Journal of Business Research 113 (2020) 13–24

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM)

Nicholas P. Danks[a], Pratyush N. Sharma[b], Marko Sarstedt[c,d,*]

[a] National Tsing Hua University, Institute of Service Science, Taiwan
[b] University of Delaware, Alfred Lerner College of Business and Economics, USA
[c] Otto-von-Guericke-University Magdeburg, Universitaetsplatz 2, 39106 Magdeburg, Germany
[d] Monash University of Malaysia, Monash University Malaysia, School of Business and Global Asia in the 21st Century Research Platform, Selangor, Malaysia

ARTICLE INFO

Keywords:
Akaike weights
Information criteria
Model selection
Partial least squares
Selection uncertainty
Model-averaged predictions

ABSTRACT

Comparing alternative explanations for behavioral phenomena is central to the process of scientific inquiry. Recent research has emphasized the efficacy of Information Theoretic model selection criteria in partial least squares structural equation modeling (PLS-SEM), which has gained massive dissemination in a variety of fields. However, selecting one model over others based on model selection criteria may lead to a false sense of confidence as differences in the criteria values are often small. To overcome this limitation researchers have proposed Akaike weights, whose efficacy however, has not been assessed in the PLS-SEM context yet. Addressing this gap in research, we analyze the efficacy of Akaike weights in PLS-SEM-based model comparison tasks. We find that Akaike weights derived from BIC and GM are well suited for separating incorrectly specified from correctly specified models, and that Akaike weights based on AIC are useful for creating model-averaged predictions under conditions of model selection uncertainty.

1. Introduction

The consideration and evaluation of alternative models for explaining certain phenomena is central to the process of scientific inquiry. Alternative models typically emerge when considering theories in new contexts with unique variables and effects, or when researchers build conceptual bridges across related streams of inquiry to provide a holistic understanding of the phenomenon (Sharma, Sarstedt, Shmueli, Kim, & Thiele, 2019). Given a set of alternative models, researchers try to identify the model that best approximates the data generation process underlying the phenomenon under study. This process of model evaluation is complicated by the fact that complex models (i.e., with more variables and paths) may overfit the data by tapping spurious patterns or noise in a sample (Myung, 2000). Because such patterns are sample-specific, the models will generalize poorly to other samples and have limited possibility of being replicated by other researchers. In contrast, parsimonious models may fit somewhat less well, but have a better chance of being scientifically replicable. Hence, any criterion for model selection needs to address the trade-off between model fit and parsimony (Burnham & Anderson, 2002).

A common way of overcoming this apparent dichotomy is to compare the alternative models using Information Theoretic model selection criteria that seek to balance model fit and parsimony, thereby incorporating Occam's razor. In their basic form, these criteria are calculated using the number of fitted model parameters, and either the estimated maximum likelihood for the model or its residual sum of squares. In general, both of these measures can easily be derived from the output of statistical software packages. Research has brought forward a broad range of Information Theoretic model selection criteria, which differ in their statistical underpinnings and the penalties applied for number of parameters, sample size, entropy, or Fisher information. Popular criteria include the Akaike Information Criterion (AIC; Akaike, 1973) and its variants such as the Consistent AIC (Bozdogan, 1987) or the AIC_3 (Bozdogan, 1994), and the Bayesian Information Criterion (BIC; Schwarz, 1978). Prior research has examined the efficacy of these criteria under various conditions in different methodological contexts such as mixtures of normal distributions (e.g., Celeux & Soromenho, 1996), mixture regression models (e.g., Becker, Ringle, Sarstedt, & Völckner, 2015), and mixture logit models (e.g., Andrews & Currim, 2003). Sharma et al. (2019) recently extended this strand of research by examining their performance in the context of partial least squares structural equation modeling (PLS-SEM), a regression-based technique that estimates relationships in path models with latent and manifest variables, and which has gained vast prominence in a variety of fields

* Corresponding author at: Otto-von-Guericke-University Magdeburg, Universitaetsplatz 2, 39106 Magdeburg, Germany.
E-mail addresses: nicholas.danks@iss.nthu.edu.tw (N.P. Danks), pnsharma@udel.edu (P.N. Sharma), marko.sarstedt@ovgu.de (M. Sarstedt).

https://doi.org/10.1016/j.jbusres.2020.03.019
Received 26 August 2019; Received in revised form 10 March 2020; Accepted 12 March 2020

(e.g., Ali, Rasoolimanesh, Sarstedt, Ringle, & Ryu, 2018; Hair, Hollingsworth, Randolph, & Chong, 2017; Ringle, Sarstedt, Mitchell, & Gudergan, 2020). These authors identify BIC and Geweke and Meese (1981) criterion (GM) as particularly suitable for PLS-SEM-based model selection tasks. In a related study, Sharma, Shmueli, Sarstedt, Danks, and Ray (2020) found that BIC and GM also achieve a sound trade-off between model fit and out-of-sample predictive power in the estimation of PLS path models. In particular, BIC and GM are useful substitutes for selecting correctly specified models with low prediction error when researchers cannot create a holdout sample—for example, due to low sample size—to assess a model's out-of-sample predictive power.

However, a possible issue in the application of model selection criteria is that—in their simple form (i.e., raw values)—they do not offer any insights regarding the relative weights of evidence in favor of models under consideration (Burnham & Anderson, 2002). More precisely, while the differences in the criteria values are useful in ranking and selecting models, such differences can often be small in practice, leading to model selection uncertainty (Preacher & Merkle, 2012). For example, in their comparative analysis of E-government performance models, Sharma, Morgeson, Mithas, and Aljazzaf (2018) empirically compare 26 models using PLS-SEM and Information Theoretic model selection criteria. The authors then rank the models based on the raw criteria values and other model evaluation statistics. However, they find the average differences in values across all 325 comparisons to be relatively small: 0.90 for AIC, 3.02 for BIC, and 5.09 for GM—see Table 4 in Sharma et al. (2018). Similarly, Sharma et al.'s (2020) empirical comparison of different configurations of Schwaiger (2004) corporate reputation model also produced relatively small average differences in AIC, BIC, and GM of 5.83, 4.19, and 3.92. Sharma et al. (2020) select the model with the best (i.e., the lowest) criteria values but do not consider the magnitude of these differences when making this decision. These two studies highlight the prevalence of model selection uncertainty when comparing models using empirical data.

Due to these reasons, numerous researchers have pointed out that selecting one model over others based on small differences in raw criteria values may lead to a false sense of confidence (e.g., Preacher & Merkle, 2012). For example, Wagenmakers and Farrell (2004, p. 193) note that "it is difficult to intuit how much statistical importance we should attach to a difference in AIC values." The Information Theoretic model selection criteria quantify the *relative* conceptual distance, or loss in information, from an unknown data generation model to the proposed candidate model. Because the criteria values are on a relative (interval) scale, an individual Information Theoretic model selection criteria value is not interpretable in isolation (Burnham & Anderson, 2002; Chapter 2.6). Instead, the metrics are interpretable only when comparing competing models to identify a model that is relatively closer to the data generating model. Therefore, instead of engaging in a binary decision process or assuming subjective cut-offs that define substantial differences in criteria values, the focus should be on documenting *how much* better a selected model is compared to others in a given sample (Symonds & Moussalli, 2011). Presenting such evidence in favor of the selected model is particularly important when the researcher is seeking to remove other candidate models from further contention (Wagenmakers & Farrell, 2004).

To handle this aspect of model selection uncertainty, statisticians have proposed using the model selection criteria's raw values to compute the *relative likelihood* of a model, given the data and set of competing models (e.g., Akaike, 1978, 1979; Bozdogan, 1987). The relative likelihood value allows a researcher to draw more robust inference by creating an additional measure that can be used to judge the relative strength of evidence in favor of each model in the set, referred to as *Akaike weights*. This measure gives researchers more information about whether to base the inference on a single superior model or consider multimodel inference. More importantly, researchers can use the Akaike weights to generate model-averaged predictions rather than relying on the predictions of a single model. That is, a prediction is made with each model, and the Akaike weights are used to compute a weighted average of these predictions (Burnham & Anderson, 2002; Chapter 5.3). Model-averaged predictions are particularly useful when there is uncertainty as to the accuracy of the predictive method or model, uncertainty in the predictive context such as high volatility or unexpected variation (e.g., in new product forecasting), and high costs associated with large predictive errors (Armstrong, 2001). These averaged predictions, or ensembles, perform best when different predictive methods are used and there is negative, near zero, or little correlation between predictions (Brown & Yao, 2001). However, even when generating averaged predictions from the same method (e.g., PLS-SEM), which generally results in positively correlated predictions, an increase in predictive accuracy can still be expected.

Yet, despite their obvious relevance for model selection tasks, the use of Akaike weights is practically absent in marketing research. Specifically, our review of all articles published since 2000 in the five leading marketing journals, *Journal of Marketing*, *Journal of Marketing Research*, *Marketing Science*, *Journals of Consumer Research*, and *Journal of the Academy of Marketing Science*, shows that only a single study (Breivik & Thorbjørnsen, 2008) uses relative likelihoods in model comparison tasks. Similarly, only a single study in *Journal of Business Research* draws on Akaike weights in multimodel inference (Daryanto, 2019). This is surprising given that the use of model probabilities has a solid standing in biology (Posada & Buckley, 2004), ecology (Symonds & Moussalli, 2011), and psychology (Wagenmakers & Farrell, 2004).

Extending recent research on model selection in PLS-SEM (Sharma et al., 2019, 2020), this study (1) analyzes the efficacy of Akaike weights to distinguish correctly specified from misspecified models in PLS-SEM-based model comparison tasks, and (2) assesses their utility to generate model-averaged predictions. To do so, we present the results of two large-scale simulation studies that analyze the behavior of Akaike weights under different model setups and experimental conditions, including sample sizes, structural model effect sizes, and indicator loading patterns. Finally, we illustrate the use of Akaike weights and model-averaged predictions by means of an empirical example.

2. Likelihood of a model

In the late 1960s and the early 1970s, several model selection criteria that penalize model complexity in the interest of the principle of parsimony were developed. These criteria are based on the framework of Information Theory, a mathematical theory of communication, which studies the transmission, processing, extraction, and utilization of information (e.g., Akaike, 1973). A large body of literature discusses the technical underpinnings and characteristics of these criteria (e.g., Akaike, 1981; Bozdogan, 1987; Sclove, 1987; Burnham & Anderson, 2002). Information Theoretic criteria (such as AIC, BIC, and GM) quantify the information lost by using a proposed model to represent an assumed data generating population model. Thus, a smaller criterion value provides evidence of a proposed model that better approximates the "true" or data generating model. Model selection criteria are typically written as a function of the maximized value of the likelihood function, but can also be expressed in terms of a target construct's sum of squared errors (Burnham & Anderson, 2002; McQuarrie & Tsai, 1998)—see Appendix B in Sharma et al. (2019) for a formal presentation of the criteria considered in this study.

The idea of using differences in model selection criteria values to compute a model's likelihood given the data was first proposed by Akaike (1978) in the context of the AIC. Given a set of K candidate models, the researcher first needs to compute, for each model i, the differences in AIC value with respect to the AIC of the best candidate model (i.e., AIC_{min}):

$$\Delta_i(AIC) = AIC_i - AIC_{min} \quad (1)$$

The differences Δ_i can then be used to obtain an estimate of the relative likelihood L of model *i*, given the data and the set of *K* models:[1]

$$L(M_i|data) \propto exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\} \tag{2}$$

Using the relative likelihoods of all *K* models as input allows computing Akaike weights defined as (Burnham & Anderson, 2002, p. 75):

$$w_i(AIC) = \frac{exp\left\{-\frac{1}{2}\Delta_i(AIC)\right\}}{\sum_{k=1}^{K} exp\left\{-\frac{1}{2}\Delta_k(AIC)\right\}} \tag{3}$$

A given weight w_i represents the relative evidence in favor of model *i*, given that one of the *K* models is the best model in a Kullback-Leibler sense (Burnham & Anderson, 2004). While the notion of weights w_i was developed in the context of the AIC, formula (3) also generalizes to other Information Theoretic model selection criteria like the BIC and GM (Wagenmakers & Farrell, 2004). In line with the prior literature (e.g., Burnham & Anderson, 2002; Chapter 2.9), however, we universally refer to the weights as Akaike weights, even when the weights stem from other Information Theoretic model selection criteria (e.g., BIC or GM).

Breiman (1996, p. 2354) shows that the selection of the best model from a set of well-fitting models—that is, models that are close in an Information Theoretic sense—can entail considerable instability in the assessment of predictive performance. In fact, "the size of the predictive loss may be a substantial fraction of the prediction error." Inspired by the Bayesian literature in which model weights have a strong standing (e.g., Raftery, Madigan, & Hoeting, 1997), researchers can then use the Akaike weights to compute a weighted average of the prediction of the focal construct's scores generated by each model in the set (Dormann et al., 2018) as follows:

$$Y = \frac{\sum_{i=1}^{R} w_i(\widehat{Y}_i)}{\sum_{i=1}^{R} w_i}, \tag{4}$$

where $\widehat{Y}_i$ is the prediction generated by model *i*. When the goal is optimal prediction of the focal construct's scores, then model-averaged predictions that are not conditional on any one model but are based on weighted averages across multiple models, can be advantageous (Posada & Buckley, 2004).

3. Simulation study I

3.1. Design, data generation, and model estimation

Our simulation study considers the BIC and GM, which Sharma et al. (2019, 2020) recently identified as clearly superior in PLS-SEM-based model selection tasks. Both these criteria are asymptotically consistent, in that their probability to select the true model in a set of alternative models approaches unity as sample size increases. When the true model is not in the set, BIC and GM asymptotically select the closest approximating model (McQuarrie & Tsai, 1998). In addition, we consider the AIC, which typically serves as the default criterion in model selection (e.g., Posada & Buckley, 2004; Symonds & Moussalli, 2011). AIC is asymptotically efficient, in that it tends to select the model with the smallest relative distance to the (unknown) true model as the sample size increases (McQuarrie & Tsai, 1998). A strength of AIC is that it tends to select models with better predictive accuracy, especially in "real-world" settings where the data generating model is complex and out of reach (Aho, Derryberry, & Peterson, 2014). However, a weakness of the AIC is its lack of accuracy and reliability for small sample sizes (Konishi & Kitagawa, 2003) and its tendency to select over-parametrized models (Burnham & Anderson, 2002; Chapter 7.7.3).[2]

[1] Please note that the symbol ∝ in formula (2) is read as "proportional to".

[2] See Burnham and Anderson (2004) and Vrieze (2012) for a detailed discussion of the similarities and differences between the criteria.

The model set-up is similar to Sharma et al. (2019, 2020) and consists of a set of seven competing models (Fig. 1). To foster our results' external validity (Paxton, Curran, Bollen, Kirby, & Chen, 2001), the models have a similar structure and complexity as those commonly encountered in fields where PLS-SEM features prominently. Examples include the American Customer Satisfaction Index (Fornell, Johnson, Anderson, Cha, & Bryant, 1996), which ranks among the most salient models in studying customer satisfaction (e.g., Anderson & Fornell, 2000), and the Unified Theory of Acceptance and Use of Technology model and its variants, which are commonly used in information systems research (e.g., Venkatesh, Morris, Davis, & Davis, 2003). Specifically, all competing structural models in our study have five reflectively measured constructs, three of which are exogenous (ξ_1, ξ_2, and ξ_3), while two are endogenous (η_1 and η_2). Each construct has four items. The focus of our investigation is on the endogenous construct η_2. Model 5 serves as the data generation model. Models 1, 3, 4, and 6 are incorrectly specified with respect to the direct path from ξ_1 to η_2; that is, they have incorrect paths that are inconsistent with the data generation process. Model 2 is a correctly specified but parsimonious variant of the data generation model with only correct (but one missing) paths, whereas Model 7 is a saturated model with all possible structural paths, including incorrect ones.

Replicating Sharma et al. (2019, 2020), we manipulated the following experimental conditions, which correspond to the conditions commonly encountered in applied research (e.g., Hair, Hollingsworth, et al., 2017; Nitzl, 2016; Ringle et al., 2020):

- Six conditions of sample size (50, 100, 150, 200, 250, and 500),
- Five conditions of effect size on the structural path $\xi_2 \rightarrow \eta_1$ (γ_2=0.1, 0.2, 0.3, 0.4, and 0.5),
- Three indicator loading patterns with different levels of average variance extracted (AVE):
  - High AVE with loadings: (0.9, 0.9, 0.9, and 0.9),
  - Moderate AVE with loadings: (0.8, 0.8, 0.8, and 0.8), and
  - Low AVE with loadings: (0.7, 0.7, 0.7, and 0.7).

In each simulation run, we compute the Akaike weights for all three model selection criteria for each of the competing structural models, and estimate model-averaged predictions for the focal outcome construct η_2. These model-averaged predictions are contrasted using the root mean squared error (RMSE), which penalizes larger prediction errors more strongly compared to, for example, the mean absolute error. As large errors are particularly undesirable in PLS-SEM applications (Shmueli et al., 2019), the RMSE is generally considered the "default" metric (e.g., Chica & Rand, 2017; Sharma et al., 2020).

The data were generated using Schlittgen's procedure available in the *cbsem* package for the R statistical software (Schlittgen, 2019). This procedure generates a model-implied covariance matrix for the parameters specified in the simulation condition, performs a Cholesky decomposition upon the covariance matrix, and imposes this upon a multivariate normal sample in order to generate data assuming a composite model population (Ringle, Sarstedt, & Schlittgen, 2014; Online Appendix II). We consider the case of normally distributed data as recent research has shown that model estimates and the behavior of model selection criteria are unaffected by data distributions when estimating data from composite model populations in PLS-SEM (Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016; Hair, Hult, Ringle, Sarstedt, & Thiele, 2017; Sharma et al., 2020). Drawing on Reinartz, Haenlein, and Henseler (2009), we ran 300 replications for each of the 90 simulation conditions, yielding a total of 27,000 cases.

3.2. Results

3.2.1. Model selection uncertainty

First, we document how Akaike weights facilitate handling uncertainty in model selection tasks. In doing so, we consider two types of

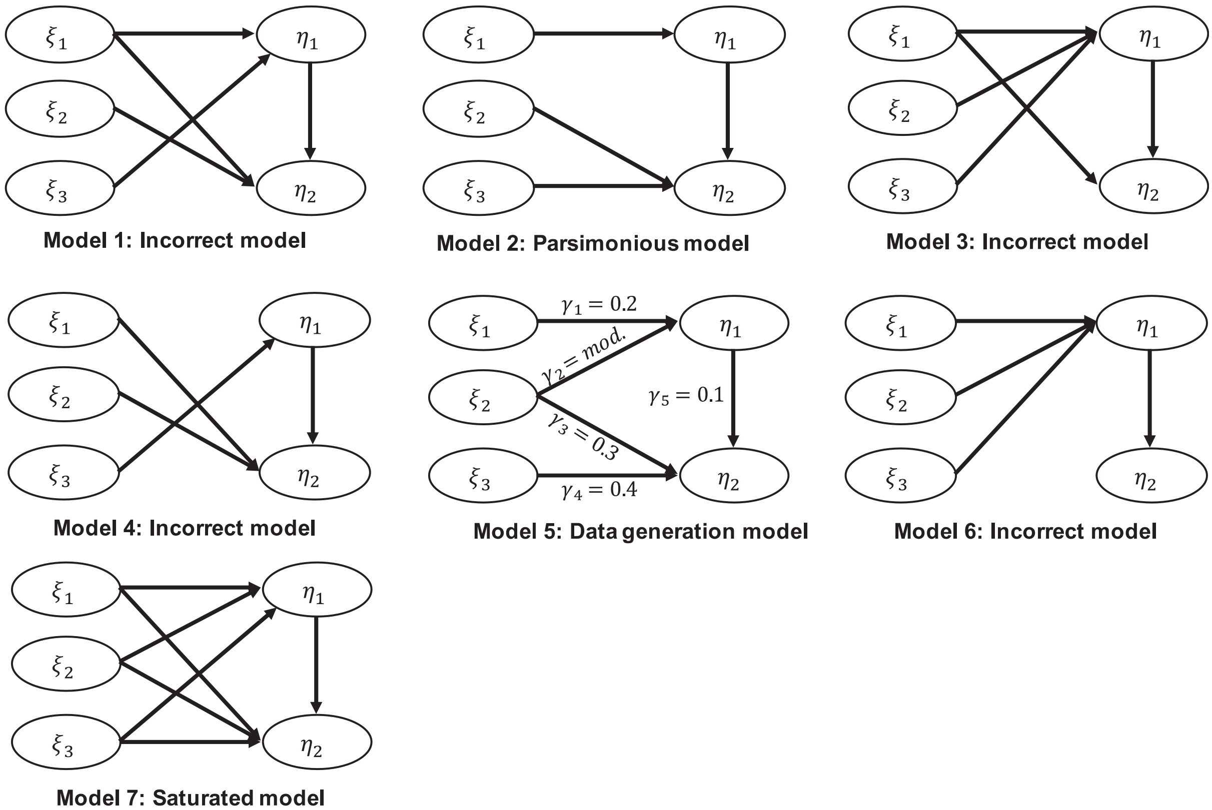

Fig. 1. Simulation study I models.

uncertainties in model selection tasks: (1) *rejection uncertainty*: where the researcher is wary of selecting incorrectly specified models (i.e., Models 1, 3, 4, and 6), and (2) *selection uncertainty*: where the researcher seeks to select a single superior model among a set of correctly, or parsimoniously, specified models which do not contain misspecified paths (i.e., Models 2 and 5). Ideally, researchers would like to utilize a model selection criterion that allows them both to reject incorrectly specified models (i.e., reduce rejection uncertainty), and select a single best model that achieves the best trade-off between model fit and parsimony among a set of models without misspecified paths (i.e., minimize selection uncertainty).

With regard to rejection uncertainty we find that all the three criteria considered in the study (i.e., AIC, BIC, and GM) perform well in rejecting incorrectly specified models as evidenced in low Akaike weights assigned to Models 1, 3, 4, and 6 (Fig. 2). BIC and GM demonstrate very similar performance with marginal differences in Akaike weights in Models 6 and 7. In light of the similarities between BIC and GM, we restrict our further discussion on the comparison between AIC and BIC.

A notable difference emerges between AIC and BIC in terms of reducing selection uncertainty among the set of correctly specified models. As shown in Fig. 2, AIC generally assigns the highest weight to the parsimonious model (Model 2), closely followed by the data generation model (Model 5). However, it also weighs the saturated model (Model 7) fairly high. This is in contrast to BIC, which assigns highest weights to the parsimonious and data generation models but assigns a

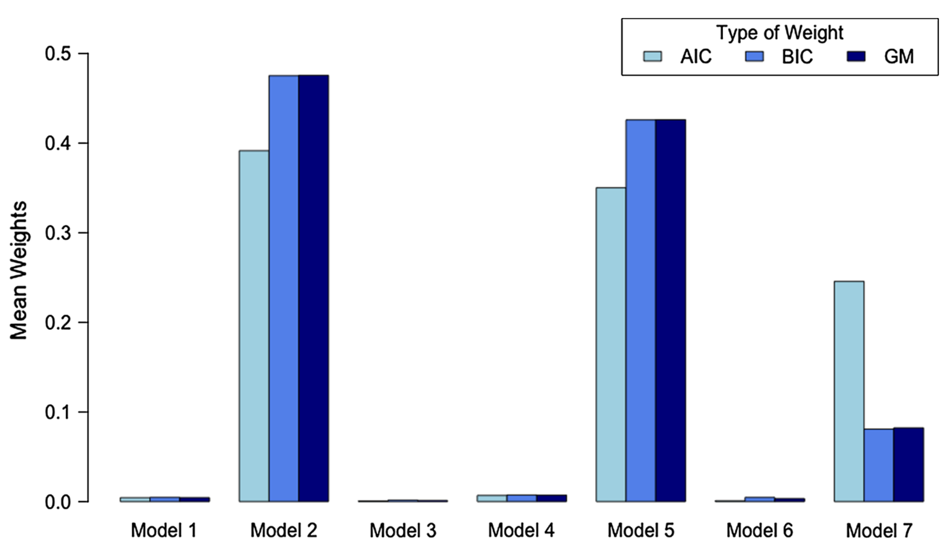

Fig. 2. Mean weights of AIC, BIC, and GM across models in simulation study I.

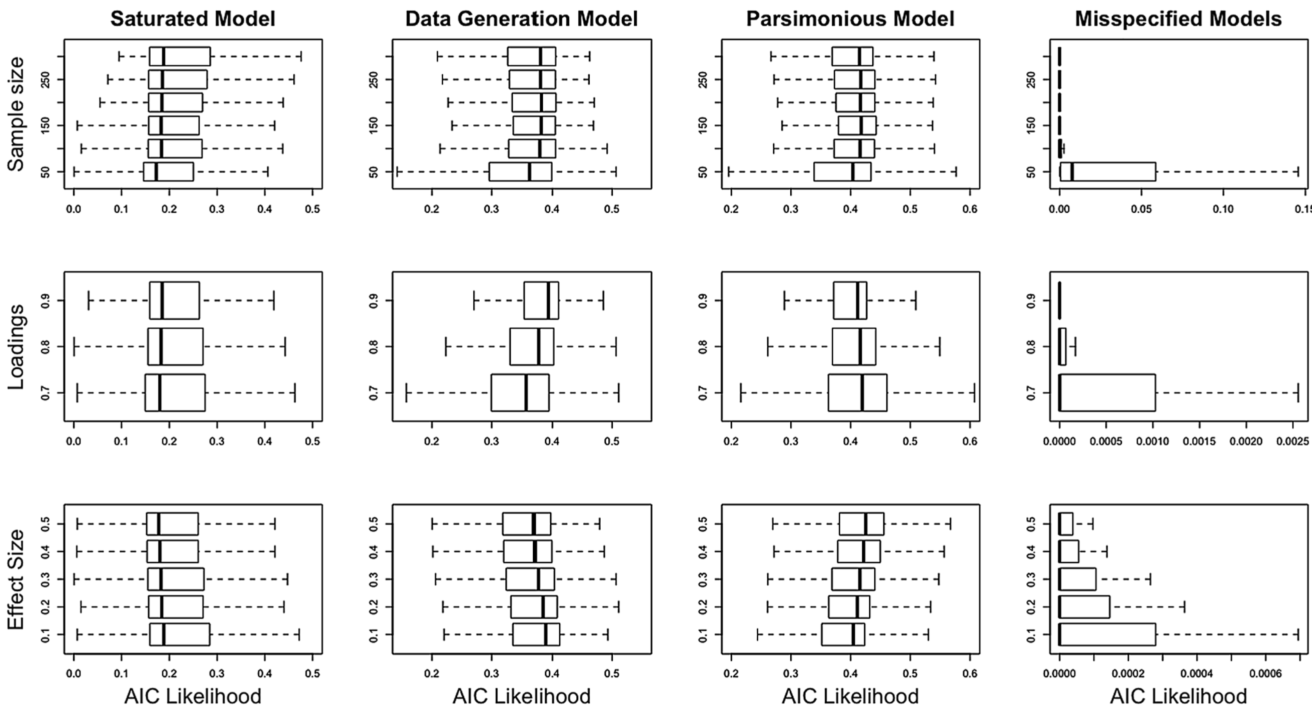

Fig. 3. Boxplots of simulation factors vs estimated AIC weights (simulation study I).

very low weight to the saturated model. Thus, overall BIC does a better job of reducing the selection uncertainty by penalizing the saturated model more strongly.

3.2.2. *Criteria performance across simulation conditions*

We analyze the impact of the simulation factors on the Akaike weights assigned to each model in order to demonstrate the performance of the model selection criteria for reducing both rejection and selection uncertainty under varying model and data conditions. Specifically, we consider the Akaike weights for four situations: (1) Model 5 (data generation model), (2) Model 2 (parsimonious model), (3) Model 7 (saturated model), and (4) sum of the weights assigned to the incorrectly specified Models 1, 3, 4, and 6 (misspecified models). We first run an analysis of variance (ANOVA) to identify which simulation conditions have a significant impact on the Akaike weights. We then plot the distributions of the AIC weights (Fig. 3) and BIC weights (Fig. 4) across the various simulation conditions to illustrate the impact of the simulation conditions on the weights. Finally, we conduct Tukey (1970) honest significant difference (HSD) tests to identify significant differences across the simulation conditions.

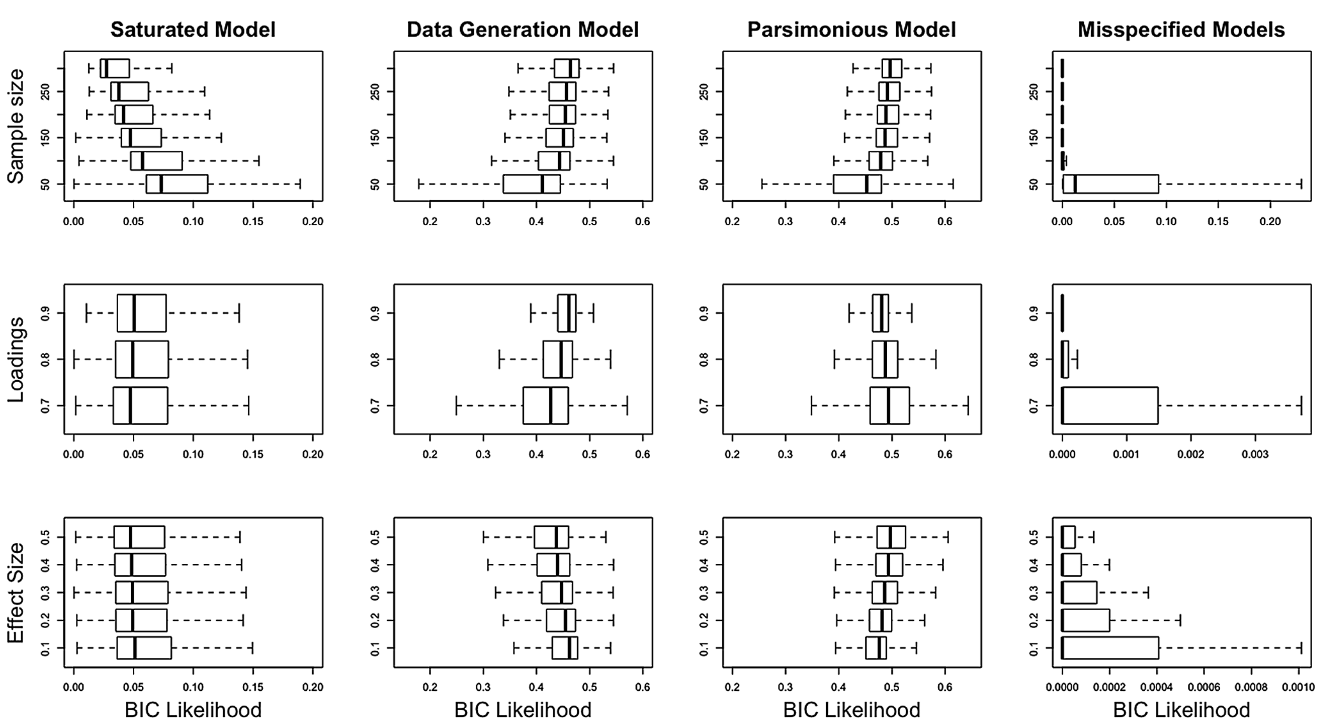

Fig. 4. Boxplots of simulation factors vs estimated BIC weights (simulation study I).

Table 1
ANOVA Results for AIC and BIC weights in simulation study I.

Factors	Df	Saturated Model				Data Generation Model				Parsimonious Model				Misspecified Models			
		Sum Sq	Mean Sq	F	Pr(> F)	Sum Sq	Mean Sq	F	Pr(> F)	Sum Sq	Mean Sq	F	Pr(> F)	Sum Sq	Mean Sq	F	Pr(> F)
AIC weights																	
Effect Size	4	0.90	0.237	9.717	< 0.001 ***	0.58	0.145	22.620	< 0.001 ***	4.04	1.010	117.870	< 0.001 ***	0.10	0.025	7.213	< 0.001 ***
Sample Size	5	2.90	0.581	23.779	< 0.001 ***	1.79	0.357	55.820	< 0.001 ***	1.86	0.371	43.340	< 0.001 ***	16.16	3.232	939.702	< 0.001 ***
Loadings	2	0.00	0.000	0.009	0.991	4.89	2.448	382.520	< 0.001 ***	1.09	0.546	63.790	< 0.001 ***	1.45	0.727	211.411	< 0.001 ***
Residuals	26,988	659.10	0.024			172.68	0.006			231.17	0.009			92.82	0.003		
BIC weights																	
Effect Size	4	0.18	0.045	4.299	0.002 **	1.54	0.384	89.980	< 0.001 ***	4.21	1.053	193.400	< 0.001 ***	0.18	0.045	7.741	< 0.001 ***
Sample Size	5	8.41	1.682	161.845	< 0.001 ***	16.14	3.227	755.880	< 0.001 ***	18.18	3.636	667.900	< 0.001 ***	32.66	6.531	1127.501	< 0.001 ***
Loadings	2	0.01	0.005	0.470	0.625	7.78	3.892	911.550	< 0.001 ***	1.13	0.565	103.800	< 0.001 ***	2.77	1.386	239.194	< 0.001 ***
Residuals	26,988	280.52	0.010			115.23	0.004			146.92	0.005			156.34	0.006		

3.2.2.1. Sample size. The ANOVA results demonstrate that sample size has a significant impact on both the AIC and BIC weights (Table 1) assigned to each category of the models (parsimonious, data generation, saturated, and misspecified). For AIC, increasing the sample size increases the weights assigned to the three correctly specified models and reduces the weight assigned to the misspecified models. Thus, increasing the sample size will increase the accuracy of AIC in selecting the correctly specified model at the risk of selecting the saturated or parsimonious model. The HSD tests indicate a significant difference in estimated weights between a sample size of 50 and sample sizes of 100, 150, 200, 250, and 500 (all $p < 0.001$); however, the differences become nonsignificant when comparing among the larger sample sizes. Thus, we find sample sizes smaller than 100 to be less amenable for reducing model selection uncertainty with AIC, and recommend sample sizes larger than this threshold to be used.

For BIC, an increase in sample size has a marked effect on weights assigned to all models. Increasing the sample size dramatically reduces the weight assigned to the saturated model, and increases the weight of the parsimonious and data generation models. The HSD tests show that this difference is significant across all sample sizes ($p < 0.001$). The decrease in weight for the saturated model is shared between the parsimonious and data generation models. This result demonstrates that BIC has a similar small sample size performance as AIC but increasingly favors the parsimonious and data generation models over the saturated model as the sample size increases.

3.2.2.2. Effect size. The ANOVA results demonstrate a significant impact of the effect sizes on all models for both AIC and BIC weights (Table 1). We find little difference in the performance of AIC and BIC in terms of this simulation condition and, thus, we will discuss these criteria jointly. The HSD test results show that the weights of the saturated, data generation, and misspecified models decrease, while the weights for the parsimonious model increase as effect sizes become larger. The differences are significant only for smaller effect sizes for the saturated model, but are significant across most of the effect sizes for the data generation and parsimonious models. A more detailed analysis of the simulation results offers a potential explanation for this surprising finding. The omission of the γ_2 relationship in the parsimonious model leads to a stronger relationship between ξ_1 and η_2, thereby increasing the focal construct's R^2 more strongly than in the data generation model as none of the variance is absorbed by the mediating effect via η_1. As the Information Theoretic model selection criteria are a function of the residual sum of squares, increasing the effect size increases the weights of the parsimonious model.

3.2.2.3. Loadings. AIC and BIC perform very similarly as loading patterns change (Table 1). The ANOVA demonstrates that the loadings have no significant effect on the Akaike weights for the saturated model, but that the effects are significant ($p < 0.001$) for the data generation, parsimonious, and misspecified models. The boxplots show that increasing the loadings has little impact on weights assigned to the saturated model (Figs. 3 and 4) with the HSD tests indicating no significant differences between loading sizes. For the data generation and parsimonious models, a different pattern emerges. Increasing the loadings increases the weights assigned to the data generation model and reduces the weights assigned to the parsimonious with the HSD tests showing all differences as significant. Increasing the loadings also has an impact on the misspecified models' weight, but the effect is negligible in size. Thus, we expect both AIC and BIC to perform more favorably in models with higher loadings, especially in terms of correctly selecting the data generation model over the parsimonious model.

3.2.3. Prediction accuracy

Next, we use the Akaike weights to generate model-averaged predictions of the construct η_2, which is the focal construct in our

Table 2
Average prediction error of candidate models and model-averaged predictions of simulation study I.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	AIC	BIC	GM	Equal
RMSE	0.834	0.749	0.895	0.959	0.749	0.891	0.752	0.750	0.750	0.750	0.785

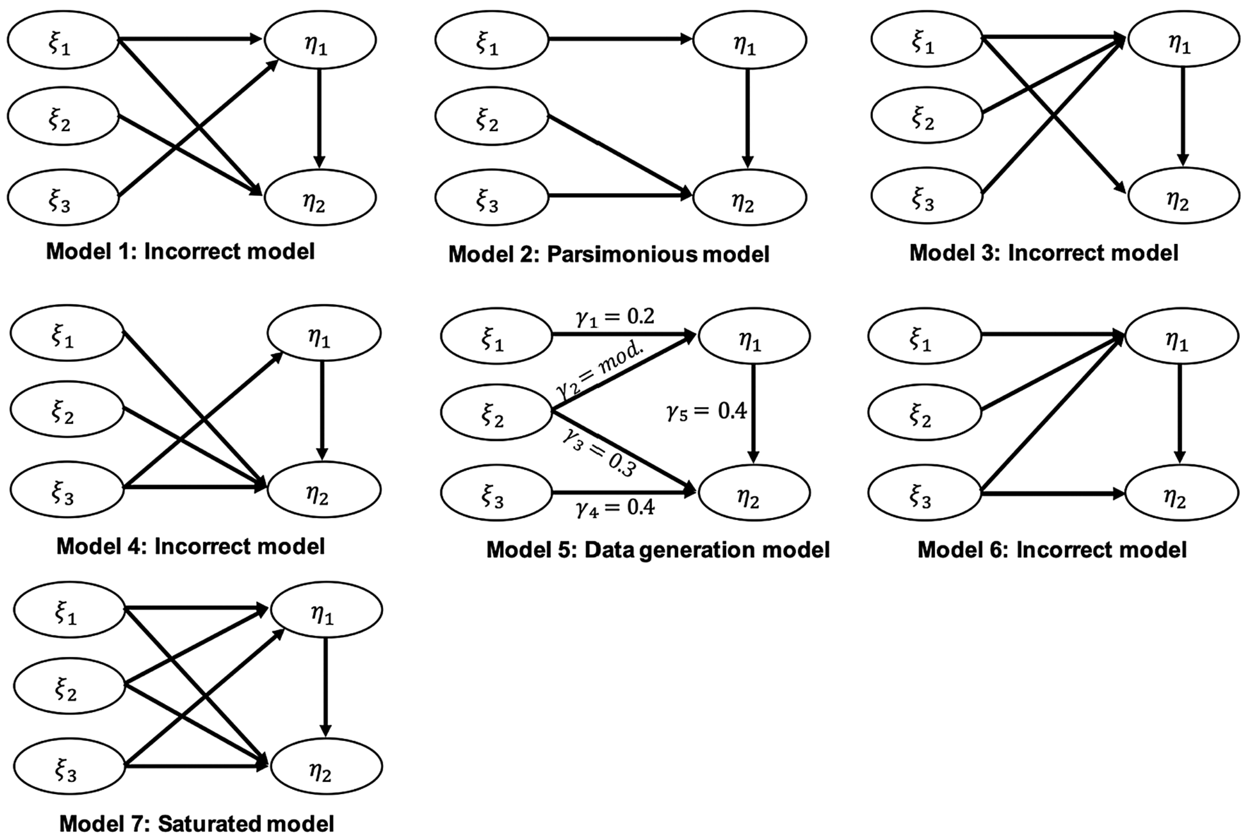

Fig. 5. Simulation study II models.

simulation study. These model-averaged predictions are calculated as a weighted sum of the predicted construct scores of all models, using Akaike weights for AIC, BIC, and GM. Table 2 demonstrates the results of our analysis of the candidate models and model-averaged predictions in terms of the predictive error (i.e., RMSE). We find that the model-averaged predictions produce an RMSE value of 0.750 when Akaike weights are estimated using AIC, BIC, and GM, which are almost the same as the prediction error produced by the data generation model (RMSE = 0.749). Thus, both AIC and BIC-based model-averaged predictions are as precise as the ones obtained by using the best predictive models in the set (Models 2 and 5), which points to their practical utility under situations of selection uncertainty where there are several equally good models (in the Information Theoretic sense). In contrast, the misspecified models yield RMSE values between 0.834 (Model 1) and 0.959 (Model 4), reflecting inferior predictive accuracy. In particular, model-averaged predictions are significantly better than the predictions obtained by misspecified models, which points to their utility under situations of rejection uncertainty where the researcher is unsure which models might be incorrectly specified.

In the final step, we contrast the efficacy of Akaike weights with the naïve approach of weighting each model equally when generating model-averaged predictions. Without the luxury of Akaike weights, equal weighting might be an appealing option for a researcher looking to generate model-averaged predictions. However, with the use of Akaike weights we expect the weighted-average to outperform this naïve benchmark. We find that this is indeed the case; creating unweighted model averages (i.e., assigning equal importance to all the models), produces an RMSE value of 0.785, which is lower than the prediction error in misspecified models but higher (i.e., worse) than when using Akaike weights. This result suggests that the Information Theoretic criteria assign low weights to the incorrectly specified models with high prediction error but give more weight to correct and parsimonious models that have low prediction errors. Overall, our analysis underlines the usefulness of Akaike weights in creating model-averaged predictions under conditions of model selection and rejection uncertainty.

4. Simulation study II

4.1. Design, data generation, and model estimation

In order to increase the generalizability of our findings, we ran a second simulation study, which extends the first in two ways: (1) we included an additional path γ_4 (between ξ_3 and η_2) in Models 4 and 6, which increases the variation in the degree of misspecification, and (2) increased the magnitude of the relationship γ_5 (between η_1 and η_2) from 0.1 to 0.4 in the data generation model in order to avoid any suppression of the impact of antecedent misspecifications (Fig. 5).[3]

[3] We would like to thank the anonymous reviewer for suggesting this extension.

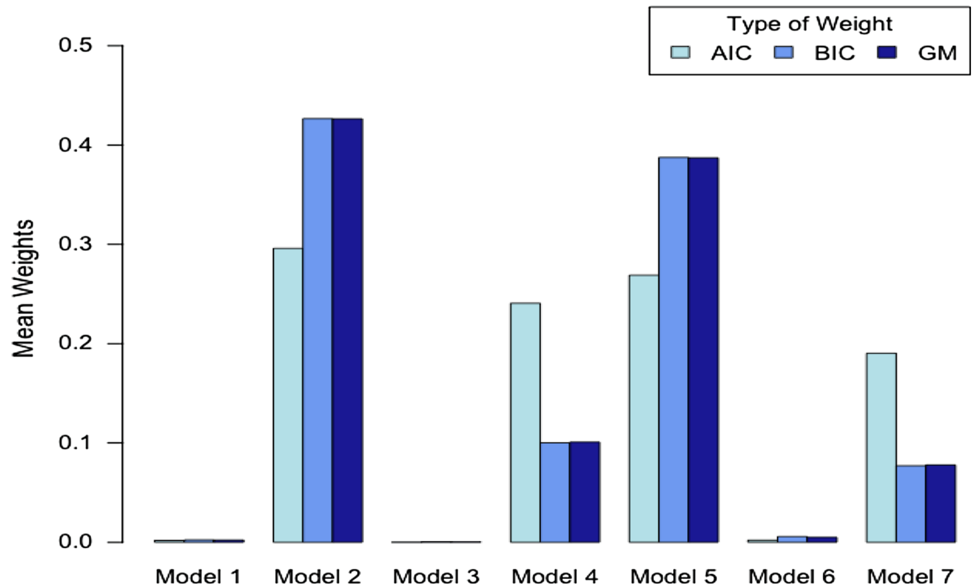

Fig. 6. Mean weights of AIC, BIC, and GM across models for simulation study II.

Table 3
Average prediction error of candidate models and model-averaged predictions of simulation study II.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	AIC	BIC	GM	Equal
RMSE	0.696	0.598	0.778	0.600	0.598	0.693	0.600	0.599	0.599	0.599	0.612

Table 4
Criteria values, deltas, and Akaike weights for candidate models.

	Model 1	Model 2	Model 3	Model 4	Model 5
AIC	−113.728	−113.461	−113.594	−120.849	−124.286
Δ (AIC - AIC_{min})	10.557	10.824	10.691	3.436	0.000
AIC weights	0.004	0.004	0.004	0.150	0.838
BIC	−102.206	−101.939	−102.073	−93.965	−97.401
Δ (BIC - BIC_{min})	0.000	0.267	0.134	8.241	4.805
BIC weights	0.343	0.300	0.321	0.006	0.031

Due to the structural similarity between Model 4 and Model 1 in terms of the direct antecedents of η_2, adding the path γ_4 serves to more strongly distinguish these two models from each other and the other five competing models. Furthermore, the addition of the γ_4 path in Model 6 increases the number of direct antecedents of η_2, and thus increases the model complexity (Fig. 1). As the path γ_4 has a magnitude of 0.4 in the data generation model, we expect that in addition to increasing the diversity of incorrectly specified models, the addition of this correctly specified path will also increase their explanatory and predictive power, thereby making them more favorable for model selection criteria. Finally, the increase in magnitude of γ_5 serves to magnify the effect of possible misspecifications in the model antecedent to, and mediated by, η_1 and to amplify the effect of modulating path γ_2.

Other than the changes described in the preceding paragraph, the experiment design, simulation setup, and model evaluation procedures are identical to those in the first simulation study. The following results discussions focus on the overall model rejection and selection uncertainties as well as the criteria's model-averaged prediction accuracies.[4]

[4] We do not report the criteria's performance across simulation conditions as these do not substantially differ from the first simulation study. The condition-specific results are available from the corresponding author upon request.

4.2. Results

With regard to rejection uncertainty, we find that both model selection criteria strongly reject the misspecified Models 1, 3, and 6, while Model 4 now receives some support (Fig. 6). The added predictive power of the γ_4 path has made this model more favourable for the model selection criteria, and thus both AIC and BIC assign a greater likelihood to this updated Model 4 than in the first simulation study. Among the two misspecified models, which are saturated in terms of η_2 (Models 4 and 7), AIC and BIC both favor the more parsimonious Model 4. However, there is a stark contrast in the way that the two model selection criteria perform in terms of rejecting these incorrectly specified models. AIC favors these two models, assigning high weights to Model 4 (Akaike weight 0.238) and Model 7 (Akaike weight 0.174) and thus doing a poorer job of reducing rejection uncertainty. On the contrary, BIC assigns much lower weights to these models and thus performs better at reducing the rejection uncertainty.

In terms of reducing model selection uncertainty, we see a similar pattern emerge as in the first simulation study. BIC performs better than AIC in terms of assigning higher weights to the data generation and parsimonious models (Fig. 6). Thus, uncertainty regarding selecting the correctly specified models can be reduced by using BIC, because there is a clear difference in the most highly weighted parsimonious model (Model 2 with BIC weight of 0.437) and data generation model (Model 5 with BIC weight of 0.392) compared the most highly weighted misspecified model (Model 4 with BIC weight of 0.095) and saturated model (Model 7 with BIC weight of 0.065).

In terms of prediction accuracy, we note that compared to study 1 (Table 2), the predictive power in this study has increased due to the increase in the effect size of the γ_5 path, as evidenced by lower RMSE values in Table 3. We find that the model-averaged predictions based on the Akaike weights produce an RMSE value of 0.599 for AIC, BIC and GM, which are practically the same as the prediction error produced by the data generation model (RMSE = 0.598). Model 4 now has similar predictive power to the saturated Model 7, due to their similarity in

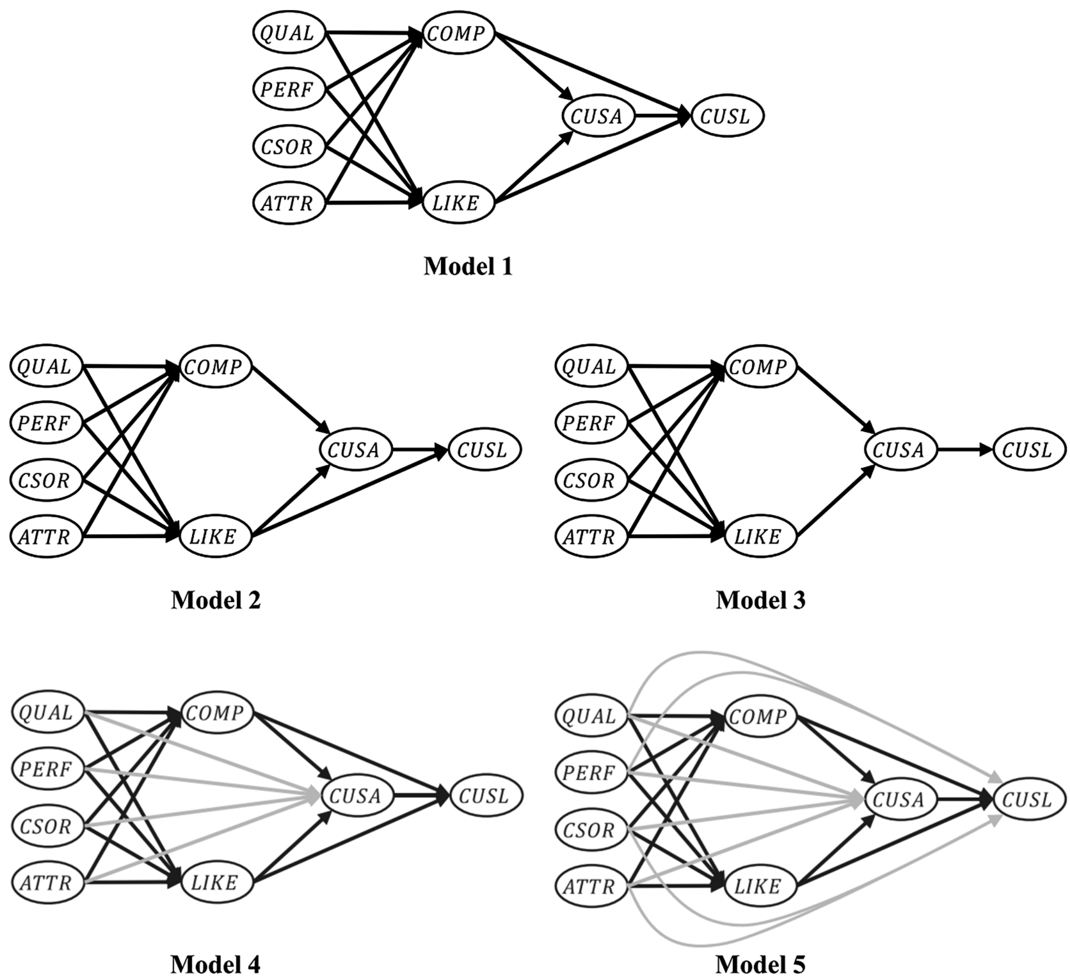

Fig. 7. Five alternate corporate reputation models. Notes: ATTR = Attractiveness, CSOR = Corporate social responsibility, COMP = Competence, CUSA = Customer satisfaction, CUSL = Customer loyalty, LIKE = Likeability, PERF = Performance, QUAL = Quality.

predictors. Additionally, Model 4 now has much better relative predictive performance than in the first simulation study, again due to the addition of path γ_5. Overall, the results support our main findings that both AIC and BIC-based model-averaged predictions are as precise as the ones obtained by using the best predictive models in the set (Models 2 and 5). Thus, when unsure of picking a single best model for predictive purposes, the researcher can utilize model-averaged predictions based on Akaike weights.

5. Empirical illustration

5.1. Model design and estimation

Our empirical illustration is similar to that of Sharma et al. (2020), with an updated focus on the efficacy of the Information Theoretic criteria for quantifying model selection and rejection uncertainty, and model-averaged prediction accuracy. The empirical set-up relies on the well-regarded corporate reputation model (Sarstedt, Wilczynski, & Melewar, 2013), which is highly relevant to the marketing literature—see Hair, Hult, Ringle, and Sarstedt (2017) for further information on the model set-up, construct measures, and dataset.[5] We present five theoretically grounded candidate model configurations as illustrated in Fig. 7, focusing on customer satisfaction (CUSA) as the immediate consequence of the two dimensions of corporate reputation (likeability, LIKE; competence, COMP). We conduct the analysis using the *SEMinR* PLS-SEM model estimation package (Ray, Danks, & Velasquez-Estrada, 2019) in the R statistical environment (R Core Team, 2019).

Extending the scope of the empirical study, we also contrast the predictive performance of the candidate models with the model-averaged predictions created using Akaike weights and equal weighting. Our analysis of predictive performance draws on an additional validation dataset collected for the corporate reputation model by Sarstedt, Hair, Cheah, Becker, and Ringle (2019). This empirical illustration is the first to evaluate predictive methodology in PLS-SEM using a true validation dataset, and not using cross-validation. We therefore avoid tapping into sample-specific patterns in the training dataset, giving a more realistic account of the generalizability of a model's predictive performance.

5.2. Practical demonstration of Akaike weights calculation

For the purposes of reproducibility, we demonstrate the calculation of Akaike weights from the model section criteria raw values.[6] The purpose of the demonstration is to provide researchers a tutorial regarding the calculations made in this research. We perform this demonstration for BIC as per Table 4.

[5] The two corporate reputation datasets can be downloaded at https://pls-sem.net.

[6] All code for the empirical example is open-source and shared in a public GitHub online repository at https://github.com/NicholasDanks/Model_selection_uncertainty. Researchers wishing to duplicate these methods for their model, can download, alter, and run the code for their unique set-ups.

Table 5
RMSE of candidate model predictions of validation data.

	Model 1	Model 2	Model 3	Model 4	Model 5	AIC	BIC	GM	Equal
RMSE	0.823	0.823	0.823	0.807	0.803	0.804	0.821	0.821	0.811

1. Select the BIC value with the smallest size amongst the competing model. We thus choose Model 1 with a BIC value of 102.206.
2. Apply Formula (1) to calculate a delta score for each BIC Model. We calculate the Model 2 delta score as follows:

   $-101.939 - (-102.206) = 0.267.$

   Repeat this calculation for Models 3, 4, and 5 to calculate 0.134, 8.241, and 4.805 respectively.
3. Apply Formula (2) to calculate the likelihoods of each model. We calculate the Model 2 likelihood as follows:

   $\exp(-0.5\cdot 0.267) = 0.875$

   Repeat this for Models 1, 3, 4, and 5 to calculate 1.0, 0.935, 0.016, and 0.090 respectively.
4. Apply Formula (3) to calculate the Akaike weights for BIC. We calculate the Model 2 BIC Akaike weight as follows:
   4.1. First, we sum the likelihoods for all models from step 3.

   $(1 + 0.875 + 0.935 + 0.016 + 0.090) = 2.916$

   4.2. Next, we apply Formula (3) to calculate the relative likelihood of Model 2 as follows:

   $\frac{0.875}{2.916} = 0.300$

   We repeat this for the four remaining models to calculate a BIC Akaike weight for each Model.

5.3. Model selection uncertainty

We concentrate our analysis of the comparative performance of AIC and BIC as our simulation results show that BIC and GM perform very similarly in terms of minimizing rejection and selection uncertainty (see Fig. 2). Table 4 demonstrates the results of our analysis of the models in terms of the Information Theoretic model selection criteria and Akaike weights. Table 5 describes the predictive performance of the candidate models and the performance of model-averaged predictions on the validation dataset.

With regard to rejection uncertainty (Table 4), we find a stark contrast in the results offered by AIC and BIC. AIC firmly rejects the parsimonious models (Models 1, 2 and 3) with very low weights of 0.004. In contrast, BIC clearly rejects the more complex Model 4 with a pronounced difference with Model 1, which yields the smallest BIC value ($BIC - BIC_{min} = 8.241$). BIC also rejects the saturated Model 5, but the difference with Model 1 is much less pronounced ($BIC - BIC_{min} = 4.805$). The low Akaike weight of 0.031 provides more compelling evidence that Model 5 is misspecified.

With regards to selection uncertainty (Table 4), AIC strongly favors the saturated Model 5 (weight = 0.838), with some support given to Model 4 (weight = 0.150). The fact that AIC has a systematic tendency to select over-parametrized models should give rise to concern regarding its choice as the preferred criterion in PLS-SEM studies for selecting a single correctly specified model. In contrast, BIC provides nearly equal support to Models 1, 2 and 3, but assigns the highest weight to Model 1 (weight = 0.343), which is the most theoretically defensible model in the set (Sarstedt et al., 2013).

5.4. Prediction accuracy

Moving to the analysis of the prediction error, we find that the complex Models 4 and 5 demonstrate the highest predictive power on the validation set with RMSE values of 0.807 and 0.803 respectively (Table 5). This is not surprising given Sharma et al.'s (2020) findings that RMSE itself often favors the saturated model. The AIC-based model-averaged prediction performs as well as the best predictive model (Model 5) on the validation data, largely due to AIC weights' heavy weighting of the saturated model. BIC model-averaged prediction performs marginally better than Models 1, 2, and 3 based on the validation data but does not outperform equal weighting. In particular, AIC based model-averaging outperforms BIC and equal weighting. This is not surprising, since Aho et al. (2014) note that when the "true" or data generating model is fairly complex and is not included in the set of models being analyzed, as is expected in most practical "real-world" situations, AIC tends to select models with better predictive accuracy, while BIC tends to select a model that best approximates the data generating model. This explains the weaker predictive performance of BIC in the empirical study than in the simulation study—where the data generation model was relatively simple and included in the candidate model set for the purpose of assessing model selection uncertainty. We thus recommend that AIC be used for generating model-averaged predictions in empirical research, while BIC should be used for reducing model selection and rejection uncertainty.

6. Discussion

For many behavioral phenomena under scientific study, there exist several possible explanations. Researchers typically formalize these competing explanations in the form of different models to facilitate their comparison using model selection criteria (Preacher & Merkle, 2012). Sharma et al. (2019, 2020) recently emphasized the efficacy of BIC and GM for model selection tasks in PLS-SEM, a multivariate method designed to analyze complex inter-relationships between observed and latent variables, which has gained massive dissemination during the last decade in various fields. Sharma et al. (2019, p. 383) conclude that "with a proper theoretical base and a strong study design, these criteria [i.e., BIC and GM] allow researchers to consider competing models within the PLS framework and select the best model among them." However, accepting a single model on the basis of only the raw criteria values might generate false sense of confidence as differences in values are often marginal (Preacher & Merkle, 2012).

Addressing this concern, researchers have proposed using Akaike weights, which can be interpreted as conditional probabilities for models (e.g., Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004), thereby offering an easier interpretation of the evidence for, or against, each model in the set. Tying in with this research, we examine the efficacy of Akaike weights derived from the AIC, BIC, and GM metrics in separating incorrectly specified from correctly specified models in the PLS-SEM context. We find that in the context of our simulation study design all three metrics are highly effective in separating incorrectly specified models from the correctly specified model, a parsimonious variant, and the saturated model. However, BIC and GM perform favorably in that both metrics put much lower weights on the saturated model compared to AIC.

We find that with an increase in sample size, measurement model loadings, and structural model effect sizes, Akaike weights become increasingly effective at reducing model *rejection uncertainty*. That is, as sample size, measurement model loadings, and structural model effect sizes increase, the weights for the misspecified Models 1, 3, 4, and 6 become smaller (i.e., these models get rejected more strongly). In terms of model *selection uncertainty*, we find that with an increase in sample size, AIC produces higher weights for the saturated model (Model 7), while BIC produces higher weights for the correctly specified model and its parsimonious variant. Thus, with an increase in sample size, BIC becomes more effective in reducing selection uncertainty than AIC. An increase in loadings induces higher weights in favor of the correctly specified model while reducing the weights of the parsimonious model for both AIC and BIC. This means that the increase in indicator loadings makes both AIC and BIC more effective in reducing selection uncertainty. Finally, an increase in structural model effect sizes increases the AIC- and BIC-based weights in favor of the parsimonious model and reduces the corresponding weights of the correctly specified and saturated models. These results indicate that some caution should be applied when using Akaike weights to select amongst models that include a strong indirect effect. The Akaike weights increasingly favor a model that is more parsimonious in the sense of paths which indirectly affect the focal outcome, when the effect sizes of these paths become larger. Future research should assess whether there are remedies to this shortcoming by, for example, first conducting model-selection by focusing on the mediator, and then proceeding to final outcome constructs.

Our analysis of prediction errors in the simulation study further supports the efficacy of the three metrics from a prediction perspective. We find that the accuracy of model-averaged predictions based on weights derived from AIC, BIC, and GM are almost identical to that of the best predictive model (i.e., the data generation model) and lower than when using equal weighting approach. In contrast to the equal weighting of the models, the prediction error based on the Akaike weights is highly robust against misspecified models included in the set.

The results of a second simulation study, which increases the differences in misspecifications, add to these conclusions. Specifically, the results highlight AIC's tendency to choose the saturated model and underline that BIC tends to favor a model that is not incorrectly specified. In addition, the results support our main findings that both AIC and BIC-based model-averaged predictions are as precise as the ones obtained by using the best predictive models in the set.

The empirical demonstration, paints a slightly different picture. When the data generating model is complex and likely not included in the set of models being analyzed, AIC generates more accurate predictions. We thus recommend that AIC be used when the goal is generating a model-averaged prediction in empirical research.

Overall, our results clearly speak in favor of using Akaike weights derived from BIC and GM in PLS-SEM-based model comparison tasks rather than (1) comparing models based on raw model selection criteria values, and (2) averaging model predictions based on equal weights. This recommendation particularly holds when comparing models that are close in the Information Theoretic sense and the researcher is unable to identify a single model as clearly superior.

Future research should extend our study by considering further simulation conditions. Specifically, while our simulation model has a similar structure and complexity as those frequently encountered in SEM simulations (Bandalos & Gagné, 2012; Paxton et al., 2001), future research could consider a broader variety of models with more pronounced differences in model specifications. Such extensions would facilitate assessing whether the Akaike weights pick up smaller Information Theoretic differences among the models. Similarly, while mimicking related simulation studies in PLS-SEM, future research could vary further simulation conditions such as the number of indicators per measurement model and the measurement model specification (reflective vs. formative). Considering different types of mediation models such as conditional process models (Sarstedt, Hair, Nitzl, Ringle, & Howard, 2020) would be particularly interesting in this regard. In addition, future simulation studies could consider a broader range of categories for each of the conditions. For example, Hair, Hollingsworth, et al.'s (2017) review of PLS-SEM use in management information systems revealed that some studies draw on much more complex models with up to 36 constructs and sample sizes of up to 1500. Examining the performance of the Akaike weights under such boundary conditions would further substantiate their general usefulness.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology, 95*(3), 631–636. https://doi.org/10.1890/13-1452.1.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csáki (Eds.). *Selected papers of Hirotugu Akaike* (pp. 199–213). New York: Springer.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*(1), 9–14. https://doi.org/10.1007/BF02480194.

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66*(2), 237–242. https://doi.org/10.1093/biomet/66.2.237.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*(1), 3–14. https://doi.org/10.1016/0304-4076(81)90071-3.

Ali, F., Rasoolimanesh, S. M., Sarstedt, M., Ringle, C. M., & Ryu, K. (2018). An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. *International Journal of Contemporary Hospitality Management, 30*(1), 514–538. https://doi.org/10.1108/IJCHM-10-2016-0568.

Anderson, E. W., & Fornell, C. G. (2000). Foundations of the American customer satisfaction index. *Total Quality Management, 11*(7), 869–882. https://doi.org/10.1080/09544120050135425.

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research, 40*(2), 235–243. https://doi.org/10.1509/jmkr.40.2.235.19225.

Armstrong, J. S. (2001). Combining forecasts. *Principles of forecasting* (pp. 417–439). Boston, MA: Springer. https://doi.org/10.1007/978-0-306-47630-3_19.

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.). *Handbook of structural equation modeling* (pp. 92–108). New York: The Guilford Press.

Becker, J.-M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. *Marketing Letters, 26*(4), 643–659. https://doi.org/10.1007/s11002-014-9299-9.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. https://doi.org/10.1007/BF02294361.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach* (pp. 69–113). Dordrecht: Springer. https://doi.org/10.1007/978-94-011-0800-3_3.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics, 24*(6), 2350–2383. https://doi.org/10.1214/aos/1032181158.

Breivik, E., & Thorbjørnsen, H. (2008). Consumer brand relationships: An investigation of two alternative models. *Journal of the Academy of Marketing Science, 36*(4), 443–472. https://doi.org/10.1007/s11747-008-0115-z.

Brown, G., & Yao, X. (2001). *On the effectiveness of negative correlation learning. Proceedings of first UK workshop on computational intelligence*.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* Heidelberg: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Research Methods, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195–212. https://doi.org/10.1007/BF01246098.

Chica, M., & Rand, W. (2017). Building agent-based decision support systems for word-of-mouth programs: A freemium application. *Journal of Marketing Research, 54*(5), 752–767. https://doi.org/10.1509/jmr.15.0443.

Daryanto, A. (2019). Avoiding spurious moderation effects: An information-theoretic approach to moderation analysis. *Journal of Business Research, 103*, 110–118. https://doi.org/10.1016/j.jbusres.2019.06.012.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., ... Guelat, J. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs, 88*(4), 485–504. https://doi.org/10.1002/ecm.1309.

Geweke, J., & Meese, R. (1981). Estimating regression models of finite but unknown order. *International Economic Review, 22*(1), 55–70. https://doi.org/10.2307/2526135.

Fornell, C. G., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing, 60*(4), 7–18. https://doi.org/10.1177/002224299606000403.

Hair, J. F., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems, 117*(3), 442–458. https://doi.org/10.1108/IMDS-04-2016-0130.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Thousand Oaks, CA: Sage.
Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017). Mirror, mirror on the wall: A comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science, 45*(5), 616–632. https://doi.org/10.1007/s11747-017-0517-x.
Konishi, S., & Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection—Functional approach. *Journal of Statistical Planning and Inference, 114*(1–2), 45–61. https://doi.org/10.1016/S0378-3758(02)00462-7.
McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection.* Singapore: World Scientific https://doi.org/10.1142/3573.
Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*(1), 190–204. https://doi.org/10.1006/jmps.1999.1283.
Nitzl, C. (2016). The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. *Journal of Accounting Literature, 37*(December), 19–35.
Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7.
Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology, 53*(5), 793–808. https://doi.org/10.1080/10635150490522304.
Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17*(1), 1–14. https://doi.org/10.1037/a0026804.
Core Team, R. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing https://www.R-project.org/.
Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615.
Ray, S., Danks, N. P., & Velasquez-Estrada, J. M. (2019). seminr: Domain-Specific Language for Building PLS Structural Equation Models. *R package version*(7).
Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing, 26*(4), 332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001.
Ringle, C. M., Sarstedt, M., & Schlittgen, R. (2014). Genetic algorithm segmentation in partial least squares structural equation modeling. *OR spectrum, 36*(1), 251–276. https://doi.org/10.1007/s00291-013-0320-0.
Ringle, C. M., Sarstedt, M., Mitchell, R., & Gudergan, S. P. (2020). Partial least squares structural equation modeling in HRM research (in press) *International Journal of Human Resource Management*. https://doi.org/10.1080/09585192.2017.1416655.
Sarstedt, M., Wilczynski, P., & Melewar, T. C. (2013). Measuring reputation in global markets—A comparison of reputation measures' convergent and criterion validities. *Journal of World Business, 48*(3), 329–339. https://doi.org/10.1016/j.jwb.2012.07.017.
Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies!. *Journal of Business Research, 69*(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007.
Sarstedt, M., Hair, J. F., Cheah, J. H., Becker, J. M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal, 27*(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003.
Sarstedt, M., Hair, J. F., Nitzl, C., Ringle, C. M., & Howard, M. C. (2020). Beyond a tandem analysis of SEM and PROCESS: Use PLS-SEM for mediation analyses!. *International Journal of Market Research* (in press).
Schlittgen, R. (2019). cbsem: Simulation, estimation and segmentation of composite based structural equation models. R package version 1.0.0. Available at: https://cran.r-project.org/web/packages/cbsem/index.html (accessed February 14, 2020).
Schwaiger, M. (2004). Components and parameters of corporate reputation: An empirical study. *Schmalenbach Business Review, 56*(1), 46–71.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.
Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360.
Sharma, P. N., Morgeson, F. V., III, Mithas, S., & Aljazzaf, S. (2018). An empirical and comparative analysis of E-government performance measurement models: Model selection via explanation, prediction, and parsimony. *Government Information Quarterly, 35*(4), 515–535. https://doi.org/10.1016/j.giq.2018.07.003.
Sharma, P. N., Sarstedt, M., Shmueli, G., Kim, K. H., & Thiele, K. O. (2019). PLS-based model selection: The role of alternative explanations in Information Systems research. *Journal of the Association for Information Systems, 40*(4), 346–397. https://doi.org/10.17705/1jais.00538.
Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., & Ray, S. (2020). Prediction-oriented model selection in partial least squares path modeling (in press) *Decision Sciences*. https://doi.org/10.1111/deci.12329.
Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J. H., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *European Journal of Marketing, 53*(11), 2322–2347. https://doi.org/10.1108/EJM-02-2019-0189.
Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology, 65*(1), 13–21. https://doi.org/10.1007/s00265-010-1037-6.
Tukey, J. W. (1970). *Exploratory data analysis* (limited preliminary ed.). Addison-Wesley Publishing Company.
Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540.
Vrieze, S. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127.
Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192–196. https://doi.org/10.3758/BF03206482.

Nicholas P. Danks is a PhD student at the Institute of Service Science in National Tsing Hua University, Taiwan. His research focuses on structural equation modeling, partial least squares, predictive methodology, and programming. Nicholas is a co-author and the primary maintainer of SEMinR, an open-source package for the R Statistical Environment for the estimation and evaluation of PLS path models.

Pratyush N. Sharma is an assistant professor in the Alfred Lerner College of Business & Economics, University of Delaware. He received his PhD in information systems from University of Pittsburgh in 2013. His research interests include online collaboration communities and networks, open source software development, and research methods used in information systems, particularly partial least squares structural equation modeling. In addition, he is also interested in issues surrounding technology use and adoption and human computer interaction. His research has been published in highly acclaimed journals such as the *Journal of the Association for Information Systems*, *Journal of Retailing, Decision Sciences*, *Journal of Information Systems*, *Journal of Business Research*, *Journal of International Marketing*, and *International Journal of Accounting Information Systems*. In addition, he has published several book chapters and presented his research at premier conferences such as the International Conference on Information Systems (ICIS), Americas Conference on Information Systems (AMCIS), INFORMS, and PLS.

Marko Sarstedt is a chaired professor of marketing at the Otto-von-Guericke-University Magdeburg (Germany) and Adjunct Professor at Monash University Malaysia. His main research interests are in the advancement of research methods and measurement to improve scholarly research and managerial decision-making. His research has been published in, for example, *Nature Human Behaviour*, *Journal of Marketing Research*, *Journal of the Academy of Marketing Science*, *International Journal of Research in Marketing*, *Organizational Research Methods*, *MIS Quarterly*, *Psychometrika*, *Journal of the Association for Information Systems*, *Decision Sciences*, *Journal of World Business*, and *Multivariate Behavioral Research*. Marko has coedited several special issues of leading journals and co-authored several widely adopted textbooks, including "A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)." He has been named member at Clarivate Analytic's 2019 Highly Cited Researcher List.