Group & Organization Management
2025, Vol. 50(4) 1145–1155


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10596011241287574
journals.sagepub.com/home/gom


A Commentary on Bazzoli (2024): Toward a Nuanced and Rigorous Model Evaluation

Yannick Griep[1,2], Wieke M. Knol[1], and William G. Obenauer[3]

Keywords
commentary, model fit, SEM

Introduction

Bazzoli's recent GOMusing (2024) highlighted critical issues in our field's application of structural equation modeling (SEM) fit indices. Specifically, it emphasized the uncritical usage—something we might be guilty of ourselves in some of our papers and teachings—of cutoff values for approximate fit indices, such as TLI (.95), CFI (.95), and RMSEA (.08). Bazzoli makes the timely argument that researchers over-rely on these indices without sufficiently considering the context within which their data were collected or the specific characteristics of the models they are testing. This commentary extends Bazzoli's critique by first substantiating the call for a more nuanced and context-specific approach to model evaluation. It then draws on recent advancements in statistical modeling to propose alternative methodologies that can enhance the rigor and reliability of SEM practices beyond Bazzoli's initial recommendations.

[1]Behavioural Science Institute, Radboud University, Nijmegen, the Netherlands
[2]School of Industrial Psychology and Human Resource Management, North-West University, Potchefstroom, South Africa
[3]Maine Business School, University of Maine, Orono, USA

Corresponding Author:
Yannick Griep, Behavioural Science Institute, Radboud University, Montessorilaan 3, Postbus 9104 Nijmegen 6500 HE, The Netherlands.
Email: yannick.griep@ru.nl

Overview of the Critique of Overreliance on Fixed Cutoff Values

Bazzoli's (2024) critique of the application of rigid cutoff values for fit indices, particularly the .08 threshold for RMSEA, reflects a growing discontent within the academic community regarding the mechanical, and often uncritical, application of these thresholds. When originally proposed by Hu and Bentler (1999), these thresholds were based on a specific set of simulation studies and were accompanied by warnings against overgeneralizing their recommendations. However, over time, the .08 or .95 cutoff value has been widely embraced as the golden standard in SEM publications without sufficient consideration of their limitations. Indeed, empirical evidence supports the notion that rigid adherence to these cutoffs can lead to flawed conclusions, such as the rejection of complex but theoretically sound models (Marsh et al., 2004). Moreover, complex models, with many parameters or latent variables, can wrongfully achieve good fit indices as their specified complexity might obscure underlying issues such as multicollinearity or unmodeled heterogeneity (cf. Kenny et al., 2015). Shi and colleagues (2019) provided empirical evidence that as model complexity increases, fit indices such as CFI and TLI become more likely to overestimate model fit. This latter issue is compounded in the context of multilevel modeling, where researchers often continue to use fit indices like CFI, TLI, and RMSEA, which do not account for the nested nature of the data. This practice can lead to misleading information about model fit at the within-person and/or between-person level. Additionally, fit indices can be influenced by sample size and data distribution, which can distort the true fit of a model (Kenny et al., 2015). For example, as sample size decreases, RMSEA values tend to increase, which can falsely suggest poor model fit in small-sample studies. Important to note is that when Hu and Bentler (1999) referred to "small" samples, they were discussing sample sizes of $N < 250$ as "small", based on the findings of their simulation. Followingly, as sample sizes with $N < 250$ are common in our field, the potential for RMSEA values above the .08 threshold to falsely indicate misfit and result in a paper being relegated to the file drawer is likely quite prevalent. Conversely, with very large samples, statistically significant $\chi 2$-values and correspondingly low RMSEA values can occur despite model misfit, leading researchers to erroneously conclude that their model fits the data well (Chen et al., 2008).

Guidance for Applying a Comprehensive Model Evaluation Framework

Bazzoli's (2024) call for a comprehensive approach to model evaluation is a crucial step toward improving the rigor of SEM practices. However, fully realizing this vision requires more specific guidelines and tools for each discussed aspect of model evaluation. First, Bazzoli emphasizes the importance of inspecting *residuals* to ensure that the model adequately captures the data. While this is a sound recommendation, it could be further strengthened by encouraging researchers to report not only the average absolute residual correlation but also the distribution of residuals, focusing on the largest residuals. Reporting the largest residuals is particularly important because it can highlight specific areas where the model fails to capture the data, providing valuable insights for model re-specification. Bentler (2007) suggested that residuals should be small both in average size and in their largest values, and any large residuals should be substantively interpretable and theoretically justifiable.

Second, the use of *modification indices* is a double-edged sword. While they can provide valuable information about potential model improvements, they also carry a significant risk of overfitting. Bazzoli (2024) rightly cautions against the uncritical use of modification indices but could offer more practical guidelines on their responsible use. Researchers should only consider modifications that align with their theoretical framework and cross-validate any changes in an independent sample or through cross-validation. Jöreskog (1993) argued modification indices should be used sparingly and only when they lead to theoretically meaningful model improvements. This cautious approach helps prevent the model from becoming overly complex and overly tailored to a specific dataset, improving generalizability.

Third, *evaluating parameter estimates* is a critical part of comprehensive model evaluation. Bazzoli (2024) suggests parameters should be statistically significant, in the expected direction, and within a plausible range, recommending non-significant parameters be fixed to zero. We advocate a more nuanced approach: non-significant or unexpected effects can still be meaningful, as falsification is key to scientific progress (Popper, 2005). If other indicators—e.g., model complexity, sample size, and residuals—don't suggest issues like multicollinearity or excessive complexity, non-significant parameters may not need removal. Simplifying a model should be done cautiously and with theoretical justification. Precision, reflected in standard errors, should also be considered, as large errors can indicate problems with identification or data quality. Researchers should examine confidence intervals and conduct sensitivity analyses to assess the robustness of estimates.

Finally, Bazzoli (2024) introduces the concept of *fit propensity*, referring to the tendency of certain models to fit data well regardless of their substantive accuracy. This is an important consideration, particularly for models that are known to be flexible or parsimonious, such as bifactor models. To address the issue of fit propensity, researchers could be encouraged to compare their model against alternative models, including more complex or theoretically distinct models, to ensure that the observed fit is not simply a result of the model's inherent flexibility. Falk and Muthukrishna (2023) developed an R package for assessing fit propensity, but for models not covered by this package, researchers can manually run fit propensity analyses by simulating data and fitting their model to these simulated datasets. This process provides a sense of whether the model's good fit is due to its structure or its tendency to fit any data well.

Beyond the recommendations provided by Bazzoli (2024), we highlight that in the case of multi-level models, SRMR offers separate estimates of model fit for both the within- and between-person levels. Therefore, while CFI, TLI, and RMSEA may indicate an overall model fit, the SRMR at the within-person level might reveal potential model misspecifications (Hsu et al., 2015). To prevent the misapplication of these indices, it is crucial to emphasize that fit indices are not absolute measures but should be interpreted within the specific context of the study (i.e., underlying assumptions, the nature of the data, and the purpose of the model).

Collectively, the concerns about rigid fit indices and guidelines for a comprehensive approach support the recommendation to report a range of fit indices and discussing their implications in the context of the study rather than relying solely on whether they meet predefined thresholds. For example, McNeish and Wolf (2023) proposed dynamic cutoff values, which are derived from simulations tailored to the specific model and data at hand. This approach allows for more accurate assessments of model fit by considering the unique properties of each study, rather than relying on universal thresholds that may not be applicable. One challenge facing management scholars, however, is that when we apply flexibility to these cutoffs, we often neglect to communicate why such flexibility has been applied (e.g., De Cannière et al., 2010; Rice et al., 2024; Trichas et al., 2017; Van Zelderen et al., 2023; Zhan et al., 2022). While this flexibility is not necessarily wrong, these examples illustrate the potential to trade one set of problems (inappropriate application of fit index cutoffs) for another (lack of clarity regarding fit index cutoffs). Therefore, in the next section, we offer an alternative approach that considers options beyond the use of our traditional fit indices.

Beyond Approximate Fit Indices: Bayesian Approach to SEM

Having discussed the limitations of Bazzoli's (2024) critique, we offer Bayesian approaches to SEM as an alternative methodology that can provide a more robust evaluation of model fit. The adoption of Bayesian approaches in SEM represents a significant shift from traditional frequentist methods, providing several critical advantages that can enhance the rigor and reliability of model evaluation. Bayesian statistics operates on the principle of updating the probability of a hypothesis as more evidence or information becomes available. Unlike frequentist approaches, which rely on fixed sample data and often use asymptotic approximations, Bayesian SEM allows researchers to incorporate prior knowledge or expert opinion into the model through prior distributions (AKA "priors") that represent the initial beliefs about the parameters before observing the data. The posterior distribution, which combines the priors with the likelihood of the observed data, provides a comprehensive summary of the parameter uncertainty. This approach is especially advantageous when dealing with small sample sizes or complex models; conditions under which, as discussed above, the frequentist approach is susceptible to yielding unstable estimates and unreliable fit indices. For example, in team research where sample sizes are often small, Bayesian methods can yield more stable and credible estimates by naturally accommodating small samples through the prior distribution. Furthermore, since Bayesian statistics do not rely on asymptotic properties, which become difficult to satisfy as model complexity increases, they are more suitable for complex models.

There are *three key Bayesian fit statistics* that provide a richer framework for evaluating model fit compared to traditional frequentist indices. First, there is the *Posterior Predictive p-value (PPP)*, a Bayesian analog to the traditional *p*-value that does not rely on asymptotic assumptions and is more robust to the small sample sizes and complex models that are common in many practical applications of SEM (Gelman et al., 2013). A PPP value close to 0.5 suggests a good fit, while values closer to zero or 1 indicate potential model misfit. In the Bayesian framework, the PPP measures how well the model predicts the observed data by comparing the observed data to data simulated from the posterior predictive distribution. Second, there is the *Deviance Information Criterion (DIC),* which extends the concept of the Akaike Information Criterion (AIC) used in frequentist statistics but adapted to the Bayesian context. DIC evaluates the trade-off between model fit and complexity, with lower DIC values indicating a better-fitting model. The DIC is useful for comparing multiple competing models, providing a clear criterion for

selecting the model that best balances fit and parsimony. This is crucial in SEM, where models of varying complexity are often compared. Finally, Bayesian SEM provides a rigorous framework for model comparison through *Bayes Factors*. Unlike traditional frequentist methods that might rely on arbitrary cutoff values for fit indices (i.e., .05 or .95), Bayes Factors provide a continuous measure of evidence, allowing for more nuanced and direct model comparisons. This can be useful when different models lead to conflicting conclusions based on traditional fit indices, as Bayes Factors offer a clearer and more consistent basis for model selection.

Empirical studies have demonstrated the effectiveness of Bayesian approaches in various contexts. For instance, a study by Depaoli and Van de Schoot (2017) illustrated how Bayesian methods could be used to improve model estimation and fit in cases where traditional methods fail, such as in models with complex hierarchical structures or when data are missing. The study showed that Bayesian methods could provide more accurate parameter estimates and a better understanding of model uncertainty, highlighting their practical utility in real-world research scenarios. Furthermore, Kaplan and Depaoli (2012) found that Bayesian methods often outperformed traditional approaches to SEM in terms of both parameter recovery and model fit, particularly in smaller samples. These findings underscore the potential of Bayesian SEM to provide more reliable and robust inferences, especially in challenging modeling scenarios.

Integrating Best Practices for Open Science to Mitigate Reporting Challenges

Despite the advantages, Bayesian SEM is not without challenges. One of the most significant challenges in Bayesian statistics is the selection of prior distributions which can bias the results if not carefully chosen, particularly in small samples. When data are scarce, the prior can dominate the posterior distribution, leading to results that are more reflective of the prior beliefs than the observed data (Van Erp et al., 2019). Generally, when there is limited data or theory to inform your priors, the advice is to select weak or uninformative priors. Such priors do not affect the posterior as much, thereby providing more reliable estimates than misinformed priors (Gelman et al., 2013). While this approach is valuable, it does not solve model fit issues pertinent to small samples and complex models. Therefore, to mitigate these issues, we encourage researchers to conduct sensitivity analyses, where different priors are tested to examine how robust the results are to these choices. This process can be time-consuming and requires careful interpretation, as different priors

might lead to different conclusions (Gelman et al., 2013), thus adding an additional layer of complexity to the Bayesian workflow.

Furthermore, Bayesian methods are computationally intensive, and often require sophisticated algorithms like Markov Chain Monte Carlo (MCMC) to estimate the posterior distributions. Ensuring that the MCMC chains have converged to the target distribution can be challenging, particularly in models with high-dimensional parameter spaces or complex hierarchical structures. Non-convergence can lead to biased estimates and incorrect inferences and interpreting convergence diagnostics (e.g., Gelman-Rubin statistic) requires expertise (Brooks & Gelman, 1998). We encourage researchers to apply Bayesian SEM, but we also strongly encourage them to gain practice and develop expertise before using and interpreting it, just as they would when first applying a frequentist approach.

However, there still are the issues of interpretation and communication of Bayesian results over those from frequentist analyses. In Bayesian statistics, results are often expressed in terms of entire distributions rather than single-point estimates. For example, instead of saying that a parameter is estimated to be 5 with a 95% confidence interval, a Bayesian analysis might report a posterior distribution with a 95% credible interval. This difference can be confusing for those more familiar with frequentist statistics, as the interpretation of credible intervals is different from that of confidence intervals (Kruschke, 2021). The unique language used in Bayesian statistics introduces challenges in effectively communicating Bayesian results to a broader audience. This can make it harder to convey results to decision-makers or to integrate Bayesian findings into standard reports and publications (Van de Schoot et al., 2021). The shift in language from frequentist to Bayesian methods should, therefore, be navigated carefully.

Whether using Bayesian models or frequentist statistics, the pressure to publish can lead researchers to prioritize portraying their research as the "one answer" and justifying model specifications over thoroughly evaluating their models. This problem is compounded by a lack of transparency in reporting, where researchers may selectively report results, omit poor fit indices, or engage in other questionable research practices to present their model in the best possible light. To combat these issues, there should be a stronger emphasis on transparency in SEM research that encourages reporting all relevant fit indices, along with residuals, modification indices, parameter estimates, and sensitivity analyses. Few management journals currently request this information explicitly in their submission guidelines.

Additionally, researchers should be encouraged to share their raw data, syntax, and code, enabling others to replicate and extend their analyses. This transparency not only improves the reliability of the findings but also fosters

a culture of openness and collaboration in the field. Tools such as the Open Science Framework (OSF; osf.io) provide repositories where these items, sensitivity analyses, and other resources can be stored free of charge. OSF also provides researchers with the ability to preserve blind review by sharing materials with an anonymized link. Despite the developed infrastructure for sharing research materials, here again we see a slow uptake in management journals requesting said information. While implementing sharing requirements seems like an easy solution, there may be complications associated with such requirements. For example, data sharing policies should explicitly account for restrictions imposed by institutional review boards, organizations providing data access, data subscription services, etc. to avoid potential unintended consequences of universal standards. However, the sharing of data and materials can help to promote reproducibility studies and replication research, which can provide further insight into findings derived from SEM research.

Finally, education and training can also contribute to the more effective use of fit indices. Many researchers may rely on approximate fit indices simply because they are the most familiar and accessible tools. By providing more comprehensive training in SEM, including Bayesian approaches, information-theoretic criteria, and advanced diagnostic tools, researchers can be better equipped to choose the most appropriate methods for their specific research contexts. Relatedly, it is imperative for reviewers and editors to reconsider the heuristics they have relied on for so long in evaluating model fit. Rigid cutoffs, such as .95 for CFI or .08 for RMSEA, can lead to misguided conclusions, especially in complex models with hierarchical structures or multilevel designs. These traditional thresholds do not always capture the full nuance of model fit in contemporary research contexts. As statistical modeling evolves, so too should the criteria for assessing it. To foster more accurate and reliable conclusions, the field must embrace flexible, data-driven approaches to model evaluation, such as those provided by Bayesian frameworks. Advancing the field is thus not just the responsibility of those assessing and reporting model fit, but a collective duty of the entire academic community.

Conclusion

Bazzoli's (2024) article and our commentary on SEM fit indices highlights the complexity and significance of model evaluation in research. As we consider evolving methodologies like Bayesian approaches, it is crucial to avoid simply replacing one set of rigid criteria with another. Instead, we must strive for a nuanced, context-sensitive evaluation that recognizes the limitations of existing indices while promoting transparency, collaboration, and rigorous scrutiny in SEM practices.

Acknowledgments

I would like to thank my daughter Fiona Ava Luz Brys-Griep for allowing me to write this commentary while she was playing outside with the neighbors' kids, every now and then peaking her head in for a hug and to criticize my slow writing "Is that all you did so far, dad? At this rate it is going to take a long time to finish it…" and then storm off again. Gotta love that childish directness!

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Yannick Griep https://orcid.org/0009-0005-3508-5635
Wieke M. Knol https://orcid.org/0000-0003-3805-3792
William G. Obenauer https://orcid.org/0000-0001-9838-1355

References

Bazzoli, A. (2024). Magic number .95? Or was it .08? A refresher on SEM approximate fit indices thresholds for applied psychologists and management scholars. In *Group & organization management*. Advance online publication. Available at: https://doi.org/10.1177/10596011241258314

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829. https://doi.org/10.1016/j.paid.2006.09.024

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Chen, F. F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462–494. https://doi.org/10.1177/0049124108314720

De Cannière, M. H., De Pelsmacker, P., & Geuens, M. (2010). Relationship quality and purchase intention and behavior: The moderating impact of relationship strength. *Journal of Business and Psychology*, *25*(1), 87–98. https://doi.org/10.1007/s10869-009-9127-z

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Falk, C. F., & Muthukrishna, M. (2023). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods*, *28*(1), 123–136. https://doi.org/10.1037/met0000422

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, *50*(2), 197–215. https://doi.org/10.1080/00273171.2014.977429

Hu, L. T., & 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

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Sage.

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). The Guilford Press.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/00491241145432

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

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

Popper, K. (2005). *The logic of scientific discovery.* Routledge.

Rice, D. B., Young, N. C. J., Taylor, R. M., & Leonard, S. R. (2024). *Politics and race in the workplace: Understanding how and when trump-supporting managers hinder black employees from thriving at work*. Advance online publication.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Trichas, S., Schyns, B., Lord, R., & Hall, R. (2017). "Facing" leaders: Facial expression and leadership perception. *The Leadership Quarterly*, *28*(2), 317–333. https://doi.org/10.1016/j.leaqua.2016.10.013

Van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1).

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Van Zelderen, A., Dries, N., & Marescaux, E. (2023). *Talents under threat: The anticipation of being ostracized by non-talents drives talent turnover. Group and Organization Management*. Advance online publication. Available at: https://doi.org/10.1177/10596011231211639

Zhan, Y., Noe, R. A., & Klein, H. J. (2022). How can organizations operating in a negative reputation industry attract job seekers? *Journal of Vocational Behavior*, *132*, 103661. https://doi.org/10.1016/j.jvb.2021.103661

Associate Editor: Thomas Zagenczyk
Submitted Date: August 23, 2024
Revised Submission Date: September 12, 2024
Acceptance Date: September 12, 2024

Author Biographies

Yannick Griep is a prolific academic who spends his time writing papers that—like most scholarly work—will likely be read by only a handful of people, half of whom are desperately trying to avoid doing real work. When he's not meticulously adding to his h-index (because obviously, that's the only number that matters), Yannick enjoys navigating the endless bureaucratic maze of academia, where the joy of discovery is buried under administrative meetings, performance metrics, and the relentless pursuit of funding. He specializes in employee inclusion and workplace dynamics, topics that are very hot right now, not because they lead to lasting change, but because they look great on university brochures. In his spare time, Yannick likes to pretend that his research will revolutionize the world, while secretly hoping for just one email that isn't about committee work or reviewer 2's latest set of inane comments. Also Yannick LOVES Jellybean!!!

Wieke M. Knol is a researcher whose work mainly centers on leadership behaviors, inclusion, cohesion, and interpersonal relationships in the workplace. She is also deeply focused on research methodologies, which is evident both in the design of her studies as well as in the content of her work.

William G. Obenauer is an Associate Professor of Management at the University of Maine and an executive board member for the Advancement of Replications Initiative in Management. He received his PhD from Rensselaer Polytechnic Institute's Lally School of Management. He is best known, however, for his grey parrot, Jellybean.