Studies in Second Language Acquisition (2024), 1–12
doi:10.1017/S0272263124000494

CRITICAL COMMENTARY

Manufactured crisis

A response to Al-Hoorie et al. (2024)

Mostafa Papi[1] and Yasser Teimouri[2]

[1]Florida State University and [2]Boğaziçi University
Corresponding author: Mostafa Papi; Email: mpapi@fsu.edu

(Received 26 February 2024; Revised 26 July 2024; Accepted 14 August 2024)

Abstract

Based on correlational and factorial analysis of data collected from 384 middle and high school students in South Korea, Al–Hoorie et al. (2024) claimed the existence of a discriminant validity crisis within the L2 motivational self-system research tradition and advocated for abandoning research in this area. In this response, we critically examined the evidence presented, re-analyzed their data, and argued that their findings actually support the discriminant validity of the target scales. We also discussed issues related to the design and implementation of their study and refuted their assertion regarding a discriminant validity crisis in this field. Finally, we emphasized the necessity of prioritizing definitional validity in the ongoing methodological reforms of L2 motivation research.

Keywords: L2 motivation; L2MSS; self-guides; discriminant validity; definitional validity

Introduction

Al-Hoorie, Hiver, and In'nami (2024) examined the discriminant validity of 18 motivational measures related to the second language (L2) motivational self system (L2MSS) research tradition. They collected data from 384 middle and high school students in South Korea and ran correlational and factorial analyses. Based on their results, they made claims about several motivational measures' lack of discriminant validity. They went so far as to call for the abandonment of the entire L2MSS research tradition, which happens to be the most prominent one in L2 motivation research (Boo, Dörnyei & Ryan, 2015). They argued that their results show that "scales used in the L2MSS suffer from severe discriminant validity concerns" (p. 18), that "there is a severe case of a jangle fallacy in the L2MSS tradition" (p. 15), and even that research on this theoretical perspective "represents a regrettable, costly detour the field has taken" (p. 18). In this response, we critically examined the merits of the authors' claims in light of the evidence provided and a reanalysis of their data. We also addressed the methodological issues that might have led to their unusual results and called for prioritizing definitional validity in L2 motivation research.

Discriminant validity evidence

Discriminant validity concerns whether measures of different constructs overlap too much to represent different constructs. If two measures correlate perfectly or almost perfectly, they *might* be considered measures of the same construct. Discriminant validity is measured using correlational and factor analytic methods (Lawson & Robins, 2021). Al-Hoorie et al. (2024) tried to provide both types of evidence. Below, we argue that the correlational results do not necessarily show a lack of discriminant validity in the measures, and the factor analytic results are either irrelevant or, in fact, supportive of the discriminant validity of the measures in focus.

Correlational results

Correlation is one of the most commonly used techniques to examine discriminant validity in social sciences. "Evidence of discriminant validity exists if other constructs do not correlate strongly enough with the construct of interest to suggest that they measure the same construct" (McKenny, Short & Payne, 2013, p. 156). The question is what correlation cutoff point is high enough to suggest unity of measures. Even though most common correlation coefficients used in social sciences to determine discriminant validity are cutoffs of .85 and .90 (Kline, 2011), Al-Hoorie and colleagues used Dörnyei's (2007) recommendation of .60, which would only explain 36% of the shared variance between two variables and is far from a value approaching unity, which is required to suggest a discriminant validity problem. They also cited Plonsky and Oswald's (2014) benchmark of .60 for large correlations. However, Plonsky and Oswald's benchmarks are not specific to motivation research, nor do they apply to discriminant validity, which "consists of demonstrating that the true correlation of [two traits] is meaningfully less than unity" (Werts & Linn, 1970, p. 208). Therefore, the commonly used cutoffs of .85 and .90, respectively explaining 72% and 81% of the shared variance—which approach unity—should be reason to suspect that two measures tap the same construct and might have insufficient discriminant validity.

From a purely statistical point of view and using the common cutoffs of .85 and .90, one could argue that a few of the measures employed by Al-Hoorie et al. (2024) might have a discriminant validity problem. However, discriminant validity is not just a statistical concept. Rather, it is a concept that depends largely on the conceptual proximity or *definitional validity* (Krause, 2012) of the measures and the specific context of the study. Theoretically, a large correlation of even above .90 does not necessarily mean two variables measure the same construct. Take sex, which is a biological attribute, and gender identity, which is a psychological construct. According to the American Psychological Association (2015), the correlation between the two can reach as high as .995. However, even this value does not suggest that measures of sex and gender identity tap the same construct and, thus, have a discriminant validity problem. Similarly, Le, Schmidt, Harter, and Lauver (2010) found a correlation of .91 between job satisfaction and organizational commitment but did not perceive discriminant validity to be an issue. In other scientific domains, Grivas, Mihas, Arapaki, and Vasiliadis (2008) found individuals' height to be strongly correlated with their left foot (r = .898) and right foot (r = .903). Yet, no one questioned the discriminant validity of

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

height and foot length measures. In short, “different variables can be perfectly correlated and so perfectly fit a common factor vector without representing the same descriptive dimension” (Krause, 2012, p. 395).

Likewise, a smaller correlation between two measures does not mean they measure different things. For example, a meta-analytic study by Hofmann, Gawronski, Gschwendner, Le, and Schmitt (2005) showed that implicit and explicit prejudice have a small correlation (r = .24) even though they represent the same underlying construct of prejudice. Similarly, the notions of happiness and sadness usually correlate moderately, although they are believed to represent the two ends of a mood continuum (Tay & Jebb, 2018). In the field of second language acquisition (SLA), Papi, Bondarenko, Wawire, Jiang and Zhou (2020) found that the growth L2 mindset had a correlation of only -.56 with the fixed L2 mindset, even though the two variables measure the same bipolar construct representing learners’ beliefs about the malleability of their intelligence.

These examples suggest that the classification of constructs and measures should be viewed on a continuum where constructs with close conceptual overlap can be considered sibling constructs. According to Lawson and Robins (2021), two constructs can be considered siblings if—among other things—they are defined in conceptually similar ways, highly correlate, together form a more general parent construct, or are causally related to each other. In other words, sibling constructs “share a close, familial relation, but are not identical” (Lawson & Robins, 2021, p. 345).

Although only five correlations in Al-Hoorie et al. (2024) exceeded the coefficient of .85, we adopt a more conservative approach by examining the correlations that are equal to or greater than .80 from a conceptual perspective. We begin with the correlation between Ideal L2 Self and Linguistic Self-Confidence (r = .80), which the authors identified as their most significant finding. Ideal L2 Self represents an image of the kind of L2 user one would like to be in the future and is measured using items such as *I can imagine myself living abroad and using English effectively for communicating with the locals* (Taguchi, Magid & Papi, 2009) or *I can imagine a day when I speak English like a native speaker of English* (Papi, Bondarenko, Mansouri, Feng & Jiang, 2019). Linguistic Self-Confidence, on the other hand, was operationalized by the researchers to represent how confident the student feels about being able to master a new language and was measured with items such as *If I make more effort, I am sure I will be able to master English (cf. Clément & Baker, 2001).* A close inspection of the concepts embedded within the items shows that whereas the Ideal L2 Self represents an imagined desirable level of L2 mastery, Linguistic Self-Confidence simply represents the confidence in one’s ability to reach that level of mastery. The two constructs are thus conceptually related, and Linguistic Self-Confidence can be argued to have a close relationship with the Ideal L2 Self (Henry & Liu, 2023). One would not normally imagine a day when they speak an L2 fluently if they did not believe in their ability to do so. However, this does not mean the two measures examine the same construct. One can easily imagine someone who believes in their ability to learn a new language but has no desire or vision for learning and speaking it fluently. A large positive correlation between the two variables could thus only mean that the stronger one believes in their ability to master a new language, the stronger their vision of a fluent L2 speaker will be. In fact, Dörnyei (2009) argued that confidence in one’s ability leads to the motivational power of the future selves by making them plausible. Therefore, the large correlation between the two constructs, their conceptual similarity, and their possible causal relationship suggest that Linguistic Self-Confidence and Ideal L2 Self are sibling constructs.

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

To statistically examine the discriminant validity of Ideal L2 Self and Linguistic Self-Confidence using Al-Hoori et al.'s (2024) publicly available data (https://osf.io/7c8qs/), we ran a semipartial correlation[1] between the two variables (see Lawson & Robins, 2021) while controlling for the correlation between Vividness of Imagery (which only represents the vision aspect of their Ideal L2 Self) and Linguistic Self-Confidence to see how much of the variance in Ideal L2 Self is explained by ability beliefs represented in the latter. The analysis showed a modest correlation of (r = .28) between L2 Self-Confidence and Ideal L2 Self, representing nearly 8% of the shared variance and leaving 92% of the variance unexplained. This clearly rejects the argument that "response to the Ideal L2 Self might be driven by a belief in ability rather than an actual–ideal discrepancy" (Al-Hoorie et al., 2024, p. 12).

Ideal L2 Self also showed large correlations with Vividness of Imagery (r = .86) and Ease of Using Imagery (r = .86). In addition, Ease of Using Imagery strongly correlated with Vividness of Imagery (r = .87) and Imagery Capacity (r = .83). If Ideal L2 Self represents an image of one using L2 fluently in the future, Vividness of Imagery, measured with items such as *When imagining how I could use English fluently in the future; I usually have a vivid mental picture of the scene*, only represents how vivid one's Ideal L2 Self is. Similarly, Ease of Imagery, measured with items such as *Sometimes images of myself using English successfully in the future come to me without the slightest effort*, only represents how easily one can imagine their Ideal L2 Self and is also related to students' Imagery Capacity. Vividness of Imagery and Ease of Using Imagery could, thus, be considered sibling constructs underlying the parent construct of the Ideal L2 Self, whereas Imagery Capacity could be argued to have a causal effect on the Ideal L2 Self. These arguments were supported by the results of semipartial correlations showing that Ideal L2 Self modestly correlated with Vividness of Imagery (r = .21), Ease of Imagery (r = .17), and Imagery Capacity (r = .06) after controlling for correlations among the three variables. These results confirm the existence of independent relationships between these variables and the Ideal L2 Self, supporting their relative discriminant validity. Therefore, the vision-related variables of Vividness of Imagery and Ease of Imagery do not represent identical constructs and can be considered siblings belonging to the parent construct of the Ideal L2 Self. In fact, we do not believe that Dörnyei intended these to be two new constructs independent of the Ideal L2 Self. Rather, he employed these measures to represent a more nuanced view of this future self for instructional purposes. In Dörnyei and Chan's (2013) words, "[o]ur focus is on the role images and senses play in shaping the motivation to learn an L2 through promoting a more vivid mental representation of one's self in future states" (p. 440).

Similarly, Ought-to L2 Self's large correlations with Family Influence (r =.87) and Instrumentality-Prevention (r = .84) make theoretical sense. Ought-to L2 Self represents the kind of L2 user the person feels obligated to become, which has social and personal dimensions (Teimouri, 2017). Instrumentality-Prevention represents the personal dimension, and Family Influence is part of the social dimension, which can also include one's teacher, friends, students, colleagues, employer, etc. (Papi et al., 2019; see also Henry & Liu, 2023). The scales for measuring this future self-guide include

[1] Similar to multiple regression analysis, semipartial (also known as part) correlation examines the unique linear relationship between a focal (e.g., L2 Self-Confidence) and an outcome variable (e.g., Ideal L2 Self) while controlling for the shared variance between the focal construct and a sibling construct (e.g., Vividness of Imagery). The semipartial correlation between the focal and outcome variable will be shared variance independent of the correlation between the sibling constructs, supporting the incremental validity of the focal variable.

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

items such as *Studying English is important to me in order to gain the approval of my peers/teachers/family/boss.* Within the specific context of South Korean middle schools and high schools in Al-Hoorie et al.'s (2024) study, it is not surprising that the Ought-to L2 Self is highly correlated with Family Influence. In fact, previous studies confirm strong correlations between Ought-to L2 Self and Family Influence (e.g., Taguchi et al., 2009) but the results do not mean that the measures represent the same construct because the former can theoretically refer to individuals other than—or in addition to — one's family. For example, in the more individualistic Western countries, where college students strive for independence and are less influenced by their families, it is easy to imagine that this population's Ought-to L2 Self may not be so highly correlated with their Family Influence (see Papi et al., 2019). By contrast, these two variables are expected to highly correlate among younger learners who are more influenced by their families, especially in collectivist societies such as South Korea.

In addition, Instrumentality-Prevention, represented in items such as *I have to learn English because I don't want to fail the English course*, strongly correlated with Ought-to L2 Self, suggesting that the negative consequences of failure in English classes are especially high among the students who are more concerned with obligations and family expectations. Papi et al. (2019) and Teimouri (2017) showed that the Ought-to L2 Self can represent either one's own perceived obligations (e.g., *If I don't improve my English, it will have a negative impact on my future.*) or one's perception of what others expect them to accomplish (e.g., *If I don't learn English, I will disappoint my parents/ teachers.*). Therefore, depending on how the Ought-to L2 Self is conceptualized (e.g., own vs. others), it could have stronger or weaker correlations with Instrumentality-Prevention. These arguments were supported by semipartial correlation results. While controlling for the correlation between Instrumentality-Prevention and Family Influence, the results of the analysis showed that Ought-to L2 Self modestly correlated with both Instrumentality-Prevention (r = .29) and Family Influence (r = .37), supporting distinctions among the measures. Therefore, Instrumentality-Prevention and Family Influence can be considered siblings to the parent construct of Ought-to L2 Self, and the large correlations in Al-Hoorie et al. (2024) could only reflect the conceptual overlap among the constructs and the characteristics of the specific population of the study.

Next, Intended Effort[2] showed large positive correlations with Attitudes to Learning English (r = .84), Ideal L2 Self (r = .83), and Positive Changes of the Future L2 Self-Image (r = .80), which have also been documented in past research (e.g., Csizér & Kormos, 2009; Henry & Cliffordson, 2017; Taguchi et al., 2009; You & Dörnyei, 2016). Intended Effort pertains to the level of effort learners intend to invest in L2 learning and is measured using items such as *I would like to spend lots of time studying English.* By contrast, Attitudes Toward Learning English gauges the positive emotions that students associate with the experience of L2 learning and is measured using items such as *I really enjoy learning English.* Attitudes toward Learning English and Intended Effort, therefore, cannot be considered the same construct because the former represents the affective dimension of one's learning experiences, whereas the latter represents one's thoughts and intentions for taking action. Attitudes toward Learning English and

[2] Al-Hoorie et al. (2024) have characterized the use of intended effort as an outcome measure in L2 motivation research as erroneous. However, L2 learners' actual behaviors cannot occur without forming such intentions. While including behavioral measures is essential in L2 motivation research, dismissing the significance of intended effort is misguided, given that it is the closest predictor of observable behavior. From a pedagogical perspective, motivation research aims to trigger changes in L2 learners' behaviors via cultivating such intentions.

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

Intended Effort can thus be considered sibling constructs because the former can have a causal effect on the latter. The more one enjoys an activity, the more strongly one intends to continue doing that activity. The same argument can be extended to the Intended Effort's correlations with the Ideal L2 Self and Positive Changes of the Future L2 Self-Image. These arguments were supported by Intended Effort's modest semi-partial correlations with Ideal L2 Self ($r = .16$), Attitudes to Learning English ($r = .29$), and Positive Changes of the Future L2 Self-Image ($r = .13$), when controlling for the correlations among these three variables.

The large correlation between Instrumentality-Promotion and Instrumentality-Prevention ($r = .81$) contradicted the results of the previous studies. For example, Taguchi et al. (2009) found a correlation of .34, .17, and .13 between the two variables in the Japanese, Chinese, and Iranian samples, respectively. This result could either be due to the conceptual overlap between promotion-oriented and prevention-oriented instrumental values in the specific context of South Korea, and/or it could be an artifact of the study's methodological issues, which will be discussed later.

In sum, the analyses presented above show that whereas common correlation values of .85 and .90 should be considered as possible symptoms for suspecting (not determining) a discriminant validity problem, even larger correlations do not necessarily mean conceptual redundancy. Such conclusions largely depend on theoretical and contextual factors, which were not considered in Al-Hoorie et al.'s (2024) study. The conceptual analysis presented above (see also Papi et al., 2019) follows the argument that any correlational analysis is misguided if it is not based on an in-depth consideration of the definitional validity of different measures. “Without this [definitional validity] having already been achieved, correlational evidence simply cannot be decisive on issues of measurement validity but can only be suggestive for further conceptual analysis leading to further definition and measure refinement” (Krause, 2012, p. 398).

Al-Hoorie et al. (2024) started their article with a thought-provoking question: “What is the value of knowing each leg's length, after already knowing the other leg's length?” (McElreath, 2020, as cited in Al-Hoorie et al., 2024, p. 164). While we concur with the authors that measuring one leg's length would probably suffice (except in cases of anisomelia or leg-length discrepancy), we tend to believe that two legs of the same length are still two different legs, each with its own unique characteristics.

Factor analytic results

Al-Hoorie et al. (2024) used both a single-model Exploratory Factor Analysis (EFA) and a model-comparison Confirmatory Factor Analysis (CFA) to test the discriminant validity of different groups of variables. The merits and results of the two analyses are examined below.

Exploratory factor analysis

The authors used a single-model EFA in the sense that they did not compare their model against alternative models with two or more factors. A single-model EFA, however, is commonly used for data reduction or testing the construct validity of a set of observed variables rather than their discriminant validity. According to Rönkkö and Cho (2022), the fitness indices of such models are not appropriate tests of discriminant validity because they are based on assessing a single scale at a time, whereas discriminant validity is about correlations between different measures. Such

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

EFA models' factor loadings are not useful either because "pattern coefficients do not provide any information on the correlation between two scales, and structure coefficients are an indirect measure of the correlation at best" (Rönkkö & Cho, 2022, p. 22).

Since such techniques are not appropriate for testing discriminant validity, we will not discuss the EFA results further and consider the results irrelevant to this issue.

Confirmatory factor analysis

The authors employed a model-comparison CFA to test the discriminant validity of the variables within six arbitrarily formed groups. For each group of variables, they compared a model with multiple factors (e.g., Group 1: Intended Effort & Attitudes to Learning English) against another model where all the indicators were merged into a single factor. They used the comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA) to assess the fitness of the model, and employed the chi-square difference test to determine which model fits the data better than the other. The results showed that, without exception, all the models with multiple factors showed better fitness indices than the single-factor models. In addition, the chi-square difference tests confirmed that the models with multiple factors were statistically superior to the models with only one factor, directly contradicting the authors' discriminant validity claims.

The authors, however, tried to dismiss these results by arguing that "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)" (Al-Hoorie et al., 2024, p. 14). However, the purpose of this chi-square difference test was to assess the discriminant validity of the scales by comparing one model against another, not to evaluate other psychometric properties in absolute terms. In addition, the CFA models with multiple factors outperformed the single-factor models in terms of other fit indices as well, contradicting the authors' claims about the unity of the scales and undermining their psychometric argument.

Due to the fact that the CFA results ran contrary to the authors' expectations and did not identify a discriminant validity problem, they rightly acknowledged that "future refinement of these scales has the potential to lead to more valid measures" (Al-Hoorie et al., 2024, p. 14), an endeavor that L2 motivation researchers have already embarked on in previous studies (e.g., Papi et al., 2019; Papi & Khajavy, 2021; Teimouri, 2017). In the very next section of their manuscript, however, Al-Hoorie and colleagues (2024), surprisingly, reverted to their original stance, renewing their original claims, decrying "a severe case of a jangle fallacy" (p. 15), "declaring a state of validation crisis" (p. 12), and even calling for the abandonment of the entire research tradition. As argued above, all of this was based on a few misunderstood correlation coefficients found in a limited dataset of questionable quality collected in a single study from a specific sample of middle and high school students in South Korea. Below, we discuss the potential methodological sources of the unusually large correlations Al-Hoorie et al. (2024) found between the measured variables.

Methodological issues

Several issues in the design and implementation of Al-Hoorie et al.'s (2024) study have probably led to the large correlations that qualify as outliers in the field of L2

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

motivation. The first issue concerns the scales used. The authors put together several items from various studies conducted in different contexts to form unique scales that had not been used in any previously published study. For instance, 14 items from three scales developed in China, Japan, and Iran (Taguchi et al., 2009) were combined to measure Instrumentality-Promotion, which is an odd decision for both theoretical and practical purposes. Discriminant validity is commonly defined as "the degree of divergence among indicators that are designed to measure different constructs" (Hamann et al., 2013, p. 72). By testing discriminant validity, researchers aim to ensure that "two measures are tapping separate constructs" (Krause, Whitler & Semadeni, 2014, p. 102). If two measures are shown to assess the same construct, they are considered to lack discriminant validity. Discriminant validity, thus, is about measures rather than constructs, even though sometimes researchers use the two interchangeably. Constructs exist independently of measures, and multiple measurements can exist for the same construct. For instance, in the field of SLA, there are several measures for examining motivation, anxiety, aptitude, etc. However, Al-Hoorie et al.'s (2024) conclusions are based on the common but misguided assumption that measures and constructs are synonymous, leading them to combine items from different scales to create decontextualized measures that had never been used in any previous study. The large correlation coefficients are, thus, specific to their scales and study, and extending their conclusions to corresponding motivational constructs is not justified.

Using a larger number of items in a scale has some benefits, such as reducing measurement error and increasing the reliability of the scales. However, items need to be skillfully crafted based on the theoretical objectives of the study, the conceptual definitions of the target measures, and the study's specific context, among other things. For example, for measuring future selves in the context of Iran, Papi and Khajavy (2021) adapted or used different items than those developed by Papi et al. (2019) in the ESL context of the US. Similarly, Tahmouresi and Papi (2021) developed new scales based on qualitative data they collected from their participants in Iran, and Taguchi et al. (2009) used three versions of the same questionnaire to collect data from Japanese, Chinese, and Iranian students. Nonetheless, Al-Hoorie and colleagues (2024) did not seriously consider their items' theoretical and conceptual relevance. For example, they used items that mismatched their corresponding constructs (e.g., *I am working hard at studying English*, to measure Intended Effort, and *The things I want to do in the future require me to use English* to measure Ideal L2 Self). Furthermore, they did not take into account the contextual appropriateness of the items. For instance, they used items such as *Studying English is important to me because I would like to spend a longer period living abroad (e.g., studying and working)*, which were not appropriate for middle and high school students living in South Korea. This data-driven approach without any consideration for a construct's underlying theoretical and contextual dimensions has led the authors to flawed, misguided, and misleading interpretations (see also, Al-Hoorie, McClelland, Resnik, Hiver & Botes, 2024).

Second, when using new scales for measuring motivational constructs or using the same scales in a new context, researchers typically run EFAs to uncover the factors underlying their observed variables (i.e., questionnaire items) and establish the construct validity of the scales before engaging in further analyses. Al-Hoorie et al. (2024), however, did the opposite and ran correlational analyses before running factor analyses: A practice that the authors ironically criticized in the same manuscript as "[a]nother example of poor validation practices" (p. 6) and resembled it to "putting the horse behind the cart" (p. 6). Thus, the construct validity of the variables that the researchers used in their correlational analyses was not established to begin with, rendering the correlational results questionable.

Third, the authors used a 7-point forced-response Likert scale with a neutral midpoint labeled *neither agree nor disagree* instead of the 6-point scale, which is often used in studies on the L2MSS (e.g., Taguchi et al., 2009). The problem with a 7-point scale with a neutral midpoint concerns the conceptual ambiguity of the midpoint. Choosing this option in a forced-response item could mean multiple things: (a) *I don't have a strong opinion*, (b) *I don't understand this question*, (c) *I don't know how to answer this question*, (d) *This does not apply to me*, and (e) *I prefer not to respond to this item.* The flexibility in the interpretation of the midpoint has probably led a large proportion of respondents—ranging from 25% (*N* = 96) to almost 40% (*N* = 152)—to choose this option in response to the majority of the questionnaire items (*N*= 66). The tendency to choose the midpoint seems to have led to the cluster of all means around the midpoint (means range: 3.71–4.64), which could have inflated the correlations between the variables.

The final point concerns the unusual length of the questionnaire. One can only imagine how exhausting it could be, especially for middle school students, to read a consent form and answer 119 questions outside of their class time and during the final weeks of their semester. It is not surprising that the response rate from the data, which was collected in person, was only 50–60% across sites. Since the items were randomized for each participant, we cannot determine the quality of the responses based on order However, the unusually long questionnaire has likely led to responses of lower quality.

In sum, the issues with the design and administration of the questionnaire, including the formation of novel scales without establishing their construct validity, lack of attention to the theoretical and contextual appropriateness of the items, the neutral midpoint of the response scale, and the unusual length of the questionnaire, have likely led to inflated correlations that are clear outliers in the literature and contradict the findings of previous studies. For example, as noted before, Taguchi et al. (2009) found a correlation of .37, .17, and .13 between Instrumentality-Promotion and Instrumentality-Prevention; Al-Hoorie et al. (2024) found a correlation of .80 between the two. Within the Iranian dataset, the correlation between Instrumentality-Prevention and Intended Effort was .12, which is very different from a correlation of .61 the researchers have found. The results of the study by Al-Hoorie et al. (2024) not only contradict the findings of numerous previous studies (e.g., Csizér & Kormos, 2009; Henry & Cliffordson, 2017; You & Dörnyei, 2016), but they also conflict with Al-Hoorie's (2018) own meta-analytic results, which included much smaller correlations for Intended Effort with L2 Learning Experience (r = .41 vs. r = .84), Ought-to L2 Self (r = .38 vs. r = .61), and Ideal L2 Self (r = .61 vs. r = .82).

Conclusions

Discriminant validity is not an all-or-nothing attribute of constructs, nor is it a one-and-done endeavor that can be determined by a single correlation coefficient; rather, it varies along a continuum and is achieved in an ongoing iterative process of theoretical and empirical exploration (Lawson & Robins, 2021; Tay & Jebb, 2018). In other words, "like any validity assessment, discriminant validity assessment requires consideration of context, possibly relevant theory, and empirical results and cannot be reduced to a simple statistical test and a cutoff no matter how sophisticated" (Rönkkö & Cho, 2022,

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

p. 33). Contrary to its claims, Al-Hoorie et al. (2024) did not present evidence for the existence of a discriminant validity crisis in relation to the target motivational measures or constructs. In fact, they provided evidence to the contrary. The correlational results were probably a byproduct of problematic methodological decisions, the EFA results were irrelevant, and the CFA results, in fact, supported the discriminant validity of the scales.

The study, however, highlighted the existence of another potential crisis in the study of L2 motivation—or perhaps more broadly in research on individual differences in SLA—that is, the issue of definitional validity, which we believe should be placed at the forefront of methodological reform in this research area. Establishing the definitional validity of different constructs and operationalizing them based on agreed-upon definitions and the context of each study must precede any statistical experimentation with data in the instrument development process (Lawson & Robins, 2021). Definitional validity is essential for identifying the conceptual scope, links, and boundaries of constructs, developing accurate measures, improving their predictive validity, and creating theoretical models that more effectively explain multifaceted phenomena, such as motivation in language learning.

There is no question that L2 motivation researchers must enhance the quality of their methodological practices, including but not limited to the psychometric properties of their instruments. Sudina (2021) rightly reported that research on this topic typically suffers from multiple drawbacks, such as lack of methodological diversity, lack of sufficient evidence for the internal consistency, and the convergent, discriminant, and predictive validity of the scales used (see also Boo et al., 2015). In fact, recent work in the field has led to the development of new scales for measuring future selves in a theoretically meaningful way and provided evidence for the construct and discriminant validity of the scales (e.g., Papi et al., 2019; Teimouri, 2017). Several studies have shown that the new scales explain theoretically meaningful differences in L2 emotions (e.g., Jiang & Papi, 2021; Tahmouresi & Papi, 2021), persistence in L2 learning (Feng & Papi, 2020), strategic inclinations in L2 use (e.g., Papi & Khajavy, 2021, Teimouri, 2017), L2 achievement (Tahmouresi & Papi, 2021), and L2 speech development (Zhou & Papi, 2023), confirming the predictive validity of the scales. Considering these developments, it can be argued that research in the L2MSS tradition has already been undergoing methodological reforms that are improving the quality of our research methods and instruments, and refining our understanding of how motivation works for language learning (Papi & Hiver, 2022).

Engaging in intellectual dialogues allows researchers to exchange ideas, challenge assumptions, explore new perspectives, and contribute to the growth and advancement of any field of study. Such exchanges can foster a collaborative environment where diverse viewpoints can contribute to the evolution of our instruments, designs, methodologies, and theoretical frameworks. However, dismissing an entire area of research built over many years can only be justified if such assertions are supported by a substantial body of robust evidence rather than the limited findings of a single study. While hypercritical and sensational positions may attract short-term attention, they can inadvertently discourage many researchers from pursuing their interest in the very field we seek to further develop. Therefore, the optimal path for moving the field forward remains grounded in systematic and theory-based engagement in the empirical process of instrument development. We invite the authors of the original study to participate in this collaborative process, fostering a climate that benefits the entire academic community.

Acknowledgements. We would like to thank the authors of the original manuscript for making their data publicly available and answering our queries. We are also grateful to SSLA Editor, Luke Plonsky, and four anonymous reviewers for their constructive feedback on the previous versions of this manuscript.

References

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*(4), 721–754.

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*, 1–23. https://doi: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

American Psychological Association. (2015). Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist*, *70*(9), 832–864. https://doi.org/10.1037/a0039906

Boo, Z., Dörnyei, Z., & Ryan, S. (2015). L2 motivation research 2005–2014: Understanding a publication surge and a changing landscape. *System*, *55*, 145–157.

Clément, R., & Baker, S. C. (2001). Measuring social aspects of L2 acquisition and use: Scalecharacteristics and administration. *Technical Report*. Ottawa: School of Psychology, University of Ottawa.

Csizér, K., & Kormos, J. (2009). Learning experiences, selves and motivated learning behavior: A comparative analysis of structural models for Hungarian secondary and university learners of English. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 98–119). Multilingual Matters.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. 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., & 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*(3), 437–462.

Feng, L., & Papi, M. (2020). Persistence in language learning: The role of grit and future self-guides. *Learning and Individual Differences*, *81*, 10194. https://doi:10.1016/j.lindif.2020.101904

Grivas, T. B., Mihas, C., Arapaki, A., & Vasiliadis, E. (2008). Correlation of foot length with height and weight in school age children. *Journal of forensic and legal medicine*, *15*(2), 89–95.

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*(5), 713–736.

Henry, A., & Liu, M. (2023). Can L2 motivation be modelled as a self-system? A critical assessment. *System*, *119*, 103158. https://doi.org/10.1016/j.system.2023.103158

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385.

Jiang, C., & Papi, M. (2021). The motivation-anxiety interface in language learning: A regulatory focus perspective. *International Journal of Applied Linguistics*, *32*(1), 25–40. https://doi.org/10.1111/ijal.12375

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford.

Krause, M. S. (2012). Measurement validity is fundamentally a matter of definition, not correlation. *Review of General Psychology*, *16*(4), 391–400. https://doi.org/10.1037/a0027701

Krause, R., Whitler, K. A., & Semadeni, M. (2014). Power to the principals! An experimental look at shareholder say-on-pay voting. *Academy of Management Journal*, *57*(1), 94–115. https://doi.org/10.5465/amj.2012.0035

Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: What are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review*, *25*(4), 344–366.

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: *An empirical investigation. Organizational Behavior and Human Decision Processes*, *112*(2), 112–125. https://doi.org/10.1016/j.obhdp.2010.02.003

McKenny, A. F., Short, J. C., & Payne, G. T. (2013). Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods*, *16*(1), 152–184. https://doi.org/10.1177/1094428112459910

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press

Papi, M., & Khajavy, G. H. (2021). Motivational mechanisms underlying second language achievement: A regulatory focus perspective. *Language Learning*, *71*(2), 537–572. https://doi.org/10.1111/lang.12443

Papi, M., Bondarenko, A. V., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research: The 2X2 model of L2 self-guides. *Studies in Second Language Acquisition*, *41*(2), 337–361.

Papi, M., Bondarenko, A. V., Wawire, B., Jiang, C., & Zhou, S. (2020). Feedback-seeking behavior in second language writing: Motivational mechanisms. *Reading and Writing*, *33*, 485–505.

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. 3–34). Routledge.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language learning*, *64*(4), 878–912.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*(3), 579–592.

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14.

Sudina, E. (2021). Study and Scale Quality in Second Language Survey Research, 2009–2019: The Case of Anxiety and Motivation. *Language Learning*, *71*(4), 1149–1193. https://doi.org/10.1111/LANG.12468

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.

Tahmouresi, S., & Papi, M. (2021). Future selves, enjoyment and anxiety as predictors of L2 writing achievement. *Journal of Second Language Writing*, *53*, 100837. https://doi.org/10.1016/j.jslw.2021.100837

Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science*, *1*(3), 375–388. https://doi.org/10.1177/2515245918775707

Teimouri, Y. (2017). L2 selves, emotions, and motivated behaviors. *Studies in Second Language Acquisition*, *39*(4), 681–709.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*(3), 193–212. https://doi.org/10.1037/h0029778

You, C., & Dörnyei, Z. (2016). Language learning motivation in China: Results of a large-scale stratified survey. *Applied linguistics*, *37*(4), 495–519. https://doi.org/10.1093/applin/amu046

Zhou, Y., & Papi, M. (2023). The role of future L2 selves in L2 speech development: A longitudinal study in an instructional setting. *System*, *119*, 103156.

Cite this article: Papi, M., & Teimouri, Y. (2024). Manufactured crisis: A response to Al-Hoorie et al. (2024). *Studies in Second Language Acquisition,* 1–12. https://doi.org/10.1017/S0272263124000494

https://doi.org/10.1017/S0272263124000494 Published online by Cambridge University Press