Studies in Second Language Acquisition (2025), **47**, 1184–1192
doi:10.1017/S0272263125100892

CRITICAL COMMENTARY

With "crisis" comes opportunity: Using invariance testing to understand and improve measurement models (A response to Al-Hoorie et al., 2024)

Charlie Nagle

College of Liberal Arts, Department of Spanish and Portuguese, The University of Texas at Austin, Austin, TX, USA
Email: cnagle@austin.utexas.edu

(Received 26 April 2025; Revised 26 April 2025; Accepted 05 May 2025)

Abstract

Al-Hoorie et al. (2024) have described L2 Motivational Self System research as having a "validation crisis." In this response, I offer a brief history of recent motivation research to contextualize how we reached this point. I then discuss measurement invariance testing, the purpose of which is to ensure that the proposed measurement model holds across groups or time. Crucially, invariance testing is a necessary precursor to subsequent analyses because if invariance is not upheld, it is impossible to know whether differences reflect true underlying differences in the latent construct or are simply the byproduct of a changing measurement model. I argue that by combining measurement invariance testing with open science practices, we can create and validate better measurement models while cultivating a better understanding of their scope of application.

Keywords: Measurement models; psychometric validity; latent constructs; motivation

Introduction

Research on language learning motivation is an interesting example of history repeating itself. In the early 1990s, Dörnyei called on motivation scholars to reevaluate the state of the art, observing that "the main problem with Gardner's social psychological approach [the dominant model at the time] appeared to be, ironically, that it was too influential" (1994, p. 273). Motivation research finds itself in a similar quagmire now. The L2 Motivational Self System (L2MSS; Dörnyei, 2009) has become the dominant model, leaving little room for conversation about alternative approaches to conceptualizing and quantifying motivation.[1] On the one hand, intense interest in the L2MSS has

[1]Quantitative approaches to motivation research can only take the field so far. Complementary qualitative research is needed, as is mixed-methods research. However, given that much motivation research, including L2MSS research, is carried out using quantitative methods, I focus on the quantitative approach here.

generated an important body of work, including numerous studies reexamining the constructs included in the model and the scales used to measure them (e.g., Papi et al., 2019). At the same time, as Al-Hoorie et al. (2024) have pointed out in their critique, the proliferation of self-report instruments and the often-ambiguous language used to describe hypothesized relationships between the core L2MSS constructs and potential sibling constructs has created a challenging research landscape where it is difficult to understand exactly what is being tested and what the application of potential findings might be.

An appropriate metaphor might be the game of telephone, where one person gives a message to the person next to them, and then that person conveys the message to the person next to them, and so on, until at the end of the chain, the original and the end-of-the-line messages are compared. Most of the time, the essence of the message is retained, but important details have been lost. In the same way, the L2MSS has gone through dozens of subtle permutations throughout its lifetime, permutations that span concept and method. I agree with Al-Hoorie et al.'s (2024) concern about jingle-jangle fallacies in the motivation literature, but I also think that Papi and Teimouri (2024) have done an excellent job of disentangling the complex idea space within which the L2MSS now operates. Ultimately, the goal of model-building is to provide clear and testable hypotheses, which, when evaluated using appropriate methods, can provide information on how a model like the L2MSS can be improved. Alternatively, if hypotheses are vague, then virtually any empirical outcome can be made to fit the model, making testing pointless. Likewise, if instruments are not psychometrically sound, then findings will be unreliable or spurious.

Regardless of whether the validity issues that Al-Hoorie et al. (2024) have described rise to the level of "crisis," undoubtedly their paper and the responses it has generated have created the necessary conditions for a broader conversation about motivation theory and method. I doubt that the L2MSS will disappear any time soon, nor should it because it does provide an important way of conceptualizing and quantifying L2 motivation. Yet, we must ensure that measurement practices are sound. In this critical commentary, I address measurement invariance, an essential but often overlooked psychometric property of latent constructs like motivation. I argue that one productive path forward is through large-scale measurement invariance testing, which can help us understand the scope of application of proposed L2MSS measurement models. To illustrate the potential of invariance testing, I draw upon the educational psychology literature, where such testing is more common than in our field.

How did we get here?

Al-Hoorie et al. (2024) have claimed that there is a validity crisis in L2MSS research, but the idea of crisis has been brewing for several years, with several key publications acting as catalysts for the present discussion. In his meta-analysis of L2MSS studies, Al-Hoorie (2018) observed that the most common outcome measure was intended effort, whereas objective measures of achievement were far rarer. When the outcome was intended effort, the pooled, meta-analytic correlation between that measure and the L2MSS constructs was in the medium to large range, but when the outcome measure was achievement-related, the pooled correlation was small for the ideal L2 self and the L2 learning experience and negligible for the ought-to L2 self. This finding reinforced a trend of questioning the predictive value or motivational potential of the ought-to L2 self.

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

Yet, Papi et al. (2019) pushed back against this characterization, noting that the lack of predictive power may have been due to conceptualization and measurement issues with respect to how the selves were formulated and tested. Building on Teimouri (2017), the authors further developed the "own versus other" distinction for the selves and tested the corresponding 2 × 2 model against Dörnyei's (2009) original two-factor model and Teimouri's (2017) three-factor model, in which the ought-to L2 self but not the ideal L2 self showed the own/other bifurcation. Papi et al. found that the 2 × 2 model showed a superior fit to data collected from a sample of 257 international students who were L2 English speakers studying at a North American university. The authors also analyzed the relationship between the four constructs and eager and vigilant L2 use, the two factors that emerged from an exploratory factor analysis of 12 self-report items related to various aspects of motivated behavior. In that analysis, the ideal L2 self/own was the only significant predictor of eager L2 use, whereas the ought L2 self/own was the strongest predictor of vigilant L2 use. In discussing the results with respect to the trajectory of motivation research, Papi et al. observed that "it was time… for the original conceptualization [of the L2MSS] to undergo modifications to better capture how motivation works for language learning" (2019, pp. 355–56). They also called for motivation research to be reanchored in the self-guide concepts and methods central to Higgins' (1987) original proposal.

In another report, Al-Hoorie and Hiver (2020) fit separate L2MSS structural equation models to data collected from 644 South Korean high school students in three disciplines: L2 English, L3 Mandarin, and math. Minor differences aside, the L2MSS model fit all three disciplines well. In the model they proposed and tested, the primary antecedents of postachievement (final grades) were prior achievement and the L2 learning experience. In contrast, the selves were hypothesized to be predictors of intended effort but not postachievement. When they fit an exploratory model including a link between the selves and postachievement, they found no evidence supporting such a link for L2 and L3 learning, but for math, the path between the ideal L2 self and achievement was small but statistically significant. Yet, in a subsequent step, model comparisons revealed no significant differences in the coefficients across disciplines, suggesting that a single motivational self system model, as opposed to a set of unique models, would be adequate to capture motivation across the three disciplines. Based on these results, the authors argued that "in the absence of empirical support for cross-subject uniqueness, it might be more constructive to move toward a unified theory of learning that is inclusive of the various psycho-social factors at play" (p. 11). In short, there did not appear to be empirical support for the fundamental difference hypothesis that L2 (or L3) learning involves unique motivational processes (or, at least, different motivational models) compared with other types of learning.

In another publication, Al-Hoorie et al. (2020) characterized L2 motivation research as having an "identity crisis," which they attributed to the "fundamental difference curse" described above and the "questionnaire curse," or the fact that L2 motivation research has historically shown an overreliance on self-report questionnaires. They also commented on an apparent tension between the interdisciplinarity of L2 motivation research and its insularity. In other words, L2 motivation research draws on concepts and methods from educational psychology, applied linguistics, and so on, but applied linguists working on L2 motivation may not be especially well-versed in developments in neighboring disciplines.

In summary, then, in the past decade, as L2MSS research has flourished, researchers have begun to critically reexamine model concepts and instruments, including questioning the uniqueness of L2 motivation compared with motivation for other types of

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

learning. Crucially, these developments do not undermine motivation science but rather speak to its strength and disciplinary maturity. Like other areas of quantitative science concerned with model building, motivation research has followed a recognizable, if not predictable, trajectory: A model is proposed and heavily investigated, resulting in a large body of work that is then synthesized and critiqued, and through that process, the model is refined and updated, and connections are sought to neighboring fields.

Alongside intense and increasing interest in L2 motivation, scholars have also grown interested in a range of related constructs, including, for example, L2 grit (e.g., Sudina & Plonsky, 2021; Teimouri et al., 2020) and L2 academic buoyancy (Yun et al., 2018). L2 grit, which is a relatively recent arrival in L2 research, at least compared with motivation, has been undergoing its own process of revision and refinement (e.g., Credé & Tynan, 2021). In terms of L2 academic buoyancy, it is worth noting that the model that Yun et al. (2018) tested included motivational antecedents such as the Ideal L2 Self, as well as persistence, which is quite close to perseverance, one of the hypothesized subcomponents of L2 grit. Conceptual broadening in terms of the range of constructs investigated is accompanied by deepening research into anxiety and other individual differences that have long stood at the core of the field. This twin pattern is adding to an already crowded multivariate space, making it critical for researchers to be as precise as possible with respect to what predicts what in the learner psychology landscape. Thus, it seems fruitful to extend Al-Hoorie et al.'s (2024) questions and concerns about L2 motivation to the field of individual difference research at large: How many constructs do we need? How do they relate to one another? What do they predict, and how well do they predict it? These questions speak to what Papi and Teimouri (2024) have referred to as the "definitional validity of different constructs" (2024, p. 10).

Indeed, at a conceptual level, it seems that one problem is that there is not a single clear and comprehensive exposition of how L2MSS constructs relate to other individual differences. Such a proposal would be invaluable because it would create a visual reference of predictors, leading to testable hypotheses. Coming up with such a visual map would be challenging for an individual researcher or research team given the sheer number of variables and possible interrelationships, but perhaps this work would be feasible for a team of experts or an organization like the International Association for the Psychology of Language Learning (https://www.iapll.com/). A fundamental goal for future work should therefore be developing not only a model of motivation and sound psychometric scales for measuring it, but also integrating that model into the overarching individual differences landscape to understand motivation relative to other key variables such as anxiety, grit, and so on. In this way, some of the jingle-jangle fallacies that Al-Hoorie et al. (2024) have commented on can be avoided.

Leveraging invariance testing to improve measurement models

The purpose of invariance testing is to ensure that the proposed model holds across groups or time. Participants at different sites might interpret items differently, in which case the constructs themselves or the scales on which they are measured might take on different meanings for different subsets of the sample. The same logic applies to time. If participants are tested over time, their interpretation of the items could change, leading to changes in the measurement model itself. In either case, if the goal is to compare groups to one another or to compare learners to themselves over time, if the model is changing, it is impossible to know if differences between or within groups reflect true

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

differences in the latent constructs and their interrelationships or are simply the byproduct of a changing measurement model. A full treatment of measurement invariance is beyond the scope of this commentary (for an overview, see Putnick & Bornstein, 2016; for information on invariance and L2 learning, see Nagle, 2023; Sudina, 2023), but some brief remarks are necessary to understand the importance of this technique for evaluating the psychometric adequacy of a model prior to subsequent group-wise and time-wise analyses.

Measurement invariance testing is carried out within the structural equation modeling (SEM) framework and involves placing a series of constraints on model form. In the configural model, it is only assumed that items load onto the same factor. Factor weights and item intercepts are freely estimated. In the weak invariance model, factor weights are then constrained to be equal across groups or time, and in the strong invariance model, item intercepts are also constrained to be equal. Because invariance is tested by specifying a series of SEMs, the typical SEM fit criteria (root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), standardized root mean square residual (SRMR)) can be applied to evaluate model adequacy. The crucial tests, however, are the model comparisons, where a nonsignificant test statistic between a more lenient and a stricter model is desirable because it suggests that imposing the relevant constraint did not significantly alter, or impair, the fit of the model to the data. If, on the other hand, model comparisons show that there is a significant change in fit, that means that invariance is not upheld, and, by extension, that the model itself could be changing.

Applied to L2 motivation work, if researchers want to compare motivation across groups (L1 Chinese, Japanese, and Iranian learners of English, as in Taguchi et al., 2009)[2] or over time (Hungarian learners of English or German, as in Dörnyei & Csizér, 2002), invariance testing is a necessary precursor. Once the necessary level of invariance is established, analysis can continue using SEM to compare latent means between groups or over time. In longitudinal research, for instance, the aim might be to estimate latent change trajectories and the relationships between them using a multivariate latent growth curve model.[3] It bears mentioning that invariance testing can also be an end in itself insofar as researchers might be interested in understanding the extent to which a single underlying model can even be specified and reliably estimated in the first place.

Despite the necessity of ensuring a psychometrically sound measurement model, invariance testing is rare in L2 research. Sudina (2021) reported only one instance of invariance testing for L2 anxiety and 19 for L2 motivation, representing 1% and 7% of the scales synthesized in the report. Yet, invariance testing is a common step in the

[2]Taguchi et al. (2009) was ground-breaking for its time. With 4,943 participants across three contexts, it was a study of unprecedent size. The authors estimated separate SEMs for the Japanese, Chinese, and Iranian subsamples, and then descriptively compared the path estimates. However, in terms of invariance testing, it could be that the measurement model is changing, in which case the observed differences might not be due to true differences in the estimates but rather to differences in the measurement model. It would be interesting to fit a single model to the complete data set, testing for measurement invariance across groups. If invariance were upheld and the path estimates were shown to differ by L1 group/context, then those differences could be assumed to reflect true underlying differences.

[3]At the same time, analysis does not need to continue via complex SEM. Instead, if invariance holds, simpler analytical techniques, including *t* test can be applied to the data to compare mean scores between groups. For an interesting example involving the development and validation of a scale to measure global orientation, including comparisons across groups and contexts, see Chen et al. (2016).

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

neighboring discipline of educational psychology. In that field, researchers have sought to understand how motivation changes over time, as learners move within and transition between academic contexts, making longitudinal invariance an essential analytical step (Arens et al., 2019; Guo, Marsh, et al., 2015; Guo, Parker, et al., 2015). Measurement invariance has also been tested for different learner groups, including, for instance, first-generation versus continuing-generation college students (Part et al., 2020).

Muenks et al. (2023) is a compelling example of the importance of invariance testing. In that study, the authors collected data from two cohorts enrolled in the same undergraduate biochemistry course, one from Spring 2020, before the onset of the COVID-19 pandemic, and one from Fall 2020, after pandemic restrictions were put into place. As a first step, they sought to establish invariance for the two cohorts. Weak invariance was upheld, but strong invariance was not, leading the authors to conclude that it would be "inappropriate to combine the two samples together as they cannot be directly compared" (p. 4). They then proceeded to analyze data from the two samples separately, demonstrating that data analysis can still move forward even when invariance is not upheld, albeit via an alternative, psychometrically licensed route. Muenks et al. (2023) also reminds us of the diverse forces that can shape how participants react to the scales used to measure latent constructs like motivation. Indeed, in the ever-evolving educational landscape, it is easy to imagine how technological innovation, such as the expansion of generative AI, could have a real impact on motivation and the models used to measure it.

In summary, invariance testing is necessary to ensure the psychometric validity of a model when comparing groups to one another or themselves over time. As Muenks et al. (2023) highlight, invariance testing reveals important information about what can be assumed of the data, helping researchers to understand what the appropriate analytical options are. I therefore urge the L2 research community not to accept group-wise or time-wise comparisons for latent constructs unless measurement invariance is tested and established. Importantly, invariance cannot be assumed or imported from one study to the next but rather must be tested each time a survey is administered, in the same way that reliability is administration-specific (McClelland & Larson-Hall, 2025). Admittedly, the SEMs used to test for invariance demand a certain sample size, which may be difficult to achieve. In that case, there is no easy solution. Instead, researchers must strike a balance between what is achievable given the affordances of the context in which they work and what is psychometrically sound. Either way, invariance testing cannot be ignored.

Open science and invariance testing

I fully agree with Al-Hoorie et al.'s position that "without valid measurement, inferences become untrustworthy—a flaw that cannot be fixed by large sample sizes, rigorous designs, or advanced statistics" (2024, p. 309). In my view, invariance testing is a necessary component of valid measurement. Yet, I cannot help but wonder if research trends are also partly to blame for the measurement problems L2 motivation research is experiencing. Historically, as in many areas of L2 research, motivation researchers have worked independently or in small teams, collecting and analyzing data from relatively homogenous single-site samples. There is nothing wrong with this mode of research, but in the aggregate, if one of the fundamental goals of L2 motivation research is to develop and test models that apply broadly across groups (contexts, cultures) and/or that are capable of

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

capturing changes in motivation, then it may be time for researchers to start engaging in big team open science (for commentary see, e.g., Marsden & Morgan-Short, 2023; Moranski & Ziegler, 2021). In this research mode, a team of researchers would work together to develop a survey to be used at several research sites. Sites could be selected purposefully to test the scope of application of a hypothesized model using the invariance testing procedure outlined in the previous section. For instance, researchers working at geographically and sociodemographically diverse institutions around the United States could collect data from samples at each site. By the same token, a group of international researchers could come together to develop a single survey, use translation and back-translation to ensure that items are correctly worded in the language(s) of the international sites, and then test for invariance across cultures, languages, or other groups relevant to the research.

This approach has several potential advantages. First, it could help combat the proliferation of idiosyncratic instruments that have been observed in the literature. Rather than, for instance, ten researchers implementing context-specific surveys, that group of ten individuals would implement the same coconstructed survey at all sites.[4] This survey would correspond to a proposed underlying measurement model, and invariance testing would reveal the extent to which this model holds across groups, however they are conceptualized (by site or by some other variable relevant to the research). If it does, then groups can be compared to one another, and broad generalizations can be reached about between-group differences in motivation as well as downstream relationships between motivation and other variables. If it does not, then the inability to establish invariance provides valuable information about the need to refine the instrument (i.e., the survey) or to consider context-specific models at one or more sites, in the same way that Muenks et al. (2023) analyzed data from each cohort separately. Furthermore, if multiple samples are recruited at each site, then invariance testing can proceed within and across sites, establishing a robust basis for the scope of application of the measurement model. Through this process, measurement models can be problematized, updated, and refined, with implications for theory. Morgan-Short et al. (2018) and The TwiLex Group (2024) are excellent examples of what such research could look like.

Conclusion

The fact that the L2MSS is generating so much debate is a sign of disciplinary progress. The rich, critical discussion that is taking place will reshape the state of the art in L2 motivation research and hopefully have a ripple effect on the conceptualization and measurement of other individual difference constructs. Ultimately, the field will be best served if the thought leaders and measurement experts in this area come together to put forward a unified proposal for the L2MSS, including what its antecedents and outcomes are hypothesized to be, and how motivation, thus conceptualized, relates to other individual differences in learner psychology. This will likely entail carrying out larger-scale, multisite studies, drawing on open science practices, and looking further afield, at motivation as conceptualized and measured in neighboring disciplines. Invariance testing should be a critical component of this endeavor.

[4]As discussed by Sudina (2023), the same items do not necessarily have to be used cross-contextually. Instead, uniquely-worded but theoretically-motivated items can be developed for each context, and cross-context comparisons can be carried out as long as the functional equivalence of items is established.

https://doi.org/10.1017/S0272263125100892 Published online by Cambridge University Press

Acknowledgments. I would like to thank the authors of the original article and the authors of the other critical commentaries for initiating a discussion on this important topic. I would also like to thank Luke Plonsky for providing feedback on an earlier version of this manuscript.

Competing interests. The author declares none.

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. https://doi.org/10.14746/ssllt.2018.8.4.2

Al-Hoorie, A. H., & Hiver, P. (2020). The fundamental difference hypothesis: Expanding the conversation in language learning motivation. *SAGE Open*, *10*(3). https://doi.org/10.1177/2158244020945702

Al-Hoorie, A. H., Hiver, P., Kim, T.-Y., & De Costa, P. I. (2020). The identity crisis in language motivation research. *Journal of Language and Social Psychology*, *40*(1), 136–153. https://doi.org/10.1177/0261927x20964507

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

Arens, A. K., Schmidt, I., & Preckel, F. (2019). Longitudinal relations among self-concept, intrinsic value, and attainment value across secondary school years in three academic domains. *Journal of Educational Psychology*, *111*(4), 663–684. https://doi.org/10.1037/edu0000313

Chen, S. X., Lam, B. C. P., Hui, B. P. H., Ng, J. C. K., Mak, W. W. S., Guan, Y., Buchtel, E. E., Tang, W. C. S., & Lau, V. C. Y. (2016). Conceptualizing psychological processes in response to globalization: Components, antecedents, and consequences of global orientations. *Journal of Personality and Social Psychology*, *110*(2), 302–331. https://doi.org/10.1037/a0039647

Credé, M., & Tynan, M. C. (2021). Should language acquisition researchers study "grit"? A cautionary note and some suggestions. *The Journal for the Psychology of Language Learning*, *3*(2), 37–44.

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal*, *78*(3), 273–284. https://doi.org/10.2307/330107

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., & Csizér, K. (2002). Some dynamics of language attitudes and motivation: Results of a longitudinal nationwide survey. *Applied Linguistics*, *23*(4), 421–462.

Guo, J., Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2015). Directionality of the associations of high school expectancy-value, aspirations, and attainment. *American Educational Research Journal*, *52*(2), 371–402. https://doi.org/10.3102/0002831214565786

Guo, J., Parker, P. D., Marsh, H. W., & Morin, A. J. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology*, *51*(8), 1163–1176. https://doi.org/10.1037/a0039440

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

Marsden, E. & Morgan-Short, K. (2023). (Why) are open research practices the future for the study of language learning?. *Language Learning*, *73*(S2), 344–387. https://doi.org/10.1111/lang.12568

McClelland, N., & Larson-Hall, J. (2025). Why you should stop using the ideal L2 self and the L2 motivational self-system tomeasure motivation (Reaction to Al-Hoorie, Hiver & In'nami, 2024). Studies in Second Language Acquisition, 1–12. https://doi.org/10.1017/S0272263124000779

Moranski, K. & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, *71*(1), 204–242. https://doi.org/10.1111/lang.12434

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B. I., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning*, *68*(2), 392–437. https://doi.org/10.1111/lang.12292

Muenks, K., Miller, J. E., Schuetze, B. A., & Whittaker, T. A. (2023). Is cost separate from or part of subjective task value? An empirical examination of expectancy-value versus expectancy-value-cost perspectives. *Contemporary Educational Psychology*, *72*. https://doi.org/10.1016/j.cedpsych.2023.102149

Nagle, C. L. (2023). A design framework for longitudinal individual difference research: Conceptual, methodological, and analytical considerations. *Research Methods in Applied Linguistics*, *2*(1), 100033. https://doi.org/10.1016/j.rmal.2022.100033

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

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

Part, R., Perera, H. N., Marchand, G. C., & Bernacki, M. L. (2020). Revisiting the dimensionality of subjective task value: Towards clarification of competing perspectives. *Contemporary Educational Psychology*, *62*. https://doi.org/10.1016/j.cedpsych.2020.101875

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

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

Sudina, E. (2023). A primer on measurement invariance in L2 anxiety research. *Annual Review of Applied Linguistics*, *43*, 140–146. https://doi.org/10.1017/s0267190523000089

Sudina, E., & Plonsky, L. (2021). Academic perseverance in foreign language learning: An investigation of language-specific grit and its conceptual correlates. *The Modern Language Journal*, *105*(4), 829–857. https://doi.org/10.1111/modl.12738

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.

Teimouri, Y. (2017). L2 selves, emotions, and motivated behaviors. *Studies in Second Language Acquisition*, *39*(4), 681–709. https://doi.org/10.1017/s0272263116000243

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2020). L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, *26*(5), 893–918. https://doi.org/10.1177/1362168820921895

The TwiLex Group (2024). First language effects on incidental vocabulary learning through bimodal input: A multisite, preregistered, and close replication of Malone (2018). *Studies in Second Language Acquisition*, *46*(5), 1413–1438. https://doi.org/10.1017/S0272263124000275

Yun, S., Hiver, P., & Al-Hoorie, A. H. (2018). Academic buoyancy. *Studies in Second Language Acquisition*, *40*(4), 805–830. https://doi.org/10.1017/s0272263118000037

Cite this article: Nagle, C. (2025). With "crisis" comes opportunity: Using invariance testing to understand and improve measurement models (A response to Al-Hoorie et al., 2024). *Studies in Second Language Acquisition,* 47: 1184–1192. https://doi.org/10.1017/S0272263125100892