Annual Review of Applied Linguistics (2023), **43**, 140–146
doi:10.1017/S0267190523000089

RESEARCH ARTICLE

A primer on measurement invariance in L2 anxiety research

Ekaterina Sudina

East Carolina University, U.S.
Email: sudinae22@ecu.edu

Abstract

Measurement invariance (MI) is essential to bolstering validity arguments behind psychometric instruments (Zumbo, 2007). Nonetheless, very few second language (L2) anxiety scales, including the most widely used L2 anxiety questionnaire—the Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986)—have been tested for MI. The present paper seeks to address this deficiency in the literature (a) by demonstrating why this procedure is key to enhancing our understanding of the latent phenomenon in question, particularly in relation to different language learning contexts, (b) by outlining the main stages of MI testing with specific recommendations for L2 scale developers and users, (c) by providing commendable examples of the application of MI in applied linguistics research in order to illustrate the potential of this technique, and (d) by making a case for employing MI in future validation studies, thereby promoting methodologically sound research practices in the context of anxiety scales and elsewhere in applied linguistics.

Keywords: measurement invariance; measurement equivalence; L2 anxiety

Researchers of L2[1] anxiety as well as other individual difference constructs (e.g., motivation, willingness to communicate) in applied linguistics are often concerned with comparing questionnaire (or scale)[2] scores across key participant characteristics (e.g., target language, language-learning context, age, gender, country) and interventions (e.g., pre- and posttest) as well as over time (e.g., as in latent growth modeling). Unfortunately, such score comparisons are not meaningful unless evidence of *construct comparability*—the situation in which scores from different groups "measure the same construct of interest on the same metric"—is convincingly demonstrated beforehand (Wu et al., 2007, p. 1). Measurement invariance (MI, also referred to as measurement equivalence; see Somaraju et al., 2022) endeavors to tackle the issue of construct comparability by embracing an empirical approach to examining group differences. This paper aims to provide a nontechnical introduction to MI for L2 anxiety researchers and applied linguists working with questionnaire data more broadly. I begin by presenting a brief history of MI and by describing this concept using nonspecialized language. I then seek to demonstrate why this procedure is key to enhancing our understanding of

L2 anxiety as well as other learner-internal characteristics, particularly with regard to various language-learning situations.

Although the history of MI goes back to the 1960s (e.g., Meredith, 1964; see Putnick & Bornstein, 2016, for more), the techniques for MI testing have been arguably clouded with overly specialized statistical terminology, thereby making this procedure less popular with applied researchers (as noted by Wu et al., 2007). Indeed, it is not surprising that one of the most widely used L2 anxiety questionnaires—a 33-item Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986)—was not tested for MI at the time of development. Until recently, only a short version of the FLCAS had been comprehensively examined for MI (see Botes et al., 2022). Other popular L2 anxiety questionnaires, including the Foreign Language Reading Anxiety Scale (Saito et al., 1999), the Second Language Writing Apprehension Test (Cheng et al., 1999), and the Foreign Language Listening Anxiety Scale (Elkhafaifi, 2005), were not assessed for MI during their initial validation either. In fact, a recent systematic review of L2 anxiety research published in twenty-two leading L2 journals in 2000–2020 revealed that only five out of 321 L2 anxiety scales in the sample were tested for MI (Sudina, 2023). What is more, no MI tests were performed on the newly developed scales, which diminishes validity arguments behind these psychometric instruments. Despite the scarcity of MI testing in L2 anxiety research, applications of this technique are becoming increasingly common in neighboring disciplines focusing on nonlanguage-related anxieties, including dating, social, and pain anxiety (see Adamczyk et al., 2022; Rogers et al., 2020; Torregrosa Díez et al., 2022). Additionally, there has been a surge in MI testing in the realm of other L2 individual differences such as self-guides, enjoyment, and engagement (see Derakhshan et al., 2022; Liu et al., in press).

So, what is MI and why is it important? The concept of MI refers to the situation in which a latent variable representing a theoretical construct[3] and consisting of one or more observed variables, such as questionnaire items, is similarly understood by respondents in different groups or by respondents in the same group over time (Putnick & Bornstein, 2016). As such, MI is a prerequisite for assessing mean scores on a latent variable across groups (e.g., L2 anxiety of students learning English in a foreign language context and students learning English in a second language context) and across time (e.g., L2 anxiety of students when they started learning English in elementary school and the same students following several years of language instruction). In a similar vein, MI should be established in experimental designs if the latent variable in question was somehow manipulated (e.g., to ensure that L2 students' anxiety decreased due to the intervention itself rather than as an artifact of respondents' interpretation of the questionnaire items following the intervention). Critically, rigorously validated questionnaires should be normed across age, gender, as well as a number of other participant characteristics to allow for group comparisons (Lee, 2018). MI can be attained by identifying and excluding scale items that carry different meanings for different groups. A hypothetical example would be to discover that nail-biting was a symptom of L2 anxiety in children but not in adults. Keeping the item inquiring about nail-biting would bias mean score comparisons across the two groups. If adults scored lower on L2 anxiety because they bite their nails less, this would be misleading because nail-biting is unrelated to anxiety in adults anyway (although it could be related to stress, for example). The importance of MI thus lies in its potential to equip researchers with the necessary tools to detect whether and to what extent scale items should be interpreted similarly or differently, depending on group membership.

https://doi.org/10.1017/S0267190523000089 Published online by Cambridge University Press

Main Stages of Measurement Invariance

I have thus far highlighted the conceptual importance of testing for MI; to allow for meaningful cross- or within-group comparisons, researchers need to ensure that questionnaire items are invariant, or similarly construed, across the groups. In reality, however, this requires a great deal of skill and decision-making on the part of the researcher. The purpose of this section is to provide a brief overview of the main stages of MI testing via multigroup confirmatory factor analysis (CFA), which is arguably the most common method for establishing MI in a structural equation modeling framework (Wu et al., 2007). Nonetheless, it is not my intention to provide a tutorial (for a comprehensive review of MI, see Putnick & Bornstein, 2016 and Somaraju et al., 2022; for a tutorial on longitudinal measurement invariance, see Nagle, 2023). For guidance on how to test for MI using item response theory by investigating differential item functioning, or DIF, in particular, see Andrich and Marais (2019) and Zumbo et al. (2015); for a gentle introduction to DIF, see Zumbo (2007).

To illustrate the MI procedure via the multigroup CFA, I will use a hypothetical L2 anxiety scale consisting of five positively keyed items that represent different facets of the construct (listening, reading, writing, speaking, and pronunciation anxiety; see Fig. 1) and are measured on a Likert scale. The composite mean score indicates the level of language-specific anxiety. Let's assume that the goal is to compare L2 English students' anxiety in the United States and Japan (i.e., in a second versus foreign language learning context).

In the first stage of MI testing, researchers examine *configural* equivalence, or invariance of the internal structure of the scale across groups. Configural invariance is tenable if the factor structure of the scale is identical in both samples (i.e., the same five items load on the same factor of L2 anxiety). If, however, in one of the groups the structure of the scale is different (e.g., L2 anxiety consists of two different factors instead of one, with three items loading on Factor 1 and two other items loading on Factor 2), this indicates configural noninvariance. To remedy the issue, researchers can either "redefine the construct (e.g., omit some items and retest the model)" (Putnick & Bornstein, 2016, p. 75) or accept the fact that the latent variable of interest is nonequivalent across groups and refrain from comparing mean group scores. Having established

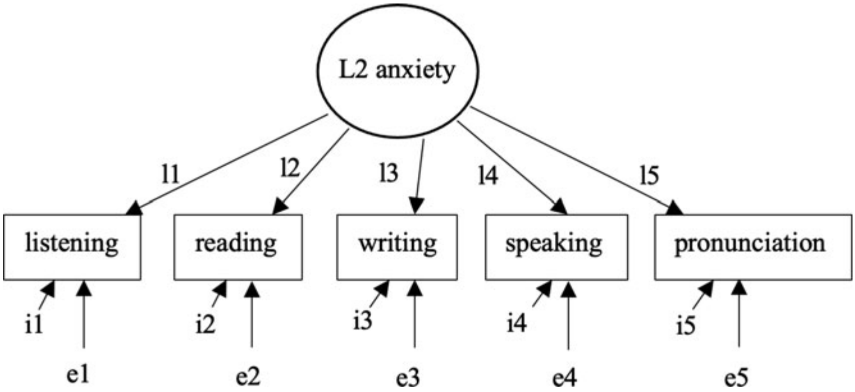

Figure 1. Hypothetical L2 anxiety scale.
Note. Factor loadings: l1-l5; item intercepts: i1-i5; item residuals: e1-e5.

https://doi.org/10.1017/S0267190523000089 Published online by Cambridge University Press

configural invariance, researchers proceed to explore *metric* equivalence, or invariance of factor loadings (also known as weights) across groups. This second stage involves examining the extent to which each item on the scale contributes to the latent variable. Statistically speaking, researchers compare the fit of the metric model with constrained factor loadings with the fit of the configural model with unconstrained factor loadings and item intercepts. Metric invariance is supported if item loadings on the factor are similar across groups, that is, the fit of the metric model should not be considerably worse. If, however, one or more items has a nonequivalent loading across the groups,[4] metric invariance is not supported. For instance, the item representing anxiety in speaking may be more strongly associated with the overall L2 anxiety in the Japanese group (foreign language context) but not in the US group (second language context). To address the issue of metric noninvariance, researchers need to determine which item(s) has a nonequivalent loading, exclude this item(s), and rerun the tests of both configural and metric invariance; alternatively, researchers can admit measurement noninvariance and refrain from further group comparisons (Putnick & Bornstein, 2016). Provided there is evidence of metric invariance, in the third stage, researchers evaluate *scalar* equivalence, or invariance of item intercepts (also referred to as means). To that end, constraints are imposed on item intercepts in both groups, and the fit of the scalar model with constrained intercepts is compared with the fit of the metric model with constrained factor loadings and unconstrained intercepts. Scalar invariance is upheld if the fit of the model has not significantly deteriorated after imposing these additional parameter constraints. If, however, the scalar model fit is considerably worse, scalar invariance should not be assumed. It would indicate that one or more item intercepts has different parameters across the groups. For example, compared to L2 English students in Japan, students in the United States may obtain a higher score on the item representing anxiety in speaking, but this amplified speaking anxiety in the U.S. group would not contribute to their overall L2 anxiety level (although it could contribute to their overall stress level, for example). If faced with scalar noninvariance, researchers can either accept it and avoid group comparisons altogether or identify the source of nonequivalence by locating problematic item intercept(s), remove this item(s), and retest all invariance models (Putnick & Bornstein, 2016).

If scalar invariance is achieved, researchers can impose constraints on item residuals (error terms) in both groups and test for *residual* equivalence, or strict invariance. However, many methodologists would argue that it is justifiable to examine mean differences across groups as long as scalar equivalence, or strong invariance, is supported because “residuals are not part of the latent factor, so invariance of the item residuals is inconsequential to interpretation of latent mean differences” (Putnick & Bornstein, 2016, p. 76; see Wu et al., 2007).

Implications of Measurement Invariance for L2 Anxiety Research and Beyond

Critically, evidence of MI allows researchers to safely make inferences about group differences in the latent variable of interest. To take our hypothetical L2 anxiety example, if MI is demonstrated, and the result of a *t*-test suggests that students have higher language anxiety in the United States than in Japan, researchers can rest assured that this finding reflects a true difference in the latent variable and is not a by-product of measurement bias. Moreover, passing MI tests allows researchers to make more advanced comparisons by examining relationships between two or more latent variables across groups (e.g., compare the relationship between L2 anxiety and achievement

https://doi.org/10.1017/S0267190523000089 Published online by Cambridge University Press

across contexts), that is, to investigate *structural* invariance. Finally, in addition to being an important step in questionnaire development and validation, MI testing can be used as a means to advance theory and "evaluate theoretical predictions" (Somaraju et al., 2022, p. 756). To illustrate, testing for MI and reporting standardized effect sizes for group differences allows for establishing construct-specific norms by comparing results across primary studies. If there is evidence of cultural dependence of L2 anxiety across target languages and language learning contexts, these findings can help language educators determine the best anxiety-reducing strategies in their language classrooms. To take it even further, the importance of MI extends beyond simple group comparisons. It enables researchers to ensure the fairness of questionnaires by helping detect and revise problematic items in order to avoid measurement bias (Jung & Yoon, 2016). It should be noted, however, that a traditional approach to MI may not account for cultural differences, particularly if questionnaire items have been developed in a WEIRD (i.e., Western, educated, industrialized, rich, democratic) context (Boehnke, 2022). Instead of developing scale items in English, which is typically used as the *lingua franca*, translating them into other languages via back-translation, and removing noninvariant culture-specific items, MI testing can be done in a more culturally sensitive manner. According to Boehnke (2022), in a culturally inclusive approach, scale developers agree on the construct of interest beforehand and develop items representing the construct in each language and context independently. Then, exploratory factor analysis is performed on each sample separately, and items with high loadings on the first factor are retained and included in MI testing, which is performed on the combined dataset. In other words, following this new approach, items measuring L2 anxiety in English learners in the United States and Japan do not have to be similarly worded "as long as functional equivalence is achieved through item intercorrelations" (p. 1164). This reduces "the bias that is brought in by relying exclusively on Western-origin items" (p. 1163).

Conclusion

By advocating to implement more MI testing in L2 anxiety research, I have sought in this paper to raise awareness of the potential of this technique to inform L2 research and practice and to present the main stages of MI via multigroup CFA within a structural equation modeling framework. Given that establishing MI is often challenging and time-consuming but nevertheless crucial for making meaningful conclusions about study findings, it is advisable to test for MI at the very least during the process of scale development and validation so that other researchers can safely use the questionnaire of interest with a similar population or in a similar context. I anticipate that the various applications of MI testing will continue to increase in L2 anxiety research as well as elsewhere in applied linguistics.

Notes

[1] Hereafter, this acronym is used to refer to both second and foreign language contexts.

[2] In language assessment, researchers typically deal with *test* score comparisons. But the principle remains the same (see Wu et al., 2007).

[3] For a distinction between a theoretical construct and a latent variable, see Wu et al. (2007).

[4] Sometimes nonequivalent items can be easily identified by looking at the data; however, various statistical procedures are also available, including the forward confidence interval method, the backward modification index method, and the factor-ratio test (see Jung & Yoon, 2016).

https://doi.org/10.1017/S0267190523000089 Published online by Cambridge University Press

References

Adamczyk, K., Morelli, N. M., Segrin, C., Jiao, J., Park, J. Y., & Villodas, M. T. (2022). Psychometric analysis of the dating anxiety scale for adolescents in samples of Polish and U.S. young adults: Examining the factor structure, measurement invariance, item functioning, and convergent validity. *Assessment*, *29*(8), 1869–1889. https://doi.org/10.1177/10731911211017659

Andrich D., & Marais I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences* (Springer Texts in Education series). Springer. https://doi.org/10.1007/978-981-13-7496-8_16

Boehnke, K. (2022). Let's compare apples and oranges! A plea to demystify measurement equivalence. *American Psychologist*, *77*(9), 1160–1168. https://doi.org/10.1037/amp0001080

Botes, E., van der Westhuizen, L., Dewaele, J., MacIntyre, P., & Greiff, S. (2022). Validating the short-form Foreign Language Classroom Anxiety Scale (S-FLCAS). *Applied Linguistics*, *43*(5), 1006–1033. https://doi.org/10.1093/applin/amac018

Cheng, Y.-S., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning*, *49*(3), 417–446. https://doi.org/10.1111/0023-8333.00095

Derakhshan, A., Doliński, D., Zhaleh, K., Enayat, M. J., & Fathi, J. (2022). A mixed-methods cross-cultural study of teacher care and teacher-student rapport in Iranian and Polish university students' engagement in pursuing academic goals in an L2 context. *System*, *106*, 102790. https://doi.org/10.1016/j.system.2022.102790

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *Modern Language Journal*, *89*, 206–220. https://doi.org/10.1111/j.1540-4781.2005.00275.x

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *Modern Language Journal*, *70*, 125–132. https://doi.org/10.1037/t60328-000

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 567–584. https://doi.org/10.1080/10705511.2015.1138092

Lee, S. T. H. (2018). Testing for measurement invariance: Does your measure mean the same thing for different participants? *APS Observer*, *31*, 32–33.

Liu, E., Wang, J., & Bai, S. (in press). Self-guides, enjoyment, gender, and achievement: A survey of Chinese EFL high school students. *Journal of Multilingual and Multicultural Development*, 1–18. https://doi.org/10.1080/01434632.2022.2153854

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. https://doi.org/10.1007/BF02289699

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

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

Rogers, A. H., Gallagher, M. W., Garey, L., Ditre, J. W., Williams, M. W., & Zvolensky, M. J. (2020). Pain Anxiety Symptoms Scale–20: An empirical evaluation of measurement invariance across race/ethnicity, sex, and pain. *Psychological Assessment*, *32*(9), 818–828. https://doi.org/10.1037/pas0000884

Saito, Y., Horwitz, E. K., & Garza, T. J. (1999). Foreign language reading anxiety. *Modern Language Journal*, *83*, 202–218. https://doi.org/10.1111/0026-7902.00016

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*, *25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, 1–29. Advance online publication. https://doi.org/10.1017/S0272263122000560.

Torregrosa Díez, M. S., Gómez Núñez, M. I., Sanmartín López, R., García Fernández, J. M., La Greca, A. M., Zhou, X., Redondo Pacheco, J., & Inglés Saura, C. J. (2022). Measurement invariance and latent mean differences between American, Spanish and Chinese adolescents using the social anxiety scale for adolescents (SAS-A). *Psicothema*, *34*(1), 126–133. https://doi.org/10.7334/psicothema2021.42

https://doi.org/10.1017/S0267190523000089 Published online by Cambridge University Press

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*, 3.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*(1), 136–151. https://doi.org/10.1080/15434303.2014.972559

Cite this article: 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