*Studies in Second Language Acquisition* (2026), **48**, 513–534
doi:10.1017/S0272263126101569

METHODS FORUM

# Fit-hacking: Questionable research practices in structural equation modeling

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

[1]Jubail English Language and Preparatory Year Institute, Royal Commission for Jubail and Yanbu, Saudi Arabia; [2]Faculty of Science and Engineering, Chuo University, Japan and [3]School of Teacher Education, Florida State University, Tallahassee, FL, USA
**Corresponding author:** Ali H. Al-Hoorie; Email: hoorie_a@rcjy.edu.sa

(Received 19 April 2025; Revised 04 January 2026; Accepted 10 January 2026)

## Abstract

Structural equation modeling (SEM) is a powerful and flexible modeling framework for testing complex relationships among observed and latent variables. However, its methodological complexity and analytical flexibility also increase the risk of questionable research practices (QRPs), especially in fields like applied linguistics, where training in advanced statistics may be limited. This article synthesizes the literature on QRPs and applies it to SEM by identifying seven categories of problematic practices: not checking assumptions, not validating a measurement model, not testing competing models, not sufficiently justifying modeling decisions, relying on post hoc model modification, overemphasizing global fit indices, and incomplete or nontransparent reporting. Each practice is described with examples and linked to broader issues in research ethics and transparency. The paper concludes with concrete recommendations for improving the credibility and reproducibility of SEM research, emphasizing the integration of best practices with the principles of open science.

**Keywords:** fit-hacking; open science; questionable research practices; research transparency; structural equation modeling

Study quality is a foundational element of credible research. Following Plonsky (2024), study quality is a multidimensional construct that includes methodological rigor, transparency, ethics, and societal relevance. These dimensions are interconnected and ensure research upholds both scientific excellence and ethical standards. Rigor and transparency are closely tied to ethical practices, as they build trust in research outcomes while reducing the risk of ethical pitfalls, such as misconduct or questionable

*Note.* AI was used to proofread and improve some parts of the manuscript.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

research practices (QRPs). Similarly, transparency plays a central role in enabling the evaluation of methods and findings, making it inseparable from rigor in promoting reproducibility and accountability (Open Science Collaboration, 2015).

When it comes to quantitative methods, structural equation modeling (SEM) has recently become a powerful and versatile modeling framework for exploring complex relationships among variables. SEM integrates factor analysis and path analysis to examine both direct and indirect effects, and mediated and moderated effects, in addition to more advanced techniques, allowing researchers to test theoretical models that include latent constructs and observed variables simultaneously (Kline, 2023). Its ability to estimate multiple relationships within a single analysis makes SEM particularly valuable in situations where phenomena are multifaceted and interdependent (Byrne, 2012), such as applied linguistics. However, SEM involves complex statistical procedures, thus increasing the potential for QRPs to inadvertently occur during its application (Flores-Kanter & Mosquera, 2023; Revelle, 2024). Researchers face numerous decisions, including selecting appropriate indicators, specifying models, and interpreting fit indices, which could lead to issues such as cherry-picking models or selectively reporting results to support desired outcomes. Importantly, these issues often arise unintentionally, driven by a lack of training or by external pressures to publish, but they have serious ethical consequences by undermining the reliability and validity of scientific findings.

In this article, we aim to address analytical transparency in the context of SEM, specifically reflective latent variable path analysis. Although best practices in SEM have been discussed for quite some time (e.g., Hancock & Mueller, 2013; Hoyle, 2023), there has been limited effort to explicitly integrate these practices with the principles of open science. We argue that the reason why many QRPs in SEM persist is precisely because of a lack of transparency. To bridge this gap, we review a range of QRPs commonly encountered in SEM and provide practical, theory-informed recommendations for improving transparency, reproducibility, and research integrity. Our goal is to promote more responsible SEM use by aligning established methodological standards with emerging open science principles.

## Definition and taxonomy of QRPs

Research practices are generally divided into three main levels of ethical integrity: responsible conduct of research, QRPs, and deliberate misconduct (Al-Hoorie & Hiver, 2024; Plonsky, Larsson, et al., 2024; Steneck, 2006). Responsible conduct of research represents the gold standard, involving practices that align with ethical principles and professional standards. On the opposite end, deliberate misconduct—fabrication, falsification, and plagiarism—is universally condemned for its damaging impact on research integrity. QRPs occupy a middle ground, including behaviors that could potentially compromise the quality and credibility of research. That is, a certain behavior (e.g., excluding an outlier) may be justifiable under specific circumstances, but it can also be seen as problematic in other situations, especially when not transparently disclosed and justified.

Adopting a Delphi method, Plonsky, Larsson, et al. (2024) attempted to provide a comprehensive taxonomy of QRPs in quantitative humanities. Their results pointed to a set of 58 practices falling under four categories: funding, design and data collection, data analysis and interpretation, and write-up and dissemination. When it comes to doing the research itself, QRPs related to design and data collection may undermine

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

methodological rigor at the early stages of a study. These include inadequate sampling methods, failure to define inclusion and exclusion criteria, and selective recruitment of participants to achieve desired outcomes. Additionally, poor documentation of data collection procedures and failure to pilot-test instruments may compromise transparency and the reliability of data. Similarly, QRPs at the data analysis and interpretation stage revolve around statistical manipulation and selective reporting. Such practices include *p*-hacking, failing to adjust for multiple comparisons, and ignoring violations of statistical assumptions. Furthermore, misinterpreting statistical findings and overlooking alternative valid explanations of data can also mislead readers and future research. QRPs in this category are particularly relevant to SEM research because specifying and estimating latent variable models entails substantial researcher degrees of freedom (i.e., flexibility) that, if abused, can undermine the substantive conclusions drawn from such models to the detriment of knowledge in the field. These practices risk skewing the scientific record and contributing to replication crises, ultimately undermining trust in research findings.

Finally, QRPs at the write-up and dissemination stage are related to the communication of research findings. Examples include selectively reporting significant results, omitting null findings, failing to disclose study limitations, and using hyperbolic language to exaggerate the importance of findings. Plonsky, Larsson, et al. (2024) argued that these practices compromise the transparency and credibility of published work. They also point out that QRP taxonomies can facilitate targeted training and the creation of clear guidelines to address them.

## Prevalence of QRPs

The prevalence of QRPs is a growing concern across disciplines, including applied linguistics. Isbell et al. (2022) conducted an exploratory study examining the prevalence of QRPs and scientific misconduct among applied linguists engaged in quantitative research. Drawing on responses from 351 participants, the study revealed that QRPs are widespread, with 94% of respondents admitting to engaging in at least one QRP, such as selectively reporting significant results, *p*-hacking, or excluding nonsignificant findings. The study also highlighted a concerning frequency of misconduct, including data fabrication and modification to achieve statistically significant results, though these behaviors were less common (approximately one in six). Interestingly, early-career researchers and those who experienced publication rejection due to non-significant results reported higher engagement in QRPs and misconduct, emphasizing the influence of systemic pressures such as the "publish-or-perish" culture. Isbell et al. (2022) also underscored the importance of improved statistical training and ethical guidelines to address these issues and promote more rigorous and transparent research practices in applied linguistics.

In the same vein, Larsson et al. (2023) investigated the prevalence and perceived severity of QRPs among researchers in the quantitative humanities. Their results revealed that 96% of respondents reported engaging in at least one QRP in the past five years. They also examined the perceived severity of these practices and found a strong negative correlation between frequency and perceived severity. Researchers tended to avoid practices they considered highly severe, such as knowingly removing data to achieve favorable results or misrepresenting qualifications in funding proposals. Differences in engagement with QRPs were also observed based on researcher characteristics. Again, early-career researchers and those with a higher publication rate were more likely to report engaging in certain QRPs, suggesting that systemic pressures, such

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

as the demand for more and more publications, may drive ethical compromises. Additionally, researchers with more quantitative training were paradoxically (weakly) associated with higher frequencies of some QRPs. Larsson et al. (2023) argued that this indicates the need for ethics-focused training in research methods courses.

Building on prior work examining QRPs in applied linguistics, Plonsky, Brown, et al. (2024) analyzed open-ended survey responses to provide richer insights into the contextual factors driving such practices. They identified nine themes from 1,226 qualitative comments, including researcher training, methodological conventions, the influence of top-down guidelines, circumstantial appropriateness, and pressures to publish. Participants frequently rationalized QRPs as context-dependent, citing space limitations, field-specific norms, and the need to simplify complex research processes for publication. For example, some respondents defended withholding methodological details to meet journal requirements or enhance accessibility for their target audience. The results also revealed significant variability in researchers' ethical stances and behaviors, with some justifying selective reporting of significant results or methodological omissions as necessary for publication success. Plonsky, Brown, et al. (2024) therefore emphasized the role of training in perpetuating QRPs, particularly among researchers relying on coauthors or statisticians for methodological decisions. The authors argued that the variability in researcher practices highlights the need for field-specific standards and ongoing professional development to address these shortcomings.

## Factors contributing to QRPs

From the above review, one running theme that has contributed to the persistence of QRPs is systemic pressures within academia and a lack of training. The relentless "publish or perish" culture plays a significant role in driving QRPs, as researchers often feel compelled to deliver impactful results that align with the expectations of high-impact journals, which tend to favor statistically significant findings over null or inconclusive outcomes. This dynamic creates incentives for behaviors like data manipulation, selective reporting, and hypothesizing after results (or HARKing) are known to achieve favorable outcomes (e.g., Plonsky, Brown, et al., 2024). Early-career researchers are especially vulnerable to these pressures, as their career progression, including securing tenure and funding, is closely tied to their publication records. These challenges are further compounded by a lack of comprehensive training in research ethics and methodology, leaving many ill-prepared to navigate the ethical dilemmas such pressures create (De Costa et al., 2021; Isbell et al., 2022; Larsson et al., 2023).

Second, cognitive biases play a role in researchers' engagement in QRPs. Even for well-meaning researchers, confirmation bias can lead them to selectively seek out and interpret results that support their favored hypotheses. This may be a spontaneous and unconscious decision, considering that QRPs exist in an ambiguous grey area and may be justifiable under specific circumstances (Andrieux et al., 2024). Similarly, motivated reasoning encourages the individual to rationalize problematic behaviors and provide plausible justifications that appear to be in line with their moral standards (Specker Sullivan, 2022). Furthermore, the psychological distance between researchers and the potential consequences of their methodological decisions, combined with various self-serving biases, can create "ethical blindness"—a temporary inability to see the ethical dimensions of one's decisions (Palazzo et al., 2012). As Specker Sullivan (2022) pointed out, ethics training may be more effective if it highlights the impact of such cognitive biases on researcher behavior.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

QRPs appear to be widespread across empirical research. Evidence for this can be seen even in relatively simple statistical procedures. For example, recent work on Cronbach's alpha has documented systematic excesses around the conventional $\alpha = .70$ threshold (Hussey et al., 2025), suggesting overfitting to in-sample data (i.e., α-hacking) or publication bias. If such patterns are observed for a basic and widely used statistic, it is reasonable to expect that more complex analytical procedures may be even more vulnerable to QRPs. In this sense, methodological complexity itself can contribute to the proliferation of questionable practices. Some methods, like SEM, require numerous analytical decisions and offer multiple paths to test hypotheses, creating opportunities for researchers' degrees of freedom that can be exploited consciously or unconsciously. When faced with complex statistical procedures, researchers must make countless decisions about model specification, variable selection, and treatment of outliers, among others—with each decision point representing a potential avenue for bias to influence the research process. This complexity is further compounded by the fact that many researchers may not have received comprehensive training in these advanced methods, leading them to rely on potentially outdated rules of thumb and published examples that may themselves perpetuate questionable practices. Additionally, the inherent flexibility in these complex methods can make it difficult for peer reviewers to identify potential issues, as the sheer number of analytical decisions can obscure problematic choices behind layers of technical detail. The fact that a model converged (i.e., a computational solution in which parameter values are stable and consistent with the data and model specifications) says little about how robust or informative that model is. Even problematic or nonsensical models (i.e., those that defy logic or theory and those with serious misspecifications) can converge perfectly. This creates a particularly challenging scenario where even well-intentioned researchers might engage in QRPs without fully recognizing the implications of their methodological decisions.

## SEM

SEM is a statistical framework that allows researchers to examine the relationships among variables, including multiple independent variables and multiple dependent variables. It can include latent and observed variables and can accommodate continuous, categorical, and count variables. Thus, it can be considered as a broader approach to analyzing the relationships among variables than general(ized) linear modeling (Jak & Jorgensen, 2025). In this paper, we use the term SEM primarily to refer to reflective latent variable path analysis, which is one of the most commonly used SEM approaches in applied linguistics research (Larsson et al., 2021, 2022). We recognize that SEM encompasses a broader family of techniques, including mixture models, growth curve models, multilevel SEM, and others. While our discussion and examples are grounded in latent variable path analysis, the underlying principles we raise—particularly those concerning assumptions, transparency, and full reporting—are also applicable to these other SEM procedures.

In SEM, models are often visually presented as path diagrams to better communicate the relationships among variables. Latent variables are represented by ovals, observed variables are represented by rectangles. Latent variables are unobservable theoretical constructs that are inferred from observed indicators, often based on prior empirical or theoretical work. Observed variables are those directly observed, such as test scores and questionnaire responses. For example, Figure 2 Panel A shows that disappointment, ideal L2 self, and demotivation are latent variables. Responses to questionnaire items (i.e., D1 and D2; I1 through I3; and De1 through De5) are observed variables. Disappointment is

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

hypothesized to influence ideal L2 self and demotivation, whereas Ideal L2 self is hypothesized to influence demotivation, as represented by one-sided arrows. See Bollen (1989) and Kline (2023) for further details on SEM, including those that have not been covered above, and Jak and Jorgensen (2025) for a tutorial.

## QRPs in SEM

In the context of SEM, we use QRPs to refer to *methodological or analytical decisions that deviate from best practices and ethical standards and that can bias the results toward the researcher's preferred hypotheses at the expense of the reliability, validity, and replicability of the results.* In this section, we discuss QRPs that we believe are particularly related to the use of SEM in research on language learning and use. These QRPs are (1) not checking assumptions, (2) not validating a measurement model, (3) not testing competing models, (4) insufficiently justified modeling, (5) modifying the model post hoc, (6) overemphasizing global model fit indices, and (7) incomplete and nontransparent reporting. We have derived this list from SEM reporting guidelines (e.g., Appelbaum et al., 2018; Morrison et al., 2017), reviews of SEM applications (Hancock & Schoonen, 2015; Ockey & Choi, 2015; Schoonen et al., 2025; Winke, 2014), foundational works on SEM (e.g., Byrne, 2012; Hancock & Mueller, 2013; Hoyle, 2023), research on QRPs (e.g., Flores-Kanter & Mosquera, 2023; Isbell et al., 2022), as well as our own experience in the field. We do not claim that this represents an exhaustive list. Our primary objective is to raise awareness of QRPs that may be unintentionally employed, consequently compromising the trustworthiness of SEM results. Finally, to enhance transparency and to explicitly integrate these practices with the principles of open science, we demonstrate in the supplementary materials how to address these QRPs (https://osf.io/xmk5b).

### *1) Not checking assumptions*

SEM relies on a range of assumptions that operate at different levels, including assumptions about model specification, estimation, data characteristics, and study design. Some of these assumptions are specific to the estimator used, whereas others are shared with more general statistical models. Because the relevance of particular assumptions depends on the type of data, the model structure, and the analytical choices made by the researcher, not all assumptions apply equally in every SEM study. In this section, we highlight some assumptions that commonly arise in applied SEM research and that have direct implications for model estimation, missing data handling, and sample size. Readers who are interested in more detailed and technical discussions of SEM assumptions are referred to standard methodological treatments (e.g., Hancock & Schoonen, 2015; Kline, 2023; West et al., 2023).

Classical SEM estimation and model-fit testing were developed under asymptotic assumptions, particularly assuming a large number of observations relative to the number of variables in the model (e.g., Marcoulides et al., 2023). Under such conditions, key statistical assumptions are more likely to be met, and commonly used test statistics behave as intended. In applied research, however, these ideal conditions are often not satisfied. Researchers frequently work with sample sizes smaller than those assumed by asymptotic theory, as well as with complex models containing many observed variables, non-normal data, and missing responses. When this mismatch between assumptions and data is not acknowledged or addressed, it can lead to

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

estimation problems and misleading conclusions, even when the model appears to fit the data well.

In light of these considerations, one of the first assumptions to consider concerns the distribution of the observed variables and the choice of estimator. When using normal-theory maximum likelihood (ML) estimation, an important assumption is that the observed continuous variables follow a multivariate normal distribution. Data normality can be examined using univariate and multivariate indicators such as skewness and kurtosis. When this assumption is violated for continuous outcomes, robust corrections such as the Satorra–Bentler scaled chi-square and robust standard errors provide a more appropriate basis for inference (e.g., Marcoulides et al., 2023). For models involving categorical observed variables, estimation approaches based on diagonally weighted least squares, such as the weighted least squares mean and variance adjusted (WLSMV) procedure, are generally preferred (e.g., Kline, 2023).

An important implication of estimator choice is that both the relevant distributional assumptions and the available options for handling missing data differ across modeling frameworks. For models estimated using WLSMV, normality is not assumed for the observed categorical variables themselves, but rather for the latent response variables that are assumed to underlie these observed responses (Li, 2015). Latent variable normality cannot be directly assessed using empirical diagnostics in the same way as observed-variable normality; instead, its implications must be evaluated indirectly through model fit, robustness checks, and transparent reporting of modeling assumptions. Estimator choice also constrains how missing data can be handled. Full information maximum likelihood (FIML) is available when using normal-theory ML estimation and provides an effective approach for handling data assumed to be missing completely at random (MCAR) or missing at random (MAR). When models involve categorical observed variables and are estimated using WLSMV, FIML is typically not available; in such cases, alternative approaches such as pairwise maximum likelihood can be used to address missing data under MCAR or MAR assumptions (Katsikatsou et al., 2022).

Beyond estimator-specific assumptions, SEM also relies on a number of general statistical assumptions that it shares with other modeling approaches based on the general linear model. One key assumption is the independence of observations, meaning that each participant's responses are assumed to be unrelated to those of other participants. This assumption is often violated in applied research, for example, when data are collected from clustered settings such as students nested within classes, schools, or institutions, or when convenience or snowball sampling is used. When such dependencies are ignored, standard errors and test statistics may be biased, leading to incorrect inferences. Another common concern is sensitivity to outliers, which can disproportionately influence parameter estimates and model fit. In addition, SEM assumes homoskedasticity, or equal error variances across observations or groups. When heteroskedasticity is present, standard errors may be inaccurate, even if the overall model appears to fit the data well. In cases where these assumptions are violated, alternative approaches such as multilevel SEM or the use of cluster-robust standard errors and test statistics can be used to address these issues.

Another consideration that directly affects estimation quality and model evaluation concerns sample size. Many SEM estimation methods rely on large-sample properties, and when sample sizes are small relative to model complexity, parameter estimates, standard errors, and fit statistics may become unstable (Appelbaum et al., 2018; Feng & Hancock, 2023; Morrison et al., 2017). In applied research, practical constraints such as limited access to participants, time, or funding often restrict achievable sample sizes. As

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

a result, researchers have relied on simple heuristics, such as minimum sample sizes of 100 or 200 (Boomsma, 1985) or the rule of ten observations per variable (Everitt, 1975). However, such rules of thumb provide only rough guidance, as the required sample size depends on multiple interacting factors. Simulation research has shown that sample size requirements vary as a function of factor loadings, the number of indicators per construct, and the presence of missing data. For example, Wolf et al.'s (2013) simulation showed that in a one-factor model with factor loadings of .50, a minimum sample size of approximately 200 was required when the factor had 3 or 4 indicators, whereas a sample of 100 was sufficient when the factor had 6 or 8 indicators. When factor loadings were .80, the required minimum sample size dropped to approximately 50 across all tested conditions. Moreover, the presence of missing data further increased the required minimum sample size, ranging from 200 to 300, depending on the extent of missingness. These findings highlight that sample size should be considered in relation to the specific model and data characteristics rather than evaluated against fixed numerical thresholds.

Given this complexity, sample size decisions are better treated as part of the overall model planning rather than evaluated using fixed cutoff values. One practical approach is to use simulation-based power analysis, which allows researchers to examine how a specific model is expected to perform under conditions that are similar to the intended study (Feng & Hancock, 2023). Through simulation, it is possible to evaluate power for overall model fit as well as for individual parameters, while taking into account features such as model structure, effect sizes, and missing data. The simsem package in R (Pornprasertmanit et al., 2021) provides one way to conduct such analyses across different SEM conditions. Using this approach, researchers can make more informed decisions about sample size and better evaluate whether the assumptions underlying their analyses are likely to hold (see also online supplementary materials).

> *Recommendation 1:* Researchers should systematically assess and report the statistical issues underlying SEM, including data normality, estimator appropriateness, and sample size adequacy. Rather than relying on fixed rules of thumb, sample size decisions should be guided by model-specific factors and informed through simulation-based power analysis when possible.

### 2) Not validating a measurement model

The use of psychometrically sound measures is vital (e.g., Kline, 2023; Ockey & Choi, 2015). In the context of L2 research, Plonsky and Derrick (2016) examined the percentage of reliability estimates reported across various L2 subdomains, revealing that reliability was documented in 6% to 64% of studies (see also Kostromitina et al., 2025; Sudina, 2021, 2023). This suggests that reliability reporting varies across subdomains. When two measures exhibit a low correlation, it is unclear whether the constructs themselves are weakly related or whether poor measurement reliability is responsible for the weak association.

In SEM, analytical models are generally composed of measurement models and structural models. Measurement models specify the relationship among observed variables and latent (unobserved) variables, thereby assessing the extent to which constructs are effectively operationalized and measured through their indicators. Structural models, in contrast, focus on the relationships among latent constructs. Given that researchers are often primarily interested in how constructs are related, structural models are often a focal point in SEM analyses.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

For structural models to be interpreted with confidence, however, the measurement model must first demonstrate good fit to the data. Poor measurement model fit suggests that the constructs may not be measured appropriately, which in turn calls into question the validity of interpreting relationships among constructs in structural models. This is analogous to ensuring that instruments such as tests and questionnaires are valid and reliable before interpreting their scores or relationships among them.

One practice that can complicate model evaluation involves modeling the measurement and structural models simultaneously rather than sequentially (the latter of which involves testing the measurement models independently before testing the structural models). This practice leads to less transparency because, if a simultaneous model shows poor fit, it is unclear whether the issue originates from the measurement models, the structural models, or both. In contrast, when researchers first ensure that the measurement models fit well, they can more confidently attribute any remaining issues to the structural models, thereby refining their theoretical understanding.

Questions about measurement model quality also extend to how reliability is assessed within SEM. Computing Cronbach's alpha is not a requirement when using SEM. In fact, it is often better to use alternative approaches to reliability that are more suitable for SEM. Some reliability measures, such as Cronbach's alpha coefficient, rely on the assumption of equal item precision (tau-equivalence), which in factor-analytic terms corresponds to equal factor loadings, an assumption that is often violated in practice (Raykov, 2004). SEM-based approaches, particularly those from confirmatory factor analysis (CFA), offer more accurate alternatives. One such approach is Composite Reliability (CR), which is a more accurate estimate of internal consistency because it accounts for the actual factor loadings of each item (Fornell & Larcker, 1981). For further details on Cronbach's alpha and CR, see Peterson and Kim (2013) and Raykov (1998). Another approach is Average Variance Extracted (AVE), which assesses the amount of variance captured by the construct relative to measurement error (Fornell & Larcker, 1981). Using CR and AVE enables researchers to better assess a scale's reliability and convergent validity within the specific context of their SEM model, and AVE is also commonly used as part of procedures for evaluating discriminant validity.

*Recommendation 2:* Because the validity of the structural model relies on the quality of the measurement model, researchers should establish and report evidence for a well-fitting measurement model before proceeding to test structural relationships.

### 3) Not testing competing models

In many cases, it is possible to think of alternative models that might explain the data equally well, if not better. Not testing these models gives the reader no choice but to accept the model hypothesized by the authors. This is problematic because the author's model is usually only one of the different models potentially supported by the data. Comparing multiple models allows researchers to critically assess the theoretical framework and its application within the model (e.g., Yun et al., 2018). Metrics such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can help compare competing models to determine which one provides the best fit to the data while maintaining parsimony. Selecting the most appropriate model after empirically and theoretically eliminating competing alternatives is more meaningful than testing and selecting a single model in isolation (Joe et al., 2017).

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

Testing competing models allows researchers to evaluate both the strength of the underlying theoretical rationale and its statistical implementation. For instance, in their review of the L2 Motivational Self System, Hiver and Al-Hoorie (2020) argued that the three components of the model—the ideal L2 self, the ought-to L2 self, and the L2 learning experience—may be modeled in various ways. Figure 1 presents four alternative models. They can be modeled as (a) a correlated three-factor model, (b) a correlated three-factor model with cross-loadings, (c) a bifactor model without cross-loadings, or (d) a bifactor model with cross-loadings. In the latter two cases, a general factor exerts influence on all items in the dataset. In some cases, cross-loadings may be theoretically expected because, for example, motivational self-constructs are conceptually related, and items may reflect overlapping aspects of learners' self-guides. Detailed treatments of competing model evaluation within SEM, including alternative measurement specifications, are found in Morin (2023) and Morin et al. (2020).

Equally important to consider is how to compare models. One key issue is the evaluation of measurement equivalence/invariance. Measurement invariance testing—conducted by comparing nested models (a less constrained vs. a more constrained model) and assessing changes in fit indices such as root mean square error of approximation (RMSEA) and comparative fit index (CFI)—is a widespread application in SEM and an important step in ensuring that constructs are measured equivalently across different conditions (e.g., Kline, 2023). Such conditions may include groups (e.g., L1 vs. L2 learners), methods (e.g., questionnaire vs. interview formats), and time points (e.g., Time 1 vs. Time 2). Failing to conduct or report invariance testing can lead to misleading conclusions. Savalei et al. (2024) review ongoing debates on this issue and recommend the use of RMSEA values associated with the chi-square difference test, rather than relying solely on the chi-square difference test or on changes in RMSEA.

*Recommendation 3:* Researchers should aim to develop multiple theoretically defensible models a priori (including invariance testing), compare them systematically, and report all results transparently.

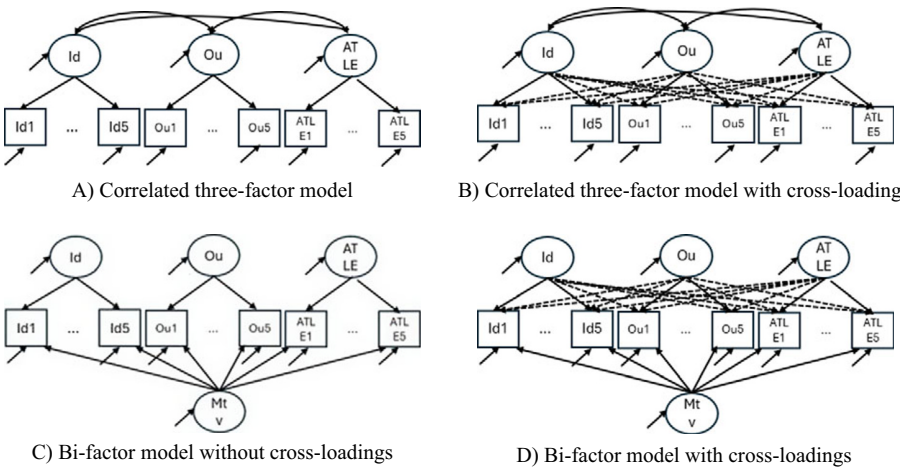



**Figure 1.** Four competing models of L2 Motivational Self System. Id = ideal L2 self, Ou = ought-to L2 self, ATLE = attitude toward L2 learning experience, Mtv = Motivation.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

### 4) Making insufficiently justified modeling decisions

SEM provides researchers with a flexible framework to model relationships among variables. While SEM has been categorized into exploratory[1] and confirmatory approaches in the literature, most applications fall somewhere in between. It is also possible to use both approaches with the same data set (Morin, 2023). The exploratory approach relies on data-driven decisions to model relationships between variables without a strong theoretical basis. For example, a researcher may specify that a latent variable influences all observed variables. They may also specify directional, causal relationships among latent variables without prior theoretical justification. In contrast, the confirmatory approach is guided by theoretical foundations. This approach may involve specifying that a latent variable influences only certain observed variables rather than all of them. Similarly, it may define directional relationships based on prior theory, ensuring one latent variable influences another but not vice versa. If the model-building process is not clearly reported, readers will be left uncertain about how constructs have been operationalized in SEM and how trustworthy the results are. Nevertheless, although SEM can be used in both an exploratory and confirmatory manner, we stress prioritizing the confirmatory approach when possible. One reason is that SEM allows users to take a theory-based, causal approach. This is particularly evident when an arrow is drawn from one variable to another, indicating a directional, cause–effect relationship that should be grounded in theory or prior research (see Al-Hoorie & Hiver, 2025, for more detail on causal inference from observational data).

It is worth noting that exploratory research, when conducted ethically, does not in itself constitute a QRP. The critical issue lies in ensuring that post hoc modeling decisions remain theory-guided and data-supported, rather than merely data-driven, while simultaneously considering how empirical evidence can inform the refinement of theory. As discussed above, research rarely falls neatly into purely exploratory or purely confirmatory categories; instead, it spans a continuum between these extremes. The most reliable scientific inferences arise not when data dictates questions or hypotheses after the fact, but when theoretically grounded questions are posed in advance and then empirically tested in order to minimize the risk of post hoc significance chasing. Any data-informed decisions about refining a theoretical model, even when conducted rigorously, require newly formulated hypotheses to be tested on new data, as evaluating such hypotheses using the same data that generated them does not constitute an independent test. This process, commonly referred to as cross-validation, is essential for drawing robust scientific inferences. One way to further increase transparency is preregistration, which allows researchers to make key modeling decisions explicit before data analysis.

As an illustration, in their analysis of studies on the L2 Motivational Self System, Hiver and Al-Hoorie (2020) reviewed how different studies approached causal modeling in SEM. For example, Taguchi et al. (2009) examined the influence of (a) ideal L2 self on (b) L2 learning experience (which they called "attitudes toward learning English"), whereas Kormos and Csizér (2009) investigated the impact of (b) L2 learning experience on (a) ideal L2 self. Despite using the same constructs, these studies specified opposite causal directions. Given that both models demonstrated data fit, this may suggest the possibility of bidirectional causality: a → b and b → a. Alternatively, as Hiver and Al-Hoorie (2020) argued, the findings may also reflect reciprocal causality,

[1]The exploratory approach is not the same as exploratory structural equation modeling (ESEM). See Wei et al. (2022) for a comparison among conventional SEM, ESEM, and Bayesian SEM.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

wherein both variables exert mutual influence (e.g., Kline, 2023): a ⇄ b. It is also possible that no causal relationship exists, and the observed association is simply caused by a third variable (see Al-Hoorie & Hiver, 2025).

Among the three types of causality, the first two cases, a → b and b → a, produce the same statistical results under specific conditions, because the models are statistically equivalent in terms of the model-implied covariance matrix and mean vector (see online supplementary materials). This equivalence reinforces the importance of robust theory over statistical results alone. For example, Hiver and Al-Hoorie (2020) demonstrated that two models describing the relationship between (a) ideal L2 self and (b) disappointment in daily life produced identical results (i.e., path coefficients and fit indices; see Figure 2), even though the direction of the path between these two variables was reversed. In one model, the ideal L2 self appeared to be influenced by disappointment, whereas in the other model, the ideal L2 self itself seemed to influence disappointment. This result arises because only the directional path between the predictor and mediator was changed, while the remainder of the model remained unchanged. In contrast, reversing the direction of the path between a mediator and an outcome would affect parameter estimates, even though the models are still statistically equivalent and produce identical model-implied covariance matrices and mean vectors. Thus, obtaining an acceptable model fit does not support a unique, causal interpretation, stressing the need for theoretical justification when specifying paths.

Stronger causal inference in SEM depends not only on statistical modeling but also on theory and research design, including careful consideration of confounders, mediators, and colliders (Al-Hoorie & Hiver, 2025). Longitudinal designs are particularly useful because they impose temporal ordering and allow researchers to examine change over time. Models such as the random-intercept cross-lagged panel model, the stable trait, autoregressive trait, and state model, and related dynamic panel approaches provide tools for evaluating directional hypotheses using repeated measures (Lucas, 2023). However, even with longitudinal data, causal interpretations cannot be based on model fit alone and must remain grounded in theory and explicit modeling assumptions.

> *Recommendation 4:* Every modeling decision—including the selection of variables, the specification of covariances, and the direction of paths—should be grounded in theory and explicitly justified.

### 5) Modifying the model post hoc

Unlike insufficiently justified modeling decisions discussed earlier, where the researcher specifies the model without theoretical justification, here the researcher modifies or respecifies the model based on the results. This is often achieved using modification indices to enhance model fit. For instance, modification indices may indicate the need to add, remove, or modify paths between variables, correlate residual errors, or incorporate cross-loadings. While these changes can enhance model fit over several iterations, implementing them without a strong theoretical basis compromises the validity of the model and may constitute a QRP. Rather, such modifications should serve as signals that encourage the researcher to reconsider the theoretical model.

A key distinction must be drawn between data-driven decisions and data-supported theoretical updates. In data-driven model modification, researchers revise the model primarily in response to statistical patterns (e.g., adding paths with large modification indices), which can encourage ad-hoc rationalizations and weaken the confirmatory interpretation of SEM results. In contrast, theory-based refinement treats statistical

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press



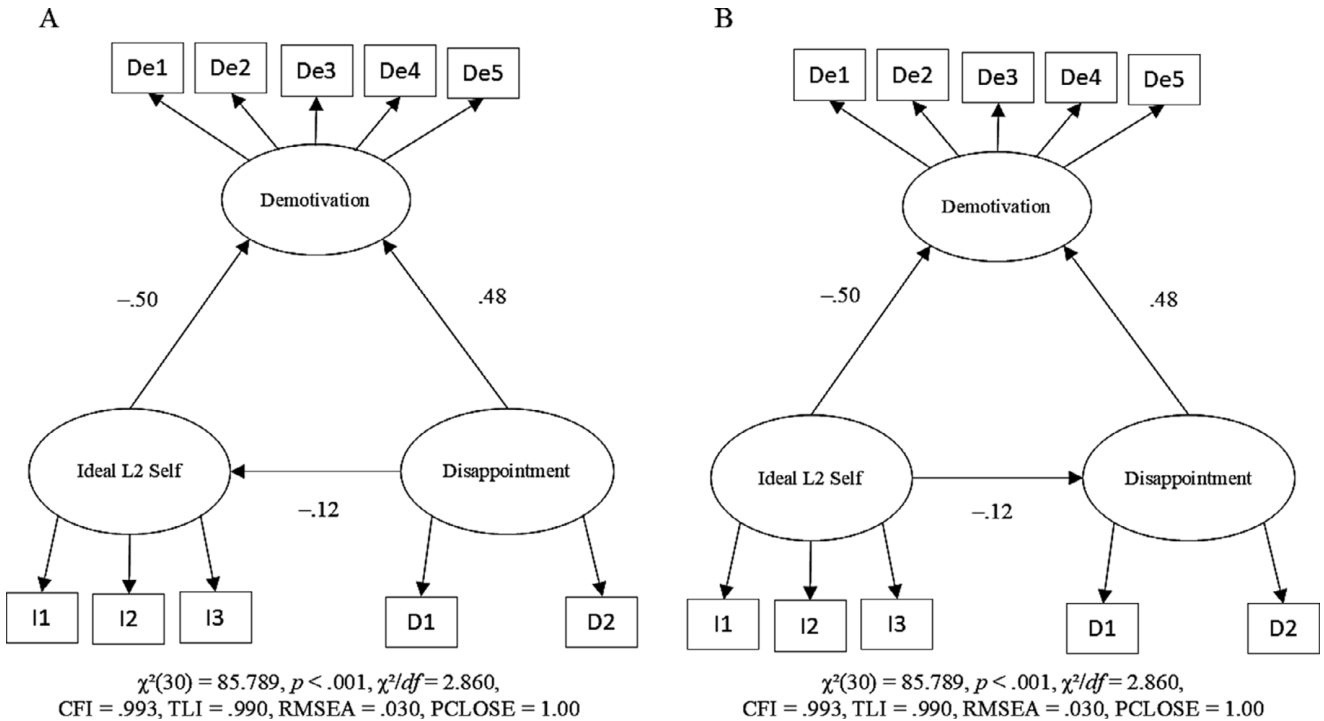



**Figure 2.** Two models with different causal specifications resulting in identical structural coefficients and identical model fit. Error terms have been removed for simplicity.

indications as prompts to reconsider the theoretical model. For example, patterns in correlation residuals may motivate a theoretically grounded explanation for why certain variables are more or less strongly related than expected. Only when such an explanation is plausible should a corresponding modification be considered. Even when modifications are introduced post hoc, this reasoning can be made more rigorous by asking whether the same path or covariance would have been specified if the model were formulated a priori. In some cases, alternative or nested models can also be specified in advance, allowing later model comparisons to be treated as confirmatory rather than purely exploratory.

Removal of parameters or variables solely on the basis of statistical non-significance or small sizes is similarly not justifiable. If these elements are theoretically relevant to the construct, as hypothesized a priori, they should be retained in order to align with the confirmatory nature of the model. Their exclusion may result in an underrepresentation of the construct, thereby diminishing the model's interpretability and theoretical validity.

Further, making post hoc modifications to models can result in overfitting, where the model fits the current dataset well but may not generalize to new data. This may raise concerns in cross-validation and the broader applicability of the substantive findings (e.g., West et al., 2023). Given that the modified model deviates from the original, theory-driven model, it should ideally be tested with an independent dataset to determine the extent to which it remains supported (e.g., Kline, 2023; Leeming et al., 2024).

Therefore, post hoc modifications should be theoretically justified. Because statistical software provides these suggestions based on statistical rather than theoretical considerations, they should be carefully reviewed and applied only when they hold substantive meaning. Otherwise, the model may achieve a strong statistical fit but lack substantive value. While we emphasize the risks associated with post hoc modifications, we also acknowledge that data often challenge theoretical expectations and that exploratory insights can play a constructive role in theory development. It is important to recognize, however, that post hoc model modifications generate new hypotheses using information from the data. As mentioned above, these hypotheses require testing on new data, as evaluating them on the same dataset does not constitute a valid confirmatory test. Adopting a transparent and self-skeptical approach helps maintain scientific integrity while still allowing theory to be refined and extended.

*Recommendation 5:* Post hoc modifications—whether additions or deletions—should be minimized, informed by theory, and reported transparently. Findings from post hoc modifications should be tested against new data.

### 6) Overemphasizing global model fit indices

Model fit in SEM is assessed at both global and local levels. Overemphasizing global fit refers to evaluating model adequacy primarily or exclusively on the basis of global fit indices, while neglecting residuals, localized misfit, and substantive theory. Global fit evaluates the overall alignment of the model with the data, relying on indices such as the chi-square test statistic (and its *p*-value for given degrees of freedom) and several indices of approximate fit: RMSEA, CFI, and standardized root mean square residual (SRMR). Local fit, by contrast, assesses the degree to which the model adequately reproduces relationships between each pair of variables. While global fit provides a

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

broad measure of model-data consistency, it does not diagnose the specific areas where misfit may occur. Local fit addresses this limitation by examining the individual associations within the model, allowing researchers to pinpoint and help address discrepancies. It is typically evaluated through standardized residuals or residual covariance matrices, which help identify specific variable pairs or parameters that contribute disproportionately to overall misfit. Given that global and local fit statistics assess models from different perspectives, both should be incorporated into model evaluation (e.g., Appelbaum et al., 2018; Kline, 2023, 2024). While global fit indices are more commonly reported, disregarding local fit may result in models that appear well-fitting at a global level but contain numerous localized inconsistencies. Models with statistically significant global misfit might be rejected without considering local fit indices, which could indicate a major discrepancy (or several of them), the accumulation of many minor (potentially ignorable) discrepancies, or both, thereby providing a more nuanced understanding of data–model fit.

Kline (2023) recommends an approach to model evaluation that incorporates both global and local fit. If the chi-square test is statistically significant, indicating that the model may not fit the data well, tentatively reject the model and examine the residuals (i.e., local fit) to identify discrepancies between the model and the data. If no substantial discrepancies are detected, evaluate the model substantively and empirically. For empirical evaluation, consider RMSEA, CFI, and SRMR. If the chi-square is statistically non-significant, suggesting that the model may fit the data, inspect the residuals. If no major inconsistencies are found, evaluate the model substantively and empirically using these fit indices.

Kline's (2023) approach to model evaluation merits further discussion in two points. First, regardless of whether the chi-square is statistically significant, it is critical to examine residuals. Residual analysis identifies differences between the model and the data at the level of individual variable pairs, allowing researchers to locate misfit and modify the model accordingly. Since global fit indices are not designed to reveal these discrepancies, a model should be rejected even if the chi-square is non-significant if large residuals are present (for more on residual analysis, see Kline, 2024). Second, for models with statistically significant chi-squares, Kline (2023) proposed to tentatively reject the model and inspect the residuals. This approach places less emphasis on statistical significance and instead encourages researchers to diagnose whether misfit is due to large residuals in specific parts of the model.

It is important to clarify the interpretation of chi-square significance. The chi-square test does not detect sample size; rather, when the null hypothesis of exact fit is false, the test becomes increasingly capable of detecting even very small discrepancies as sample size increases (e.g., West et al., 2023). This behavior is not unique to SEM: for example, an independent-samples *t*-test gains power to detect increasingly small differences as sample size increases. In other contexts, researchers accompany statistical tests with effect-size measures to distinguish statistically significant but practically negligible results. In SEM, approximate fit indices were originally intended to serve this function by providing information about the magnitude of model misfit, rather than replacing statistical tests or local fit diagnostics. A common misunderstanding—that a significant chi-square "reflects sample size more than misfit"—may have led researchers to dismiss chi-square significance entirely. Recognizing that chi-square significance provides meaningful information about model misfit, while approximate fit indices and residuals provide complementary perspectives on its magnitude and location, helps avoid this misunderstanding and supports more defensible model evaluation.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

Relatedly, evaluating models rigidly based on fit index cutoff criteria is not a recommended approach. This is because the studies that established these thresholds may not be directly applicable to all research contexts. For instance, Hu and Bentler's (1999) recommended cutoffs—CFI close to .95, RMSEA close to .06, and SRMR close to .08—are frequently used to evaluate models somewhat uncritically. However, these cutoffs were derived from a simulation study using confirmatory factor models with 3 factors and 15 observed variables, tested under specific conditions. Unless a researcher's model and study conditions are similar to those in Hu and Bentler's (1999) study, these cutoffs should not be applied uncritically. In fact, Hu and Bentler (1999) viewed these values as general guidelines rather than absolute rules. Furthermore, the cutoffs were not supported in later studies, questioning their applicability across different modeling conditions (e.g., Fan & Sivo, 2005; Yuan, 2005).

To address these challenges, one approach is to examine dynamic, study-specific cutoffs rather than strictly adhering to Hu and Bentler's (1999) criteria. Simulation-based approaches can be used not only for power and sample size planning, but also to evaluate how fit indices are expected to behave under a given model, sample size, and degree of misspecification. This strategy does not diminish the importance of local fit inspection or theoretical justification in model evaluation (see McNeish & Wolf, 2023, for further discussion; see also online supplementary materials).

> *Recommendation 6:* Researchers should avoid overreliance on global fit indices and instead adopt a comprehensive model evaluation strategy that includes both global and local fit, careful inspection of residuals, and theoretical justification. Fixed cutoff values should be interpreted cautiously and in light of model complexity, degrees of freedom, and sample size, rather than applied rigidly across contexts.

### 7) *Incomplete or nontransparent reporting*

As discussed above, SEM provides a flexible framework for modeling relationships among variables. This flexibility extends to the wide range of output options available, which depend on the type of model analyzed and the additional details requested. For example, models with a large number of variables produce extensive lists of parameter estimates. If the dataset contains non-normal data, the output may include results both with and without adjustments for nonnormality. While these analytical capabilities allow for more comprehensive analyses, they also introduce challenges in presenting results clearly. This complexity increases the risk of selective reporting, where findings that align with the researcher's hypothesis are prioritized while contradictory results are omitted. Selective reporting, as well as incomplete or nontransparent reporting of the issues discussed above, compromises the transparency and rigor of model evaluation and interpretation.

Researchers should therefore strive for complete and transparent reporting. Rather than relying solely on data-driven decisions, they should provide theoretical justifications for model specification. Modifying the model based on results is appropriate only if the changes are substantively supported and transparently documented. Testing measurement models is an important step in model development and should be systematically reported. Evaluating competing models allows researchers to test hypotheses, and documenting this process contributes to theoretical refinement. Reporting both global and local fit statistics indices provides a comprehensive assessment of model–data fit. Because fixed cutoffs for fit indices may not always be

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

appropriate in a given study context, incorporating dynamic cutoffs can offer additional insights. Lastly, reporting whether the necessary assumptions for SEM were met —and describing how any violations were addressed—enhances the credibility of the research.

Ensuring complete and transparent reporting is a fundamental principle of research (Al-Hoorie et al., 2024; Liu et al., 2023). However, practical constraints, such as word limits, may restrict the amount of information that can be included in published studies. Even when supplementary files are provided or data are made available through repositories such as IRIS (https://www.iris-database.org) and Open Science Framework (OSF; https://osf.io/), essential details may still be missing, or additional information may be required for future analyses. Moreover, some readers may seek to reproduce the reported models. To enhance reproducibility, researchers are encouraged to share analysis code, data, and output files whenever feasible (though see Chapelle & Ockey, 2024, for potential drawbacks associated with this practice).

Ethical concerns may prevent researchers from sharing raw data. However, they can still provide summary statistics, such as means, standard deviations, and correlations, which allow other researchers to reproduce the reported model. SEM software calculates these values during analysis, meaning that summary statistics can sometimes be sufficient for reproducing models. Thus, reproducing SEM results without raw data is possible under certain conditions, such as when the dataset is normally distributed and complete (i.e., contains no missing responses). If these conditions are not met, raw data are necessary to accurately reproduce the original model. To determine whether a model can be reproduced using summary statistics alone, researchers can compare the consistency of results derived from both raw data and summary statistics-based analyses. Regardless of whether raw data can be shared, reporting summary statistics is strongly recommended, as it provides readers with a clear overview of the dataset's characteristics.

A recent example is provided by Nagle (2023), who shared data, analysis code, and output files. In examining longitudinal change in individual difference variables, the study addressed the issue of longitudinal measurement invariance. To illustrate this approach in motivational research, the author tested invariance for the ought-to L2 self using data from 141 U.S. university students collected at three time points across a semester (beginning, midterm, and end). Results supported strong invariance, indicating the same factor structure, equal loadings, and equal intercepts across time. These findings suggest that the ought-to L2 self was measured consistently, and that observed changes reflected true change in the construct rather than changes in measurement. We ran the R code and successfully reproduced the results reported in the study.

Another example is Suzuki and Kormos (2023), who examined the relationship between cognitive fluency and utterance fluency in second language oral proficiency among 128 Japanese learners of English. After examining measurement models for both constructs, they tested a structural model linking them. One key finding was that speed fluency—one component of utterance fluency—was related to linguistic resources—one component of cognitive fluency—only in certain task types (e.g., reading-to-speaking), highlighting the moderating role of task characteristics. Using the shared materials, we successfully reproduced the reported analyses and results.

> *Recommendation 7:* Researchers should ensure complete and transparent reporting of SEM procedures and results, including model specifications, assumptions, fit indices, and any post hoc modifications. Whenever feasible, they should share data, analysis code, and output files to enhance reproducibility.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

**Table 1.** SEM QRPs and Best Practice Recommendations

| QRP | Corresponding Best Practices |
|---|---|
| 1. Not checking assumptions | Researchers should systematically assess and report the statistical issues underlying SEM, including data normality, estimator appropriateness, and sample size adequacy. Rather than relying on fixed rules of thumb, sample size decisions should be guided by model-specific factors and informed through simulation-based power analysis when possible. |
| 2. Not validating a measurement model | Because the validity of the structural model relies on the quality of the measurement model, researchers should establish and report evidence for a well-fitting measurement model before proceeding to test structural relationships. |
| 3. Not testing competing models | Researchers should aim to develop multiple theoretically defensible models a priori (including invariance testing), compare them systematically, and report all results transparently. |
| 4. Making insufficiently justified modeling decisions | Every modeling decision—including the selection of variables, the specification of covariances, and the direction of paths—should be grounded in theory and explicitly justified. |
| 5. Modifying the model post hoc | Post hoc modifications—whether additions or deletions—should be minimized, informed by theory, and reported transparently. Findings from post hoc modifications should be tested against new data. |
| 6. Overemphasizing global model fit indices | Researchers should avoid overreliance on global fit indices and instead adopt a comprehensive model evaluation strategy that includes both global and local fit, careful inspection of residuals, and theoretical justification. Fixed cutoff values should be interpreted cautiously and in light of model complexity, degrees of freedom, and sample size, rather than applied rigidly across contexts. |
| 7. Incomplete or nontransparent reporting | Researchers should ensure complete and transparent reporting of SEM procedures and results, including model specifications, assumptions, fit indices, and any post hoc modifications. Whenever feasible, they should share data, analysis code, and output files to enhance reproducibility. |

Table 1 provides an overview of the main QRPs discussed in this article, along with corresponding recommendations.

## Conclusion

In this article, we have addressed analytical transparency in the context of SEM. Specifically, we have reviewed best practices in SEM by synthesizing insights from relevant methodological literature. While these best practices have been well documented (e.g., Hancock & Mueller, 2013; Hoyle, 2023), limited attempts have been made to explicitly integrate these practices with the principles of open science.

We chose to center the term fit-hacking (Flores-Kanter & Mosquera, 2023) in the title because model fit remains the final arbiter in determining whether SEM results are deemed acceptable or publishable. While QRPs in SEM may take many forms—ranging from post hoc modifications to selective reporting and under-disclosure—they are often motivated by a desire to achieve acceptable model fit indices. In this broader sense, fit-hacking is not a single technique but a unifying concept that captures the *end goal* of many questionable modeling behaviors. Placing fit-hacking at the conceptual center of our discussion underscores the need to reconsider the overreliance

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

on model fit as a proxy for research quality and calls for a more balanced and transparent approach to model evaluation.

Although we have discussed seven QRPs related to SEM (Table 1), they represent only some of the key considerations for ensuring the responsible use of SEM. Other QRPs not discussed in this paper may be just as important and deserve further exploration as they become more widely utilized in the field. They should be investigated in future research to advance transparency, enhance best practices in SEM, and improve the trustworthiness of insights derived from SEM-based studies.

Furthermore, we emphasize that our goal is not to discourage researchers from using SEM, but to discourage its use *without sufficient training* and careful consideration of its assumptions and reporting requirements. As SEM becomes increasingly accessible through user-friendly software and brief tutorials (and now AI), there is a growing risk that powerful modeling frameworks are applied without adequate understanding or safeguards. At the same time, many of the open science concerns discussed in this piece have only gained prominence in recent years and are therefore often not addressed explicitly in classic methodological references, whether in SEM or in other statistical traditions. Experiences from neighboring fields suggest that a lack of transparency, together with growing flexibility in statistical analyses, contributed to later concerns about reproducibility and trust in published research.

Finally, in offering such guidance above, we do not intend to imply that researchers in the field engage in poor SEM practices or that findings from previous SEM-based studies are necessarily questionable. Some techniques we reviewed have only recently been developed or become widely accessible. Furthermore, the emphasis on open science, research misconduct, and QRPs has itself only recently gained prominence within the field. Indeed, we acknowledge that some of the recommendations discussed in this paper were not implemented in our own past research for these very reasons. While QRPs, including those discussed here, may be unlikely to disappear in the immediate future, we hope they will become gradually rarer over time.

**Data availability statement.** The experiment in this article earned Open Data and Open Materials badges for transparent practices. The data are available at https://osf.io/xmk5b.

**Competing interests.** The authors declare that there are no competing interests.

## References

Al-Hoorie, A. H., Cinaglia, C., Hiver, P., Huensch, A., Isbell, D. R., Leung, C., & Sudina, E. (2024). Open science: Considerations and issues for TESOL research. *TESOL Quarterly*, *58*(1), 537–556. https://doi.org/10.1002/tesq.3304

Al-Hoorie, A. H., & Hiver, P. (2024). Open science in applied linguistics: An introduction to metascience. In L. Plonsky (Ed.), *Open Science in Applied Linguistics* (pp. 17–43). Applied Linguistics Press.

Al-Hoorie, A. H., & Hiver, P. (2025). Causality and ability beliefs: An introduction to confounders and colliders. *Studies in Second Language Learning and Teaching*, *15*(2), 227–249. https://doi.org/10.14746/ssllt.48231

Andrieux, P., Leonard, S., Simmering, V., Simmering, M., & Fuller, C. (2024). How cognitive biases influence problematic research methods practices. *Electronic Journal of Business Research Methods*, *22*(1). https://doi.org/10.34190/ejbrm.22.1.3212

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. https://doi.org/10.1007/BF02294248

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Chapelle, C. A., & Ockey, G. J. (2024). Open Science in language assessment research contexts: A reply to Winke. *Language Testing*, *41*(4), 882–885. https://doi.org/10.1177/02655322241239377

De Costa, P. I., Sterling, S., Lee, J., Li, W., & Rawal, H. (2021). Research tasks on ethics in applied linguistics. *Language Teaching*, *54*(1), 58–70. https://doi.org/10.1017/S0261444820000257

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*(3), 237–240. https://doi.org/10.1192/bjp.126.3.237

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 343–367. https://doi.org/10.1207/s15328007sem1203_1

Feng, Y., & Hancock, G. R. (2023). Power analysis within a structural equation modeling framework. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed., pp. 163–183). Guilford Press.

Flores-Kanter, P. E., & Mosquera, M. (2023). How do you behave as a psychometrician? Research conduct in the context of psychometric research. *The Spanish Journal of Psychology*, *26*, e13, Article e13. https://doi.org/10.1017/SJP.2023.14

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39–50. https://doi.org/10.2307/3151312

Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course* (2nd ed.). IAP Information Age Publishing.

Hancock, G. R., & Schoonen, R. (2015). Structural equation modeling: Possibilities for language learning researchers. *Language Learning*, *65*(S1), 160–184. https://doi.org/10.1111/lang.12116

Hiver, P., & Al-Hoorie, A. H. (2020). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*(1), 48–102. https://doi.org/10.1111/lang.12371

Hoyle, R. H. (Ed.). (2023). *Handbook of structural equation modeling* (2nd ed.). Guilford Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. (2025). An aberrant abundance of Cronbach's alpha values at .70. *Advances in Methods and Practices in Psychological Science*, *8*(1), 1–18. https://doi.org/10.1177/25152459241287123

Isbell, D. R., Brown, D., Chan, M., Derrick, D., Ghanem, R., Gutiérrez Arvizu, M. N., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*(1), 172–195. https://doi.org/10.1111/modl.12760

Jak, S., & Jorgensen, T. D. (2025). *A lavaan compendium for structural equation modeling in educational research*. https://tdjorgensen.github.io/SEM-in-Ed-compendium/index.html

Joe, H.-K., Hiver, P., & Al-Hoorie, A. H. (2017). Classroom social climate, self-determined motivation, willingness to communicate, and achievement: A study of structural relationships in instructed second language settings. *Learning and Individual Differences*, *53*, 133–144. https://doi.org/10.1016/j.lindif.2016.11.005

Katsikatsou, M., Moustaki, I., & Jamil, H. (2022), Pairwise likelihood estimation for confirmatory factor analysis models with categorical variables and data that are missing at random. *British Journal of Mathematical and Statistical Psychology*, *75*(1), 23–45. https://doi.org/10.1111/bmsp.12243

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Kline, R. B. (2024). How to evaluate local fit (residuals) in large structural equation models. *International Journal of Psychology*, *59*(6), 1293–1306. https://doi.org/10.1002/ijop.13252

Kormos, J., & Csizér, K. (2009). Learning experiences, selves and motivated learning behaviour: 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.

Kostromitina, M., Sudina, E., & Baghlaf, E. (2025). Study and instrument quality in perception-based L2 pronunciation research: A methodological synthesis. *Studies in Second Language Acquisition*, Advance online publication. https://doi.org/10.1017/S027226312500018X

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

Larsson, T., Plonsky, L., & Hancock, G. R. (2021). On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, *17*(3), 683–714. https://doi.org/10.1515/cllt-2020-0051

Larsson, T., Plonsky, L., & Hancock, G. R. (2022). On learner characteristics and why we should model them as latent variables. *International Journal of Learner Corpus Research*, *8*(2), 237–260. https://doi.org/10.1075/ijlcr.21007.lar

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics*, *2*(3), 100064. https://doi.org/10.1016/j.rmal.2023.100064

Leeming, P., Vitta, J. P., Hiver, P., Hicks, D., McLean, S., & Nicklin, C. (2024). Willingness to communicate, speaking self-efficacy, and perceived communicative competence as predictors of second language spoken task production. *Language Learning*, Advance online publication. https://doi.org/10.1111/lang.12640

Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

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

Lucas, R. E. (2023). Why the cross-lagged panel model is almost never the right choice. *Advances in Methods and Practices in Psychological Science*, *6*(1), 1–22. https://doi.org/10.1177/25152459231158378

Marcoulides, K. M., Yuan, K.-H., & Deng, L. (2023). Structural equation modeling with small samples and many variables. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed., pp. 525–542). Guilford Press.

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling and bifactor-ESEM. In G. Tenenbaum, & R.C. Eklund (Eds.), *Handbook of Sport Psychology* (4th ed., Vol. *2*, pp. 1044–1073). Wiley.

Morin, A. J. S. (2023). Exploratory structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed., pp. 503–524). Guilford Press.

Morrison, T. G., Morrison, M. A., & McCutcheon, J. M. (2017). Best practice recommendations for using structural equation modelling in psychological research. *Psychology*, *8*(9), 1326–1341. https://doi.org/10.4236/psych.2017.89086

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

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*(3), 305–319. https://doi.org/10.1080/15434303.2015.1050101

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Palazzo, G., Krings, F., & Hoffrage, U. (2012). Ethical blindness. *Journal of Business Ethics*, *109*(3), 323–338. https://doi.org/10.1007/s10551-011-1130-4

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, *98*, 194–198. https://doi.org/10.1037/a0030767

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, Advance online publication. https://doi.org/10.1017/S0267190524000059

Plonsky, L., Brown, D., Chen, M., Ghanem, R., Arvizu, M. N. G., Isbell, D. R., & Zhang, M. (2024). "Significance sells": Applied linguists' views on questionable research practices. *Research Methods in Applied Linguistics*, *3*(1), 100099. https://doi.org/10.1016/j.rmal.2024.100099

Plonsky, L., & Derrick, D. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538–553. https://doi.org/10.1111/modl.12335

Plonsky, L., Larsson, T., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2024). A taxonomy of questionable research practices in quantitative humanities. In P. I. De Costa, A. Rabie-Ahmed, & C. Cinaglia (Eds.), *Ethical Issues in Applied Linguistics Scholarship* (pp. 10–27). John Benjamins.

Pornprasertmanit, S., Miller, P., Schoemann, A. M., Jorgensen, T. D., & Quick, C. (2021). *simsem: SIMulated structural equation modeling.* (Version 0.5-16) [R package]. http://simsem.org

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*, 375–385. https://doi.org/10.1177/014662169802200407

Raykov, T. (2004). Point and interval estimation of reliability for multiple-component measuring instruments via linear constraint covariance structure modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 342–356. https://doi.org/10.1207/s15328007sem1103_3

Revelle, W (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences*, *221*, 112552. https://doi.org/10.1016/j.paid.2024.112552

Savalei, V., Brace, J. C., & Fouladi, R. T. (2024). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*, *29*(3), 480–493. https://doi.org/10.1037/met0000537

Schoonen, R., Cheung, M. W. L., & In'nami, Y. (2025). Structural equation modeling. In C.A. Chapelle, R. Suvorov, & Z. Li (Eds.), *The Encyclopedia of Applied Linguistics (Quantitative Methods)*. Wiley.

Specker Sullivan, L. (2022). Motivated reasoning and research ethics guidelines. *Journal of Applied Philosophy*, *39*(3), 519–535. https://doi.org/10.1111/japp.12577

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, *12*(1), 53–74. https://doi.org/10.1007/PL00022268

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). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455. https://doi.org/10.1017/s0272263122000560

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. https://doi.org/10.1017/S0272263121000899

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.

Wei, X., Huang, J., Zhang, L., Pan, D., & Pan, J. (2022). Evaluation and comparison of SEM, ESEM, and BSEM in estimating structural models with potentially unknown cross-loadings. *Structural Equation Modeling*, *29*(3), 327–338. https://doi.org/10.1080/10705511.2021.2006664

West, S. G., Wu, W., McNeish, D., & Savord, A. (2023). Model fit in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed., pp. 184–205). Guilford Press.

Winke, P. (2014). Testing hypotheses about language learning using structural equation modeling. *Annual Review of Applied Linguistics*, *34*, 102–122. https://doi.org/10.1017/S0267190514000075

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115–148. https://doi.org/10.1207/s15327906mbr4001_5

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

**Cite this article:** Al-Hoorie, A. H., In'nami, Y., & Hiver, P. (2026). Fit-hacking: Questionable research practices in structural equation modeling. *Studies in Second Language Acquisition,* 48: 513–534. https://doi.org/10.1017/S0272263126101569

https://doi.org/10.1017/S0272263126101569 Published online by Cambridge University Press