Qual Life Res (2013) 22:2595–2601
DOI 10.1007/s11136-013-0465-y

COMMENTARY

Strengthening the assessment of factorial invariance across population subgroups: a commentary on Varni et al. (2013)

Cameron N. McIntosh

Accepted: 21 June 2013 / Published online: 9 July 2013

Abstract

Objectives This article provides a commentary in response to "Varni et al. (Qual Life Res. doi:10.1007/s11136-013-0370-4, 2013).“

Methods and results The commentary argues that the approximate model fit indexes commonly used in maximum-likelihood confirmatory factor analysis and factorial invariance testing are seriously flawed, as they overlook potentially serious model misspecifications that could bias parameter estimates and compromise inference.

Conclusions Flexible and convenient Bayesian estimation approaches are presented that can substantially aid in: (1) resolving commonly encountered specification errors in confirmatory factor models and (2) locating specific measurement parameters that are non-invariant across population subgroups. It is recommended that these methods should be more widely adopted for evaluating the factorial invariance of patient-reported outcome measures and other types of instruments.

Keywords Factorial invariance · Fatigue · Patient-reported outcome measures · Quality of life · Bayesian analysis

List of Abbreviations

AFI	Approximate fit index
CFA	Confirmatory factor analysis
HRQoL	Health-related quality of life
MCMC	Markov chain Monte Carlo
SEM	Structural equation modeling
PedsQL^{TM} MFS	Pediatric Quality of Life Inventory^{TM} Multidimensional Fatigue Scale
PPP	Posterior predictive *p* value
PROM	Patient-reported outcome measure

Introduction

With the ongoing proliferation of patient-reported outcome measures (PROMs) for assessing health-related quality of life (HRQoL) in addition to more clinical and objective indicators of intervention impact [1–4], it is essential to continually verify the reliability and validity of these tools within different populations and contexts [5–7]. In particular, establishing that the content of a given PROM is interpreted consistently by respondents across a wide variety of patient subgroupings (e.g., different chronic diseases, socioeconomic strata, ethnicity, age, and gender) and modes of administration (i.e., paper-based vs. electronic) is necessary for performing accurate intergroup comparisons of both baseline and post-treatment levels of HRQoL [8, 9]. In this regard, the recent study by Varni et al. [10] fills an important gap in the psychometric literature on PROMs by examining the factorial invariance of the Pediatric Quality of Life Inventory^{TM} Multidimensional Fatigue Scale (PedsQL^{TM} MFS) across gender and three age groupings (5–7, 8–12, and 13–18 years of age). Beginning with the best-fitting confirmatory factor analysis (CFA) model for the PedsQL^{TM} MFS within each age and gender group—namely a bi-factor model consisting of a global fatigue factor and three domain-specific factors of General, Sleep/Rest, and Cognitive Fatigue—Varni et al. conducted the following sequence of cross-group invariance tests in order to detect potential interactions of age and gender with the measurement properties of the scale:

C. N. McIntosh (✉)
National Crime Prevention Centre, Public Safety Canada,
269 Laurier Avenue West, Ottawa, ON K1A 0P8, Canada
e-mail: cameron.mcintosh@ps-sp.gc.ca

(1) *metric* (i.e., equality constraints on factor loadings); (2) *strong* (i.e., equality constraints on factor loadings and intercepts); (3) *strict* (i.e., equality constraints on factor loadings, intercepts, and measurement errors); and (4) *homogeneity of variance* (i.e., equality constraints on factor variances). Strict and strong invariance were found to hold, respectively, across the gender and age groups, and therefore, it was concluded that the meaning of the PedsQL^{TM} MFS items was sufficiently similar across these subpopulations to allow the intergroup comparison of scores and justify the interpretation of any differences as true rather than artifactual.

Despite these encouraging findings, however, there are some serious limitations in the methodological strategy employed in this study to both locate the optimal baseline model structure for the PedsQL^{TM} MSF and perform the series of factorial invariance tests. Therefore, the present commentary consists of two parts: (1) a discussion of the main statistical problems with Varni et al.'s [10] approach and (2) the presentation of a flexible and innovative resolution for these issues, based on a recently developed Bayesian framework for conducting CFA.

Evaluating model fit in CFA: two perspectives

Varni et al. [10] make a distinction between "traditional" and "practical" perspectives on assessing global model fit and factorial invariance in CFA. The traditional perspective entails using the maximum-likelihood chi-square (χ^2) test of exact fit for establishing a cleanly fitting baseline model [11], and then evaluating the tenability of successive, nested factorial invariance constraints using χ^2 difference ($\Delta\chi^2$) tests [12, 13]. On the other hand, the practical perspective focuses on the use of supplemental or approximate fit indices (AFIs) with associated "cut-off" thresholds to locate a well-fitting baseline model [14–16], followed by assessing the magnitude of the differences in fit indices across the series of equality constraints imposed on the measurement parameters [17–19]. In keeping with what is still popular mainstream CFA practice, Varni et al. opt for the practical perspective, using AFIs for all of their model assessments and dismissing the χ^2 and $\Delta\chi^2$ tests due to a presumed "oversensitivity" to sample size. However, there is somewhat of a sleight of hand with this oft-used claim. More specifically, in any given CFA or structural equation modeling (SEM) application where a significant χ^2 statistic is obtained, there is always the possibility that model misspecification is the culprit, regardless of how high the sample size is. Unfortunately, AFIs cannot be used to safely "override" a failed χ^2 test because, as highlighted by several CFA/SEM methodologists for more than a decade, these indexes cannot shed further light on the precise sources of model misfit [20–27]. Thus, relying on AFIs could potentially lead to grossly incorrect models being misjudged as good representations of the phenomena or behaviors under study. And by a straightforward extension, differences in AFIs cannot speak to the tenability of factorial invariance across population subgroups, in the event that the $\Delta\chi^2$ tests do not support the various sets of equality constraints.

In the Varni et al. [10] study, the χ^2 statistic obtained for the overall bi-factor model (all participants combined) for the PedsQL^{TM} MSF was 250.23 ($N = 837$, $df = 117$, $p < 0.001$). Concerning the configural models for the age and gender invariance testing, the χ^2 values were reported, respectively, as 535.49 ($N = 837$, $df = 357$, $p < 0.001$) and 379.91 ($N = 756$, $df = 234$, $p < 0.001$). These findings indicate a strong possibility of model misspecification, which could have also resulted in distortions in parameter estimates and their associated significance tests [28, 29]. Moreover, while a number of the exact $\Delta\chi^2$ tests obtained by Varni et al. were actually non-significant and thus appeared to support the invariance constraints (particularly for the gender-based analysis), these can only be taken as accurate if the initial baseline or configural model shows a non-significant χ^2 [12]. Therefore, a more focused diagnosis and resolution of model misspecification would have been much more prudent than simply deferring to AFIs for establishing the baseline models.

Where might CFA/SEM model misspecifications lie?

Encountering some degree of specification error tends to be the norm rather than exception in virtually all CFA/SEM applications, given that the relationships among certain variables are typically set precisely at zero. For example, it is conventionally specified that each observed variable loads on one and only one latent factor [30]; in the case of a bi-factor model, as used by Varnie et al. [10] for the PedsQL^{TM} MFS, it is specified that each variable loads on the global factor and only one of the domain-specific factors [31]. Furthermore, it is also typically assumed that the observed items are conditionally independent given the factor model; in other words, absolutely no additional covariation should exist among the measurement error terms (i.e., the portion of the variance in the items not explained by the factor). In practice, however, the observed variables will load to some degree on factors other than their hypothesized "parent" latent variables [32, 33], and the presence of minor unmodeled factors will often lead to residual covariances among the items [34–36]. The latter finding is often due to what Meehl [37] aptly dubbed as "the crud factor" (p. 204), which essentially means that to some extent, all variables are intercorrelated in the world of social science research. Even though these types of misspecifications may be small in a practical sense when considered individually, they will have a cumulative deleterious effect on both model fit and parameter estimates if they are not accounted for. To be sure, more serious varieties of CFA misspecification are also entirely possible (e.g., too few factors in the model), but the ubiquity of cross-loadings and correlated errors means that they should always be the focus of an initial diagnosis of the sources of ill fit.

Conventionally, specification checks on constrained parameters are performed post hoc and one-at-a-time, using modification indices that show the improvement in model fit that would result from freeing a single fixed parameter. Trying to achieve good model fit using this strategy can involve a long series of model modifications that carry a substantial risk of type I error, which makes it potentially dangerous to generalize a substantially revised model to the population under study [38]. Given these issues, researchers engaged in estimating and testing CFA models might be tempted to just simply free all measurement errors and cross-loadings a priori in order to maximize fit. However, under Frequentist estimation methods such as ML or least squares-based procedures, such an approach will lead to the model being statistically underidentified, that is, having too little numerical information in the raw data for estimating the unknown parameter values [39]. However, a new Bayesian framework for CFA/SEM allows practitioners to incorporate the possibility of nonzero cross-loadings and measurement error covariances in advance, minimizing the potential for post hoc data snooping while still ensuring that the model parameters are identified [40–42]. The following section provides an overview of the key basic principles of Bayesian estimation, followed by a discussion of their application to both establishing baseline CFA models and testing whether factorial invariance holds across population subgroups.

Bayesian statistical analysis: an overview

While still certainly not yet as commonplace as conventional Frequentist approaches to statistical analysis, the use of Bayesian methods in applied research has grown dramatically over the last two decades, largely due to an increase in the number of high-quality introductory and advanced textbooks, as well as the development of versatile software packages for implementing Bayesian estimation [43–45]. Briefly, the two fundamental, overarching differences between the Bayesian and Frequentist paradigms are that Bayesians: (1) view model parameters as variables with probability distributions, not as constants with one and only one true population value and (2) advocate combining empirical data with a researcher's prior beliefs about model parameters in order to ultimately arrive at a posterior distribution for those parameters, rather than estimating parameters based strictly on the observed data only. More formally, these principles can be compactly expressed using Bayes' rule [46–48]:

$$p(\theta|\text{Data}) = \frac{p(\text{Data}|\theta) * p(\theta)}{p(\text{Data})}, \qquad (1)$$

where $p(\theta|\text{Data})$ is the posterior probability (p) distribution of the model parameters (θ) given the empirical data, $p(\text{Data}|\theta)$ is the likelihood of observing the empirical data given the set of model parameters, $p(\theta)$ is the prior probability distribution of the model parameters, and p(Data) is simply the probability distribution of the observed data. Therefore, the posterior distribution is essentially a statistical combination of observed data and prior beliefs about model parameters; in other words, the posterior reflects the extent to which the empirical evidence modifies the researcher's initial convictions about the distribution of the model parameters.

It is important to point out that the prior distribution—which is constructed based on the theory, expert opinion, or empirical findings from previous research—can vary widely in *informativeness*, or the degree of influence on the results of a Bayesian analysis [48, 49]. As a simple example, consider that a normal distribution for any model parameter q can be represented as $q \sim \text{Normal}(\mu_q, {\sigma_q}^2)$, where μ_q and ${\sigma_q}^2$ are the population mean and variance of q, respectively. Whatever prior value is selected for μ_q in a Bayesian application, the value chosen for ${\sigma_q}^2$ reflects how certain the researcher is about the chosen value for μ_q. For instance, a researcher with no available prior knowledge of μ_q could simply assign it an infinite prior variance, such that the likelihood portion of Eq. [1] would dominate the prior in determining the form of the posterior distribution, rendering the analysis essentially Frequentist in nature. On the other hand, smaller prior values for ${\sigma_q}^2$ reflect greater degrees of certainty about μ_q and would therefore give the prior more weight in shaping the posterior. Note also that priors need not be normal or even part of the family of parametric distributions at all, if the data being analyzed typically depart from these standard forms [50, 51]. Thus, the researcher has considerable flexibility in dealing with non-normality up front and ensuring that it is appropriately reflected in the posterior distribution, rather than applying post hoc non-normality corrections to standard errors and test statistics.

In practice, the posterior distribution is almost always too complex to be calculated directly or even estimated iteratively using familiar ML or least squares techniques. Instead, sophisticated Markov chain Monte Carlo (MCMC) algorithms are required for gradually locating and mapping out the entire posterior distribution, using a long sequence of random samples of parameter values [52]. Point estimates for parameters (e.g., mean, median, and mode) can then be computed directly using the samples from the posterior, while inference is done using *credible intervals*, which are fundamentally different from Frequentist confidence intervals [53, 54]. In particular, a credible interval is a range of parameter values in the posterior distribution that cover a certain percentage of the probability. For example, a 95 % credible interval means that given the observed data, there is a 95 % chance that the interval contains the true value of the parameter in question; if this interval does not include zero, then the estimate is considered substantively meaningful. On the other hand, the more familiar 95 % confidence interval means that if we repeated the exact same study 100 times, 95 % of those replications will contain the true value of the parameter. Thus, Bayesian credible intervals focus on the credibility of parameter values given the data at hand (and the prior), not probabilities based on the theoretical replications and data that were never actually observed. For this reason, credible intervals are regarded as more useful and informative than confidence intervals [55, 56].

Bayesian CFA for a single model

The flexibility of Bayesian priors can help CFA practitioners to pre-empt some of the most commonly encountered varieties of model misspecification. A helpful initial step in understanding and using Bayesian CFA is to first view the conventional Frequentist approach to model specification through a Bayesian lens. In particular, the conventional omission of cross-loadings and measurement error covariances in CFA is essentially equivalent to specifying a prior distribution where both the mean and variance of the parameter are exactly zero [40, 41], which is extremely unlikely to be true in the population. While freeing all of these parameters *en masse* would lead to an underidentified model from a Frequentist standpoint, the use of informative Bayesian priors brings additional statistical information into the analysis that can render the model identified [40, 41, 57]. More specifically, instead of exact zeros where relationships are hypothesized to be absent, “approximate” zero priors can be used in which the distribution for a given parameter is centered at zero, but allowed to vary within the bounds of what might be considered as non-substantive values.

For instance, guidance on what constitutes a nonzero yet trivial range for a cross-loading can be taken from Comrey and Lee’s [58] widely used classification of loading magnitudes, according to which loadings $\geq$0.71 are considered excellent; $\geq$0.63 very good; $\geq$0.55 good; $\geq$0.45 fair; and $\geq$0.32 poor. Therefore, a prior that allows a cross-loading to take on values up to even 0.44 might be reasonable, providing that all major loadings of conceptual interest were ultimately estimated to be in the good to excellent range. Thus, assuming that all variables have been standardized for convenience, an approximate zero prior for a given cross-loading λ_{cross} could be specified as $\lambda_{\text{cross}} \sim$ Normal(0.0, 0.05), which implies a prior 95 % credibility interval of -0.44 to $+0.44$. This strategy gives the posterior estimate for λ_{cross} considerable leeway to depart its zero prior mean—thereby reducing the potential for model misfit—but at the same time helps to restrict it from attaining a more substantively important value. Of course, if one believes the cross-loadings to be even more trivial in magnitude, lower prior variances could be used, for example, $\lambda_{\text{cross}} \sim$ Normal(0.0, 0.005) or $\lambda_{\text{cross}} \sim$ Normal(0.0, 0.01) yield 95 % prior credible intervals of -0.14 to $+0.14$ and -0.20 to $+0.20$, respectively [40, 41]. Approximate zero priors could be set for measurement error covariances in a similar manner, with the exception that instead of the normal distribution, inverse-Wishart or inverse-gamma priors are required for these parameters [40, 41, 57].

Of course, the option of using informative Bayesian priors should not be taken as giving CFA practitioners free rein to fill their models with superfluous parameters, simply for the purpose of mathematically improving model fit. Prior distributions should always be carefully specified and justified based on the content area knowledge and any available findings from past research, which still renders the Bayesian approach highly confirmatory. Further, the adequacy of overall model fit should be evaluated using posterior predictive p values (PPPs) [59], which are based on the repeated comparisons of the observed dataset with a series of simulated or “model-implied” datasets created using MCMC-based random samples from the posterior distribution of the model parameters. More formally, in the CFA/SEM context, one computes the following discrepancy measure for each iteration of the MCMC algorithm:

$$D_i = T_i^{\text{sim}}(Y_i^{\text{sim}}, \theta_i) - T_i^{\text{obs}}(Y^{\text{obs}}, \theta_i) \qquad (2)$$

where T_i^{sim} is a summary test statistic (typically, χ^2) produced by fitting the ith posterior sample of model parameters θ_i to the simulated dataset Y_i^{sim}, and $T_i^{\text{obs}}$is an analogous test statistic generated by fitting θ_i to the observed dataset Y^{obs}. If the model is correct, the observed and simulated data (and thus $T_i^{\text{sim}}$and T_i^{obs}) should match very closely, yielding a distribution of the D_i (across the MCMC samples) that is symmetric around 0, and therefore a PPP = 0.5. Thus, PPP values close to 0 (or 1) would suggest that the model is not a good representation of the data, and one could potentially use familiar “alpha” cutoffs such as $p < 0.05$ (or $p > 0.95$). However, it is also recommended to inspect the bivariate scatterplot of T_i^{sim} and T_i^{obs} as an additional aid in evaluating overall model quality [59]. Furthermore, cross-loadings and measurement error covariances with approximate zero priors should have posterior credible intervals that include zero; if not, then the researcher might need to consider larger prior variances, or perhaps rethink the entire model structure. However, a recent application of Bayesian CFA found that the small variance prior strategy produces excellent fit, without requiring any further model modifications [42].

Bayesian CFA for factorial invariance testing

After establishing a well-fitting baseline model, the flexibility of Bayesian priors can be extended to factorial invariance testing across population subgroups, where they can also be used to relax some conventional yet typically untenable parameter constraints [41]. More specifically, in addition to the usual practice of setting cross-loadings and measurement error covariances to exact zeros in single group CFA, Frequentist-based factorial invariance testing involves setting measurement parameters to be precisely equal to each other across groups, which is also not likely to hold in applications. Therefore, instead of imposing such stringent equality constraints, practitioners can use Bayesian priors to allow a given parameter some degree of variability across groups. For example, a small, likely non-substantive prior variance for the difference in a factor loading between two groups could be specified as $\Delta\lambda_{g1,g2} \sim$ Normal(0.0, 0.005), which gives a narrow 95 % prior credible interval of -0.14 to $+0.14$ for the cross-group discrepancy. (Whatever prior variance values are chosen should be based on the knowledge of the conceptual background for and psychometric properties of the scale). In addition to improving the capacity of the multigroup model to fit the data, the strategy of allowing some cross-group variability in the measurement parameters greatly simplifies the identification and correction of the sources of non-invariance. Whereas the Frequentist approach to factorial invariance involves the rather cumbersome specification and testing of several nested models, accompanied by a series of specification searches on individual invariance constraints, the Bayesian method entails two simple steps, which are implemented in a user-friendly manner in the M*plus* software package [41].

In the first step, the model is estimated using small variance priors on the differences in the measurement parameters across groups, and then the average cross-group differences and accompanying 95 % posterior credible intervals for each parameter are displayed in the output for inspection. A parameter is considered non-invariant if the credible interval for its cross-group differences excludes zero. In the second step, all non-invariant parameters identified in step one are set free (i.e., given infinite prior variances), all invariant parameters are set exactly equal across groups (i.e., zero mean and zero variance priors), and then the model is re-estimated. While one could of course still retain the small variance priors for the parameters that were found to be invariant across groups in the first step, it has been demonstrated that switching to exact equalities on these parameters in the second step actually leads to superior overall fit in terms of PPPs [41]. Achieving a good fit in the second step allows factor means to be compared without the risk of confounding by factorial non-invariance across groups.

Conclusions

Establishing factorial invariance is an essential component of the psychometric evaluation of a measuring instrument, as it provides a sound statistical basis for the comparison of scores across different population subgroups. This commentary sheds further light on some common yet still underappreciated problems encountered when using conventional Frequentist estimation methods and AFIs for developing a well-fitting baseline factor model, as well as conducting a follow-up series of factorial invariance tests. In response to these issues, a flexible and convenient Bayesian framework was presented that can substantially aid in: (1) effectively tackling the most ubiquitous sources of model misspecification in CFA and (2) locating specific measurement parameters that are non-invariant across population subgroups. It is recommended that these promising techniques be more widely adopted by HRQoL researchers for evaluating the factorial invariance of PROMS and other types of self-report instruments.

References

1. Doward, L. C., Gnanasakthy, A., & Baker, M.G. (2010). Patient reported outcomes: looking beyond the label claim. *Health and Quality of Life Outcomes*, *8*(89), 1–9.
2. Dinan, M. A., Compton, K. L., Dhillon, J. K., Hammill, B. G., Dewitt, E. M., Weinfurt, K. P., et al. (2011). Use of patient-reported outcomes in randomized, double-blind, placebo-controlled clinical trials. *Medical Care, 49*(4), 415–419.
3. Sprangers, M. A. (2010). Disregarding clinical trial-based patient-reported outcomes is unwarranted: Five advances to substantiate the scientific stringency of quality-of-life measurement. *Acta Oncologica, 49*(2), 155–163.
4. Snyder, C. F., Aaronson, N. K., Choucair, A. K., Elliott, T. E., Greenhalgh, J., Halyard, M. Y., et al. (2012). Implementing patient-reported outcomes assessment in clinical practice: A review of the options and considerations. *Quality of Life Research, 21*(8), 1305–1314.

5. McKenna, S. P. (2011). Measuring patient-reported outcomes: moving beyond misplaced common sense to hard science. *BMC Medicine*, *9*(86), 1–12.
6. Jones, P., Miravitlles, M., van der Molen, T., & Kulich, K. (2012). Beyond FEV1 in COPD: A review of patient-reported outcomes and their measurement. *International Journal of Chronic Obstructive Pulmonary Disease, 7*, 697–709.
7. Swartz, R. J., Schwartz, C., Basch, E., Cai, L., Fairclough, D. L., McLeod, L., et al. (2011). The king's foot of patient-reported outcomes: Current practices and new developments for the measurement of change. *Quality of Life Research, 20*(8), 1159–1167.
8. Cook, K. F., Bamer, A. M., Amtmann, D., Molton, I. R., & Jensen, M. P. (2012). Six patient-reported outcome measurement information system short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. *Archives of Physical Medicine and Rehabilitation, 93*(7), 1289–1291.
9. Coons, S. J., Gwaltney, C. J., Hays, R. D., Lundy, J., Sloan, J. A., Revicki, D. A., et al. (2009). Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value in Health, 12*(4), 419–429.
10. Varni, J. W., Beaujean, A. A., & Limbers, C. A. (2013). Factorial invariance of pediatric patient self-reported fatigue across age and gender: A multigroup confirmatory factor analysis approach utilizing the PedsQL^{TM} Multidimensional Fatigue Scale. *Quality of Life Research*, Online First. doi:10.1007/s11136-013-0370-4.
11. Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 183–202.
12. Yuan, K.-H., & Bentler, P. M. (2004). On Chi square difference and *z* tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*(5), 737–757.
13. Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate distribution of sequential Chi square statistics. *Psychometrika, 50*(3), 253–264.
14. Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McCardle (Eds.), *Contemporary psychometrics: A festschrift to Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.
15. Bagozzi, R. P., & Yi, Y. (2012). 'Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science, 40*, 8–34.
16. Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modelling. *Psychological Methods, 17*(1), 1–14.
17. Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872–882.
18. Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504.
19. Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592.
20. McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859–867.
21. McIntosh, C. N. (2007). Improving the evaluation of model fit in confirmatory factor analysis: A commentary on Gundy, C. M., Fayers, P. M., Groenvold, M., Petersen, M. Aa., Scott, N. W., Sprangers, M. A. J., Velikov, G., Aaronson, N. K. (2011). Comparing higher-order models for the EORTC QLQ-C30. Quality of Life Research. *Quality of Life Research*, *21*(9), 1619–1621. doi:10.1007/s11136-011-0082-6.
22. Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three—testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841–850.
23. Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7*(1), 1–35.
24. Shipley, B. (2002). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference* (2nd ed.). Cambridge, UK: Cambridge University Press.
25. Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.
26. Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561–582.
27. Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086–1120.
28. Yuan, K.-H., Kouros, C. D., & Kelley, K. (2008). Diagnosis for covariance structure models by analyzing the path. *Structural Equation Modeling, 15*, 564–602.
29. Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology, 41*(1), 119–157.
30. McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, New Jersey: Erlbaum.
31. Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.
32. Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438.
33. Morin, A. J. S., Marsh, H. W., & Nagengast, B. (in press). Exploratory structural equation modeling. To appear in G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) Charlotte, NC: Information Age Publishing.
34. Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent variable covariance structure analysis. *Psychological Methods, 12*(4), 381–398.
35. Saris, W. E., & Aalberts, C. (2003). Different explanations for correlated disturbance terms in MTMM studies. *Structural Equation Modeling, 10*(2), 193–213.
36. Reddy, S. K. (1992). Effects of ignoring correlated measurement error in structural equation models. *Educational and Psychological Measurement, 52*(3), 549–570.
37. Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.
38. MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.
39. Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
40. Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335.
41. Muthen, B., & Asparouhov, T. (January 11, 2013). *BSEM measurement invariance analysis*. Mplus Web Notes: No. 17. Accessed 15 March 2013 at: http://www.statmodel.com/examples/webnotes/webnote17.pdf.
42. Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2012). Further insights on the French WISC-IV factor structure through

Bayesian structural equation modeling. *Psychological Assessment*, Online First. doi:10.1037/a0030676.

43. Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology, 66*(1), 1–7.
44. Wetzels, R., & Wagenmakers, E.-J. (2010). Exemplary introduction to Bayesian statistical inference. (book review of "Bayesian modeling using WinBUGS"). *Journal of Mathematical Psychology, 54*, 466–469.
45. Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis, 1*(4), 969–980.
46. Alston, C. L., Mengersen, K. L., & Pettitt, A. N. (Eds.). (2013). *Case studies in Bayesian statistical modelling and analysis*. Chichester, UK: Wiley.
47. Jackman, S. (2009). *Bayesian analysis for the social sciences*. West Sussex, UK: Wiley.
48. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC Press.
49. Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other models. *The Annals of Applied Statistics, 2*(4), 1360–1383.
50. Chow, S. M., Tang, N., Yuan, Y., Song, X., & Zhu, H. (2011). Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior. *British Journal of Mathematical and Statistical Psychology, 64*(1), 69–106.
51. Kyung, M., Gill, J., & Casella, G. (2011). New findings from terrorism data: Dirichlet process random-effects models for latent groups. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 60*(5), 701–721.
52. Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (Eds.). (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton, FL: Chapman and Hall/CRC.
53. Eberly, L. E., & Casella, G. (2003). Estimating Bayesian credible intervals. *Journal of Statistical Planning and Inference, 112*, 115–132.
54. Curran, J. M. (2005). An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, Probability and Risk, 4*, 115–126.
55. Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722–752.
56. Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14*(4), 301–322.
57. Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: with applications in the medical and behavioral sciences*. Chichester, UK: Wiley.
58. Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). San Diego, CA: Academic Press.
59. Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling, 18*(4), 663–685.