Studies in Second Language Acquisition (2025), 1–15
doi:10.1017/S0272263125000166

CRITICAL COMMENTARY

Resolving the crisis in L2 motivational self-system research: Constructive dialogue and argument-based validation

Meng Liu[1] and Alastair Henry[2,3]

[1]Beijing Foreign Studies University, School of English and International Studies, Beijing, China; [2]Lund University, Centre for Languages and Literature, Lund, Sweden and [3]University West, Department of Social and Behavioural Studies, Trollhättan, Sweden
Corresponding author: Alastair Henry; Email: alastair.henry@englund.lu.se

(Received 05 October 2024; Revised 03 February 2025; Accepted 05 February 2025)

Abstract

Al-Hoorie, Hiver, and In’nami (2024) offer compelling arguments for why L2 motivational self-system research is currently in a state of validation crisis. Seeking a constructive resolution to the crisis, in this response we argue that two fundamental conditions are needed for the field to emerge stronger: psychological readiness and methodological maturity. For psychological readiness, we call for a reframing of the “crisis” narrative. We highlight the need to value controversy, to normalize failure and (self-)correction, and to resist the allure of novelty. For methodological maturity, we suggest that an argument-based approach to validation can provide a constructive solution to current controversies. We present an integrated framework that can guide systematic validation efforts, and we demonstrate its application using a recent validation study as an example.

Keywords: argument-based validation; credibility revolution; L2MSS; validation crisis

Introduction

Al-Hoorie, Hiver and In’nami (2024; henceforth Al-Hoorie et al.) make a compelling case for why L2 motivational self-system (L2MSS) research is currently in a state of validation crisis. It is a crisis evidenced by (a) a lack of systematic validation efforts, and (b) initial examinations that reveal discriminant validity issues regarding core L2MSS scales. Both call into question the credibility of the field.

In a previous response to Al-Hoorie et al., we (Henry & Liu, 2024) argued that even though a validation crisis in L2MSS research might be manifested in jangle fallacy problems at the measurement level, its roots lie at the construct level. While we argued that jingle fallacy problems at the construct level are of a magnitude such that it may no longer be meaningful to continue the investigation of the ideal L2 self and the ought L2

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

self as they are currently conceptualized, other scholars have focused their responses on controversies at the measurement level.

In this paper, we engage with the current debate on validation by identifying approaches that can lead toward a constructive resolution of the crisis, and through which the field can emerge in a stronger position. For this to occur, two fundamental conditions are needed: psychological readiness and methodological maturity. While discussions have so far mostly concentrated on methodology, we believe that the human element—researcher emotions and mindsets—is equally important if the crisis is to be resolved. With this in mind, we examine the emotional responses that a crisis can trigger, and consider how a state of crisis can be transformed into opportunities for growth. Moving on to methodological maturity, we discuss contemporary views of validity and advocate for an argument-based approach to validation that can support more exacting efforts in validation research.

The emotional weight of a crisis

As L2 motivation researchers, we recognize that the concerns raised by Al-Hoorie et al. (2024) can trigger complex emotions. Initial reactions can be underpinned by "fight or flight" instincts that range from denial to despair. While many of us will recognize the need to separate the research from the researcher, the research that we do and the discipline in which we operate can feel deeply personal. A crisis that calls into question the credibility of a field that we are part of cannot be other than unsettling. It can be tempting to downplay or dismiss issues that can compromise confidence in our work and to shy away from critiques that have been leveled. Anxiety, or unease over the implications that a crisis can have on ongoing or future projects is an understandable reaction. Frustration is also a natural response, not least in our field when researchers can have made career-spanning investments in work based on L2MSS theory. The realization that their research may have been built on less than stable foundations can be disheartening. Fear can also be a prevalent emotion. Concern about the stigma of association with a potentially discredited field is easy to fathom. Whatever the emotion—or combination of emotions that are generated—there is a risk of paralyzing effects. Action needed to address the crisis in constructive ways may not be taken.

While rarely a part of open discussion, it is important to recognize that these emotions stem from and reflect a commitment to the field. However, if unaddressed, they can present psychological barriers that can hinder meaningful progress. Denial can lead to a reluctance to engage with critical perspectives. Anxiety can result in unduly conservative research practices that decelerate progress. While frustration might escalate into a loss of motivation or the cynical dismissal of an entire field, fear can prevent the researcher from opening their work to critical scrutiny.

Given these complex emotions and their implications, it becomes important to consider how the psychological readiness needed to navigate the crisis in productive ways can be achieved. While emotions are ultimately personal (and private), we believe that the creation of a safe space for critical reflection and scrutiny is crucial if change is to happen. In the following section, we explore strategies that can foster such an environment, and which can bring about a transformation from a state of "crisis" to a state of credibility.

From crisis to credibility

Reframing the "crisis" narrative

An immediate and readily implementable step toward the creation of an environment for constructive dialogue is to consider Al-Hoorie et al.'s narrative framing. As a rhetorical device, the declaration of a "validation crisis" has been effective. It has directed attention to issues that have been systematically underappreciated, and it has sparked much-needed debate. However, as with any powerful framing, there is a possibility that the term can be misinterpreted and that the entire field is perceived as fundamentally unsound.

Of particular concern is the potential conflation of a *validation* crisis and a *validity* crisis. While Al-Hoorie et al. purposefully used the term "validation crisis," it can easily be misunderstood as a "validity crisis." The distinction is nuanced but crucial. A validation crisis primarily refers to insufficient or absent validation efforts, a point that the authors emphasize. In contrast, a validity crisis would indicate a demonstrated lack of validity resulting from systematic validation attempts. Currently, the field faces the former. While initial evidence has now been supplied, field-wide systematic validation efforts would be needed before convincing claims can be made about the latter.

Another concern involves the defeatist connotations associated with the term "crisis." As Vazire (2018) has put it, "crisis implies that we are at a loss for solutions, when in fact we have identified many ways to improve science's credibility" (p. 411). While Vazire (2018) was commenting on why she preferred to refer to psychology's replicability crisis as a "credibility revolution" (see Liu, 2023 for a recent discussion of this in relation to applied linguistics), the same principle applies to the validation crisis in our field. Given that systematic validation efforts have yet to be undertaken, perpetuation of a "crisis" discourse carries an additional risk—the premature stigmatizing an entire field. Association with a field in crisis can prompt researchers to steer clear of this line of inquiry. This, in turn, could lead to a premature dismissal of valuable research directions, and stifle progress in addressing the very issues that require resolution.

As motivation researchers working in applied linguistics, we are well aware of the importance of semantics and the power of (re)framing. How we think about our challenges can shape how we approach them. While Al-Hoorie et al.'s portrayal of a "validation crisis" has been effective in drawing attention to critical issues that have long been neglected, and has been the catalyst for productive thinking (e.g., Henry & Liu, 2024; Oga-Baldwin, 2024), to move to the next phase it can be useful to talk in terms of a "credibility revolution." Calling this a "credibility revolution" instead of a "crisis" is not just a play on words. Rather, it signifies a strategic shift in orientation. By thinking productively about methodological innovation and conceptual revision, we stand not only to enhance our agency in driving positive change; we can also mitigate the risk of allowing complex negative emotions to cloud our judgment.

Valuing controversy

Unlike research in other applied disciplines of motivation science, L2 motivation research has a history largely built around models developed *within* the field, and where the influence of mainstream theories and frameworks has been limited. Among the many problems that insularity has brought, the "endowment effect" has created particular challenges. While in any field researchers may "end up within a certain theoretical camp

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

for reasons other than pure science" (King & Fryer, 2024, p. 10), in a field as insular as ours, flags are easily tied to the masts of particular models. However, it is not merely the case that researchers can have a natural predisposition to afford greater value to the theories upon which their careers have been founded. The "endowment effect" can easily lead to camp-ism, defensiveness, and a shuttering off of productive communication with researchers who take opposing views (King & Fryer, 2024).

As motivation researchers, we need to recognize that controversy is important. Because it can highlight "the importance of exacting definitions of constructs," and can encourage researchers from different camps to engage in debate, controversy can be the driver of development (Ryan, 2024, p. 6). Beyond the need to approach controversy in a non-defensive manner, it is important to engage with *challenging* controversies—those that have the potential to be productive and which can require us "to take a step back and rigorously evaluate the theories we use" (King & Fryer, 2024, p. 10).

While controversies abound in motivational science, not all will necessarily be productive. Controversy per se is not a trigger for development:

> To make progress in motivation research, it may be useful to focus on resolving existing controversial issues. However, it is also important to consider under what conditions *productive controversies* arise. Two especially important conditions are (a) precision of theories and (b) precision of measures and empirical study designs to test them. Precise theoretical propositions and precise measurements are needed, otherwise, contradictions may not be detectable.
>
> (Pekrun, 2024, p. 7, emphasis added)

In relation to Pekrun's first point—the precision of theories—we (2024) have argued that jingle fallacy problems at the construct level have created significant problems in the construal and operationalization of the original L2MSS constructs, and in subsequent iterations where L2 self-guides have been bifurcated to reflect promotion and prevention motives (Dörnyei, 2009; Papi, Bondarenko, Mansouri, Feng, & Jiang, 2019). In our response to Al-Hoorie et al.'s (2024, p. 10) initiative in "opening a discussion" around validity in L2MSS research, we were at pains to not only address the theoretical imprecision in the L2 self-guide construct and the consequences that follow when it is operationalized. In a spirit of productive engagement and potential cross-fertilization (King & Fryer, 2024), we also explained how self-guides and other standards (Higgins, Strauman, & Klein, 1986) can be theoretically incorporated into frameworks of L2 motivation that draw on self-determination theory (e.g., Noels et al., 2019). While it needs to be recognized that "whenever scholars forward new theoretical models or attempt to reframe or restructure what already exists, they are taking risks" (Alexander, 2024, p. 11), we believe that a commitment to theoretical precision can facilitate integration across frameworks and can shift L2 motivation research into a more productive orbit.

Normalizing failure and (self-)correction

Another step that we believe to be important in developing a constructive environment for reform is the normalization of failure and the encouragement of (self-)correction. Here, we can again look to our colleagues in psychology for valuable lessons.

One lesson involves evaluation of the odds of failure. In recent decades, and in response to the replicability crisis, researchers in psychology have worked hard to

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

improve the replicability of their research findings. However, as they have come to realize, a 100% replication rate is neither realistic nor desirable. As Nosek et al. (2022) have pointed out, achieving a near 100% replicability would require "adopting an extremely conservative research agenda that studies phenomena that are already well understood or have extremely high prior odds. Such an approach would produce nearly zero research progress" (p. 730). In fact, Nosek et al. argue that a healthy, theoretically generative research enterprise will inevitably include some nonreplicable findings. As they put it, "science exists to expand the boundaries of knowledge. In this pursuit, false starts and promising leads that turn out to be dead ends are inevitable" (p. 730). Here we can extrapolate this lesson to the case of validation. Just as we should not expect 100% replication rates, neither should we anticipate perfect validation results across all measures and constructs in all contexts. The process of validation is iterative, ongoing, and contextual (AERA, APA, & NCME, 2014). Rather than providing a binary "valid" or "invalid" verdict, the key is to reveal areas for improvement or refinement, a point to which we return in our discussion of validity and validation.

Another lesson involves the importance of intellectual humility in navigating research challenges. As Nosek et al. (2022) have emphasized, researchers should "[get] used to being wrong – a lot" (p. 733). They need to develop mindsets that prioritize *getting it right* over *being right.* In the context of the current crisis (Al-Hoorie et al., 2024), this would involve a willingness to critically examine our own work and an openness to revising our perspectives in the light of emerging evidence. Here, the "loss-of-confidence project" in psychology (Rohrer et al., 2021) can be a source of inspiration. This project invited researchers to publicly share instances where they had lost confidence in their own published findings. By creating a platform for such disclosures, the project aimed to destigmatize self-correction and promote it as a normal and valuable part of the research process. As Bishop (2018) has argued, "the reputations of scientists will depend not on whether there are flaws in their research, but on how they respond when those flaws are noted" (p. 437). By shifting our cultural norms to value critical self-reflection and correction, we can create an environment where rigorous scrutiny of one's own work constitutes a hallmark of scientific integrity.

Resisting the allure of novelty

If we are serious about normalizing failure and encouraging (self-)correction, an obsession with novelty also needs to be confronted. Normalizing failure is not about lowering standards. Rather, it involves creating an environment for continuous improvement through critical self-examination. By focusing on quality, we reduce the temptation to conduct hasty, speculative, or careless research in the pursuit of novelty.

Here, Plonsky's (2024b) framework for study quality provides valuable guidance. High-quality research is described as "(a) methodologically rigorous, (b) transparent, (c) ethical, and (d) of value to society" (p. 1). Notably, novelty is not a criterion. This absence is particularly relevant to validation challenges in L2MSS research, where the pursuit of novel findings has often overshadowed rigorous validation efforts.

The omission of novelty is important. By removing novelty as a parameter for high-quality research, validation, and replication studies gain equal footing with the original research. Moreover, the emphasis on methodological rigor and transparency as hallmarks of quality also aligns with open science movements in applied linguistics. Recent discussions on the topic highlight the importance of these aspects (e.g., Al-Hoorie, Cinaglia, et al., 2024; Liu et al., 2023; Marsden & Morgan-Short, 2023; Plonsky, 2024a). Given the scarcity of open science practices in L2MSS research (Liu, 2024), a credibility revolution would also set the field on a path where it could catch up with ongoing developments in applied linguistics. Adopting open science practices, such as pre-registration, data/code/materials sharing, and transparent reporting would help ensure that validation efforts also meet the standards of quality research.

Toward systematic validation research

Having examined several preconditions for successful navigation of the "validation crisis" in L2MSS research, we now turn to methodological maturity. Here, we define "methodological maturity" as a field's collective capacity to consistently implement, evaluate, and refine rigorous research methods. In the context of validity and validation, this would mean the ability (a) to design and implement robust and systematic validation studies, (b) to critically evaluate the results, and (c) to continuously refine measures and theories based on the findings. In the following sections, we briefly review validity concerns in applied linguistics and explain how contemporary views of validity can help to address them. We then present an argument-based approach to validation and describe how it can provide a promising framework for guiding systematic validation efforts.

Prevalent concerns for validity

While researchers involved in L2MSS research may have been the first to officially declare a state of validation crisis (Al-Hoorie, et al., 2024), concerns regarding validity issues are not new in applied linguistics. Over a decade ago, Norris and Ortega (2012) problematized a "tendency to assume – rather than build an empirical case for – the validity for whatever assessment method is adopted" (p. 575) regardless of the learner population studied or the theoretical interpretations that a researcher employs. Ellis (2021) went further, noting that while there is general recognition of validity issues, researchers have "largely ignored" them (p. 197). To evaluate the extent to which concerns such as these are warranted, Plonsky (2024) showed that in a corpus analysis of 23,142 articles from 22 mainstream applied linguistics journals, only 4% made explicit mention of construct validity. In a similar vein, Teimouri, Sudina, and Plonsky (2021) observed that researchers often "rely on conventions and/or to report reliability and validity evidence from other studies, for example, rather than doing so themselves" (p. 378). These findings underscore the urgent need for more rigorous validation practices and transparent reporting in applied linguistics research. As Plonsky (2024) has argued, "it is incumbent upon researchers to provide explicit evidence of the validity of their measures" (p. 7). This is necessary not simply to fulfill the criteria for methodological rigor, but also to meet the ethical obligation of producing trustworthy findings. In this sense, the validation crisis extends beyond L2MSS research. Lessons drawn from the crisis can resonate with the wider applied linguistics community and can contribute to the further improvement of research quality.

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

Contemporary views of validity

To address the validation crisis effectively, it is important that we align our understanding of validity and validation with views currently held in measurement science. The Standards for Educational and Psychological Testing (henceforth the Standards; AERA et al., 2014) represent the current consensus and state-of-the-art guidelines in measurement research. As Purpura, Brown, and Schoonen (2015) stated in their call for greater validity of quantitative measures in applied linguistics, "the development, use, and evaluation of *all* measured constructs… should be guided by professional standards for "good" practices such as those recommended in the Standards for Educational and Psychological Testing" (p. 39, original emphasis). A comprehensive guide for best practices in test development, use, and interpretation, the Standards can offer some guidance for more robust validation efforts.

Notwithstanding the ongoing debates on validity theories, such as differing views on the role of consequential validity (Cizek, 2020), there is a broad consensus regarding the key characteristics of validity. This set of characteristics—mainly derived from Cronbach (e.g., 1971) and Messick (e.g., 1989)—constitutes the core of the contemporary view of validity as reflected in the Standards. Table 1 shows the six foundational tenets as summarized by Cizek (2020, p. 37).

The Standards define *validity* as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests," and *validation* as "accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" (p. 11). In other words, validity is not an inherent property of the instrument. As Cronbach (1971) noted, "one validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447). This point is further emphasized by Messick (1989), who has observed that "what is validated is not the test or observation device as such but the inferences derived from test scores" (p. 13).

Furthermore, the current version of the Standards (2014) also favors Messick's (1989) notion of validity as a unitary concept, rather than the different types of validity (e.g., content validity, predictively validity) originally specified in the first edition of the Standards back in the 1950s. The different "types" of validity have now been replaced by differing "sources" of validity evidence used to evaluate the adequacy of the inferences from a set of test scores. An important implication of this unitary view of validity is that it resists a binary verdict: "validity is a matter of degree, not all or none" (Messick, 1989, p. 13). To place validity on a continuum is necessary because, in practice, we rarely have all evidence pointing unequivocally to a dichotomous evaluation of the inference as valid or invalid. Similarly, the same evidence may bear different weight, depending on the intended inferences, the context, or the person making the judgment. Such variability means that validation cannot be a one-time activity. Rather, it involves a continuous process to ensure ongoing support for a test's intended inferences, qualification of those inferences, or discovery that the intended inferences are no longer adequately supported (Cizek, 2020).

Table 1. Key tenets of contemporary validity theories (Cizek, 2020, p. 37).

1. Validity pertains to test score inferences.
2. Validity is not a characteristic of an instrument.
3. Validity is a unitary concept.
4. Validity is a matter of degree.
5. Validation involves gathering and evaluating evidence bearing on intended test/measurement score inferences.
6. Validation is an ongoing endeavor.

From instrument validity to inference validity

In line with the contemporary view of validity, we believe an important step toward addressing the validation crisis requires a shift in the way we think about validity from *instrument*-focused, to *inference*-focused. A shift in thinking is needed for multiple reasons. First, it can help researchers move away from the problematic assumptions that have contributed to the current validation crisis. In conceptualizing validity as about inferences (and not instruments), it challenges the false assumption that once an instrument has been "validated" in one context, it can be uncritically applied to another. While not every operational use of an instrument would require a full validation study, researchers must move beyond simply citing prior validation evidence without thoughtful consideration (e.g., merely citing that a scale has been "validated" in other research, as cautioned by Teimouri et al., 2021). This shift in perspective foregrounds the need to carefully evaluate whether existing evidence adequately supports the intended interpretation or use in a new context or population.

Second, it can encourage researchers to be more measured in the claims that they make. Rather than relying on generic statements about a scale's validity, researchers would need to specifically articulate the inferences that they seek to make and provide evidence to support them. As Plonsky (2024b) has noted, "transparency is what allows us to evaluate – and is therefore a prerequisite for – every other facet of quality" (p. 4). In the context of validation, transparency extends beyond sharing research instruments. It involves making transparent the theoretical assumptions that underpin our measures and the inferences that we seek to draw. This, we argue, is a crucial component in a critical evaluation of validity.

A further advantage of this approach is that it aligns better with the complex and context-dependent nature of a psychological construct such as motivation. Just as we would not expect a construct to function similarly across all contexts, neither should we assume uniform measurement quality (regardless of context). This resonates with the emphasis on validation as an ongoing process that seeks continuous improvement through the adaptation of measures. As theories evolve, and as contexts shift, it ensures that measurement remains relevant and meaningful. Finally, this perspective can help mitigate the "jingle-jangle" fallacies prevalent in our field (Al-Hoorie et al., 2024; Henry & Liu, 2024), where constructs with the same name may be conceptualized differently (and thereby entail different inferences), or where constructs with different names overlap substantially in their intended interpretations. By focusing on specific inferences, rather than general claims of validity, we will be able to more clearly delineate and evaluate what, exactly, our measures are capturing.

Argument-based approach to validation

Now that we have established the importance of shifting focus from instrument validity to inference validity, the next logical question is: How do we go about validating these inferences? To address this question, we turn to the *argument-based approach to validation*, an approach that can provide a systematic framework for evaluating the validity of score interpretations and uses. The argument-based approach to validation was primarily developed by Kane (1992, 2006, 2013) who drew on Toulmin's (2003) model of argument to structure and evaluate test score inferences. This approach aligns well with contemporary views of validity and supplies the conceptual tools needed for applying the Standards (2014) in practice.

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

Importantly, the focus on inferences and evidential support renders the argument-based validation framework applicable to any type of scores, whether derived from performance tasks or self-report measures. As a systematic framework, the argument-based approach to validation has been applied in varying forms of psychological and educational research, including language testing and applied linguistics. In the field of language testing, the approach gained traction back in the 2000s. In validating the TOEFL iBT, Chapelle, Enright, and Jamieson (2008) provided one of the first comprehensive applications of this approach. Moving beyond large-scale language tests, Purpura et al. (2015) and Révész and Tineke (2020) made a compelling case for how Kane's framework could be utilized to justify the interpretation of scores obtained through L2 elicitation devices for research purposes, and thus expanded the scope of application to second language acquisition and applied linguistics research in general. Over the years, edited volumes and monographs on validity arguments in language testing and beyond (e.g., Chapelle, 2021; Chapelle & Voss, 2021; Cizek, 2020) have been produced to facilitate wider adoption of the approach.

Given the successful application of argument-based validation in neighboring fields, there is significant potential in applying this framework to L2MSS research (and indeed other areas of L2 psychology). By adapting the principles to the specific context of L2 motivation, a more robust and systematic approach to addressing the crisis can be developed.

An integrated framework for argument-based validation

Drawing on insights from the Standards (2014) and key works on argument-based validation (e.g., Chapelle, 2021; Cizek, 2020; Kane, 2013), we present a schematic representation of an integrated framework for argument-based validation (Figure 1).

The figure illustrates the workflow for argument-based validation, which begins with theory and ends with validated score interpretation and use. In an argument-based approach, theory plays a fundamental role throughout the process—from informing the initial construction of the argument and the instrument to guiding the validation process and interpretation of generated evidence (Chappelle, 2021). From theory, we move to the core, argument-based validation process. This process consists of three key

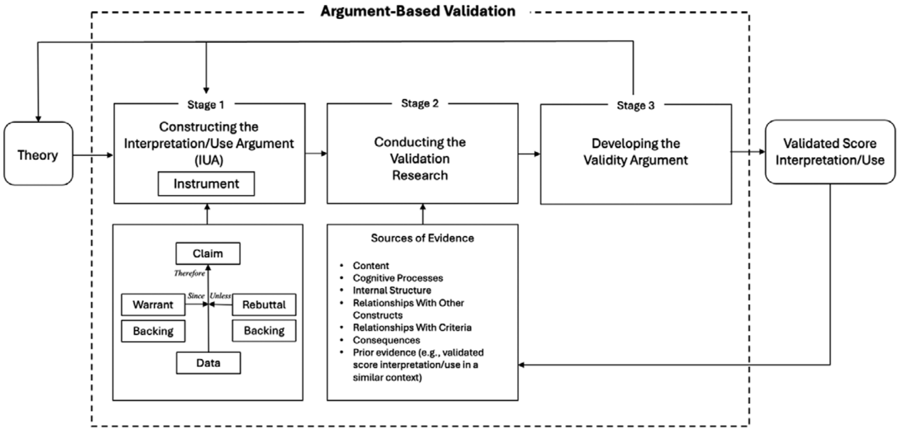

Argument-Based Validation
Theory
Stage 1
Constructing the Interpretation/Use Argument (IUA)
Instrument
Stage 2
Conducting the Validation Research
Stage 3
Developing the Validity Argument
Validated Score Interpretation/Use
Claim
Therefore
Since
Unless
Warrant
Rebuttal
Backing
Backing
Data
Sources of Evidence
Content
Cognitive Processes
Internal Structure
Relationships With Other Constructs
Relationships With Criteria
Consequences
Prior evidence (e.g., validated score interpretation/use in a similar context)

Figure 1. An integrated framework for argument-based validation.

stages. Stage 1 involves constructing the *interpretation/use argument* (IUA). This concerns the intended, theory-informed interpretations or uses of the test/instrument. Since the focus of research will vary, different types of inferences are required. For instance, while language testing often focuses on assessment-based inferences (e.g., generalization, extrapolation), in motivation research explanation inference is likely to be of particular relevance, that is inferences that articulate how the scale scores relate to the underlying construct. Regardless of the type, these inferences must be carefully specified in alignment with both the theory and the evidence (see Chapelle, 2021 for detailed instructions on how to build and combine complex chains of inferences). A basic argument structure (Toulmin, 2003) is illustrated in the figure: the data (or grounds, i.e., the scores), claim (i.e., the intended interpretation/use), warrant (i.e., justification for the claim), backing (i.e., supporting evidence), rebuttal (i.e., counterclaim). The IUA serves as the guide for the next stage.

Stage 2 focuses on conducting validation research. In this stage, evidence is collected to support the claim. Here we draw on the Standards for a comprehensive list of sources of evidence: content-oriented evidence (i.e., analysis of the instrument and its relevance to the construct being measured), evidence-based on response processes (i.e., theoretical or empirical evidence about the psychological processes or cognitive operations of the respondents), evidence based on internal structure (i.e., analysis of relationships among scale items or parts of instrument), evidence based on relations to other variables (i.e., analysis of relationships with other related variables), evidence based on relations to criteria (i.e., analysis of how the scores relate to criterion variables), and evidence-based on consequences of testing (i.e., evaluation of intended and unintended consequences of the test). The type and combination of evidence gathered will depend on the specific claims or (chain of) inferences to be validated (see Chapelle, 2021 for a list of evidence corresponding to various types of inferences), as well as practical considerations such as resources and feasibility (Purpura et al., 2015).

Stage 3 involves developing the validity argument: an integrated evaluative judgment that assesses how well the collected evidence supports or challenges the IUA. The backward arrows in the diagram represent the iterative nature of this process. If the validity argument does not adequately support the intended score interpretation or use, researchers may need to repeat the process. This could involve constructing a new interpretive argument, collecting additional or different types of evidence, or even revising the measurements or underlying theory. As evidence accumulates over time, secondary research/synthesis will be required to provide more informed guidance on refining the measurements and/or the theory. The figure also illustrates how previous cases of score interpretation and uses can serve as supporting evidence when constructing similar IUAs in future research. This iterative approach ensures that the validation process is: (a) cumulative, (b) responsive to new evidence, (c) continuously improving in its ability to support meaningful score interpretations and uses, and (d) supporting the refinement of measurements and theories in L2 motivation research.

An example application

To showcase the utility of this framework, we draw on Al-Hoorie, McClelland, et al. (2024) as an example of how argument-based validation might work in practice. The authors conducted two studies examining the validity of the ideal L2 self-construct. In Study 1, they experimentally manipulated ideal L2 self-items to explicitly refer to ability

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

beliefs and tested for discriminant validity across three countries. Both exploratory and confirmatory factor analysis suggested ideal L2 self and L2 ability beliefs were not distinct. In Study 2, the authors used cognitive interviewing to examine participants' thought processes when responding to ideal L2 self-items and found that responses to the ideal L2 self-scales were dominated by references to current ability beliefs.

These studies represent an excellent example of the type of validation efforts needed in the field. In conducting this work, and by incorporating evidence based on response processes, the authors moved beyond the conventional sole focus on the internal structure of the construct and its relationship with other variables. While Al-Hoorie, McClelland, et al. (2024) did not explicitly align their study with a formal validation framework, we would advocate the use of the argument-based approach to make it easier for future synthesis and cumulative work. By conducting studies in accordance with this approach, several advantages stand to be gained: (a) the specific inferences being drawn can be more clearly articulated, (b) the evidence that accumulates can be systematically evaluated, and (c) findings (validity arguments) can be situated within a broader validation program for L2MSS research.

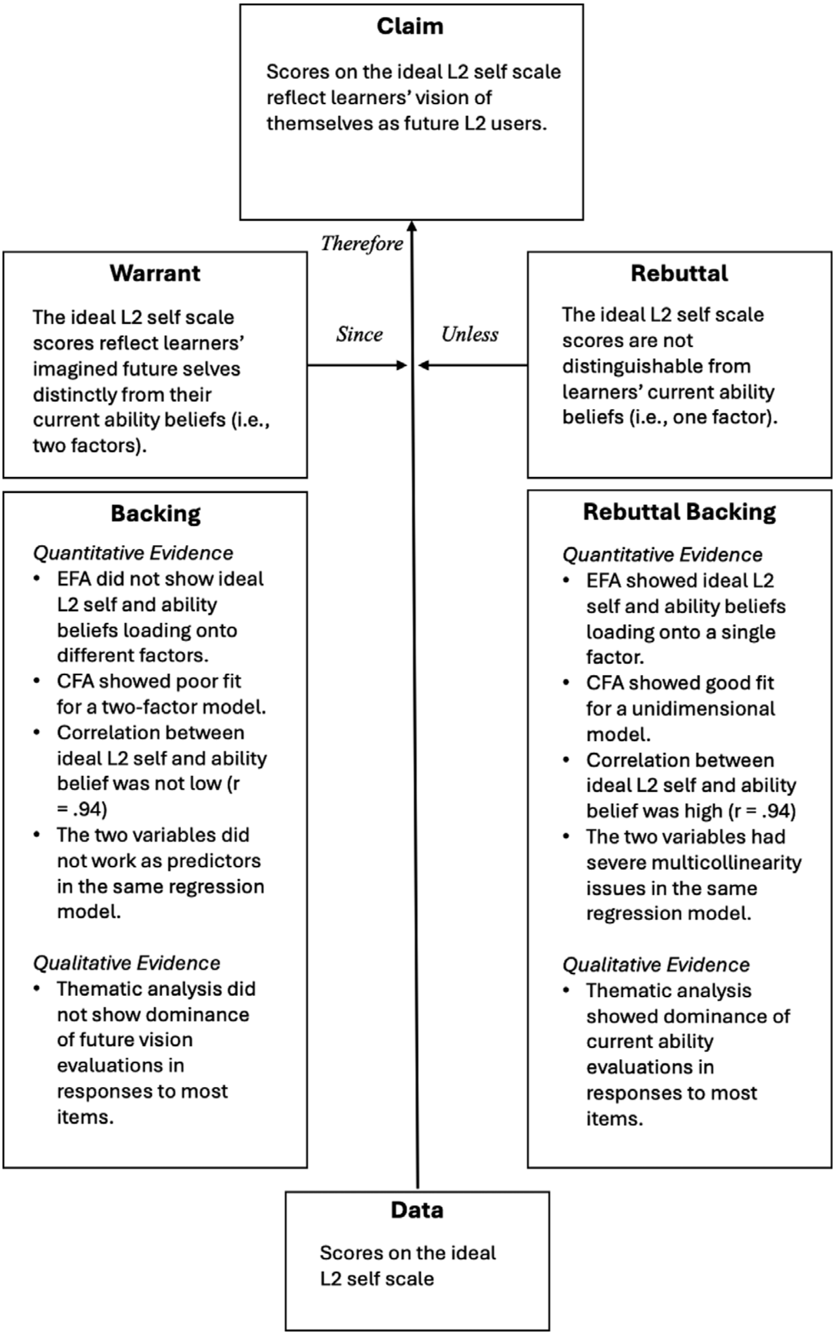
Applying the argument-based validation framework, a structure is provided for the central arguments of Al-Hoorie, McClelland, et al.'s research (Figure 2). The core of the IUA is that scores on the ideal L2 self-scale reflect the intended construct, i.e., learners' vision of themselves as future L2 users (Dörnyei, 2009). This claim is supported by the warrant that the scale scores reflect an imagined future L2 self that is distinct from beliefs involving current L2 abilities. However, the study also considers a potential rebuttal—that the ideal L2 self-scale scores are not empirically distinguishable from those of learners' current ability beliefs. It should be noted that while as an illustrative example, this IUA only focuses on one (explanation) inference, a comprehensive IUA typically involves a chain of interconnected inferences (Chapelle, 2021).

To evaluate this argument, the researchers collected both quantitative and qualitative evidence. They conducted factor analyses and regression analysis to supply evidence of the construct's internal structure and relationships with other variables, largely supporting the rebuttal. They also conducted cognitive interviews to gather evidence based on the response processes, which also predominantly aligned with the rebuttal.

Based on the evidence, we can construct the following validity argument: the intended interpretation of the ideal L2 self-scale scores as reflecting future visions is not adequately supported. The evidence collected suggests that the scale scores were not empirically distinguishable from those of current ability beliefs, thus challenging the intended interpretation of the ideal L2 self-scale scores. At this juncture, researchers will have two basic options. They can either modify the ideal L2 self-scale to better capture future visions, or they can reconsider how the construct can be conceptualized within L2MSS theory. Both would necessitate follow-up studies to validate new inferences.

From this application, we can see how the argument-based validation framework translates abstract validity concepts into a concrete/actionable steps. By providing a clear structure for articulating and evaluating validity claims, the framework "forces" us to think more intentionally, and to critically consider the inferences that we make from measurement scores. It also moves us beyond traditional psychometric analyses to consider multiple sources of evidence. Most tellingly, the approach can serve as a common language and methodology that would enable more systematic and programmatic validation efforts across the field. To facilitate its wider adoption, we have developed a free and open-access tool (https://validarg.netlify.app/) that makes it easier

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press



Claim

Scores on the ideal L2 self scale reflect learners' vision of themselves as future L2 users.

Warrant

The ideal L2 self scale scores reflect learners' imagined future selves distinctly from their current ability beliefs (i.e., two factors).

Rebuttal

The ideal L2 self scale scores are not distinguishable from learners' current ability beliefs (i.e., one factor).

Backing

Quantitative Evidence

- EFA did not show ideal L2 self and ability beliefs loading onto different factors.
- CFA showed poor fit for a two-factor model.
- Correlation between ideal L2 self and ability belief was not low (r = .94)
- The two variables did not work as predictors in the same regression model.

Qualitative Evidence

- Thematic analysis did not show dominance of future vision evaluations in responses to most items.

Rebuttal Backing

Quantitative Evidence

- EFA showed ideal L2 self and ability beliefs loading onto a single factor.
- CFA showed good fit for a unidimensional model.
- Correlation between ideal L2 self and ability belief was high (r = .94)
- The two variables had severe multicollinearity issues in the same regression model.

Qualitative Evidence

- Thematic analysis showed dominance of current ability evaluations in responses to most items.

Data

Scores on the ideal L2 self scale

Figure 2. Argument structure for Al-Hoorie, McClelland, et al. (2024).

for researchers to construct and visualize argument structures like the one illustrated in Figure 2. Ultimately, this approach can serve as a unifying framework for more rigorous and cumulative research, both for the L2MSS and beyond.

Looking forward

As L2 motivation researchers, we know how the will to learn and the skill to do so are equally crucial for successful language acquisition. The same principle applies to resolving the current crisis in L2MSS research. To move forward constructively, we need both psychological readiness and methodological maturity.

For psychological readiness, there is a need to reframe the "crisis" narrative, to value controversy, to normalize failure and (self-)correction, and to resist the allure of novelty. Methodologically, we suggest that an argument-based approach to validation can provide a promising direction. The integrated framework outlined in this article offers an anchor for systematic validation and for structured thinking about how validation is approached.

Moving forward, both individual local validation studies and field-wide syntheses of validity arguments will be crucial if the validation crisis is to be successfully navigated. Enhanced rigor is needed at multiple levels. Increased theoretical precision in construct definitions, more rigorous construct operationalization, and more systematic validation efforts, can each go some way toward resolving the controversies now plaguing L2MSS research. Finally, we believe that the validation crisis in L2MSS research can have field-wide implications. Efforts to address the crisis can form the foundations for a credibility revolution that can place L2 motivation research at the forefront of methodological rigor and constructive self-scrutiny in applied linguistics.

Acknowledgements. This research was conducted within the research program Transdisciplinary Approaches to Learning, Acquisition, Multilingualism (TEAM) funded by Riksbankens Jubileumsfond, grant number M23-0052. We gratefully acknowledge this support. We would also like to express our gratitude to Dan Isbell, Ali H. Al-Hoorie, and Phil Hiver for their valuable feedback on earlier versions of this manuscript.

References

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

Alexander, P. A. (2024). Hybridizing psychological theories: Weighing the ends against the means. *Educational Psychology Review*, *36*(1), 23. https://doi.org/10.1007/s10648-024-09856-3

Al-Hoorie, A. H., Cinaglia, C., Hiver, P., Huensch, A., Isbell, D. R., Leung, C., & Sudina, E. (2024). Open science: Considerations and issues for TESOL research. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3304

Al-Hoorie, A. H., Hiver, P., & In'nami, Y. (2024). The validation crisis in the L2 motivational self system tradition. *Studies in Second Language Acquisition*, *46*(2), 1-307–329. https://doi.org/10.1017/S0272263123000487

Al-Hoorie, A. H., McClelland, N., Resnik, P., Hiver, P., & Botes, E. (2024). The ideal L2 self versus ability beliefs: Are they really distinct? *Journal of Multilingual and Multicultural Development*, 1–19. https://doi.org/10.1080/01434632.2024.2401103

Bishop, D. V. M. (2018). Fallibility in Science: Responding to Errors in the Work of Oneself and Others. *Advances in Methods and Practices in Psychological Science*, *1*(3), 432–438. https://doi.org/10.1177/2515245918776632

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE.

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign LanguageTM*. Routledge.

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity Argument in Language Testing Case Studies of Validation Research*. Cambridge University Press.

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge, Taylor and Francis Group.

Cronbach, L. J. (1971). Test validation. *Educational Measurement.*

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9–42). Multilingual Matters.

Ellis, R. (2021). A short history of SLA: where have we come from and where are we going? *Language Teaching*, *54*(2), 190–205.

Henry, A., & Liu, M. (2024). Jingle–jangle fallacies in L2 motivational self system research: A response to Al-Hoorie et al. (2024). *Applied Linguistics*, amae041. https://doi.org/10.1093/applin/amae041

Higgins, E. T., Strauman, T., & Klein, R. (1986). Standards and the process of self-evaluation. *Handbook of Motivation and Cognition: Foundation of Social Behavior*, *1*, 23–63.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger and Greenwood Publishing.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

King, R. B., & Fryer, L. K. (2024). Hybridizing motivational strains: How integrative models are crucial for advancing motivation science. *Educational Psychology Review*, *36*(2), 38. https://doi.org/10.1007/s10648-024-09850-9

Liu, M. (2023). Whose open science are we talking about? From open science in psychology to open science in applied linguistics. *Language Teaching*, *56*(4), 443–450. https://doi.org/10.1017/S0261444823000307

Liu, M. (2024). Mapping the landscape of research on the L2 motivational self system theory (2005–2021): A bibliometric and text network analysis. *System*, *120*, 103180. https://doi.org/10.1016/j.system.2023.103180

Liu, M., Chong, S. W., Marsden, E., McManus, K., Morgan-Short, K., Al-Hoorie, A. H., …, & Hui, B. (2023). Open scholarship in applied linguistics: What, why, and how. *Language Teaching*, *56*(3), 432–437. https://doi.org/10.1017/S0261444822000349

Marsden, E., & Morgan-Short, K. (2023). (Why) are open research practices the future for the study of language learning? *Language Learning*, 1–44. https://doi.org/10.1111/lang.12568

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

Noels, K. A., Lou, N. M., Lascano, D. I. V., Chaffee, K. E., Dincer, A., Zhang, Y. S. D., & Zhang, X. (2019). Self-determination and motivated engagement in language learning. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *The Palgrave handbook of motivation for language learning* (pp. 95–115). Springer.

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.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., …, & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821114157

Oga-Baldwin, W. L. Q., (2024). *Validation crisitunity: A response to Al- Hoorie, Hiver, and In'nami (2024). Studies in Second Language Acquisition*, 1–13. https://doi.org/10.1017/S0272263124000597

Papi, M., Bondarenko, A. V., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research the 2 x 2 model of L2 self-guides. *Studies in Second Language Acquisition*, *41*(2), 337–361. https://doi.org/10.1017/S0272263118000153

Pekrun, R. (2024). Overcoming fragmentation in motivation science: Why, when, and how should we integrate theories? *Educational Psychology Review*, *36*(1), 27. https://doi.org/10.1007/s10648-024-09846-5

Plonsky, L. (2024a). *Open science in applied linguistics*. Applied Linguistics Press.

Plonsky, L. (2024b). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, 1–15.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics Research1. *Language Learning*, *65*(S1), 37–75. https://doi.org/10.1111/lang.12112

https://doi.org/10.1017/S0272263125000166 Published online by Cambridge University Press

Révész, A., & Brunfaut, T. (2020). Validating assessments for research purposes. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (1st ed., pp. 21–32). Routledge. https://doi.org/10.4324/9781351034784

Rohrer, J. M., Tierney, W., Uhlmann, E. L., DeBruine, L. M., Heyman, T., Jones, …, & Yarkoni, T. (2021). Putting the self in self-correction: Findings from the loss-of-confidence project. *Perspectives on Psychological Science*, *16*(6), 1255–1269. https://www.doi.org/10.1177/174569162096410

Ryan, R. M. (2024). Comments on integration, theory conflicts, and practical implementations: Some contrarian ideas for consideration. *Educational Psychology Review*, *36*(1), 16. https://doi.org/10.1007/s10648-024-09858-1

Teimouri, Y., Sudina, E., & Plonsky, L. (2021). What counts as evidence? In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (1st ed.). Routledge. https://doi.org/10.4324/9780429321498

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Cite this article: Liu, M., & Henry, A. (2025). Resolving the crisis in L2 motivational self-system research: Constructive dialogue and argument-based validation. *Studies in Second Language Acquisition,* 1–15. https://doi.org/10.1017/S0272263125000166