Studies in Second Language Acquisition (2024), **46**, 307–329
doi:10.1017/S0272263123000487

RESEARCH ARTICLE

The validation crisis in the L2 motivational self system tradition

Ali H. Al-Hoorie[1], Phil Hiver[2] and Yo In'nami[3]

[1]Royal Commission for Jubail and Yanbu, Saudi Arabia; [2]Florida State University, USA; [3]Chuo University, Japan
Corresponding author: Phil Hiver; Email: phiver@fsu.edu

(Received 28 February 2023; Revised 27 August 2023; Accepted 29 August 2023)

Abstract

Concerns have recently been raised about the validity of scales used in the L2 motivational self system tradition, particularly in relation to sufficient discriminant validity among some of its scales. These concerns highlight the need to systematically examine the validity of scales used in this tradition. In this study, we therefore compiled a list of 18 scales in widespread use and administered them to Korean learners of English (N = 384). Testing the factorial structure of these scales using multiple exploratory and confirmatory factor-analytic criteria revealed severe discriminant validity issues. For example, the ideal L2 self was not discriminant from linguistic self-confidence, suggesting that participant responses to such ideal L2 self items is not driven by actual–ideal discrepancies as previously presumed but more likely by self-efficacy beliefs. We discuss these results in the context of the need to encourage systematic psychometric validation research in the language motivation field.

> "What is the value of knowing each leg's length, after already knowing the other leg's length?"
> —McElreath (2020, p. 164)

In 1994, Dörnyei commented on the validity of the Attitude/Motivation Test Battery, which underlies the socioeducational model (Gardner, 1979, 1985, 2010). Dörnyei raised discriminant validity concerns regarding the three motivation variables in that model, arguing that "if we mixed the thirty items, it would be rather difficult to reconstruct the three scales" (Dörnyei, 1994, p. 518). He went on to suggest using factor analysis to refine the scales: "I suspect that an exploratory factor analysis of all items in the AMTB [Attitude/Motivation Test Battery] would not come up with the exact scales the test contains" (Dörnyei, 1994, p. 519; see also Dörnyei, 2005). In response, Gardner and Tremblay (1994) emphasized that these three components actually represent a single latent variable, "motivation" in the model, and thus this concern was unwarranted. They additionally explained that extensive research had been conducted on the construct, predictive, convergent, and discriminant validity aspects of variables in the socioeducational model. In their words, "All of the published research associated with the socio-educational model is concerned with the issue of construct validation" (Gardner & Tremblay, 1994, p. 252).

In the past 2 decades, another model of second language (L2) motivation has seen widespread, though often uncritical, adoption. The L2 motivational self system (L2MSS; Dörnyei, 2005, 2009) has, for some, become synonymous with L2 motivation and its instruments and data-elicitation scales isomorphic with learners' actual levels of motivation. The L2MSS has three primary elements: the ideal L2 self, the ought-to L2 self, and the L2 learning experience. The ideal L2 self, which reflects the actual–ideal discrepancy, has received the most attention. In contrast, the ought-to L2 self, concerned with the less internalized actual–ought discrepancy, did not receive equal enthusiasm from the start in that it "does not lend itself to obvious motivational practices" (Dörnyei, 2009, p. 32), ultimately yielding disappointing results (Al-Hoorie, 2018). The L2 learning experience has the least theoretical relevance to this "self"-based model, and its inclusion seems to be primarily to maintain correspondence to its historical antecedent, the socioeducational model (see Dörnyei, 2009, p. 30). Research adopting this model ranges from qualitative case study designs to large scale cross-sectional designs relying on quantitative analysis. The characteristics of this body of scholarship, along with its methodological and conceptual shortcomings, have been described and summarized in recent work taking stock of the past decades of L2 motivation research (Al-Hoorie, 2018; Al-Hoorie et al., 2021; Hiver & Al-Hoorie, 2020a). This work spotlights very similar concerns and indicates that criticisms leveled at the AMTB measurement could apply equally well to the L2MSS. Given the large, and still growing, body of work on the L2MSS, it is perhaps alarming that little work exists to validate its widely adopted measurement scales.

In line with the resurgence of interest in the psychometric validity of measures used in the L2 motivation field (e.g., Arndt, 2023; Sudina, 2021, 2023), the purpose of the present study was to investigate the validity of the scales used in the L2MSS tradition. Specifically, our purpose was to examine whether discriminant validity concerns, as those described above, would apply to scales associated with the L2MSS. We compiled a list of widely used scales and tested their factorial structure using multiple exploratory and confirmatory factor-analytic criteria.

The validation crisis

Validity, simply put, refers to whether a test actually measures what it is intended to measure (Kelley, 1927). Because tests and measures typically assess an intangible construct, such as attitudes or proficiency, it is important to examine whether these tests do so validly. The idea of construct validity, the umbrella term for validity, has evolved over time. Originally, Cronbach and Meehl (1955) introduced construct validity to refer to test scores as part of a nomological network and as being consistent with empirical and theoretical relations within that network. Messick (1989) subsequently extended construct validity to the justifiability of actions based on test score interpretation, not exclusively in relation to scientific evidence but also in relation to the social and ethical consequences of that interpretation. Advancing a more modern conception of validity, Borsboom et al. (2004) argued for a realist, causality-based interpretation where variation in the attribute under investigation causes variation in test scores. This conceptualization is akin to (reflective) latent variable modeling in that response to items is posited to be an effect of the respective latent variable(s). More

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

recently, Stone (2019) distinguished between construct validity and construct legitimacy, with the latter referring not to measurement but to theoretical clarity of the construct and its usefulness to the larger theory or model.

A validation crisis occurs when a discipline neglects the validity of its instruments, and as a result skepticism about a whole body of literature starts to spread. A validation crisis has been argued to be a precursor to a replication crisis, making even replicable results uninformative (Schimmack, 2021). If the instruments used to come up with a finding are questionable, this will naturally overshadow the credibility of that finding. This is because, without valid measurement, inferences become untrustworthy—a flaw that cannot be fixed by large sample sizes, rigorous designs, or advanced statistics. Besides neglecting the validity of its instruments, a discipline may experience a validation crisis when questionable measurement practices become widespread. Questionable measurement practices refer to researcher degrees of freedom that, intentionally or unintentionally, are nontransparently exploited to obtain a desired result (Flake & Fried, 2020). Examples of these questionable measurement practices include not justifying why a certain measure was selectively chosen when there are other options, not transparently acknowledging arguments against the validity of the chosen measure and thus implying that the measure is more valid than it actually is, not reporting the measure in full (e.g., giving only one example item), and not disclosing measure treatment in detail (e.g., coding, transformation, item removal) especially when this measure was developed on the fly.

One consequence of a validation crisis is the jingle and jangle fallacies (Kelley, 1927; Pedhazur & Schmelkin, 1991). The jingle fallacy is the belief that, because two or more measures have the same name, they must measure the same construct. In contrast, the jangle fallacy is the belief that, because the measures have different names, they must also refer to different constructs. The jangle fallacy—which is probably more relevant to our field—risks muddying theoretical clarity and introducing theoretical clutter due to the proliferation of abstract constructs that, in reality, all measure more or less the same thing (Al-Hoorie, 2018; Shaffer et al., 2016). One example is emotional intelligence, which has been argued to actually underlie well-established aspects of personality (Joseph et al., 2015) and when these aspects are controlled for, the effect of emotional intelligence disappears (van der Linden et al., 2017). One might assume that, in line with the recent methodological turn, research advances with more sophisticated analyses and more rigorous designs will provide the antidote to theoretical shortcomings of work in the field. After all, hypothesis testing and exploratory research can help to further develop incomplete theories. Ironically, however, robust theorizing along with coherent conceptual thinking is a prerequisite to accurate measurement. Proliferation of measures that actually underlie the same construct usually reflects the emergence of fads that a field latches on to uncritically coupled with the neglect to systematically validate measures before substantive research, culminating in a validation crisis.

Possible overlap among constructs that measures tap into may be investigated through the assessment of discriminant validity (Campbell & Fiske, 1959). Although the definition of discriminant validity has varied across studies, it can be grouped into four categories (Rönkkö & Cho, 2022). These categories pertain to the presence or extent of correlations (a) between constructs (e.g., how Constructs A and B are correlated), (b) between measures (e.g., how scores in Tests A and B are correlated), (c) between constructs and measures (e.g., how Construct A and a test score measuring Construct B is correlated), and (d) combinations of these categories. When two

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

constructs fail discriminant validity testing, three possible interpretations arise (Shaffer et al., 2016). One is that the constructs are empirically redundant. In this case, the multiplicity of constructs is spurious and clouds theoretical clarity, indicating the need to clean up the field’s landscape. A second interpretation is that the constructs are indeed distinct but share a causal relationship. Here, researchers should carefully consider the need for a construct that is entirely or almost entirely caused by another one. A final possibility is that the two constructs are unique but the measures developed and used to assess them are still unable to tap into their uniqueness. Recent advances have offered different techniques to address these discriminant validity concerns (Rönkkö & Cho, 2022).

Generally speaking, discriminant validity requires that two measures demonstrate a sufficiently low absolute correlation to regard them as representing two meaningfully distinct constructs (Rönkkö & Cho, 2022). According to Dörnyei (2007), “if two tests correlate with each other in the order of 0.6, we can say that they measure more or less the same thing” (p. 223). Although a correlation of .60 per se is not a hard and fast rule, the basic idea is that “scores having high correlations ‘measure the same thing’” (Cronbach, 1990, p. 372), whereas “if the correlations are low to moderate, this demonstrates that the measure has discriminant validity” (Carless, 2004, p. 272). That a high correlation constitutes a risk to discriminant validity has been repeatedly voiced by psychometricians (e.g., Kenny, 1976, p. 251; McDonald, 1985, p. 220; Nunnally & Bernstein, 1994, p. 93). The correlation referred to here is not the one typically reported (e.g., Pearson correlation coefficient) but the disattenuated or error-corrected correlation that removes the effect of unreliability. Thus, this “true” correlation is invariably *higher* than the one typically reported. On the other hand, reliability (e.g., coefficient α or ω) is not an appropriate measure for discriminant validity, or any other type of validity, besides internal consistency under the assumption of unidimensionality (Al-Hoorie & Vitta, 2019). If two measures fail to exhibit adequate discriminant validity, this suggests the need to “clean up” theoretical constructs to minimize empirical redundancy (Hiver & Al-Hoorie, 2020a; Shaffer et al., 2016).

In the second language acquisition field more broadly, although validity is already recognized as being “of central importance for the credibility of research results” (Chapelle, 2021, p. 11), language researchers have long lamented the status of validation in the field. Researchers have “largely ignored it, often happy to talk about acquisition with no consideration of the type of data they had collected” (Ellis, 2021, p. 197), and may simply “assume … the validity of whatever assessment method is adopted” (Norris & Ortega, 2012, p. 575). In fact, there seems to be “an unwritten agreement” that readers will accept instrument validity “at face value” (Cohen & Macaro, 2013, p. 133) without asking for explicit evidence for it. When it comes to language learning motivation specifically, the most popular instrument by far is Likert-type self-report questionnaires (Dörnyei & Ushioda, 2021). In a survey of motivation and anxiety questionnaire scales between 2009 and 2019, Sudina (2021) found that severe validity concerns exist in studies published in the field’s top five journals (*Applied Linguistics*, *Language Learning*, *The Modern Language Journal*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*). For example, less than half of motivation studies used factor analysis to investigate the psychometric properties of their scales and only 4% specifically examined discriminant validity. In addition, more than 90% of the studies did not provide readers with any validity information for the scales used. These findings are particularly concerning in light of the review we present in the next section.

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Validation in the L2MSS tradition

The above review shows that validity concerns are widespread both within second language acquisition and beyond. When it comes to the L2MSS, it is noteworthy that the term "validation" is sometimes used inappropriately to refer to whether the results support the theory itself instead of the robustness of the measures used (e.g., Dörnyei & Ryan, 2015, p. 91). This incorrect application of the term validation has been copied widely by the L2MSS literature. As explained above, validation is a procedure related to measurement. Theories may be proposed, tested, developed, challenged, and refuted based on falsifiable hypotheses they should put forward. A measure may or may not be valid depending on the extent to which the construct intended is the underlying cause of item covariation in that measure (DeVellis, 2017).

Scales in the L2MSS tradition have been variously described as exhibiting a high level of rigor. For example, Vlaeva and Dörnyei (2021) explained that they did not conduct a pilot because they relied on "established instruments widely used in the literature" (p. 952). You et al. (2016) similarly claimed that the scales they used were "tried and tested in previous studies" (p. 103). Such generic, referenceless statements are widespread in the L2MSS literature. Busse (2013) described the scales she used as "validated research instruments" (p. 382), Chan (2014a) as "taken from an established motivation inventory" (p. 363); Danesh and Shahnazari (2020) remark that the "validity of the questionnaire has been affirmed by different studies" (p. 4), and Du (2019) depicts the ideal L2 self as "validated as a key motivational source" (p. 135). Validation-related terminology in such instances is used rather loosely and does not mean that these scales have been subjected to rigorous psychometric analyses. The same applies to Dörnyei and Ryan's (2015) claim that "Virtually all the validation studies reported in the literature found the L2 motivation self system providing a good fit for the data" (p. 91). Generally, these so-called validation studies did not actually conduct any psychometric validation of scales used or include meaningful learning-related behaviors, apart from relating "one measure based on verbal report to another measure based on verbal report" (Gardner, 2010, p. 73)—introducing common-method bias into the entire empirical literature.

As an illustration of the absence of measurement validation, consider the study by Al-Shehri (2009). Al-Shehri reported high correlations between several L2MSS scales without investigating the validity of these scales or even reporting their reliabilities. Nevertheless, the author went on to argue that the pattern of the results "confirms" the role of the ideal L2 self and "proves" his hypothesis (Al-Shehri, 2009, p. 167).[1] Despite these obvious weaknesses, this study has been widely cited as "validating" the L2MSS in the Saudi context (e.g., Hessel, 2015; Muir, 2020; You & Chan, 2015; You & Dörnyei, 2016). (See also Waninge et al., 2014, for another illustration of the disregard of best measurement practices under the guise of complex dynamic systems theory [Hiver & Al-Hoorie, 2020b].)

In some instances, validation was attempted but clearly inappropriately so. For example, Dörnyei and Chan (2013) used a Visual Style scale that originally had had five items, but it showed a very poor reliability (exact reliability not declared). The authors therefore had to drop two items, but the remaining items still had a very low reliability of .49. Ideally, as this scale was adapted and applied to the L2 context for the first time, it should have been dropped and not used in any further inferential analyses. However,

[1] It is worth noting that it is inappropriate to describe findings as "confirming" or "proving" a theory, as results supporting a theory merely make it empirically adequate until future research refutes it (see Hiver & Al-Hoorie, 2020b, p. 253).

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

the authors decided to retain this scale, interpret its results as meaningful, and then argue that this very poor reliability "did produce important significant results" (Dörnyei & Chan, 2013, p. 448), and that the low reliability simply "restricted the scale's sensitivity" (p. 456). A more parsimonious explanation that the authors did not acknowledge is simply that the scale lacks adequate psychometric properties—and results based on it are consequently questionable. Another example of poor validation practices is applying factor analysis *after* completing inferential analysis. Validation should precede any inferential statistical analysis, not be relegated to the end of a study, as this would be tantamount to putting the horse behind the cart. In the study by Dörnyei and Chan (2013), the authors conducted exploratory factor analysis (on two scales only, and separately) after completing their analyses. The results also revealed problematic patterns including Heywood cases (loadings over 1.0) and scales that were clearly multidimensional. Both of these issues should have raised red flags about the scales. (Another example of inappropriately using exploratory factor analysis at the end of the study is Tseng et al., 2006.)

Since the first empirical anthology on the L2MSS (Dörnyei & Ushioda, 2009) appeared, a number of scales have been introduced to the field. These scales, however, did not undergo psychometric validation before they were extensively used to make substantive claims. Close examination of the wording of some of these scales reveals significant overlap. For example, in the context of learning L2 English, "*My future career is closely related to speaking English*" and "*My future goal needs English*" have a very similar underlying concept (almost a tautology) that it would be hard to imagine that these two items could belong to two distinct constructs. Similarly, "*If an English course was offered in the future, I would like to take it*" and "*I'm always looking forward to my English classes*" seem very closely related and placing them under separate scales would most likely lead to discriminant validity issues. Examples of such cases abound in the L2MSS tradition, with alarmingly high correlations almost the default finding (see Table 1).

As explained in the previous section, a high correlation raises red discriminant validity flags. The correlations reported in Table 1, which have yet to be corrected for disattenuation, are rather high—especially in light of the apparent overlap in items purported to belong to different scales. Indeed, these correlations seem too high to the extent that Dörnyei and associates have acknowledged that a correlation of this magnitude is "an exceptionally high figure" (Dörnyei, 2009, p. 31) and "an unusually high figure in motivation studies" (Dörnyei & Ryan, 2015, p. 91). Unsurprisingly, in studies where a factor-analytic procedure was implemented, the correlation typically drops significantly (Al-Hoorie, 2018). This suggests that researchers who conducted a factor-analytic procedure were able to locate problematic items (e.g., loading highly on two variables) and exclude them before inferential analyses. This in turn suggests that the high correlations reported in the literature might be an artifact of a lack of discriminant validity.

In a direct empirical test of these discriminant validity concerns, Hiver and Al-Hoorie (2020a) conducted a replication of You et al. (2016). You et al. (2016) conducted a large-scale study involving over 10,000 Chinese learners of English. A number of scales were used in that study, including the Ideal L2 Self, Vividness of Imagery, Ease of Using Imagery, and Positive Changes of the Future L2 Self-Image. The magnitude of the path coefficients in their structural equation model was rather high in some instances. For example, the Vividness of Imagery predicted the Ideal L2 Self at $\beta = .81$. In turn, Visual Style predicted Vividness of Imagery at $\beta = .67$, and Attitudes toward L2 Learning predicted Intended Effort at $\beta = .68$. Examination of the items used to

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Table 1. Examples of high correlations between observed variables involving items with apparent overlap

	Scale name (as used in the study)	Sample items	*r*
Al-Shehri (2009)	Ideal Self	Whatever I do in the future, I think I will be needing English.	.78
	Motivated Behavior and Effort	Learning English is one of the most important aspects in my life.	
Kim and Kim (2011)	Ideal L2 Self	I can imagine a time when I can speak English with native speakers from other countries.	.718
	Motivated L2 Behavior	If I had the opportunity to speak English outside of school, I would do it as much as I can.	
Kim (2009)	Ideal L2 Self	The things I want to do in the future require me to speak English.	.575
	Motivated Behavior	Learning English is very important in my life.	
Ryan (2009)	Intended Learning Effort	If an English course was offered in the future, I would like to take it.	.86
	Attitudes to Learning English	I'm always looking forward to my English classes.	
Papi (2010)	English Learning Experience	Would you like to have more English lessons at school?	.72
	Intended Effort	If an English course was offered in the future, I would like to take it.	
Yang and Kim (2011)	Ideal L2 Self	My future career is closely related to speaking English	.53–.70
	Motivated L2 Behavior	My future goal needs English	
Taguchi et al. (2009)	Ideal L2 Self	I can imagine myself studying in a university where all my courses are taught in English.	.44–.68
	Criterion Measures	If an English course was offered in the future, I would like to take it.	

Table 2. Items presumed to belong to different scales but appear to have significant wording overlap

Scale	Example item
Ideal L2 Self	I can imagine myself in the future giving an English speech successfully to the public in the future.
Vividness of Imagery	If I wish, I can imagine how I could successfully use English in the future so vividly that the images and/or sounds hold my attention as a good movie or story does.
Ease of Using Imagery	It is easy for me to imagine how I could successfully use English in the future.
Positive Changes of the Future L2 Self-Image	In the past I couldn't imagine of myself using English in the future, but now I do imagine it.

operationalize these constructs reveals a great deal of overlap (Table 2). In Hiver and Al-Hoorie's (2020a) replication, these scales in question failed to exhibit adequate discriminant validity, indicating that these items represent one latent variable. Thus, the high correlations obtained are more likely to be spurious due to lack of discriminant validity.

Table 3. Scales used in the present study

	Scale	Items	Source
Group 1	Intended Effort	10	(Taguchi et al., 2009)
	Attitudes to Learning English	5	
Group 2	Attitudes to L2 Community	4	(Taguchi et al., 2009)
	Cultural Interest	4	
Group 3	Auditory Style	5	(You et al., 2016)
	Visual Style	5	
Group 4	Feared L2 Self	4	(Chan, 2014b)
	Negative Changes of the Future L2 Self-image	2	(You et al., 2016)
Group 5	Ought-to L2 Self	10	(Taguchi et al., 2009)
	Instrumentality–Promotion	14	
	Instrumentality–Prevention	8	
	Family Influence	11	
Group 6	Ideal L2 Self	10	(Taguchi et al., 2009)
	Linguistic Self-Confidence	4	(Dörnyei, 2010)
	Vividness of Imagery	5	(You et al., 2016)
	Ease of Using Imagery	5	(You et al., 2016)
	Imagery Capacity	5	(Dörnyei & Chan, 2013)
	Positive Changes of the Future L2 Self-image	3	(You et al., 2016)

The present study

In light of the above literature review, the purpose of this study was to conduct an empirical investigation of the discriminant validity of a collection of scales used in the L2MSS tradition. We surveyed the literature and compiled a list of 18 scales used in this tradition. These scales were also included in Dörnyei’s (2010) authoritative guide to questionnaires in L2 research. We categorized these scales into six groups (see Table 3) based on construct similarity (see also Instruments later). The overall research question guiding the present study was whether and to what extent the scales in each group would show sufficient discriminant validity.

Method

Participants

The participants were 384 (female = 214) secondary school learners of English as a foreign language in Korean public schools. Participants ranged in age from 13–18 (M_{age} = 15.6, *SD* = 3.31) and all used Korean as their L1. Using quota sampling, we recruited a roughly equal number of students from Seoul and the most densely populated urban province immediately surrounding the capital (*n* = 204) and from schools in provinces located further to the south and center of the country (*n* = 180). These two different geographic regions represent divergent socioeconomic strata and levels of educational investment and competitiveness. The ratio of high school students sampled (*n* = 232) relative to middle school students (*n* = 152) was approximately 60:40. Typical of many other foreign language contexts, all these participants reported engaging in regular independent L2 study outside of the compulsory classroom setting; however, none indicated study-abroad experience in the L2. Ability levels, obtained from participants’ most recent secondary school standardized exam reports ranged from novice-high to intermediate-mid according to ACTFL proficiency guidelines.

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Instruments

As explained above, we used 18 scales that represent major constructs in the L2MSS tradition (see Table 3). We grouped the first element of the L2MSS, the Ideal L2 Self, along with vision- and imagery-related scales because of the close connection between the Ideal L2 Self and vision, with the theory sometimes referred to as *vision theory* (Muir & Dörnyei, 2013, p. 362; see also Al-Hoorie & Al Shlowiy, 2020). We also included Linguistic Self-Confidence in this group because the Ideal L2 Self Items do not refer to any actual–ideal discrepancy per se, so we hypothesized that ability beliefs might drive response to items in this scale. The second element of the L2MSS, the Ought-to L2 Self, was grouped with other scales related to external motivational forces. These include Family Influence, Promotional Instrumentality, and Preventive Instrumentality. The third element, the L2 Learning Experience (i.e., attitudes toward learning English) was grouped with intended effort (see Hiver & Al-Hoorie, 2020a). Other scales were likewise grouped based on thematic similarly: Attitudes Toward L2 Community with Cultural Interest, the Feared L2 Self with negative changes of the Future L2 Self-Image, Auditory Style with Visual Style. All items and their original Korean wording are listed in the online supplementary materials.

Procedure

The questionnaire items were translated into the students' L1 (Korean) by a researcher familiar with the principles of questionnaire construction and both languages in question. These 114 items across 18 scales were then back-translated by a professional translator for consistency. Once we obtained ethical approval, we approached school administration and teaching staff for institutional consent and participant assent to collect data. Students from the schools that agreed to participate completed the questionnaire outside of their regular class hours in the final weeks of the academic year. The regular L2 teacher and a research assistant were both present to inform participants about the purpose of the survey and administer it. Students were reminded that participation was voluntary and were assured their responses would remain confidential. Participation was voluntary and uncompensated. All data-elicitation measures were administered in person using the SurveyMonkey platform that randomized and intermixed the 114 items such that no two items from a single scale were adjacent or presented to any one respondent in the same order. Response rates ranged from 50%–60% across sites. Respondents were given 30 min to complete the scales but on average spent less than 20 min to complete their response to all items. Throughout data collection, participants were treated in accordance with APA ethical guidelines.

Data analysis

We conducted exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). For EFA, we set the number of factors to retain as a range from one to the maximum number of scales. For example, Group 1 had two scales: (1) Intended Effort and (2) Attitudes to Learning English. We examined whether one factor would be retained or two factors (based on the maximum number of the scales in Group 1). We retained factors (a) with eigenvalues greater than 1 computed from the current data, (b) with scree plots showing factors appearing before the point of inflexion, and (c) with parallel analysis using eigenvalues greater than the corresponding eigenvalues computed from many data sets randomly generated from the current data (e.g.,

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Widaman, 2012). We acknowledge that the eigenvalues-greater-than-1 rule risks overdimensionalization (van der Eijk & Rose, 2015), though, and we report it here for completeness and transparency. As parallel analysis is not available for categorical variables, we treated item responses as continuous and used the robust maximum-likelihood estimator.

For CFA, the data were multivariate nonnormal as judged by Mardia's multivariate normality test available in the MVN R package (Korkmaz et al., 2014), Mardia Skewness = 3396.713, $p < .001$, Mardia Kurtosis = 48.067, $p < .001$. Further, the data were categorical. For these reasons, we used the WLSMV estimator. We compared the fit of a CFA model with both scales under one latent variable and a model with each scale as a separate latent variable. The CFA models were considered to fit the data as judged by a comparative fit index (CFI) and a Tucker–Lewis index (TLI) of .90 or higher, a standardized root mean square residual of .08 or lower, and root mean square error of approximation (RMSEA) values of .08 or lower (Browne & Cudeck, 1993). The models were additionally compared using a chi-square difference test (Brown, 2015), though we note that this is a relative procedure showing which model fits the data better than providing an absolute indication of appropriateness of scale psychometric properties.

Throughout these processes, each model was evaluated individually as to whether it met these statistical criteria and was substantively interpretable. The analysis was implemented in Mplus 8.6 (Muthén & Muthén, 1998–2012), and there were no missing values. Supplementary materials are available from the OSF project page (https://osf.io/7c8qs/). They are also available on the IRIS Digital Repository.

Results

Table 4 shows the observed correlations among the scales (the 95% confidence intervals are available in the online supplementary materials). The table shows rather high correlations, in many cases exceeding .70. For the sake of convenience, the triangles in Table 4 indicate correlations between scales within one group. The weakest correlation was observed in Group 4, with a correlation of .41 between (7) Feared L2 Self and (8) Negative Changes of the Future L2 Self-Image. Conversely, the strongest correlation was observed in Group 5, with a correlation of .87 between (9) Ought-to L2 Self and (12) Family Influence. The magnitudes of these correlations aligned with medium (.40) and large (.60), respectively, according to Plonsky and Oswald's (2014) benchmarks for correlation magnitude. Although these are the observed correlations, their latent counterparts exceeded .90 (not reported here). All of this underscores the need to scrutinize the discriminant validity of variables underlying these scales.

In subsequent analyses, three scales had to be excluded due to a nonpositive definite error: Family Influence, Vividness of Imagery, and Ease of Using Imagery. Nonpositive definiteness refers to matrices that are unsuitable for analysis, for example, due to high correlations among variables (e.g., Wothke, 1993). Our data indeed showed extremely high latent correlations. For example, Family Influence exhibited a latent correlation of .97 with Ought-to L2 Self, whereas Vividness of Imagery and Ease of Using Imagery correlated with Ideal L2 Self at .94 and .96, respectively. With such high correlations, it is highly unlikely that these scales represent distinct constructs. These problematic patterns may not be particularly surprising for the latter two scales considering that they were first created by You et al. (2016) and did not undergo extensive psychometric

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Table 4. Observed variable correlations

G^{a}	S^{b}	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
1	1)	—																	
1	2)	**.84**	—																
2	3)	**.73**	.65	—															
2	4)	**.70**	.68	**.74**	—														
3	5)	**.73**	.67	.63	.54	—													
3	6)	.65	.60	.59	.55	**.71**	—												
4	7)	.62	.47	.44	.44	.54	.52	—											
4	8)	.20	.12	.15	.16	. 23	.33	.41	—										
5	9)	.61	.48	.43	.42	.48	.52	**.71**	.46	—									
5	10)	**.79**	.61	.62	.59	.66	.63	**.70**	.33	**.75**	—								
5	11)	.61	.43	.40	.39	.52	.53	**.72**	.44	**.84**	**.81**	—							
5	12)	.54	.44	.37	.38	.39	.46	.63	.44	**.87**	.66	**.74**	—						
6	13)	**.82**	**.74**	**.76**	**.72**	.66	.56	.55	.16	.54	**.70**	.50	.48	—					
6	14)	**.77**	**.70**	.67	.66	.64	.53	.45	.06	.43	.67	.47	.36	**.80**	—				
6	15)	**.77**	**.72**	.68	.64	.63	.60	.49	.17	.52	.60	.43	.49	**.86**	**.71**	—			
6	16)	**.75**	**.73**	.66	.68	.68	.62	.48	.13	.48	.61	.44	.43	**.86**	**.76**	**.87**	—		
6	17)	.69	.63	.66	.68	**.70**	**.71**	.54	.22	.49	.61	.47	.43	**.76**	.69	**.77**	**.82**	—	
6	18)	**.80**	**.73**	.62	.60	.62	.57	.62	.16	.58	.69	.55	.52	**.80**	**.70**	**.79**	**.75**	.67	—
	M	42	20	19	18	23	22	17	7	41	64	35	43	43	18	20	21	22	12
	SD	14	7	6	5	6	6	5	3	13	18	11	15	14	6	7	7	6	5
	α	.92	.88	.84	.77	.77	.78	.73	.54	.90	.93	.91	.91	.93	.83	.89	.84	.81	.84

Note. Triangles represent correlations within each group. G^{a} = Group; S^{b} = Scale. (1) Intended Effort; (2) Attitudes to Learning English; (3) Attitudes to L2 Community; (4) Cultural Interest; (5) Auditory Style; (6) Visual Style; (7) Feared L2 Self; (8) Negative Changes of the Future L2 Self-image; (9) Ought-to L2 Self; (10) Instrumentality–Promotion; (11) Instrumentality–Prevention; (12) Family Influence; (13) Ideal L2 Self; (14) Linguistic Self-Confidence; (15) Vividness of Imagery; (16) Ease of Using Imagery; (17) Imagery Capacity; (18) Positive Changes of the Future L2 Self-image. Means and standard deviations have been rounded to the nearest integer to save space. For instance, the mean value of (1) Intended Effort was 42.21, but was rounded to 42. Results prior to rounding can be found in the online supplementary material. Correlations of .70 and stronger are in bold.

Table 5. Exploratory factor analysis results

Group	No. of scales	Eigenvalue	Scree plot	Parallel analysis
1	2	1 factor	1 factor	1 factor
2	2	1 factor	1 factor	1 factor
3	2	2 factors	1 factor	1 factor
4	2	1 factor	1 factor	1 factor
5	3	4 factors	2 factors	2 factors
6	4	3 factors	1 factor	1 factor

Note. Group 1: Intended Effort and Attitudes to Learning English. Group 2: Attitudes to L2 Community and Cultural Interest. Group 3: Auditory Style and Visual Style. Group 4: Feared L2 Self and Negative Changes of the Future L2 Self-image. Group 5: Ought-to L2 Self, Instrumentality–Promotion, and Instrumentality–Prevention. Group 6: Ideal L2 Self, Linguistic Self-Confidence, Imagery Capacity, and Positive Changes of the Future L2 Self-image.

validation. Exclusion of these scales does not affect the main purpose of this study as these scales are not central to the L2MSS.

Table 5 presents the EFA results. The scree plot and parallel analysis always agreed, showing that—with the exception of Group 5—the results for all groups indicated only one meaningful underlying variable. For Group 5, there were two factors, which is less than the three expected factors. As expected, the eigenvalue-over-1 rule was frequently prone to overdimensionalization (van der Eijk & Rose, 2015). One surprising aspect in these results is that the Ideal L2 Self and Linguistic Self-Confidence (Group 6) did not turn out to be distinct factors. Additional examination using scree plot as well as parallel analysis simulations similarly supported one factor only (see Figure 1; scree plots for other groups are available in the online supplementary materials). As shown in Table 4, the observed correlation between the Ideal L2 Self and the Linguistic Self-Confidence scales was .80, whereas the latent correlation reached .90. These are problematically high magnitudes. This suggests that response to the Ideal L2 Self might be driven by belief in ability rather than an actual–ideal discrepancy. As for Group 5 (Ought-to L2 Self, Instrumentality–Promotion, and Instrumentality–Prevention), the results suggested two factors only. Table 6 presents the factor loadings for these two factors. The factor loading pattern suggests that the Ought-to L2 Self is not distinct from the

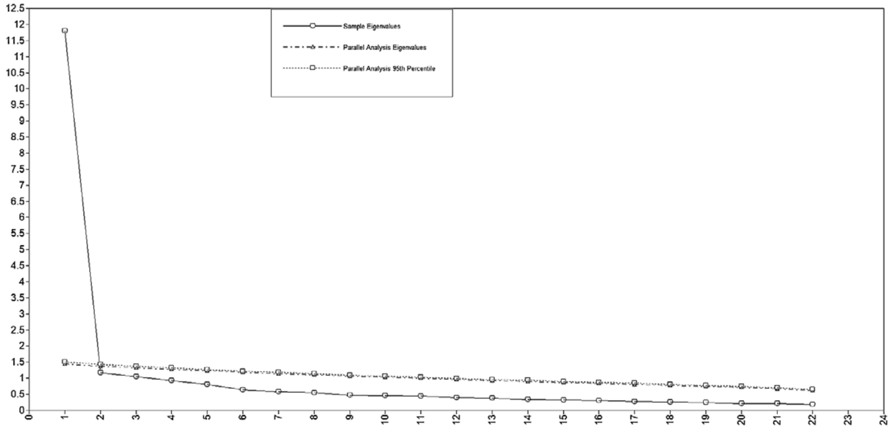

Figure 1. Scree plot of Group 6 (Ideal L2 Self, Linguistic Self-Confidence, Imagery Capacity, and Positive Changes of the Future L2 Self-Image). The results suggest that all these scales represent only one latent variable.

Table 6. Exploratory factor analysis for Group 5 (Ought-to L2 Self, Instrumentality–Promotion, and Instrumentality–Prevention)

Item	Factor 1	Factor 2
Ought6	**.92**	–.22
Ought7	**.79**	–.06
Prev7	**.75**	.01
Ought1	**.75**	–.13
Prev6	**.73**	.05
Ought2	**.70**	.07
Ought3	**.68**	.05
Prev8	**.64**	.20
Ought4	**.61**	–.06
Ought9	**.61**	.17
Prev4	**.57**	**.31**
Prom10	**.57**	–.13
Ought5	**.52**	**.31**
Prom12	**.49**	**.36**
Prev3	**.49**	**.35**
Prev1	**.46**	.28
Prev2	**.39**	**.40**
Ought8	**.39**	**.33**
Prom9	**.32**	.22
Prom1	–.03	**.84**
Prom4	–.01	**.82**
Prom7	–.13	**.82**
Prom14	.02	**.80**
Prom3	–.02	**.78**
Prom13	.01	**.72**
Prom6	.06	**.71**
Prev5	.12	**.68**
Prom5	.15	**.67**
Prom2	.18	**.65**
Prom11	.22	**.58**
Ought10	.24	**.39**
Prom8	.25	**.34**

Note. Geomin rotated loadings. Loadings over .30 in bold.

Instrumentality–Prevention, but it tends to be distinct from the Instrumentality–Promotion scale.

As Table 5 shows, there were cases where even the eigenvalue-over-1 rule suggested only one factor and agreed with the other two criteria. Intended Effort and Attitudes to Learning English (Group 1) turned out to represent only one factor using all three criteria, indicating that using these two scales in one study may lead to a jangle fallacy. The same applies to Attitudes to L2 Community and Cultural Interest (Group 2) and to Feared L2 Self and Negative Changes of the Future L2 Self-Image (Group 4).

Table 7 presents the CFA results. All models showed signs of problematic model fit. Specifically, the RMSEA values were larger than .08, most of the time even exceeding .10. In fact, even the *lower* confidence interval was usually larger than the standard .08 threshold. The CFI and TLI values were occasionally too low, especially for Group 5. Following Lai and Green (2016), who reported on the conditions under which CFI and RMSEA did not agree with each other, we examined whether and how our results satisfied those conditions. We found that the necessary condition was met in 12 of the 16 cases and that the sufficient condition was met in five of the 16 cases (see online supplementary material). Thus, the fit indices of all of these models were considered suboptimal.

Table 7. Model fit and chi-square different test results using confirmatory factor analysis results

Model	χ^2(*df*)	CFI	TLI	RMSEA [90% CI]	SRMR	χ^2 Diff. test (*df*)	Support
Group 1: Intended Effort and Attitudes to Learning English							
1 factor	912.024* (90)	.921	.908	**.154 [.145, .163]**	.042	61.166*** (1)	2 factors
2 factors	809.031* (89)	.931	.918	**.145 [.136, .154]**	.040		
Group 2: Attitudes to L2 Community and Cultural Interest							
1 factor	133.249* (20)	.968	.955	**.121 [.102, .141]**	.029	9.637*** (1)	2 factors
2 factors	106.874* (19)	.975	.963	**.110 [.090, .130]**	.027		
Group 3: Auditory Style and Visual Style							
1 factor	284.222* (35)	.915	**.891**	**.136 [.122, .151]**	.043	25.274*** (1)	2 factors
2 factors	253.546* (34)	.925	.901	**.130 [.115, .145]**	.040		
Group 4: Feared L2 Self and Negative Changes of the Future L2 Self-image							
1 factor	108.330* (9)	.899	**.831**	**.170 [.142, .199]**	.038	47.946*** (1)	2 factors
2 factors	37.678* (8)	.970	.943	**.098 [.068, .131]**	.023		
Group 5: Ought-to L2 Self, Instrumentality–Promotion, and Instrumentality–Prevention							
1 factor	3,166.454* (464)	**.862**	**.853**	**.123 [.119, .127]**	.070	271.375*** (3)	3 factors
3 factors	2,591.560* (461)	**.891**	**.883**	**.110 [.106, .114]**	.062		
Group 6: Ideal L2 Self, Linguistic Self-Confidence, Imagery Capacity, and Positive Changes of the Future L2 Self-image							
1 factor	1241.217* (209)	.935	.928	**.113 [.107, .120]**	.042	267.356*** (6)	4 factors
4 factors	922.180* (203)	.955	.949	**.096 [.091, .102]**	.036		

Note. Bold denotes problematic values. SRMR = standardized root mean square residual.
*$p < .05$;
***$p < .001$.

For model comparison, the results supported the model with separate factors. However, as explained previously, the chi-square test is a relative measure of fit and does not provide an absolute indication of psychometric properties (e.g., the better fitting model might still have poor psychometric properties). This might be promising, implying that—despite the poor model fit—future refinement of these scales has the potential to lead to more valid measures.

Discussion

The epigraph of this article is a quote from McElreath (2020, p. 164) asking about the value of knowing the length of a leg if we already know the length of the other leg. Obviously, there is little to be learned from knowing the length of the other leg. The same notion applies to measurement, variables, and inferential analysis. The question that linear multiple regression asks is this: What can we learn from a variable if we already know the value of another variable (or variables)? If these variables are highly correlated, each will explain little unique variance in the outcome variable (Al-Hoorie & Hiver, 2022). Clearly, designs such as those used in the L2MSS tradition rest on the assumption that the variables in question indeed underlie distinct latent variables. The results of this study suggest otherwise, showing that there is a severe case of a jangle fallacy in the L2MSS tradition. This is rather surprising considering that this model was first introduced almost 2 decades ago. This indicates a clear neglect of psychometrics and scale validation in this tradition. It also suggests the need to pause substantive studies until issues with validity can be ironed out. Results are only as good as a model's measures are.

Revisiting our results

Virtually all scales we examined in this study showed problematic psychometric properties. Further, none of the three main elements of the L2MSS (Ideal L2 Self, Ought-to L2 Self, and L2 Learning Experience) exhibited adequate discriminant validity when tested alongside related scales. This is concerning.

One interesting finding from this study is the lack of discriminant validity between arguably the crown jewel of this model, the Ideal L2 Self, and linguistic self-confidence. Linguistic self-confidence, ability beliefs, self-efficacy, and related concepts have been around for decades (Goetze & Driver, 2022) and, in theory, should have little to do with actual–ideal discrepancies. In fact, Higgins (1987, p. 336) speculated whether self-efficacy beliefs might moderate the effect of self-discrepancies (whether ideal or ought), a clear acknowledgment of the theoretical distinction between these constructs.

As for the second component of the L2MSS, the Ought-to L2 Self, it also failed to demonstrate discriminant validity in relation to Instrumentality–Prevention or Family Influence. On one hand, it makes sense for the Ought-to L2 Self to be related to a prevention focus, but these results suggest that treating these two scales as separate is problematic. On the other hand, although it also makes sense for the Ought-to L2 Self to be related to family influence, the extremely high latent correlation between them (.97) suggests that—just like the Ideal L2 Self—the Ought-to L2 Self does not represent actual–ought discrepancies per se as originally proposed. This is in turn part of the jangle fallacy where multiple scales with different names refer basically to the same thing.

When it comes to the third component of the L2MSS, again, there was little evidence that it was distinct from Intended Effort. Intended Effort is erroneously yet commonly used as the sole dependent variable in substantive research (Henry, 2021; Papi & Hiver, 2022) while the L2 Learning Experience is perhaps the least theorized construct in the model (Hiver et al., 2019) despite 2 decades of intense interest in the L2MSS. Claims that the L2 learning experience is the best predictor of intended effort may therefore be explained away by the lack of discriminant validity between the scales used.

Nor was there a clear empirical distinction between Cultural Interest and Attitudes Toward the L2 Community, again despite clear links to 6 decades of work by Robert Gardner and the ostensibly desire to move from integrative motivation to a more global outlook (Al-Hoorie & MacIntyre, 2020). The same applies to the various still unvalidated imagery-related constructs whose theoretical foundations have been problematized in motivational science for decades (see Baumeister et al., 2016; Oettingen & Reininger, 2016; Oettingen et al., 2018).

Recent work has attempted to revise the L2MSS model with information about the sources of these self-related perceptions and images in order to achieve greater

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

conceptual clarity and measurement accuracy (e.g., see Papi et al., 2019; Papi & Khajavy, 2021). As we elaborate below, our results suggest that this work does not go far enough, and more fundamental thinking is needed that revisits the validity of the ideal L2 self construct itself and the original eclectic theoretical base underlying it (Henry & Cliffordson, 2017). What is the theoretical rationale for the conflation of self-discrepancy theory with possible selves theory? What does this self construct comprise (i.e., over and above other self-related constructs)? What, if anything, makes it specific to additional language use and learning? What does an actual–ideal discrepancy entail in concrete terms? Do multilingual learners have these, and if so in which domains or in reference to what? These and many other important questions remain unanswered (Henry, 2023).

In more practical terms, the ideal L2 self is conceptually claimed to represent an actual–ideal discrepancy that the learner holds and consequently wants to bridge (Dörnyei, 2005, 2009). As Al-Hoorie (2018) argued, the actual items of this scale do not reflect any actual–ideal discrepancy as they explicitly refer to imagining oneself in the future using the language competently, leading him to suggest relabeling it to the *imagined self*. Our results suggest that response to Ideal L2 Self items might be driven by ability beliefs. Indeed, for an item like "*I can imagine a situation where I am speaking English with foreigners*," it is not clear why a respondent should first think of their current state, then their future state, and subsequently the gap between the two and finally respond to this item by saying "strongly agree" for example. It is more likely that the potential respondent will instead think about their ability and their belief in the extent to which they can achieve the task in question. This suggests that the ideal L2 self suffers not just from a discriminant validity problem but also a content validity one. Items in this scale are simply measuring ability. If this is the case, then findings from the ideal L2 self should probably be attributed to self-efficacy and related ability belief constructs instead. Even if the point of reference is a vision that is anchored to an achievement of a specific task, this is likely to not be qualitatively different from a goal (Al-Hoorie & Al Shlowiy, 2020), and such achievement-oriented perceptions have been studied under the rubric of self-efficacy for many decades (Bandura, 1986, 1997; Goetze & Driver, 2022).

In parallel with the above, what does it mean that the results did not show evidence that the Intended Effort scale is discriminant from Attitudes Toward Learning the L2? Looking at the actual items, it would be hard to argue that "*I would like to study English even if I were not required*" (intended effort) and "*I really enjoy learning English*" (attitudes toward learning the L2) should represent distinct latent variables. If a learner reports that they would like to expend voluntary effort in language learning despite it not being required, it would sound very strange if simultaneously they do not enjoy that class. These seem like two sides of the same coin or sharing a consistent causal relationship (Shaffer et al., 2016). The problematic validity (both discriminant and predictive) of intended effort and its ambiguity in models of L2 motivation are well established (see Al-Hoorie, 2018, for one review). In fact, this lack of discriminant validity replicates a previous report by Hiver and Al-Hoorie (2020a), who failed to replicate You et al. (2016).

Implications for the field

Part of the problem our results revealed lies in the exclusive reliance in the literature on subjective self-report measures, particularly Likert-type items (Al-Hoorie, 2018). For

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

historical reasons (Al-Hoorie, 2017) and for convenience (Al-Hoorie et al., 2021), the L2 motivation field generally, and the L2MSS tradition more specifically, has overemphasized self-report measures of variables that, as a baseline, are conceptually hazy. Instead of rating statements or even specific attributes, in self-discrepancy research participants are asked to *list* attributes they think they actually, ideally, and ought to possess. Researchers then have to compare these attributes and determine whether there are matches and mismatches (Higgins et al., 1985). The problem of overreliance on subjective self-report has extended to outcome measures, where these too for convenience are measured with a quick self-report scale (i.e., Intended Effort, Motivated Behavior). This dependence on proxy measures for important outcomes may not reflect learners' actual achievement ($r = .12$; Al-Hoorie, 2018) or tangible engagement (see Hiver, Al-Hoorie, & Mercer, 2021; Hiver, Al-Hoorie, et al., 2021).

Furthermore, validation research should not be limited to conventional correlational approaches. As construct validity essentially entails causality (Borsboom et al., 2004), research should investigate whether and to what extent measures are responsive to the manipulation of intended constructs. For example, if a measure is intended to assess state anxiety, it is expected that response to this measure should change depending on the specific conditions in which the instrument is administered, such as when respondents are placed in anxiety-inducing conditions (Gregersen et al., 2014). As it turns out, the picture might not be straightforward and individual differences may play a role (Gardner et al., 1992).

Another consideration is the usefulness of the construct for broader theoretical understanding in the field. Construct legitimacy (Stone, 2019) encourages researchers to consider the risk of conceptual clutter when too many overlapping constructs are introduced to the field. In some cases, it is possible to reduce this clutter by focusing on a smaller, more parsimonious set of "core" variables. Existing variables may sometimes be hardly distinguishable at a theoretical level from newly introduced ones even if the new ones are measured and conceptualized somewhat differently. An example is the distinction, or lack thereof, between a goal and a vision (Al-Hoorie & Al Shlowiy, 2020). Curbing this atheoretic proliferation may help researchers pay closer attention to the dynamic relationships among these constructs and factors influencing these relationships.

Regardless of the validation paradigm deemed fit for a specific area of inquiry, there is a need to place validation research on the map of applied linguistics research. After all, "validation research is research" (Borsboom et al., 2004, p. 1063), making valuable contributions to the rigor and validity of substantive research. When such validation research is recognized as a legitimate area of investigation in its own right, it should ideally be separate, ongoing, and a precursor to substantive research using the measures in question. This will help the field move beyond the mire of a repetitive exploratory research stage and the nefarious researcher degrees of freedom and other questionable measurement practices accompanying it. It will additionally minimize the need to create measures on the fly for substantive purposes as well as the dubious practice of unsystematically selecting a subset of existing scales merely for practical reasons (Claro, 2020).

Our study is not without limitations. Our sample is limited to participants in one specific learning context, sharing one L1, and having a rather narrow age range. Nevertheless, we do not anticipate these sample characteristics to be a major factor in our results given the theoretical complications, conceptual ambiguities, and previous replications showing similar results. Readers may also question our scale-grouping choices or point out other scales we did not include in this study. These concerns may

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

have some merits, and in this study we had to make strategic decisions to keep the analysis manageable and to minimize capitalizing on chance. We see this study as one step toward recognizing the validation crisis, opening a discussion around it, and eventually hopefully overcoming it.

Conclusion

Just like Dörnyei (1994) argued in relation to the Attitude/Motivation Test Battery, if we were mix items in the L2MSS, it would be rather difficult to reconstruct their scales using factor-analytic procedures. This applies to all three elements of the L2MSS including its uncontested crown jewel, the Ideal L2 Self. These findings, coupled with sundry other validity limitations and concerns even in top-tier journals in the field (e.g., Sudina, 2021, 2023), justifies declaring a state of validation crisis.

From a psychometrics perspective, replicability is not enough (Schimmack, 2021). Empirical demonstration of measure validity must be proffered, as “rationalization is not construct validation” (Cronbach & Meehl, 1955, p. 291). When measures turn out to be problematic afterward, this realization will inevitably overshadow the legitimacy of findings generated using these measures. Our results indicate that scales used in the L2MSS suffer from severe discriminant validity concerns. Future research should therefore prioritize addressing this validation crisis before undertaking further substantive research. Still, we caution against running down the rabbit hole of more and more “selves,” as “the multitude of overlapping concepts in the literature on the self is more confusing than integrativeness ever could be” (MacIntyre et al., 2009, p. 54). A focus on the self rather than on identity, which has a long and established tradition, represents a regrettable, costly detour the field has taken.

Another future direction is revisiting findings from studies that employed scales with suspect discriminant validity. Once a study is published and becomes widespread, it may be a slow and uphill process to promote research showing contrary results, or even publish it, in light of the general atmosphere in academia that is bias toward “innovative” and newsworthy research. Nevertheless, based on our experience and available evidence, it appears that channeling efforts and resources into self-determination theory (Al-Hoorie et al., 2022; Oga-Baldwin et al., 2019, 2022) might present more promising research avenues.

Finally, just like other disciplines in applied linguistics, it is important for validation research to embrace transparency. Researchers should be transparent about the limitations of their measures and consider the extent to which these limitations might have affected their results. It is equally important to reflect such limitations in any claims and recommendations stemming from findings based on these measures. Furthermore, it is imperative that researchers engaged in validation research adopt open science practices. This includes replication, sharing data and code, and preregistration. This research should also recruit samples representative of the language learner population. Doing so will ensure the rigor and quality of research findings and their relevance to policy and practice.

Data availability statement. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials and data are available at https://osf.io/7c8qs/.

References

Al-Hoorie, A. H. (2017). Sixty years of language motivation research: Looking back and looking forward. *SAGE Open*, *7*, 1–11. https://doi.org/10.1177/2158244017701976

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*, 721–754. https://doi.org/10.14746/ssllt.2018.8.4.2

Al-Hoorie, A. H., & Al Shlowiy, A. S. (2020). Vision theory vs. goal-setting theory: A critical analysis. *Porta Linguarum*, *33*, 217–229.

Al-Hoorie, A. H., & Hiver, P. (2022). Complexity theory: From metaphors to methodological advances. In A. H. Al-Hoorie & F. Szabó (Eds.), *Researching language learning motivation: A concise guide* (pp. 175–184). Bloomsbury.

Al-Hoorie, A. H., Hiver, P., Kim, T.-Y., & De Costa, P. I. (2021). The identity crisis in language motivation research. *Journal of Language and Social Psychology*, *40*, 136–153. https://doi.org/10.1177/0261927x20964507

Al-Hoorie, A. H., & MacIntyre, P. D. (Eds.). (2020). *Contemporary language motivation theory: 60 years since Gardner and Lambert (1959).* Multilingual Matters.

Al-Hoorie, A. H., Oga-Baldwin, W. L. Q., Hiver, P., & Vitta, J. P. (2022). Self-determination mini-theories in second language learning: A systematic review of three decades of research. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688221102686

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. https://doi.org/10.1177/1362168818767191

Al-Shehri, A. S. (2009). Motivation and vision: The relation between the ideal L2 self, imagination and visual style. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, Language Identity and the L2 Self* (pp. 164–171). Multilingual Matters.

Arndt, H. L. (2023). Construction and validation of a questionnaire to study engagement in informal second language learning. *Studies in Second Language Acquisition.* Advance online publication. https://doi.org/10.1017/s0272263122000572

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* Freeman.

Baumeister, R. F., Vohs, K. D., & Oettingen, G. (2016). Pragmatic prospection: How and why people think about the future. *Review of General Psychology*, *20*, 3–16. https://doi.org/10.1037/gpr0000060

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

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 136–162). SAGE.

Busse, V. (2013). An exploration of motivation and self-beliefs of first year students of German. *System*, *41*, 379–398. https://doi.org/10.1016/j.system.2013.03.007

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. https://doi.org/10.1037/h0046016

Carless, S. A. (2004). Discriminant validity. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The SAGE encyclopedia of social science research methods* (pp. 272). SAGE.

Chan, L. (2014a). Effects of an imagery training strategy on Chinese university students' possible second language selves and learning experiences. In K. Csizér & M. Magid (Eds.), *The impact of self-concept on language learning* (pp. 357–376). Multilingual Matters.

Chan, L. (2014b). *Possible selves, vision, and dynamic systems theory in second language learning and teaching* [Unpublished doctoral dissertation]. University of Nottingham, UK.

Chapelle, C. A. (2021). Validity in language assessement. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

Claro, J. (2020). Identification with external and internal referents: Integrativeness and the ideal L2 self. In A. H. Al-Hoorie & P. MacIntyre (Eds.), Contemporary language motivation theory: 60 years since Gardner and Lambert (*1959*) (pp. 233–261). Multilingual Matters.

Cohen, A. D., & Macaro, E. (2013). Research methods in second language acquisition. In E. Macaro (Ed.), *The Bloomsbury companion to second language acquisition* (pp. 107–136). Bloomsbury.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). HarperCollins.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Danesh, J., & Shahnazari, M. (2020). A structural relationship model for resilience, L2 learning motivation, and L2 proficiency at different proficiency levels. *Learning and Motivation*, *72*, Article 101636. https://doi.org/10.1016/j.lmot.2020.101636

DeVellis, R. F. (2017). *Scale development: Theory and applications*. SAGE.

Dörnyei, Z. (1994). Understanding L2 motivation: On with the challenge! *The Modern Language Journal*, *78*, 515–523. https://doi.org/10.1111/j.1540-4781.1994.tb02071.x

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

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.

Dörnyei, Z. (2010). *Questionnaires in second language research: construction, administration, and processing* (2nd ed.). Routledge.

Dörnyei, Z., & Chan, L. (2013). Motivation and vision: An analysis of future L2 self images, sensory styles, and imagery capacity across two target languages. *Language Learning*, *63*, 437–462. https://doi.org/10.1111/lang.12005

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

Dörnyei, Z., & Ushioda, E. (Eds.). (2009). *Motivation, language identity and the L2 self*. Multilingual Matters.

Dörnyei, Z., & Ushioda, E. (2021). *Teaching and researching motivation* (3rd ed.). Routledge.

Du, X. (2019). The impact of semester-abroad experiences on post-sojourn L2 motivation. *Studies in Second Language Learning and Teaching*, *9*, 117–155. https://doi.org/10.14746/ssllt.2019.9.1.6

Ellis, R. (2021). A short history of SLA: Where have we come from and where are we going? *Language Teaching*, *54*, 190–205. https://doi.org/10.1017/s0261444820000038

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*, 456–465. https://doi.org/10.1177/2515245920952393

Gardner, R. C. (1979). Social psychological aspects of second language acquisition. In H. Giles & R. N. St. Clair (Eds.), *Language and social psychology* (pp. 193–220). Blackwell.

Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. Edward Arnold.

Gardner, R. C. (2010). *Motivation and second language acquisition: The socio-educational model*. Peter Lang.

Gardner, R. C., Day, J. B., & Maclntyre, P. D. (1992). Integrative motivation, induced anxiety, and language learning in a controlled environment. *Studies in Second Language Acquisition*, *14*, 197–214. https://doi.org/10.1017/S0272263100010822

Gardner, R. C., & Tremblay, P. F. (1994). On motivation: Measurement and conceptual considerations. *The Modern Language Journal*, *78*, 524–527. https://doi.org/10.1111/j.1540-4781.1994.tb02073.x

Goetze, J., & Driver, M. (2022). Is learning really just believing? A meta-analysis of self-efficacy and achievement in SLA. *Studies in Second Language Learning and Teaching*, *12*, 233–259. https://doi.org/10.14746/ssllt.2022.12.2.4

Gregersen, T., Macintyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic case studies of learners’ foreign language anxiety. *The Modern Language Journal*, *98*, 574–588. https://doi.org/10.1111/modl.12084

Henry, A. (2021). Motivational connections in language classrooms: A research agenda. *Language Teaching*, *54*, 221–235. https://doi.org/10.1017/s0261444820000026

Henry, A. (2023). Multilingualism and persistence in multiple language learning. *The Modern Language Journal*, *107*, 183–201. https://doi.org/10.1111/modl.12826

Henry, A., & Cliffordson, C. (2017). The impact of out-of-school factors on motivation to learn English: Self-discrepancies, beliefs, and experiences of self-authenticity. *Applied Linguistics*, *38*, 713–736. https://doi.org/10.1093/applin/amv060

Hessel, G. (2015). From vision to action: Inquiring into the conditions for the motivational capacity of ideal second language selves. *System*, *52*, 103–114. https://doi.org/10.1016/j.system.2015.05.008

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*, 319–340. https://doi.org/10.1037/0033-295x.94.3.319

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Higgins, E. T., Klein, R. A., & Strauman, T. (1985). Self-concept discrepancy theory: A psychological model for distinguishing among different aspects of depression and anxiety. *Social Cognition*, *3*, 51–76. https://doi.org/10.1521/soco.1985.3.1.51

Hiver, P., & Al-Hoorie, A. H. (2020a). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*, 48–102. https://doi.org/10.1111/lang.12371

Hiver, P., & Al-Hoorie, A. H. (2020b). *Research methods for complexity theory in applied linguistics*. Multilingual Matters.

Hiver, P., Al-Hoorie, A. H., & Mercer, S. (Eds.). (2021). *Student engagement in the language classroom*. Multilingual Matters.

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2021). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688211001289

Hiver, P., Obando, G., Sang, Y., Tahmouresi, S., Zhou, A., & Zhou, Y. (2019). Reframing the L2 learning experience as narrative reconstructions of classroom learning. *Studies in Second Language Learning and Teaching*, *9*, 83–116. https://doi.org/10.14746/ssllt.2019.9.1.5

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology*, *100*, 298–342. https://doi.org/10.1037/a0037681

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, *12*, 247–252. https://doi.org/10.1016/0022-1031(76)90055-X

Kim, T.-Y. (2009). Korean elementary school students' perceptual learning style, ideal L2 self, and motivated behavior. *Korean Journal of English Language and Linguistics*, *9*, 461–486.

Kim, Y.-K., & Kim, T.-Y. (2011). The effect of Korean secondary school students' perceptual learning styles and ideal L2 self on motivated L2 behavior and English proficiency. *Korean Journal of English Language and Linguistics*, *11*, 21–42.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, *6*, 151–162. https://doi.org/10.32614/RJ-2014-031

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*, 220–239. https://doi.org/10.1080/00273171.2015.1134306

MacIntyre, P. D., Mackinnon, S. P., & Clément, R. (2009). The baby, the bathwater, and the future of language learning motivation research. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 43–65). Multilingual Matters.

McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11. https://doi.org/10.3102/0013189x018002005

Muir, C. (2020). *Directed motivational currents and language education: Exploring implications for pedagogy*. Multilingual Matters.

Muir, C., & Dörnyei, Z. (2013). Directed motivational currents: Using vision to create effective motivational pathways. *Studies in Second Language Learning and Teaching*, *3*, 357–375.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Oettingen, G., & Reininger, K. M. (2016). The power of prospection: Mental contrasting and behavior change. *Social and Personality Psychology Compass*, *10*, 591–604. https://doi.org/10.1111/spc3.12271

Oettingen, G., Sevincer, A. T., & Gollwitzer, P. M. (Eds.). (2018). *The psychology of thinking about the future*. Guilford Press.

Oga-Baldwin, W. L. Q., Fryer, L. K., & Larson-Hall, J. (2019). The critical role of the individual in language education: New directions from the learning sciences. *System*, *86*, Article 102118. https://doi.org/10.1016/j.system.2019.102118

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press

Oga-Baldwin, W. L. Q., Parrish, A., & Noels, K. (2022). Taking root: Self-determination in language education. *Journal for the Psychology of Language Learning*, *4*, 1–7. https://doi.org/10.52598/jpll/4/1/9

Papi, M. (2010). The L2 motivational self system, L2 anxiety, and motivated behavior: A structural equation modeling approach. *System*, *38*, 467–479. https://doi.org/10.1016/j.system.2010.06.011

Papi, M., Bondarenko, A., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research: The 2 × 2 model of L2 self-guides. *Studies in Second Language Acquisition*, *41*, 337–361. https://doi.org/10.1017/s0272263118000153

Papi, M., & Hiver, P. (2022). Motivation. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 113–127). Routledge.

Papi, M., & Khajavy, G. H. (2021). Motivational mechanisms underlying second language achievement: A regulatory focus perspective. *Language Learning*, *71*, 537–572. https://doi.org/10.1111/lang.12443

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Taylor & Francis.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*, 6–14. https://doi.org/10.1177/1094428120968614

Ryan, S. (2009). Self and identity in L2 motivation in Japan: The ideal L2 self and Japanese learners of English. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 120–143). Multilingual Matters.

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, *5*, Article 1645. https://doi.org/10.15626/MP.2019.1645

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, *19*, 80–110. https://doi.org/10.1177/1094428115598239

Stone, C. (2019). A defense and definition of construct validity in psychology. *Philosophy of Science*, *86*, 1250–1261. https://doi.org/10.1086/705567

Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning*, *71*, 1149–1193. https://doi.org/10.1111/lang.12468

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/s0272263122000560

Taguchi, T., Magid, M., & Papi, M. (2009). The L2 motivational self system among Japanese, Chinese and Iranian learners of English: A comparative study. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 66–97). Multilingual Matters.

Tseng, W.-T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary Acquisition. *Applied Linguistics*, *27*, 78–102. https://doi.org/10.1093/applin/ami046

van der Eijk, C., & Rose, J. (2015). Risky business: factor analysis of survey data: Assessing the probability of incorrect dimensionalisation. *PLoS One*, *10*, Article e0118900. https://doi.org/10.1371/journal.pone.0118900

van der Linden, D., Pekaar, K. A., Bakker, A. B., Schermer, J. A., Vernon, P. A., Dunkel, C. S., & Petrides, K. V. (2017). Overlap between the general factor of personality and emotional intelligence: A meta-analysis. *Psychological Bulletin*, *143*, 36–52. https://doi.org/10.1037/bul0000078

Vlaeva, D., & Dörnyei, Z. (2021). Vision enhancement and language learning: A critical analysis of vision-building in an English for Academic Purposes programme. *Language Teaching Research*, *25*, 946–971. https://doi.org/10.1177/13621688211014551

Waninge, F., Dörnyei, Z., & de Bot, K. (2014). Motivational dynamics in language learning: Change, stability, and context. *The Modern Language Journal*, *98*, 704–723. https://doi.org/10.1111/modl.12118

Widaman, K. F. (2012). Exploratory factor analysis and confirmatory factor analysis. In H. Cooper (Ed.), APA handbook of research methods in psychology: Vol. *3*. Data analysis and research publication (pp. 361–389). American Psychological Association.

Wothke, W. (1993). Nonpositive definite matrices in structural equation modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). SAGE.

Yang, J.-S., & Kim, T.-Y. (2011). The L2 motivational self-system and perceptual learning styles of Chinese, Japanese, Korean, and Swedish students. *English Teaching*, *66*, 141–162.

You, C., & Chan, L. (2015). The dynamics of L2 imagery in future motivational self-guides. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 397–418). Multilingual Matters.

You, C., & Dörnyei, Z. (2016). Language learning motivation in China: Results of a large-scale stratified survey. *Applied Linguistics*, *37*, 495–519. https://doi.org/10.1093/applin/amu046

You, C., Dörnyei, Z., & Csizér, K. (2016). Motivation, vision, and gender: A survey of learners of English in China. *Language Learning*, *66*, 94–123. https://doi.org/10.1111/lang.12140

Cite this article: 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: 307–329. https://doi.org/10.1017/S0272263123000487

https://doi.org/10.1017/S0272263123000487 Published online by Cambridge University Press