (overall M for practicality = 5.67; overall M for validity = 4.32).

RQ 3: To what extent do SLA researchers' proficiency assessment and reporting practices align with methodological standards recommended by the field?

Nine survey items targeted researchers' own proficiency assessment and reporting practices in SLA research (items 3-1 to 3-9 in the survey). Five items reflected recommended practices, while four indicated less desirable practices. Participants rated the frequency of each practice on a 1–6 Likert scale, ranging from "almost never true" to "almost always true." Descriptive statistics for all items are presented in Table 3, and histograms for each item are available in Appendix E (online supplementary material). McDonald's omega, used as a proxy for internal consistency, was estimated at .71. Additional item-level statistics, including factor loadings, communalities, and uniqueness values, are reported in Appendix F.

As can be observed in Table 3, respondents reported frequently engaging in practices recommended by the field for conducting SLA research (overall M of five items = 4.27). These practices included testing L2 proficiency (item 3-1, M = 4.77), defining the concept of language proficiency (item 3-2, M = 4.30), checking and reporting the quality of the proficiency test(s) in use (item 3-3, M = 4.20), and assessing L2 proficiency using an independent test (item 3-5, M = 4.39). They also reported occasionally using more than one proficiency assessment tool (item 3-9, M = 3.69). It should be noted, however, that responses varied widely as evidenced by relatively high *SDs* and IQRs. For each practice-related item, respondents used the full range of the scale (that is, with at least one respondent selecting "almost never true" and at least one respondent selecting "almost always true"). Regarding less desirable practices (overall M of four items = 3.11), respondents reported infrequent engagement in two practices that contrasted with items 3-1 and 3-2 (item 3-4, M = 2.75; item 3-6, M = 2.79). However, they reported more frequent use of indirect assessment methods (item 3-7, M = 3.52) and skipping quality checks or reporting if high quality had been previously established (item 3-8, M = 3.38).

RQ 4. Which assessment techniques are frequently used by SLA researchers to assess L2 proficiency in their research?

For each of the seven assessment techniques, respondents indicated how often they used it to assess L2 proficiency in their studies published over the past 5 years, using 1–6 Likert-type scale ("never" to "always [or very frequently]"; item 4 in the survey).

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Table 3. Descriptive statistics for participants reported proficiency assessment practices

Item	Researcher practices	*n*	M	*SD*	95% CI	Median	IQR
Practices consistent with field recommendations							
3–1	I test learners' L2 proficiency when I conduct an SLA study.	108	4.77	1.32	[4.52, 5.02]	5	2.00
3–2	I define in my research the concept of language proficiency prior to assessing learners' proficiency.	109	4.30	1.50	[4.02, 4.59]	5	3.00
3–3	I check and report the quality (e.g., validity, reliability) of the proficiency test(s) that I use for my research.	108	4.20	1.39	[3.94, 4.47]	4	2.00
3–5	I assess L2 proficiency using an independent test (e.g., standardized tests, cloze test).	107	4.39	1.43	[4.12, 4.67]	5	1.00
3–9	I use more than one proficiency assessment tool to assess L2 proficiency.	109	3.69	1.56	[3.39, 3.98]	4	3.00
Practices not consistent with field recommendations							
3–4	I do not measure learners' L2 proficiency unless I use L2 proficiency either as a dependent or an independent variable for my research.	106	2.75	1.69	[2.42, 3.07]	2	3.00
3–6	I do not provide a definition of language proficiency in my research because the term is self-explanatory.	107	2.79	1.56	[2.49, 3.08]	2	2.00
3–7	I assess L2 proficiency using an indirect method (e.g., classroom or institutional level, learner self-ratings, researcher impressionistic judgments).	108	3.52	1.56	[3.22, 3.82]	4	3.00
3–8	If the high quality of the proficiency test was reported elsewhere, I do not check nor report it in my own research.	107	3.38	1.65	[3.07, 3.70]	4	3.00

Note: Item codes correspond to the numbering in the survey provided in Appendix A in the online supplementary material. Six-point Likert scale for all items: 1 = almost never true, 6 = almost always true.

Figure 3 summarizes the reported frequencies, and detailed descriptive statistics are provided in Appendix G (online supplementary material).

The method researchers, as a group, reported using most frequently was institutional level. Following institutional level, standardized tests, reduced redundancy tests, self-ratings, and "other" techniques were reported at similar frequencies. The "other" category encompassed a range of methods, including course grades, teacher evaluations, accent ratings, grammaticality judgment tests, researcher-developed independent measures, oral picture descriptions, and the Language History Questionnaire (LHQ) 3.0. Impressionistic judgment, oral interviews, and vocabulary tests were reported as the least frequently used techniques.

RQ 5. What are researchers' reasons for choosing particular proficiency assessment techniques?

To explore researchers' reported reasons for choosing particular proficiency assessment methods, we examined responses to the open-ended item (item 5 in the survey;

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

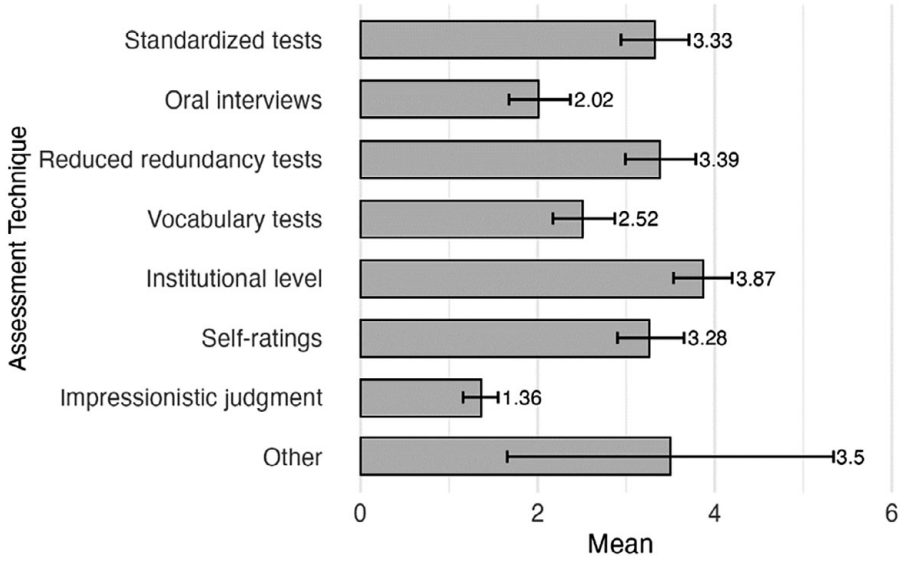

Figure 3. Mean frequency of employing various proficiency assessment techniques.

k = 90): "What is your preferred method of proficiency assessment for research purposes and why?" We used a flexible content analysis procedure (Krippendorff, 2004) whereby we developed a codebook based on recurring reasons for preferred assessment methods in the response data, and each response was coded accordingly. The second author coded all responses using this codebook (provided in Appendix H and on the Open Science Framework: https://osf.io/fv8ug/?view_only=f34b7ed61f2e4b8887ebf87661c9604d), with multiple codes allowed per response. To establish interrater reliability, the first author independently coded 30% of the data; percent agreement across codes ranged from 84% to 100%. Following recommendations by Feng (2015) and given our nominal coding, uneven marginal distributions, and low task difficulty, we calculated interrater reliability using Gwet's agreement coefficient 1 (AC1) (2014) and the irrCAC R package (Gwet, 2019). AC1 values for all categories attested in the double-coded data ranged from 0.72 to 1, indicating substantial to very good agreement (Landis & Koch, 1977), with only one category falling below 0.81 (Landis & Koch's benchmark cutoff for "almost perfect" agreement, p. 165). We then grouped categories together and identified overarching patterns in the data.

For ease of exposition and interpretation, codes were grouped into seven overarching categories based on respondents' reasons for preferred assessment methods: (1) validity/rigor, (2) precedent/comparability, (3) practicality, (4) participant considerations, (5) research/project-dependent considerations, (6) proficiency measure considerations, and (7) other. Responses that named a preferred method without providing a rationale (k = 16) were classified under "Method Only" and excluded from further analysis. Figure 4 shows the percentage of codes applied to each of the seven categories. For a detailed distribution of codes within each category and the percentage of responses that received each code, refer to Appendix I (online supplementary material).

As shown in Figure 4, approximately 36% of the codes applied to responses to this open-ended item had to do with practicality. Included within this category were responses that described various practical concerns including the ease of administration

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

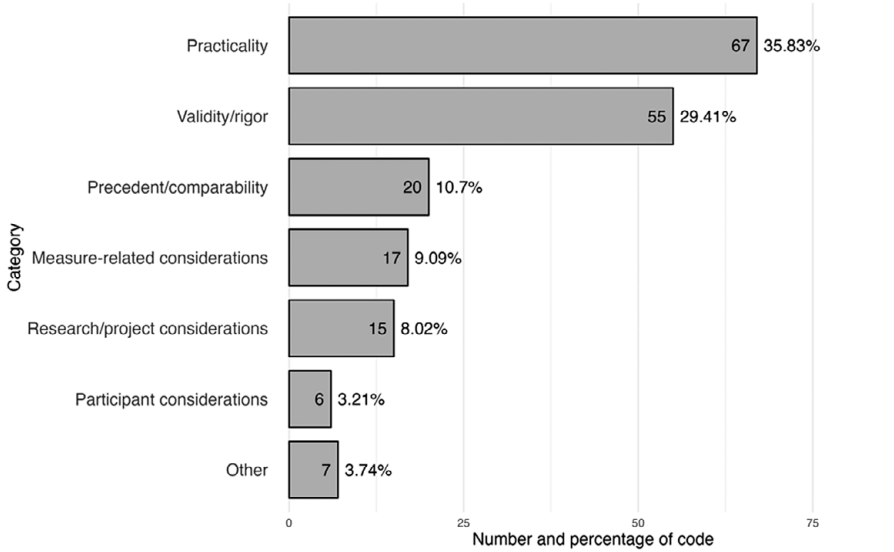

Figure 4. Percentage of codes applied corresponding to each of seven overarching categories.

of the measure (mentioned in 21% of total responses), time considerations, availability or accessibility of the measure to the researcher, cost or money considerations, scoring ease, and others (see Appendix I in the online supplementary material). For example, one participant responded, "I often use the Lextale-ESP test and a portion of the Spanish DELE test, because then I can test both grammar and vocabulary, both have been widely used, have been validated and shown to discriminate well between learner levels and between even very advanced non-native speakers and native speakers (although the DELE test discriminates less well between lower to intermediate levels) **and because they are easy to administer, not time consuming, available, and easy to score**" [emphasis added]. As illustrated in this response and shown in Figure 4, respondents also commonly mentioned considerations related to the validity, reliability, and/or discriminability of a proficiency measure, with 29% of codes applied having to do with validity and related constructs. Approximately 11% of the codes applied were related to following precedent or ensuring that measures were comparable across studies, languages, or populations. For example, one participant responded, "Standardized tests scores, as they allow for a comparison between populations across studies when needed," and another responded "Currently, I prefer using an independent measure created by a group of researchers who study Spanish second language acquisition alongside an indirect measure such as course/institutional level. This practice is largely based on using practices that already occur in the field so that some level of comparison across studies can be made." Nearly 17% of total responses indicated that the choice of a particular proficiency assessment method depends on the type and/or goals of the research being conducted (constituting over 8% of all codes applied). Additionally, researchers mentioned measure-related considerations (e.g., choosing a method because it is well recognized or interpretable) and participant-related considerations (e.g., that a particular approach does not place undue pressure on participants) as reasons for selecting certain methods. About 7% of total responses included reasons categorized as "other." Some of these reasons included "being able

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

to edit the instrument for specific learners," "the test being more natural (less test-like)," and "being suitable for the context of teaching/learning."

Discussion

While previous research syntheses (Park et al., 2022; Thomas, 1994, 2006; Tremblay, 2011) have made important contributions by documenting overall trends in proficiency assessment and reporting practices, they have primarily focused on describing what researchers have done without examining the underlying reasoning behind researchers' choices. The present study extends this line of inquiry by surveying SLA researchers about their beliefs, practices, and decision-making, offering a new layer of insight that complements the findings of these earlier syntheses. The findings reveal several key insights into the current state of proficiency assessment practices and suggest pathways for improvement.

First, our investigation shows that SLA researchers generally recognize the importance of defining, assessing, and reporting L2 proficiency as part of their research design (RQ 1). This is evident from the high level of agreement with statements endorsing recommended practices, such as defining learners' proficiency and using direct assessment methods. However, respondents displayed some ambivalence regarding the soundness of indirect measures. This ambivalence is not surprising as the appropriateness of indirect measures partly depends on the context of their use and thus cannot be simply dismissed as a questionable practice. While indirect measures are generally considered less methodologically rigorous than direct measures for estimating learners' L2 proficiency, their use may be justified and more appropriate when combined with proficiency data obtained from direct measures or when L2 proficiency is not central to the RQ.

When asked to evaluate various proficiency assessment techniques according to key test characteristics (RQ 2), researchers rated standardized tests and oral interviews highest for validity-related traits (e.g., accuracy, trustworthiness) but lower for practicality-related traits (e.g., ease of administration, cost-effectiveness). In contrast, self-ratings and impressionistic judgments were seen as more practical but less valid. This pattern highlights an enduring trade-off between methodological rigor and practical feasibility in proficiency assessment. Notably, reduced redundancy tests and vocabulary tests were rated relatively high on both validity and practicality dimensions, suggesting that these methods may offer a balanced solution for researchers seeking to navigate this trade-off. The high ratings for the validity-related traits of reduced redundancy tests likely reflect researchers' improved understanding of assessment tools such as elicited imitation tasks. These tasks have long been debated and frequently criticized as measures of language proficiency, but recent validation efforts—through empirical studies (e.g., Park, Solon, Henderson, & Dehghan–Chaleshtori, 2020; Wu, Tio, & Ortega, 2022) and meta-analyses (Kostromitina & Plonsky, 2021; Yan, Maeda, Lv, & Ginther, 2016)—may have contributed to a more favorable shift in researchers' perception.

In line with their beliefs about proficiency assessment, researchers' reported practices were generally consistent with current methodological recommendations for defining, assessing, and reporting L2 proficiency (RQ 3). However, reported practices showed greater variability than beliefs, as reflected in larger *SDs*. When practice items modeled after belief items (Table 2) were compared side by side with corresponding belief items (see Table 4), it became evident that researchers' practices did not always

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

Table 4. Comparison between mean scores on similarly worded belief and practice items

n	Belief item		M	Practice item		M
Beliefs and practices consistent with field recommendations						
108	1–10	It is important for SLA researchers to check and report the quality (e.g., validity, reliability) of the L2 proficiency test that they use.	5.06 (agree)	3–3	I check and report the quality (e.g., validity, reliability) of the proficiency test(s) that I use for my research.	4.20 (occasionally true)
107	1–5	It is better practice to assess L2 proficiency directly with an independent test (e.g., standardized tests, cloze test) than indirectly (e.g., through classroom or institutional level, learner self-ratings, researcher impressionistic judgments).	4.55 (agree)	3–5	I assess L2 proficiency using an independent test (e.g., standardized tests, cloze test).	4.39 (occasionally true)
108	1–3	It is important for SLA researchers to assess participants' proficiency in the L2 as part of their research design.	5.05 (agree)	3–1	I test learners' L2 proficiency when I conduct an SLA study.	4.77 (usually true)
Beliefs and practices not consistent with field recommendations						
105	1–4	Knowing learners' L2 proficiency is not important if L2 proficiency is not a variable of interest (either dependent or independent) in the research design.	2.41 (disagree)	3–4	I do not measure learners' L2 proficiency unless I use L2 proficiency either as a dependent or an independent variable for my research.	2.74 (rarely true)
105	1–7	If the high quality of the proficiency test was reported elsewhere, it is unnecessary for SLA researchers to check and report it in their own research.	2.85 (slightly disagree)	3–8	If the high quality of the proficiency test was reported elsewhere, I do not check or report it in my own research.	3.40 (rarely true)

Note: The comparison of belief and practice items was conducted only with respondents who provided responses to both items. Consequently, the sample size varies across item comparisons, which may result in mean scores for belief and practice items differing from those in Tables 2 and 3.

align with their strongly endorsed beliefs about best practices. For instance, although most respondents agreed that checking and reporting test quality is important (item 1-10, M = 5.06), they reported doing so only occasionally in practice (item 3-3, M = 4.20). Likewise, while respondents tended to disagree with the notion that reporting test quality is unnecessary if it has been established elsewhere (item 1-7, M = 2.85), their corresponding practice scores (item 3-8, M = 3.40) suggest that they sometimes skip this step. Bearing in mind the obvious caveat that the scales used to measure beliefs and practices were different (beliefs: agreement scale; practices: frequency scale), these comparisons nevertheless highlight a consistent belief–practice gap. These discrepancies between what researchers believe and what they actually practice may foreshadow the practical constraints reported more explicitly in response to RQ 5, where researchers cited issues, such as time, accessibility, and ease of administration as key considerations in selecting proficiency assessment tools. More broadly, they underscore a tension within the field: Even when researchers are methodologically aware, practical barriers may limit the extent to which recommended practices are fully implemented.

When asked about the proficiency assessment techniques they have used over the past 5 years (RQ 4), researchers reported using institutional level most frequently, despite its well-documented limitations, a finding that corresponds with trends reported in recent syntheses of proficiency assessment methods and reporting (e.g., Park et al., 2022; Tremblay, 2011). However, independent measures, such as standardized tests and reduced redundancy tests, as well as self-ratings and other techniques were also reported to be used with similar frequency. This diversity in assessment techniques, together with the frequent use of multiple assessment techniques (see Table 3), suggests a growing recognition among researchers of the value of using multiple and varied tools to capture L2 proficiency more comprehensively.

In exploring the rationale behind researchers' preferred assessment approaches (RQ 5), practicality emerged as the most frequently cited factor, accounting for 36% of all applied codes. This included concerns such as time, cost, ease of administration, accessibility, and scoring—supporting prior claims that practical considerations often play a central role in methodological decision-making (Leclercq & Edmonds, 2014). In addition to practicality, however, researchers emphasized the importance of validity, reliability, and overall rigor in guiding their choices. While the ways researchers conceptualize these constructs warrants further investigation, the dual emphasis on feasibility and methodological soundness suggests a strong interest in tools that strike a balance between these priorities. This may help explain the relatively favorable evaluations and frequent use of reduced redundancy tests (see Figure 3) as well as the growing interest in shortcut proficiency tools—an interest reflected in a growing number of validation studies (e.g., Alpizar, Li, Norris, & Gu, 2023), meta-analyses (e.g., Kostromitina & Plonsky, 2021), and special issues (e.g., Solon & Park, 2024b). Tools such as elicited imitation tasks and C-tests have not only been recognized for their validity, reliability, and practicality (Solon & Park, 2024a) but also are highly adaptable —capable of targeting specific linguistic features or assessing global proficiency (Kostromitina & Plonsky, 2021; Yan et al., 2016). Recent innovations such as online administration and automated scoring (e.g., Isbell, Kim, & Chen, 2023; Kim, Chen, & Liu, 2024; McGuire & Larson-Hall, 2025) have further enhanced their accessibility and scalability, helping to reduce the logistical burdens traditionally associated with rigorous proficiency assessment. Continued development and refinement of such accessible, validated tools hold considerable promise for advancing methodological rigor in SLA research. By lowering logistical barriers without sacrificing validity, this line of work

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

may enable more researchers to implement best practices across a wider range of research contexts.

Conclusions

The present study sought to better understand ongoing trends in L2 proficiency assessment and reporting practices by exploring SLA researchers' beliefs and reported practices related to measuring L2 proficiency. The findings suggest that researchers' beliefs and reported practices are generally in line with current methodological standards recommended in the field, although room for improvement remains. Notably, the discrepancy between researchers' beliefs and their actual practices highlights the challenges researchers face in implementing recommended standards, possibly due to practical constraints, which according to the present survey play a central role in the decision-making process for proficiency assessment (sometimes at the expense of methodological rigor). Although this study did not directly examine what those constraints are, Park et al. (2022) noted that certain research contexts may inherently discourage the use of more rigorous assessment methods, such as the implementation of independent tests. Future research may explore ways to improve the adaptability and accessibility of existing proficiency assessment tools to better meet the diverse needs and limitations of research contexts. This could include developing more flexible, context-specific instruments that can be easily integrated into diverse research settings, potentially supported by emerging technologies. For example, recent work on automated scoring and validation of elicited imitation tasks (e.g., Kim et al., 2024; McGuire & Larson-Hall, 2025) demonstrates promising advances that could help lower the logistical barriers traditionally associated with administering independent proficiency measures.

Ultimately, improving L2 proficiency assessment practices will require not only greater methodological awareness but also the creation of practical support mechanisms that help close the gap between recommended methodological standards and feasible implementation. This may involve increasing access to ready-made assessment tools through data-sharing platforms such as the IRIS repository (Marsden, Mackey, & Plonsky, 2016), offering targeted training and resources to build assessment literacy, and fostering a research culture that values transparency in methodological decision-making. Institutional stakeholders—such as journal editors and large-scale proficiency test providers—can play a key role in supporting these efforts. Journal editors can encourage transparent reporting and justification of proficiency assessment choices, while test developers might consider embracing a form of civic responsibility by making select, validated materials available for research purposes. Increasing access to such tools would reduce financial and logistical barriers, particularly for researchers working in under-resourced contexts, and promote more equitable adoption of rigorous assessment practices. By addressing both knowledge-based and access-based barriers, the field can move toward more consistent, context-specific, and methodologically sound approaches to L2 proficiency assessment.

While the current study provides valuable insights into the beliefs and practices of SLA researchers regarding L2 proficiency assessment, we recognize that it is not without limitations. Our sample (n = 111) was relatively small, and a majority of respondents were from North America, potentially limiting the generalizability of our findings to the broader SLA community. Furthermore, more than half of respondents (n = 88) graduated after 2010—that is, after the "methodological turn" (Byrnes, 2013, p. 825)

https://doi.org/10.1017/S0272263125101058 Published online by Cambridge University Press

started to gain momentum in the field—and this may have influenced their approaches to proficiency assessment and skewed the findings toward more contemporary methodologies and perspectives. We also recognize the possibility of self-selection bias in that researchers particularly attentive to methodological rigor may have been more inclined to participate in the survey. Beyond limitations of the sample, the survey instrument itself represents an initial effort to explore researchers' rationales behind proficiency assessment practices. As such, the questions provide preliminary insights but cannot offer fully detailed explanations of the underlying reasons for participants' choices. Moreover, as is typical of survey research, we could not fully control for variation in how respondents interpreted survey terms (e.g., vocabulary test, oral interview, independent test, indirect method), despite efforts to provide clear definitions and examples where appropriate. Given the wide range of research contexts, assessment tools, and researchers, it was not possible to define or delimit all such terms. Such variability may have influenced responses and should be considered when interpreting the findings.

While these limitations may restrict the generalizability and/or depth of the study's findings, they do not diminish the importance of the insights gained in the present study. Future research may complement these findings by exploring SLA communities in different geographic regions to capture a broader range of perspectives and practices within the global SLA research community. The present results also signal the need for further research into, for example, how researchers match a proficiency assessment instrument to the goals of their research or how the availability of assessment tools impacts choice of technique. Future research that pursues more detailed exploration of particular aspects of the present findings and that triangulates survey findings with other measures will aid in our understanding of methodological choices, thereby enhancing our ability to promote greater methodological rigor and ethical integrity in SLA research.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263125101058.

Data availability statement. The experiment in this article earned Open Material badge for transparent practices. The data are available at https://url.avanan.click/v2/r02/ and https://osf.io/fv8ug.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.

Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality, and attitudes*. John Wiley & Sons.

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*, 727–744.

Alpizar, D., Li, T., Norris, J. M., & Gu, L. (2023). Psychometric approaches to analyzing C-tests. *Language Testing*, *40*, 107–132.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F. (1988). Language testing-SLA research interfaces. *Annual Review of Applied Linguistics*, *9*, 193–209.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge University Press.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press.

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*, 825–827.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 46–69). Oxford University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Cummins, J. (1980). The construct of language proficiency in bilingual education. In J. E. Alatis (Ed.), *Current issues in bilingual education: Georgetown University Round Table on Languages and Linguistics 1980* (pp. 81–103). Georgetown University Press.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

De Costa, P. I., Cinaglia, C., & Rabie-Ahmed, A. (2024). *Ethical issues in applied linguistics scholarship*. John Benjamins.

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*, 132–153.

Dodeen, H. (2023). The effects of changing negatively worded items to positively worded items on the reliability and the factor structure of psychological scales. *Journal of Psychoeducational Assessment*, *41*, 298–310.

Dörnyei, Z. (2010). *Questionnaires in second language re- search: Construction, administration, and processing* (2nd ed.). Routledge.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*, *11*, 13–22.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *ITL – International Journal of Applied Linguistics*, *168*, 4–32.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics, LLC.

Gwet, K. L. (2019). Calculating chance-corrected agreement coefficients (CAC). https://cran.r-project.org/web/packages/irrCAC/vignettes/overview.html.

Harding, L., & Kremmel, B. (2021). SLA researcher assessment literacy. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 54–65). Routledge.

Hulstijn, J. H. (2010). Measuring second language proficiency. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 185–200). Benjamins.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*, 422–433.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Hulstijn, J. H. (2019). An individual-differences framework for comparing nonnative with native speakers: Perspectives from BLC theory. *Language Learning*, *69*(S1), 157–183.

Hulstijn, J. (2024). Predictions of individual differences in the acquisition of native and non-native languages: An update of BLC theory. *Languages*, *9*(5), 173.

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*, 172–195.

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics*, *2*, 100060.

Isbell, D. R., Kim, K. M., & Chen, X. (2023). Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test. *Research Methods in Applied Linguistics*, *2*, 100076.

Kim, K. M., Chen, X., & Liu, X. (2024). Accuracy scoring of elicited imitation: A tutorial of automating speech data with commercial NLP support. *Research Methods in Applied Linguistics*, *3*, 100127.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Publications.

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*, 886–911.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Lado, R. (1961). *Language testing*. McGraw–Hill.

Larson–Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics 1. *The Modern Language Journal*, *101*, 244–270.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159.

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics*, *2*, 100064.

Leclercq, P., & Edmonds, A. (2014). How to assess L2 proficiency? An overview of proficiency assessment research. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency perspectives from SLA* (pp. 3–23). Multilingual Matters.

Loewen, S., Gönülal, T., Isbell, D. R., Ballard, L., Crowther, D., Lim, J., Maloney, J., & Tigchelaar, M. (2020). How knowledgeable are applied linguistics and SLA researchers about basic statistics? Data from North America and Europe. *Studies in Second Language Acquisition*, *42*, 871–890.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*, 360–388.

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design*. Routledge.

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge.

McGuire, M., & Larson-Hall, J. (2025). Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. *Research Methods in Applied Linguistics*, *4*, 100197.

McManus, K. (2022). Are replication studies infrequent because of negative attitudes? Insights from a survey of attitudes and practices in second language research. *Studies in Second Language Acquisition*, *44*, 1410–1423.

Norris, J. M. (2010). *Understanding instructed SLA: Constructs, contexts, and consequences* [Plenary address]. 20th Annual Conference of the European Second Language Association, Università di Modena e Reggio Emilia Reggio Emilia, Italy, 1–4 September 2010.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 716–761). Blackwell Publishing Ltd.

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). Routledge.

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). Proficiency reporting practices in research on second language acquisition: Have we made any progress? *Language Learning*, *72*, 198–236.

Park, H. I., Solon, M., Henderson, C., & Dehghan–Chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *The Modern Language Journal*, *104*, 133–151.

Plonsky, L. (Ed.). (2015). *Advancing quantitative methods in second language research*. Routledge.

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591–610.

Plonsky, L., Larsson, T., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2024). A taxonomy of questionable research practices in quantitative humanities. In P. I. De Costa, A. Rabie-Ahmed, & C. Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship* (pp. 10–27). John Benjamins.

Porte, G. (2010). *Appraising research in second language learning: A practical approach to critical analysis of quantitative research* (2nd ed.). John Benjamins.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal*, *100*(S1), 190–208.

Sato, M. (2022). Mixed methods research in ISLA. In L. Gurzynski-Weiss & Y. Kim (Eds.), *Instructed second language acquisition research methods* (pp. 79–102). John Benjamins.

Schoonen, R. (2011). How language ability is assessed. In E. Hikel (Ed.), *Handbook of research in second language teaching and learning: Vol. II*. Routledge.

Solon, M., & Park, H. I. (2024a). Elicited imitation in second language acquisition research: New insights to advance methodological rigor (Introduction to the special issue). *Research Methods in Applied Linguistics*, *3*, 100112.

Solon, M., & Park, H. I. (Eds.) (2024b). Elicited imitation in second language acquisition research [Special issue]. *Research Methods in Applied Linguistics*, *3*.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). John Benjamins.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372.

Winke, P., & Brunfaut, T. (2021). Perspectives on "knowing" a second language: What are we seeking to measure? In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 1–8). New York: Routledge.

Wu, S.-L., Tio, Y. P., & Ortega, L. (2022). Elicited imitation as a measure of L2 proficiency: New insights from a comparison of two L2 English parallel forms. *Studies in Second Language Acquisition*, *44*, 271–300.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*, 497–528.

Cite this article: Park, H. I., Solon, M., & Lee, K. (2025). Understanding proficiency assessment practices in SLA research: Insights from researcher beliefs and practices. *Studies in Second Language Acquisition,* 1–25. https://doi.org/10.1017/S0272263125101058