Prediction-oriented model selection in partial least squares path modeling

Pratyush Nidhi Sharma
University of Delaware
pnsharma@udel.edu

Galit Shmueli
National Tsing Hua University
galit.shmueli@iss.nthu.edu.tw

Marko Sarstedt
Otto-von-Guericke-University Magdeburg
marko.sarstedt@ovgu.de

Nicholas Danks
National Tsing Hua University
nicholas.danks@iss.nthu.edu.tw

Soumya Ray
National Tsing Hua University
soumya.ray@iss.nthu.edu.tw

Citation: Sharma, P.N., Shmueli, G., Sarstedt, M., Danks, N. & Ray, S. 2018. "Prediction-oriented Model Selection in Partial Least Squares Path Modeling," *Decision Sciences*, forthcoming.

Prediction-oriented model selection in partial least squares path modeling

Abstract

Partial least squares path modeling (PLS-PM) has become popular in various disciplines to model structural relationships among latent variables measured by manifest variables. To fully benefit from the predictive capabilities of PLS-PM, researchers must understand the efficacy of predictive metrics used. In this research, we compare the performance of standard PLS-PM criteria and model selection criteria derived from Information Theory, in terms of selecting the best predictive model among a cohort of competing models. We use Monte Carlo simulation to study this question under various sample sizes, effect sizes, item loadings, and model setups. Specifically, we explore whether, and when, the in-sample measures such as the model selection criteria can substitute for out-of-sample criteria that require a holdout sample. Such a substitution is advantageous when creating a holdout causes considerable loss of statistical and predictive power due to an overall small sample. We find that when the researcher does not have the luxury of a holdout sample, and the goal is selecting correctly specified models with low prediction error, the in-sample model selection criteria, in particular the Bayesian Information Criterion (BIC) and Geweke-Meese Criterion (GM), are useful substitutes for out-of-sample criteria. When a holdout sample is available, the best performing out-of-sample criteria include the root mean squared error (RMSE) and mean absolute deviation (MAD). Finally, we recommend against using standard the PLS-PM criteria (R^2, Adjusted R^2, and Q^2), and specifically the out-of-sample mean absolute percentage error (MAPE) for prediction-oriented model selection purposes. Finally, we illustrate the model selection criteria's practical utility using a well-known corporate reputation model.

Keywords: *Partial Least Squares Path Modeling (PLS-PM), Prediction, Model Selection, Model Selection Criteria, Monte Carlo Simulation*.

1. Introduction

Structural equation modeling (SEM) has become the quasi-standard in the social sciences to analyze cause-effect relationships between latent variables. Its ability to model latent variables while simultaneously taking into account various forms of measurement error makes it useful for a plethora of research questions (e.g., Babin, Hair & Boles, 2008; Steenkamp & Baumgartner, 2000). While factor-based SEM has been considered the standard approach to estimate structural equation models (Jöreskog, 1973), recently, partial least squares path modeling (PLS-PM), a composite-based approach to SEM, has gained vast dissemination in a variety of disciplines such as accounting (Nitzl, 2016), international management (Richter, Sinkovics, Ringle & Schlägel, 2016), operations management (Peng & Lai, 2012), management information systems (Hair, Hollingsworth, Randolph, & Chong, 2017a), and marketing (Hair, Sarstedt, Ringle & Mena, 2012a).

One reason for PLS-PM's attractiveness is that it allows researchers to estimate complex models with many constructs and indicator variables, even at low sample sizes (e.g., Rigdon, 2016; Henseler et al., 2014; Sarstedt, Ringle & Hair, 2017). In this light, it is not surprising that review studies of PLS-PM use show that path models estimated by this method are much more complex compared to those used in factor-based SEM studies, and that there is a general trend toward more complex PLS path models (e.g., Hair et al. 2012a; Hair, Sarstedt, Pieper & Ringle, 2012b; Ringle, Sarstedt & Straub, 2012). Using more complex model to map causal process is adequate when the primary goal of the analysis is to test or quantify the underlying causal relationship between cause and effect that can be generalized from the sample to the population of interest (i.e., explanatory modeling; Shmueli, 2010). However, when the goal of the analysis is to predict the output value of *new* cases by applying the model parameters estimated from one data sample (i.e., predictive modeling; Shmueli, 2010), complex models often perform poorly (e.g. Forster & Sober, 1994;

Hitchcock & Sober, 2004) as they tend to tap spurious patterns in the data (Myung, 2000). Because such patterns are sample-specific, an overly complex (i.e., overfitted) model will predict poorly and may not be generalizable or replicable by other researchers. In contrast, models with fewer parameters stand a better chance of having higher predictive power and being scientifically replicable (Bentler & Mooijaart, 1989). Thus, PLS-PM users should be aware of the trade-off between model complexity and the predictive accuracy.

However, fully grasping this trade-off requires researchers to have a sound understanding of PLS-PM's prediction capabilities, which research has only recently started to systematically explore. For example, Becker, Rai, & Rigdon (2013) examined the predictive accuracy of different PLS-PM estimation modes with models including formatively specified constructs, using a modified R^2 criterion that involves a comparison of sample and population composite scores. These authors show that the reflex-like use of Mode B estimation for formatively specified constructs is not optimal from a predictive modeling perspective under all conditions. Evermann & Tate (2016) recently extended this work by showing that PLS-PM has better predictive accuracy than factor-based SEM across a broad range of conditions commonly encountered in applied research. While these studies make valuable contribution to the literature on PLS-PM, they rely solely on the out-of-sample prediction criteria for judging predictive accuracy, which require the construction of a holdout sample for model comparison and selection. However, limitations in data availability often prevent the creation of a holdout sample in many studies. Furthermore, collecting additional data from the same population to be used as holdout can be prohibitively expensive or even impossible.

The regression literature offers a way of overcoming this dilemma by providing the means to evaluate a model's predictive accuracy using criteria that do *not* require the use of a holdout sample. Specifically, Akaike (1973) has shown that taking into account a model's fit to the data as

well its parsimony allows obtaining an unbiased estimate of its predictive accuracy. Based on this notion, research has brought forward a range of model selection criteria derived from Information Theory that optimize predictive accuracy by striking a balance between model fit and complexity (e.g., Akaike, 1973; Burnham & Anderson, 2002; Myung, 2000). While several studies have assessed the model selection criteria's predictive accuracy for models estimated using maximum likelihood (Faraway & Chatfield, 1998; Kuha, 2004), in the non-maximum likelihood context of PLS-PM, their performance has only been studied in terms of model selection accuracy for choosing a specific model among a set of alternative models (Sharma & Kim, 2012; Sharma, Sarstedt, Shmueli, Kim & Thiele, 2018). Their efficacy for predictive modeling, however, has remained unaddressed.

Addressing this gap in research, we systematically explore whether the model selection criteria can substitute for out-of-sample predictive criteria that require the use of a holdout sample when comparing and selecting models from a predictive modeling perspective and, if so, under which conditions. Such a substitution is especially advantageous because splitting datasets into training and holdout samples may cause substantial loss of statistical and predictive power, particularly when the overall dataset is not large. Utilizing the information contained in the entire dataset (rather than a subset) to derive the best predictive model could be particularly advantageous in PLS-PM as researchers routinely justify the use of this technique on the grounds of small sample sizes (e.g., Goodhue, Lewis & Thompson, 2012; Henseler et al., 2014; Rigdon, 2016).

The results from our Monte Carlo study indicate that when researchers do not have the luxury of a holdout sample and the goal is selecting correctly specified models with low prediction error, the model selection criteria, in particular BIC and GM, are useful substitutes for out-of-sample criteria that require a holdout sample. When the holdout sample is available, we find that

the best performing criteria for prediction are the RMSE and MAD, followed by SMAPE. Our results also advise against the use of the standard PLS-PM criteria (i.e. R^2, Adjusted R^2, and Q^2), and, in particular, the out-of-sample MAPE for prediction-oriented model selection purposes. Finally, using a well-known corporate reputation model, we illustrate the criteria's practical application by means of empirical data.

2. Model selection criteria

Regression literature has brought forward a range of criteria that allow comparing the predictive performance of alternative models in the PLS-PM context. The simplest criterion that could be used is the R^2. Given that R^2 will increase as predictors are added to the model and hence will select a more complex model, regression researchers have widely used the Adjusted R^2, which attempts to correct for model complexity by including a penalty proportional to the number of predictors in the model. However, the Adjusted R^2 lacks formal justification and is not suitable for assessing a model's predictive accuracy (Berk, 2008).

As an alternative to the R^2 metrics, researchers can revert to model selection criteria derived from Information Theory, which began to appear in the literature in the late 1960s and the early 1970s (McQuarrie & Tsai, 1998). These criteria strike a balance between model fit and complexity to avoid over-fitting so that the model generalizes beyond the particular sample (Myung, 2000). One of the first metrics to be proposed was Akaike's Final Prediction Error (FPE; Akaike, 1969) from which two widely used model selection criteria emerged: Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

The AIC and BIC differ somewhat in their conceptual underpinnings and assumptions. Specifically, BIC provides an estimate of the posterior probability of a model being true, and chooses the model that maximizes this probability on a given dataset. That is, it strives to select a model that is most likely (in the Bayesian sense) to coincide with the underlying data generation model. In

contrast, AIC is designed to estimate the relative amount of information lost (using the Kullback-Leibler divergence measure between distributions) when a given model estimated from data is compared to a "true" but unknown data generating process (Burnham & Anderson, 2002).

AIC's strength as a model selection criterion in terms of predictive accuracy has been shown empirically as well as theoretically (Burnham & Anderson, 1998). For example, Stone (1977) showed that the AIC and leave-one-out cross-validation are asymptotically equivalent. One disadvantage of AIC is that it is asymptotically inconsistent, in that if the set of models includes the "true" model (i.e., the data generation model in the case of a simulation set-up), then the probability of selecting the true model does not converge to one as the sample size approaches infinity (Shao, 1993). In contrast, BIC is consistent and, at the same time, puts a heavier penalty than AIC on model complexity (Vrieze, 2012). BIC is also related to cross-validation and has been shown to be asymptotically equivalent to leave-ν-out cross-validation. Regardless of their differences, however, there is no general agreement whether AIC or BIC should be given preference in empirical applications (Shi & Tsai, 2002).

Several variations of the original AIC and BIC criteria have also been proposed over the last decades including the Mallow's Cp Criterion, the Unbiased AIC (AICu), the Corrected AIC (AICc), the Geweke-Meese Criterion (GM), the Hannan-Quinn Criterion (HQ), and the corrected Hannan-Quinn Criterion (HQc) (Mcquarrie & Tsai, 1998). These criteria are typically written as a function of the maximum likelihood value. However, they can also be expressed as a function of the model residuals when the error distribution is normal with a constant variance (Burnham & Anderson, 2002, p. 63). This characteristic makes them suitable for PLS-PM estimation, which relies on an iterative estimation of piecewise linear regression models (e.g., Hair, Hult, Ringle & Sarstedt, 2017b).

All the criteria discussed above are considered *in-sample* in that their computation draws on

the entire data. That is, their computation requires estimating the parameters of a PLS model and then using the model to predict values for cases from the same sample. However, researchers using PLS-PM can also draw on out-of-sample criteria that require the use of a holdout sample. Specifically, these metrics analyze a model's predictive accuracy by using different types of summaries of prediction errors. Out-of-sample criteria that feature prominently in the regression literature include the mean absolute deviation (MAD), root mean square error (RMSE), mean absolute percentage error (MAPE), and symmetric mean absolute percentage error (SMAPE).

Finally, many PLS-PM studies draw on the Q^2 to assess a model's predictive accuracy (Geisser, 1974; Stone, 1974). This metric builds on the blindfolding procedure, which omits single points in the data matrix, imputes the omitted elements, and estimates the model parameters. Using these estimates as input, the blindfolding procedure predicts the omitted data points. This process is repeated until every data point has been omitted and the model re-estimated. As its computation does not draw on holdout samples, but on single data points (as opposed to entire observations) being omitted and imputed, the Q^2 can only be partly considered a measure of out-of-sample prediction (Nitzl & Chin, 2017). Therefore, in line with Shmueli, Ray, Estrada & Chatla (2016), we will treat the Q^2 as an in-sample criterion.

3. Monte Carlo Study

3.1 Data and experimental conditions

The Monte Carlo study analyzes the predictive performance of standard in-sample PLS-PM criteria (R^2, Adjusted R^2 and Q^2), and the model selection criteria that Sharma & Kim (2012) and Sharma et al. (2018) evaluated in the PLS-PM context: FPE, Cp, GM, AIC, AICu, AICc, BIC, HQ, and HQc.[1] Table A1 (in Appendix A) presents more details about the model selection criteria and their specific

[1] Note that we did not consider Tenenhaus et al.'s (2005) GoF index as prior research identified this metric as ineffective for model selection tasks (Henseler & Sarstedt, 2013).

formulations. We compare the performance of the above-mentioned criteria with the following out-of-sample criteria: MAD, RMSE, MAPE, and SMAPE (Shmueli et al., 2016). Table A2 (in Appendix A) presents more details about the predictive metrics.

Our choice of manipulated factors and their factor levels follows prior research (e.g., Hwang, Malhotra, Kim, Tomiuk & Hong, 2010; Sharma & Kim, 2012; Vilares & Coelho, 2013). These conditions compare well with those seen in the empirical applications using PLS as evidenced in prior reviews of the method's use (e.g., Hair et al., 2012a, 2017a; Ringle et al., 2012). Specifically, we manipulate the following factors:

- Six conditions of sample size (50, 100, 150, 200, 250, and 500).
- Five conditions of varying effect size on a structural path (0.1, 0.2, 0.3, 0.4, and 0.5).[2]
- Three factor loading patterns with different levels of average variance extracted (AVE):
    - High AVE with loadings: (0.9, 0.9, 0.9),
    - Moderate AVE with loadings: (0.8, 0.8, 0.8), and
    - Low AVE with loadings: (0.7, 0.7, 0.7).

The simulation study only considers the case of normally distributed data as recent research has shown that PLS-PM provides consistent estimates across data distributions when the underlying population is composite model-based (Hair et al., 2017c).[3] We generated composite model data using the procedure available in the SEGIRLS package for the R statistical software (Schlittgen, 2015)—see Ringle, Sarstedt & Schlittgen (2014) for more details on the data generation approach. All simulations were run in the R computing environment (R Development Core Team, 2014) using

[2] Note that a path coefficient of 0.1 is generally undesirable as it points to an insufficient degree of explanatory power. Considering such a condition in a simulation study, however, is important to understand the performance of the criteria under boundary conditions (e.g., Paxton, Curran, Bollen, Kirby, & Chen, 2001), particularly since many authors using PLS-PM report such small effect sizes in their analyses. In line with this argument, prior simulation research on PLS-PM has routinely considered similar effect sizes (e.g., Reinartz, Haenlien, & Henseler, 2009; Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016; Hair, Hult, Ringle, Sarstedt, & Thiele, 2017c).

[3] Nevertheless, we conducted additional robustness checks in non-normal distribution settings by utilizing the log-normal distribution that is characterized by extremely high levels of skewness and kurtosis (Hair et al., 2017c). These results (in Appendix C) show that our conclusions are robust across the normal and log-normal distributions.

the *sempls* package (Monecke, 2012). We ran 200 replications for each of the 90 simulation conditions, yielding a total of 18,000 runs.

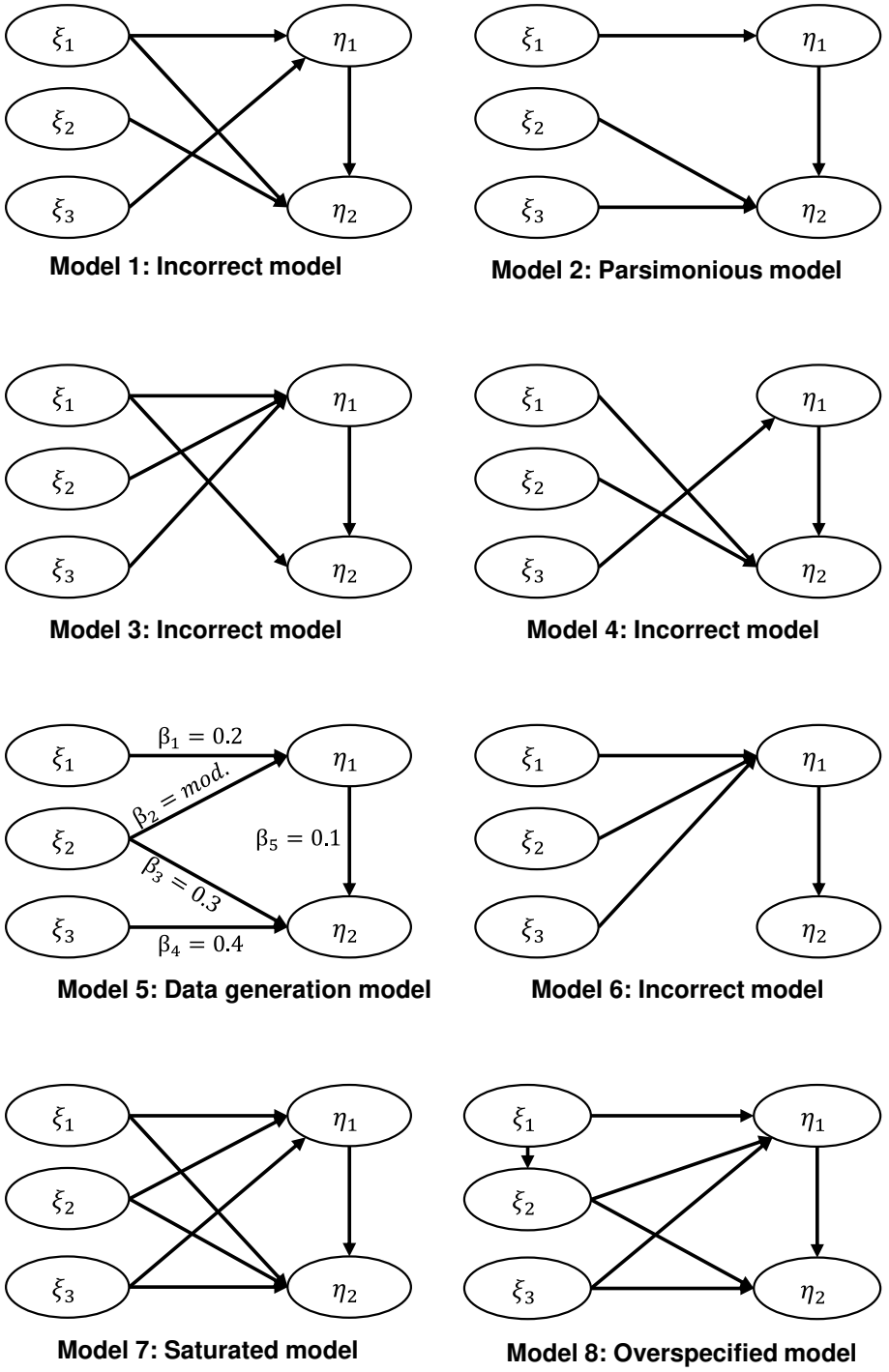
3.2 Model Estimation and Measurement

Drawing on the recommendations by Paxton et al. (2001), we utilize models of similar structure and complexity as those commonly encountered in management information systems (MIS) research. We chose this discipline as PLS-PM was first adopted (Chin, 1998) and still features very prominently in MIS (e.g., Hair et al. 2017a; Ringle et al., 2012). As a result of this review, our simulation model has a similar structure and complexity as those commonly encountered in MIS research, such as the unified theory of acceptance and use of technology (UTAUT) model (Venkatesh, Morris, Davis & Davis, 2003; Venkatesh, Brown, Maruping & Bala, 2008) or other models of information systems success (e.g., Polites & Karahanna, 2012; Iyengar, Sweeney & Montealegre, 2015; Park, Sharman & Rao, 2015). Furthermore, the model is similar to those used in prior PLS-PM-based simulation studies (e.g., Reinartz, Haenlein & Henseler, 2009; Ringle et al., 2014), most notably Dijkstra & Henseler's (2015) study on consistent PLS.

A set of eight potential models that differed from each other on certain paths formed the competing set (Figure 1). Each model had three reflectively measured exogenous variables (ξ_1, ξ_2, and ξ_3) and two reflectively measured endogenous variables (η_1 and η_2). The focal endogenous variable of interest was η_2. Model 5 served as the data generation model. Models 1, 3, 4, and 6 were incorrectly specified with respect to direct paths into η_2; that is, they had incorrect paths that were not consistent with the data generation process. Model 2 was correctly specified with respect to the direct paths into η_2, but was underspecified relative to the data generation process.[4] Model 7 was a fully saturated model with all possible paths into η_2, including one incorrect path into η_2 relative to the data generating model. Finally, Model 8 was correctly specified with respect to direct paths into

[4] In SEM terminology, Model 2 would be considered a nested (or restricted) version of the data generating Model 5.

η_2, but was overspecified relative to the data generation process. We describe the choice of competing models under different scenarios in more detail in section 3.3.



Note: The effect size on β_2 was one of the modified design factors with values .1, .2, .3, .4 and .5.

Figure 1: The eight competing models.

To evaluate the criteria's predictive accuracy, we created a holdout set (n = 1,000) for each experimental condition to mimic the population that the training sample originated from. We chose a large holdout sample to obtain the most precise estimates for the out-of-sample criteria (e.g., RMSE). In other words, the size of the holdout set was set for evaluation purposes and for best calibration against "out of sample performance." Table 1 presents the procedure used for assessing the predictive model selection performance of in-sample criteria vis-à-vis the out-of-sample criteria.

Table 1: Procedure for assessing predictive model selection performance.

Step #	Details
1	Generate training data according to the data generating model (e.g., Model 5 in Scenario 1) by manipulating the different experimental conditions (sample size, effect size, and loadings). Using the same population parameters as the training set, generate the holdout sample (n=1,000).
2	Estimate all the eight competing models using the PLS-PM algorithm on the training data.
3	Compute the in-sample criteria for all eight competing models using the training data, including the PLS-PM criteria and the model selection criteria.
4	Compute out-of-sample criteria for all eight competing models using the training sample PLS-PM parameters and holdout sample items as outlined by Shmueli et al. (2016).
5	Record which of the eight models is chosen as the best by each of the in-sample and out-of-sample criteria.
6	Compare the best model selected by each in-sample criterion to the RMSE-selected model.

Our analysis considers the following dependent variables: The value for each criterion for each of the eight models (PLS-PM, model selection, and out-of-sample criteria) along with a binary variable that assumed the value 1 if a criterion selected the model with the best predictive accuracy (measured in terms of the model with the lowest out-of-sample RMSE), 0 otherwise.

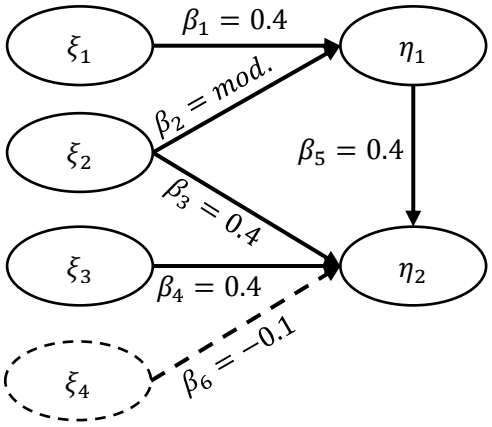
3.3 Different model setups: Two scenarios

Our simulation design considers two scenarios, one where the data generation model was included in the set of models to select from, and another where this model was excluded from the consideration set.

Scenario 1: The first scenario assumes that the researcher possesses all the context-specific

variables that generated the data and that there were no hidden (or inaccessible) variables. In addition, this scenario also assumes that the researcher correctly theorizes and includes the data generation model in the cohort of competing models. While this scenario is unlikely to occur in practice, it allows us to benchmark the performance of the in-sample criteria, vis-à-vis the out-of-sample criteria. In this case, Model 5 was the data generating model and was included in the set of the eight competing models (Figure 1). We selected these eight models because they allowed comparing incorrect, underspecified, overspecified, and saturated models. The underspecified models may outperform more complex models in their predictive accuracy (Shmueli, 2010). The inclusion of saturated model in this set allowed us to benchmark the in-sample explanatory power (R^2) and assess the relative improvements in prediction achieved by other models.

Scenario 2: The second, and more realistic, scenario expanded the first by explicitly accepting the existence of a hidden, or otherwise unavailable, latent variable that helped generate the data. Under most exploratory research scenarios, it is practical to assume that researchers may not possess all the variables that took part in the data generation process. Therefore, we incorporated an extraneous hidden variable (ξ_4) that was unavailable to the researcher but that directly impacted the focal endogenous variable of interest (η_2). As illustrated in Figure 2, we created a Model X that generated the data by relying on the hidden variable (ξ_4). Note that in this case the data generation Model X is automatically excluded from the competing set of models because the hidden variable (ξ_4) rendered it out of researcher's reach.



Model X: Data generation model
(NOT INCLUDED IN CASE 3)

Figure 2: Data generating Model X for Scenario 2 with hidden variable ξ_4

In each of the two scenarios above, prediction-oriented model comparison may be performed using one of the two possible lenses that reflect the nature of the underlying theory and the goal of the study: (1) Prediction only (***P***), and (2) Balanced explanation and prediction (***EP***). When the researcher is utilizing the prediction only lens (***P***), the model selected is expected to have the best predictive accuracy among the cohort but is not required to have well-developed causal explanations (Gregor, 2006). In this case the role of theory is limited, and the out-of-sample criteria (e.g. RMSE) are the "gold-standard" by which the models must be judged. A model with the best predictive accuracy is chosen regardless of whether it is correctly specified or not (Shmueli & Koppius, 2011). In the context of this study, we ask the question: which in-sample criteria can help select the best predictive model (e.g., per RMSE) regardless of whether the selected model is correctly specified or not?

Alternatively, when the researcher is utilizing the ***EP*** lens, the goal of prediction-oriented model comparison is to select a model that "provides predictions and has both testable propositions and causal explanations" (Gregor, 2006, p. 620). This perspective corresponds to Jöreskog and Wold's (1982, p. 270) understanding of PLS-PM who labeled the method as a "causal-predictive" technique, meaning that when structural theory is strong, path relationships can be interpreted as

causal. Correspondingly, many authors using PLS-PM emphasize the predictive nature of their analyses (e.g., Hair et al., 2012a, b; Ringle et al., 2012), while, at the same time, testing a path model derived from causal theory (e.g., Rönkkö & Evermann, 2013; Sarstedt et al., 2016, 2017). Our paper follows this common practice by assuming that the researcher's primary constraint is that the model selected should be correctly specified first and foremost—that is, the model should be consistent with the data generation process. This is in contrast to many machine learning forecasting methods, such as artificial neural networks, where prediction is critical but theoretical consistency may be of secondary concern (e.g., Shmueli & Koppius, 2011). Thus, a more relevant question in the context of PLS-PM is, which criteria can be used to select a suitably predictive yet correctly specified model? This question assumes that the researcher is willing to accept some trade-off in predictive power to remain within the correctly specified set. In the following section, we analyze the relative strengths of in-sample criteria using the two lenses (i.e. ***P*** and ***EP***), but with stronger focus on balanced explanation and prediction.

4. Results

4.1 Scenario 1: Data generation model included in the consideration set

4.1.1 Overall results and benchmarking

The first set of results is for the case where the generating model (Model 5) was part of the consideration set. We calculated each criterion for each of the eight models and chose the model that achieved the best score for each criterion. In the case of PLS-PM criteria, a model with the highest value was considered the best. In contrast, a model with the lowest value on all model selection and out-of-sample criteria was considered the best (McQuarrie & Tsai, 1998; Burnham & Anderson, 2002; Shmueli et al., 2016). Table 2 shows the average choice percentages for each model per criterion (PLS-PM, model selection, and out-of-sample criteria) across all conditions of sample sizes, item loadings, and effect sizes.

The results show that except for MAPE, all out-of-sample criteria show a very similar performance, choosing Models 2, 5, 7, or 8, with greater preference for Model 2. MAPE's choice shares ranged across all possible models, with heavier preference towards the incorrect and underspecified Models 3 and 6. The model selection criteria show a similar performance as the out-of-sample criteria, but with stronger preferences for Model 2 with choice shares ranging between 63.8% (AIC and FPE) to 74.3% (GM). However, unlike most model selection criteria, GM and BIC rarely chose the saturated Model 7. In contrast, among the PLS-PM criteria, except for Q^2 that behaved somewhat similar to the out-of-sample criteria, all other criteria chose either Models 2 or 7, with R^2 heavily preferring Model 7.

Table 2: Overall proportion of model choice by each criterion (Scenario 1)

	Model #	1	2	3	4	5	6	7	8
PLS-PM Criteria	**R^2**	0.000	0.273	0.000	0.003	0.019	0.000	0.695	0.009
	Adjusted R^2	0.000	0.537	0.000	0.005	0.074	0.000	0.303	0.081
	Q^2	0.003	0.305	0.000	0.004	0.224	0.002	0.179	0.281
Model Selection Criteria	**FPE**	0.000	0.638	0.000	0.006	0.091	0.000	0.163	0.101
	CP	0.000	0.686	0.000	0.006	0.100	0.001	0.096	0.111
	GM	0.000	0.743	0.000	0.006	0.109	0.007	0.011	0.123
	AIC	0.000	0.638	0.000	0.006	0.091	0.000	0.164	0.101
	AICu	0.000	0.688	0.000	0.006	0.099	0.002	0.093	0.112
	AICc	0.000	0.649	0.000	0.006	0.093	0.001	0.146	0.104
	BIC	0.000	0.731	0.000	0.006	0.107	0.005	0.032	0.120
	HQ	0.000	0.695	0.000	0.006	0.100	0.001	0.085	0.112
	HQc	0.000	0.705	0.000	0.006	0.102	0.002	0.070	0.114
Out-of-Sample Criteria	**MAD**	0.000	0.351	0.000	0.000	0.183	0.000	0.236	0.229
	RMSE	0.000	0.365	0.000	0.000	0.186	0.000	0.218	0.230
	MAPE	0.094	0.044	0.247	0.076	0.044	0.347	0.090	0.058
	SMAPE	0.000	0.365	0.000	0.000	0.123	0.000	0.343	0.168

To further compare the criteria's patterns of model choices, we examined the distributions of each criterion's values across the eight models (see Figure B1 in Appendix B). This analysis shows that the RMSE values have nearly identical distribution for Models 2, 5, and 8. The same was true

for out-of-sample MAD, SMAPE, the in-sample model selection criteria, and the Q^2. We also find slightly worse values for Model 7 for all these criteria. In contrast, R^2 achieved the highest value for Model 7, while the Adjusted R^2 had identical value distributions for Models 2, 5, 7, and 8.

As RMSE is generally preferred as the "default" in predictive modeling over other criteria (e.g., Chica and Rand, 2017; Nau, 2016), our subsequent analyses draw on this criterion as the predictive "gold standard" by which we judge the relative performances of in-sample criteria in selecting a predictive model. However, we note that all out-of-sample criteria—except MAPE—behaved very similarly.

4.1.2 Assessing performance of in-sample criteria using the prediction-only lens (P)

Next, we compared the choices of the different criteria (PLS-PM and model selection criteria) to the choice made by RMSE. Table 3 shows the percentage of cases where each criterion agreed with RMSE's choice (strict agreement); that is, the percentage of cases where both the RMSE and the in-sample criteria chose the same model in the same run, regardless of whether the model chosen was correctly or incorrectly specified. This table also breaks down the results by loading value, sample size, and effect size. We see that the model selection criteria agreed with the "best model" chosen by RMSE between 16-40% of the time (more so with larger effect size and loading value). The PLS-PM criteria behaved differently, with much lower and more variable agreement rates with RMSE (between 9-44%), and were at their highest agreement with RMSE at higher sample sizes. Among the PLS-PM criteria, Q^2 was the least variable.

Table 3: Percentage agreement between RMSE and each criterion on "best model" (Scenario 1).

Experimental Condition			PLS Criteria			Model Selection Criteria								
Loading Values	Effect Size	Sample Size	R^2	Adjusted R^2	Q^2	FPE	CP	GM	AIC	AICu	AICc	BIC	HQ	HQc
0.7	*0.1*	*50*	*0.125*	*0.190*	*0.185*	*0.210*	*0.225*	*0.215*	*0.205*	*0.230*	*0.225*	*0.220*	*0.230*	*0.230*
0.7	*0.1*	*100*	*0.170*	*0.225*	*0.230*	*0.265*	*0.280*	*0.300*	*0.265*	*0.280*	*0.275*	*0.295*	*0.280*	*0.285*
0.7	*0.1*	*150*	*0.235*	*0.190*	*0.225*	*0.210*	*0.225*	*0.230*	*0.210*	*0.220*	*0.215*	*0.225*	*0.225*	*0.215*
0.7	*0.1*	*200*	*0.225*	*0.165*	*0.155*	*0.165*	*0.170*	*0.190*	*0.165*	*0.180*	*0.160*	*0.180*	*0.185*	*0.185*
0.7	*0.1*	*250*	*0.305*	*0.210*	*0.205*	*0.210*	*0.225*	*0.230*	*0.210*	*0.215*	*0.210*	*0.225*	*0.210*	*0.205*
0.7	*0.1*	*500*	*0.415*	*0.255*	*0.185*	*0.220*	*0.210*	*0.185*	*0.220*	*0.200*	*0.225*	*0.180*	*0.195*	*0.195*
0.7	*0.2*	*50*	*0.240*	*0.230*	*0.210*	*0.245*	*0.260*	*0.260*	*0.240*	*0.260*	*0.260*	*0.260*	*0.260*	*0.265*
0.7	*0.2*	*100*	*0.205*	*0.225*	*0.185*	*0.255*	*0.290*	*0.320*	*0.255*	*0.295*	*0.275*	*0.315*	*0.295*	*0.305*
0.7	*0.2*	*150*	*0.220*	*0.230*	*0.195*	*0.260*	*0.275*	*0.310*	*0.260*	*0.275*	*0.260*	*0.290*	*0.275*	*0.275*
0.7	*0.2*	*200*	*0.250*	*0.185*	*0.190*	*0.200*	*0.210*	*0.265*	*0.200*	*0.215*	*0.200*	*0.240*	*0.220*	*0.225*
0.7	*0.2*	*250*	*0.240*	*0.210*	*0.200*	*0.230*	*0.255*	*0.280*	*0.230*	*0.265*	*0.235*	*0.275*	*0.265*	*0.265*
0.7	*0.2*	*500*	*0.375*	*0.255*	*0.265*	*0.235*	*0.230*	*0.255*	*0.235*	*0.235*	*0.230*	*0.260*	*0.235*	*0.245*
0.7	*0.3*	*50*	*0.190*	*0.265*	*0.180*	*0.305*	*0.310*	*0.310*	*0.305*	*0.315*	*0.315*	*0.300*	*0.310*	*0.315*
0.7	*0.3*	*100*	*0.180*	*0.220*	*0.235*	*0.270*	*0.270*	*0.330*	*0.270*	*0.285*	*0.270*	*0.325*	*0.285*	*0.290*
0.7	*0.3*	*150*	*0.255*	*0.260*	*0.160*	*0.300*	*0.315*	*0.335*	*0.300*	*0.315*	*0.315*	*0.335*	*0.320*	*0.325*
0.7	*0.3*	*200*	*0.245*	*0.245*	*0.220*	*0.260*	*0.255*	*0.290*	*0.260*	*0.260*	*0.260*	*0.285*	*0.275*	*0.275*
0.7	*0.3*	*250*	*0.225*	*0.200*	*0.160*	*0.200*	*0.205*	*0.250*	*0.200*	*0.205*	*0.200*	*0.230*	*0.210*	*0.215*
0.7	*0.3*	*500*	*0.385*	*0.335*	*0.260*	*0.345*	*0.370*	*0.340*	*0.345*	*0.355*	*0.345*	*0.340*	*0.350*	*0.345*
0.7	*0.4*	*50*	*0.155*	*0.260*	*0.240*	*0.270*	*0.285*	*0.290*	*0.265*	*0.285*	*0.280*	*0.285*	*0.280*	*0.285*
0.7	*0.4*	*100*	*0.185*	*0.220*	*0.170*	*0.260*	*0.285*	*0.310*	*0.260*	*0.290*	*0.270*	*0.305*	*0.290*	*0.305*
0.7	*0.4*	*150*	*0.170*	*0.200*	*0.205*	*0.225*	*0.235*	*0.245*	*0.225*	*0.230*	*0.235*	*0.235*	*0.230*	*0.230*
0.7	*0.4*	*200*	*0.260*	*0.290*	*0.210*	*0.330*	*0.325*	*0.350*	*0.330*	*0.330*	*0.325*	*0.340*	*0.340*	*0.340*
0.7	*0.4*	*250*	*0.310*	*0.305*	*0.250*	*0.320*	*0.325*	*0.350*	*0.320*	*0.320*	*0.320*	*0.345*	*0.325*	*0.330*
0.7	*0.4*	*500*	*0.275*	*0.265*	*0.220*	*0.265*	*0.280*	*0.325*	*0.265*	*0.280*	*0.265*	*0.320*	*0.285*	*0.285*
0.7	*0.5*	*50*	*0.240*	*0.250*	*0.245*	*0.265*	*0.300*	*0.305*	*0.265*	*0.305*	*0.280*	*0.310*	*0.280*	*0.320*
0.7	*0.5*	*100*	*0.170*	*0.255*	*0.235*	*0.290*	*0.305*	*0.345*	*0.285*	*0.320*	*0.300*	*0.340*	*0.320*	*0.335*
0.7	*0.5*	*150*	*0.235*	*0.270*	*0.220*	*0.295*	*0.325*	*0.350*	*0.295*	*0.330*	*0.305*	*0.345*	*0.335*	*0.345*
0.7	*0.5*	*200*	*0.260*	*0.270*	*0.210*	*0.270*	*0.280*	*0.330*	*0.270*	*0.295*	*0.270*	*0.325*	*0.300*	*0.300*
0.7	*0.5*	*250*	*0.195*	*0.220*	*0.210*	*0.270*	*0.270*	*0.305*	*0.270*	*0.280*	*0.270*	*0.295*	*0.280*	*0.280*
0.7	*0.5*	*500*	*0.320*	*0.265*	*0.195*	*0.255*	*0.245*	*0.275*	*0.255*	*0.240*	*0.255*	*0.280*	*0.245*	*0.255*
0.8	*0.1*	*50*	*0.150*	*0.230*	*0.225*	*0.260*	*0.290*	*0.275*	*0.260*	*0.280*	*0.270*	*0.280*	*0.275*	*0.285*
0.8	*0.1*	*100*	*0.215*	*0.215*	*0.200*	*0.250*	*0.275*	*0.310*	*0.250*	*0.270*	*0.260*	*0.300*	*0.270*	*0.285*
0.8	*0.1*	*150*	*0.230*	*0.165*	*0.145*	*0.180*	*0.205*	*0.220*	*0.180*	*0.205*	*0.190*	*0.215*	*0.205*	*0.205*
0.8	*0.1*	*200*	*0.200*	*0.135*	*0.170*	*0.165*	*0.170*	*0.200*	*0.165*	*0.170*	*0.160*	*0.200*	*0.180*	*0.180*
0.8	*0.1*	*250*	*0.250*	*0.195*	*0.225*	*0.235*	*0.265*	*0.305*	*0.235*	*0.260*	*0.240*	*0.300*	*0.275*	*0.275*
0.8	*0.1*	*500*	*0.390*	*0.325*	*0.245*	*0.260*	*0.255*	*0.250*	*0.260*	*0.260*	*0.260*	*0.250*	*0.265*	*0.265*
0.8	*0.2*	*50*	*0.175*	*0.240*	*0.220*	*0.305*	*0.335*	*0.340*	*0.305*	*0.330*	*0.310*	*0.340*	*0.325*	*0.335*
0.8	*0.2*	*100*	*0.110*	*0.265*	*0.250*	*0.295*	*0.310*	*0.330*	*0.295*	*0.310*	*0.300*	*0.330*	*0.310*	*0.315*
0.8	*0.2*	*150*	*0.215*	*0.240*	*0.185*	*0.270*	*0.270*	*0.300*	*0.270*	*0.275*	*0.270*	*0.295*	*0.275*	*0.280*
0.8	*0.2*	*200*	*0.225*	*0.225*	*0.175*	*0.230*	*0.240*	*0.275*	*0.230*	*0.240*	*0.230*	*0.270*	*0.245*	*0.245*
0.8	*0.2*	*250*	*0.255*	*0.230*	*0.185*	*0.220*	*0.210*	*0.245*	*0.215*	*0.210*	*0.220*	*0.235*	*0.215*	*0.215*
0.8	*0.2*	*500*	*0.320*	*0.235*	*0.190*	*0.200*	*0.210*	*0.210*	*0.200*	*0.215*	*0.200*	*0.215*	*0.215*	*0.215*
0.8	*0.3*	*50*	*0.150*	*0.250*	*0.220*	*0.285*	*0.315*	*0.310*	*0.285*	*0.315*	*0.310*	*0.315*	*0.315*	*0.320*
0.8	*0.3*	*100*	*0.200*	*0.240*	*0.195*	*0.300*	*0.335*	*0.360*	*0.300*	*0.330*	*0.305*	*0.355*	*0.330*	*0.345*
0.8	*0.3*	*150*	*0.200*	*0.270*	*0.215*	*0.300*	*0.325*	*0.370*	*0.300*	*0.335*	*0.300*	*0.350*	*0.330*	*0.335*
0.8	*0.3*	*200*	*0.240*	*0.235*	*0.160*	*0.250*	*0.265*	*0.305*	*0.250*	*0.265*	*0.255*	*0.295*	*0.270*	*0.270*
0.8	*0.3*	*250*	*0.220*	*0.240*	*0.245*	*0.275*	*0.290*	*0.310*	*0.275*	*0.295*	*0.275*	*0.305*	*0.285*	*0.285*
0.8	*0.3*	*500*	*0.315*	*0.250*	*0.200*	*0.240*	*0.255*	*0.265*	*0.240*	*0.255*	*0.250*	*0.265*	*0.265*	*0.260*
0.8	*0.4*	*50*	*0.210*	*0.295*	*0.250*	*0.335*	*0.370*	*0.380*	*0.335*	*0.365*	*0.355*	*0.370*	*0.360*	*0.375*
0.8	*0.4*	*100*	*0.170*	*0.255*	*0.210*	*0.305*	*0.325*	*0.370*	*0.305*	*0.330*	*0.315*	*0.370*	*0.330*	*0.360*

0.8	*0.4*	*150*	*0.200*	*0.240*	*0.160*	*0.290*	*0.300*	*0.325*	*0.290*	*0.305*	*0.290*	*0.325*	*0.305*	*0.310*
0.8	*0.4*	*200*	*0.245*	*0.230*	*0.205*	*0.240*	*0.255*	*0.280*	*0.240*	*0.255*	*0.240*	*0.275*	*0.260*	*0.260*
0.8	*0.4*	*250*	*0.225*	*0.220*	*0.155*	*0.235*	*0.260*	*0.290*	*0.235*	*0.260*	*0.240*	*0.280*	*0.260*	*0.270*
0.8	*0.4*	*500*	*0.235*	*0.220*	*0.190*	*0.220*	*0.230*	*0.280*	*0.220*	*0.235*	*0.215*	*0.275*	*0.240*	*0.240*
0.8	*0.5*	*50*	*0.190*	*0.290*	*0.235*	*0.350*	*0.365*	*0.375*	*0.350*	*0.370*	*0.360*	*0.370*	*0.365*	*0.370*
0.8	*0.5*	*100*	*0.165*	*0.240*	*0.215*	*0.290*	*0.300*	*0.340*	*0.290*	*0.310*	*0.295*	*0.320*	*0.310*	*0.310*
0.8	*0.5*	*150*	*0.170*	*0.245*	*0.310*	*0.270*	*0.310*	*0.335*	*0.270*	*0.310*	*0.280*	*0.325*	*0.315*	*0.315*
0.8	*0.5*	*200*	*0.295*	*0.335*	*0.225*	*0.345*	*0.360*	*0.385*	*0.345*	*0.360*	*0.345*	*0.375*	*0.365*	*0.365*
0.8	*0.5*	*250*	*0.275*	*0.290*	*0.240*	*0.310*	*0.330*	*0.345*	*0.310*	*0.335*	*0.310*	*0.345*	*0.345*	*0.345*
0.8	*0.5*	*500*	*0.285*	*0.245*	*0.130*	*0.245*	*0.280*	*0.320*	*0.245*	*0.280*	*0.250*	*0.315*	*0.285*	*0.290*
0.9	*0.1*	*50*	*0.145*	*0.205*	*0.220*	*0.240*	*0.245*	*0.250*	*0.240*	*0.230*	*0.240*	*0.240*	*0.235*	*0.235*
0.9	*0.1*	*100*	*0.125*	*0.180*	*0.190*	*0.235*	*0.260*	*0.300*	*0.235*	*0.260*	*0.245*	*0.295*	*0.260*	*0.260*
0.9	*0.1*	*150*	*0.220*	*0.170*	*0.240*	*0.235*	*0.260*	*0.270*	*0.235*	*0.260*	*0.255*	*0.270*	*0.260*	*0.260*
0.9	*0.1*	*200*	*0.220*	*0.190*	*0.195*	*0.240*	*0.250*	*0.275*	*0.240*	*0.250*	*0.240*	*0.260*	*0.250*	*0.255*
0.9	*0.1*	*250*	*0.195*	*0.165*	*0.220*	*0.225*	*0.275*	*0.305*	*0.225*	*0.255*	*0.230*	*0.300*	*0.285*	*0.285*
0.9	*0.1*	*500*	*0.275*	*0.205*	*0.160*	*0.190*	*0.200*	*0.235*	*0.190*	*0.190*	*0.190*	*0.235*	*0.205*	*0.210*
0.9	*0.2*	*50*	*0.120*	*0.240*	*0.260*	*0.310*	*0.355*	*0.340*	*0.310*	*0.360*	*0.345*	*0.355*	*0.350*	*0.360*
0.9	*0.2*	*100*	*0.225*	*0.265*	*0.160*	*0.290*	*0.325*	*0.345*	*0.290*	*0.320*	*0.295*	*0.340*	*0.320*	*0.325*
0.9	*0.2*	*150*	*0.185*	*0.240*	*0.230*	*0.290*	*0.330*	*0.350*	*0.290*	*0.335*	*0.295*	*0.345*	*0.335*	*0.340*
0.9	*0.2*	*200*	*0.250*	*0.160*	*0.155*	*0.205*	*0.250*	*0.260*	*0.205*	*0.245*	*0.205*	*0.255*	*0.245*	*0.245*
0.9	*0.2*	*250*	*0.240*	*0.200*	*0.170*	*0.235*	*0.270*	*0.295*	*0.235*	*0.270*	*0.250*	*0.280*	*0.275*	*0.275*
0.9	*0.2*	*500*	*0.255*	*0.245*	*0.165*	*0.235*	*0.255*	*0.280*	*0.235*	*0.250*	*0.235*	*0.275*	*0.260*	*0.260*
0.9	*0.3*	*50*	*0.150*	*0.295*	*0.255*	*0.335*	*0.370*	*0.395*	*0.335*	*0.360*	*0.350*	*0.395*	*0.350*	*0.380*
0.9	*0.3*	*100*	*0.190*	*0.200*	*0.220*	*0.255*	*0.285*	*0.305*	*0.255*	*0.280*	*0.260*	*0.300*	*0.280*	*0.285*
0.9	*0.3*	*150*	*0.215*	*0.260*	*0.190*	*0.315*	*0.340*	*0.375*	*0.310*	*0.340*	*0.325*	*0.375*	*0.345*	*0.350*
0.9	*0.3*	*200*	*0.205*	*0.255*	*0.250*	*0.310*	*0.325*	*0.340*	*0.310*	*0.310*	*0.305*	*0.335*	*0.315*	*0.330*
0.9	*0.3*	*250*	*0.255*	*0.285*	*0.215*	*0.295*	*0.335*	*0.360*	*0.295*	*0.335*	*0.310*	*0.350*	*0.330*	*0.330*
0.9	*0.3*	*500*	*0.230*	*0.240*	*0.170*	*0.280*	*0.305*	*0.325*	*0.280*	*0.305*	*0.290*	*0.330*	*0.295*	*0.295*
0.9	*0.4*	*50*	*0.145*	*0.250*	*0.200*	*0.355*	*0.380*	*0.405*	*0.350*	*0.370*	*0.370*	*0.405*	*0.370*	*0.380*
0.9	*0.4*	*100*	*0.180*	*0.225*	*0.275*	*0.280*	*0.315*	*0.335*	*0.280*	*0.310*	*0.300*	*0.325*	*0.310*	*0.315*
0.9	*0.4*	*150*	*0.210*	*0.310*	*0.220*	*0.375*	*0.395*	*0.405*	*0.375*	*0.390*	*0.375*	*0.405*	*0.395*	*0.400*
0.9	*0.4*	*200*	*0.200*	*0.175*	*0.190*	*0.210*	*0.240*	*0.280*	*0.210*	*0.245*	*0.220*	*0.275*	*0.250*	*0.265*
0.9	*0.4*	*250*	*0.225*	*0.270*	*0.170*	*0.295*	*0.320*	*0.335*	*0.295*	*0.315*	*0.305*	*0.340*	*0.325*	*0.330*
0.9	*0.4*	*500*	*0.335*	*0.290*	*0.255*	*0.330*	*0.345*	*0.380*	*0.330*	*0.345*	*0.335*	*0.370*	*0.355*	*0.360*
0.9	*0.5*	*50*	*0.130*	*0.205*	*0.195*	*0.295*	*0.335*	*0.335*	*0.295*	*0.335*	*0.325*	*0.340*	*0.330*	*0.345*
0.9	*0.5*	*100*	*0.170*	*0.220*	*0.205*	*0.275*	*0.290*	*0.330*	*0.275*	*0.290*	*0.280*	*0.320*	*0.290*	*0.300*
0.9	*0.5*	*150*	*0.185*	*0.250*	*0.205*	*0.300*	*0.320*	*0.320*	*0.300*	*0.315*	*0.300*	*0.320*	*0.315*	*0.320*
0.9	*0.5*	*200*	*0.210*	*0.295*	*0.210*	*0.335*	*0.380*	*0.420*	*0.335*	*0.390*	*0.340*	*0.410*	*0.395*	*0.400*
0.9	*0.5*	*250*	*0.235*	*0.270*	*0.200*	*0.295*	*0.330*	*0.350*	*0.295*	*0.325*	*0.305*	*0.345*	*0.330*	*0.330*
0.9	*0.5*	*500*	*0.230*	*0.225*	*0.195*	*0.225*	*0.245*	*0.295*	*0.225*	*0.250*	*0.225*	*0.295*	*0.250*	*0.255*
OVERALL STRICT AGREEMENT			***0.224***	***0.238***	***0.207***	***0.266***	***0.285***	***0.308***	***0.266***	***0.285***	***0.272***	***0.303***	***0.287***	***0.292***

Notes: Darker shading represents higher percentages.

To assess which specific models the RMSE and in-sample criteria agreed on, we also analyzed the percentage of agreement broken down by each model (Table 4). This analysis allowed us to understand which models were being chosen by RMSE and in-sample criteria at the same time. The highest agreement percentages for all model selection criteria, as well as Adjusted R^2 and Q^2, were over Model 2. Recall that Model 2 was the correct but underspecified version of the data generation

model (Model 5). In contrast, these criteria agreed to a much a lesser extent on the data generation model (Model 5) and the correct but overspecified version (Model 8). In terms of R^2, we found that the agreement with RMSE was on Model 7 because of their tendency to heavily prefer the saturated model. Among all the criteria, BIC and GM found agreement with RMSE more often than others. However, because the overall strict agreement rates with RMSE were fairly low for all criteria (last row of Table 3), none of the in-sample criteria are suitable replacements for out-of-sample criteria (RMSE) when the focus is on prediction-only (***P***). In such a case, the availability of a holdout set and the computation of out-of-sample criteria is necessary and cannot be avoided.

Table 4: Percentage agreement with RMSE by model number (Scenario 1)

Model #		1	2	3	4	5	6	7	8	Total
PLS-PM Criteria	**R^2**	0.000	0.092	0.000	0.000	0.003	0.000	0.128	0.001	0.224
	Adjusted R^2	0.000	0.183	0.000	0.000	0.011	0.000	0.031	0.014	0.238
	Q^2	0.000	0.101	0.000	0.000	0.034	0.000	0.018	0.054	0.207
Model Selection Criteria	**FPE**	0.000	0.223	0.000	0.000	0.013	0.000	0.011	0.018	0.266
	CP	0.000	0.244	0.000	0.000	0.015	0.000	0.006	0.021	0.285
	GM	0.000	0.267	0.000	0.000	0.016	0.000	0.000	0.024	0.308
	AIC	0.000	0.223	0.000	0.000	0.013	0.000	0.011	0.018	0.266
	AICu	0.000	0.244	0.000	0.000	0.015	0.000	0.005	0.022	0.285
	AICc	0.000	0.229	0.000	0.000	0.014	0.000	0.011	0.019	0.272
	BIC	0.000	0.263	0.000	0.000	0.016	0.000	0.001	0.023	0.303
	HQ	0.000	0.247	0.000	0.000	0.015	0.000	0.003	0.022	0.287
	HQc	0.000	0.252	0.000	0.000	0.015	0.000	0.003	0.022	0.292

4.1.3 Assessing performance of in-sample criteria using the explanation-prediction lens (EP)

The analyses above assumed strict agreement with RMSE to choose a best model regardless of whether the selected model was correctly specified or not. However, because PLS-PM puts theoretical consistency at a premium, we asked whether the in-sample criteria can help select a suitably predictive model that is also consistent with the data generation process (correctly specified). This necessitates some trade-off in out-of-sample predictive power to ensure that the model selected lies in the set of correctly specified models. To shed light on this question, we

analyzed the agreement of in-sample criteria with RMSE over the three types of models (in terms of specification relative to η_2) included in our experimental set-up: correctly specified (Models 2, 5, and 8), incorrectly specified (Model 1, 3, 4, and 6), and saturated (Model 7). Recall that Models 2 and 8 were correct but under- and over-specified versions of the data generation model (Model 5) respectively, with respect to η_2. Ideally, we would like to see whether RMSE and the in-sample criteria agreed more over the set of correctly specified models, and disagreed over the misspecified and saturated sets.[5] Table 5 presents the agreement percentages broken down by model types.

Table 5: Percentage agreement with RMSE by model type (Scenario 1)

Model Type		Correctly Specified (Model 2 or 5 or 8)	Incorrectly Specified (Model 1 or 3 or 4 or 6)	Saturated (Model 7)
PLS-PM Criteria	**R^2**	0.211	0.000	0.128
	Adjusted R^2	0.504	0.000	0.031
	Q^2	0.611	0.000	0.018
Model Selection Criteria	**FPE**	0.623	0.000	0.011
	CP	0.684	0.000	0.006
	GM	0.757	0.000	0.000
	AIC	0.623	0.000	0.011
	AICu	0.685	0.000	0.005
	AICc	0.639	0.000	0.011
	BIC	0.740	0.000	0.001
	HQ	0.692	0.000	0.003
	HQc	0.705	0.000	0.003

Among the PLS-PM criteria, we see that Q^2 achieved the highest agreement with RMSE over the correctly specified set (61.1%), followed by Adjusted R^2 (50.4%). In contrast, R^2 had low levels of agreement (21.1%) and also agreed over the saturated model (12.8%). All the model selection criteria found strongest agreement with RMSE on the correctly specified set, with GM and BIC achieving over 74% agreement—higher by a significant margin compared to other criteria. In

[5] In the following, we focus on agreement with RMSE over a set of models rather than a specific model as in Table 4. Thus, agreement over the correctly specified set captures the percentage of cases when *both* RMSE and model selection (or PLS-PM) criteria selected a model in the corresponding set at the same time.

particular, the agreement level of model selection criteria over Model 7 were low because of their tendency to penalize the saturated model more than other models. In addition, the agreement over the set of incorrectly specified models were zero for all the criteria. These results suggest that there are significant gains to be had by preferring the use of the model selection criteria (in particular BIC and GM) over PLS-PM criteria (including the Q^2). If the primary goal of the researcher is to select a suitably predictive and correctly specified model in the absence of a holdout sample, BIC and GM are promising substitutes for RMSE.

4.1.4 Impact of simulation design factors

After establishing the suitability of model selection criteria as potential substitutes when utilizing the ***EP*** lens, we next analyzed the question: How do the individual experimental conditions affect the agreement levels between RMSE and in-sample criteria over the correctly specified set? Asking this question can help us understand the conditions under which the researcher may expect more agreement with RMSE than others. We created three marginal means tables that break down the results presented in Table 5 by individual experimental condition.

Table 6 presents the percentage agreement with RMSE by model type broken by sample size[6]. An immediate pattern to note is that the model selection criteria showed significantly higher agreement levels with RMSE than any PLS-PM criteria, including Q^2 which was the best performing PLS-PM criterion. For example, while GM showed 79.2% agreement over the correctly specified set at sample size 50, R^2 was able to manage only 26.6%. With an increase in sample size, the "general" trend of agreement with RMSE over correctly specified models was of a gradual decrease for all in-sample criteria. In contrast, the agreement with the saturated model showed a pattern of gradual increase. However, there were certain exceptions. For example, BIC and GM (in

[6] We note that the agreement values over the incorrectly specified set are close to zero for all criteria and are hence not presented in Table 6.

addition, HQ and HQc) "peaked" in agreement over correctly specified models at sample size 100 (BIC: 79.9% and GM: 82.2%) but trailed off gradually after that, reaching around 65% agreement at sample size 500. The "sweet spot" for BIC and GM lies between sample sizes 50-200, precisely the conditions under which splitting the sample into training and holdout samples becomes impractical! At higher sample sizes (say 500 or more), the researcher has the luxury of splitting the sample into training and holdout and may not need to rely on in-sample criteria at all.

Table 6: Percentage agreement with RMSE by model type by Sample Size (Scenario 1)

	Criterion	Model Type	50	100	150	200	250	500	Pattern
PLS-PM Criteria	**R^2**	Correctly Specified	0.266	0.212	0.226	0.199	0.201	0.162	↓
		Saturated	0.047	0.083	0.108	0.142	0.153	0.235	↑
	Adjusted R^2	Correctly Specified	0.589	0.544	0.528	0.479	0.477	0.409	↓
		Saturated	0.003	0.011	0.018	0.034	0.035	0.085	↑
	Q^2	Correctly Specified	0.685	0.663	0.636	0.599	0.583	0.497	↓
		Saturated	0.001	0.008	0.011	0.014	0.021	0.052	↑
Model Selection Criteria	**FPE**	Correctly Specified	0.704	0.676	0.661	0.605	0.591	0.504	↓
		Saturated	0.000	0.001	0.005	0.009	0.014	0.039	↑
	Cp	Correctly Specified	0.761	0.742	0.720	0.663	0.653	0.564	↓
		Saturated	0.000	0.000	0.002	0.002	0.007	0.023	↑
	GM	Correctly Specified	0.792	0.822	0.788	0.750	0.736	0.655	↓
		Saturated	0.000	0.000	0.000	0.000	0.000	0.000	↑
	AIC	Correctly Specified	0.702	0.675	0.659	0.605	0.591	0.504	↓
		Saturated	0.000	0.001	0.005	0.009	0.014	0.039	↑
	AICu	Correctly Specified	0.755	0.743	0.721	0.669	0.656	0.566	↓
		Saturated	0.000	0.000	0.002	0.002	0.005	0.020	↑
	AICc	Correctly Specified	0.737	0.697	0.675	0.612	0.603	0.509	↓
		Saturated	0.000	0.001	0.004	0.007	0.013	0.039	↑
	BIC	Correctly Specified	0.773	0.799	0.771	0.731	0.720	0.645	↓
		Saturated	0.000	0.000	0.000	0.000	0.000	0.003	↑
	HQ	Correctly Specified	0.742	0.743	0.726	0.682	0.674	0.589	↓
		Saturated	0.000	0.000	0.002	0.002	0.003	0.013	↑
	HQc	Correctly Specified	0.765	0.765	0.737	0.689	0.679	0.593	↓
		Saturated	0.000	0.000	0.001	0.002	0.002	0.013	↑

Table 7 presents the percentage agreement with RMSE by model type per effect size. With an increase in effect size, all criteria show stronger agreement on the correctly specified models, due to stronger signal strength. However, the base rates of agreement of model selection criteria are

significantly higher than those of the PLS-PM criteria. For example, BIC and GM show high agreement with RMSE over correctly specified models even at a low effect size of 0.1 (71.4% and 73.3% respectively), while R^2 had 14.8% agreement. Again, Q^2 is the best performing PLS-PM criterion but still lags behind top performing model selection criteria. Conversely, as effect size increases, the agreement level over the saturated model decreases for all criteria. Here, BIC and GM again show very low (almost zero) base rates of agreement with the saturated model.

Table 7: Percentage agreement with RMSE by model type by Effect Size ($\xi_2 \rightarrow \eta_1$) (Scenario 1)

	Criterion	Model Type	0.1	0.2	0.3	0.4	0.5	Pattern
PLS-PM Criteria	**R^2**	Correctly Specified	0.148	0.182	0.220	0.239	0.265	↑
		Saturated	0.165	0.150	0.119	0.105	0.101	↓
	Adjusted R^2	Correctly Specified	0.458	0.494	0.509	0.519	0.541	↑
		Saturated	0.039	0.034	0.035	0.023	0.023	↓
	Q^2	Correctly Specified	0.589	0.603	0.616	0.620	0.624	↑
		Saturated	0.019	0.017	0.021	0.016	0.015	↓
Model Selection Criteria	**FPE**	Correctly Specified	0.587	0.611	0.630	0.637	0.652	↑
		Saturated	0.016	0.010	0.014	0.009	0.007	↓
	Cp	Correctly Specified	0.653	0.677	0.689	0.697	0.703	↑
		Saturated	0.009	0.005	0.008	0.003	0.004	↓
	GM	Correctly Specified	0.733	0.746	0.764	0.767	0.775	↑
		Saturated	0.000	0.000	0.000	0.000	0.000	↓
	AIC	Correctly Specified	0.586	0.610	0.630	0.636	0.652	↑
		Saturated	0.016	0.010	0.014	0.009	0.007	↓
	AICu	Correctly Specified	0.652	0.678	0.688	0.700	0.708	↑
		Saturated	0.008	0.004	0.007	0.002	0.004	↓
	AICc	Correctly Specified	0.603	0.627	0.646	0.651	0.666	↑
		Saturated	0.015	0.009	0.014	0.009	0.007	↓
	BIC	Correctly Specified	0.714	0.727	0.747	0.751	0.760	↑
		Saturated	0.001	0.001	0.000	0.001	0.001	↓
	HQ	Correctly Specified	0.663	0.684	0.696	0.706	0.713	↑
		Saturated	0.006	0.003	0.004	0.002	0.002	↓
	HQc	Correctly Specified	0.673	0.695	0.708	0.722	0.728	↑
		Saturated	0.005	0.003	0.003	0.002	0.002	↓

Finally, Table 8 presents the percentage agreement with RMSE by model type broken down by loading value (AVE). Again, the base rates of agreement of model selection criteria are significantly higher than those of the PLS-PM criteria. As item loadings (AVE) increase, model

selection criteria yield higher levels of agreement with RMSE over the correctly specified model set, due to the reduction in noise. Q^2's performance is similar to the model selection criteria and improves with an increase in AVE; however, its base level agreement is much less than BIC and GM. The other PLS-PM criteria display drastically different behavior. With an increase in AVE, the R^2 and the Adjusted R^2 show a decrease in agreement over the correctly specified set. These findings suggest that as the measurement model quality improves, the model selection criteria become much more reliable in terms of agreeing with RMSE over the correctly specified set, while R^2 and Adjusted R^2 do not.

Table 8: Percentage agreement with RMSE by model type by Loading Values (AVE) (Scenario 1)

	Criterion	Model Type	0.7	0.8	0.9	Pattern
PLS-PM Criteria	**R^2**	Correctly Specified	0.264	0.218	0.152	↓
		Saturated	0.126	0.124	0.135	↑
	Adjusted R^2	Correctly Specified	0.504	0.510	0.499	↓
		Saturated	0.039	0.033	0.021	↓
	Q^2	Correctly Specified	0.603	0.610	0.618	↑
		Saturated	0.022	0.019	0.012	↓
Model Selection Criteria	**FPE**	Correctly Specified	0.606	0.626	0.639	↑
		Saturated	0.018	0.011	0.006	↓
	Cp	Correctly Specified	0.648	0.688	0.716	↑
		Saturated	0.012	0.004	0.002	↓
	GM	Correctly Specified	0.726	0.762	0.784	↑
		Saturated	0.000	0.000	0.000	↓
	AIC	Correctly Specified	0.605	0.625	0.639	↑
		Saturated	0.018	0.011	0.006	↓
	AICu	Correctly Specified	0.658	0.689	0.708	↑
		Saturated	0.008	0.005	0.002	↓
	AICc	Correctly Specified	0.619	0.641	0.656	↑
		Saturated	0.017	0.010	0.006	↓
	BIC	Correctly Specified	0.708	0.744	0.767	↑
		Saturated	0.001	0.000	0.001	↓
	HQ	Correctly Specified	0.666	0.696	0.716	↑
		Saturated	0.005	0.004	0.001	↓
	HQc	Correctly Specified	0.678	0.708	0.729	↑
		Saturated	0.005	0.003	0.001	↓

4.2 Scenario 2: Data generation model not included in the consideration set

We repeated the prior analyses, but this time using a new generating model (Model X) that was excluded from the set of competing models (Figure 2). The results help us assess whether the conclusions drawn from the earlier analyses generalize to more practical situations where hidden (unobserved or unavailable) variables may exist. Table 9 shows the average choice shares for each model per criterion (PLS-PM, model selection, and out-of-sample criteria) across all conditions of sample sizes, item loading, and effect sizes.

Table 9: Overall proportion of model choice by each criterion (Scenario 2)

	Model #	1	2	3	4	5	6	7	8
PLS-PM Criteria	**R^2**	0.000	0.274	0.000	0.004	0.019	0.000	0.688	0.015
	Adjusted R^2	0.000	0.539	0.000	0.004	0.073	0.000	0.306	0.078
	Q^2	0.004	0.310	0.000	0.002	0.216	0.002	0.181	0.284
Model Selection Criteria	**FPE**	0.000	0.642	0.000	0.005	0.093	0.000	0.165	0.095
	CP	0.000	0.690	0.000	0.005	0.099	0.000	0.098	0.107
	GM	0.000	0.755	0.000	0.005	0.109	0.006	0.007	0.117
	AIC	0.000	0.642	0.000	0.005	0.093	0.000	0.166	0.095
	AICu	0.000	0.693	0.000	0.005	0.101	0.001	0.092	0.108
	AICc	0.000	0.653	0.000	0.005	0.094	0.000	0.149	0.098
	BIC	0.000	0.740	0.000	0.005	0.107	0.003	0.030	0.114
	HQ	0.000	0.699	0.000	0.005	0.102	0.001	0.084	0.109
	HQc	0.000	0.707	0.000	0.005	0.103	0.002	0.072	0.111
Out-of-Sample Criteria	**MAD**	0.001	0.347	0.000	0.001	0.193	0.000	0.234	0.224
	RMSE	0.000	0.352	0.000	0.000	0.206	0.000	0.214	0.227
	MAPE	0.086	0.047	0.258	0.075	0.039	0.347	0.087	0.061
	SMAPE	0.001	0.370	0.000	0.000	0.122	0.000	0.341	0.166

The results closely match those derived in Scenario 1 (see also Figure B2 in Appendix B). Again, we find that, except for MAPE, all out-of-sample criteria choose Models 2, 5, 7, or 8, with a greater preference for Model 2. A similar preference holds for all model selection criteria, but showing more pronounced preferences for Model 2 (e.g., BIC: 74% and GM: 75.5%) with GM and BIC rarely choosing the saturated Model 7. Among the PLS-PM criteria, Q^2 exhibits a similar

performance as the out-of-sample criteria, while R^2 shows a strong preference for Model 7. Table B1 (Appendix B) presents the percentage of agreement broken down by model number.

Analogous to Scenario 1, we again assessed the agreement of in-sample criteria with RMSE over the three types of models (in terms of specification relative to η_2) included in our experimental set-up: correctly specified (Models 2, 5, and 8), incorrectly specified (Model 1, 3, 4, and 6), and saturated (Model 7). Table 10 presents the agreement percentages broken down by model types for Scenario 2.

Table 10: Agreement with RMSE's choice of best model by model type (Scenario 2)

		Correctly Specified (Model 2 or 5 or 8)	Incorrectly Specified (Model 1 or 3 or 4 or 6)	Saturated (Model 7)
PLS-PM Criteria	**R^2**	0.219	0.000	0.126
	Adjusted R^2	0.506	0.000	0.031
	Q^2	0.612	0.001	0.016
Model Selection Criteria	**FPE**	0.625	0.000	0.011
	CP	0.686	0.000	0.005
	GM	0.766	0.000	0.000
	AIC	0.625	0.000	0.011
	AICu	0.692	0.000	0.005
	AICc	0.641	0.000	0.010
	BIC	0.747	0.000	0.000
	HQ	0.697	0.000	0.003
	HQc	0.708	0.000	0.003

The results almost perfectly mimic those in Table 5 with percentage values differing in the third decimal place.[7] These results suggest that the model selection criteria's performance is immune to whether the data generating model is included or excluded from the set of competing models. In other words, the results show that the conclusions drawn from Scenario 1 generalize to cases where there may be hidden variables that may directly impact the focal endogenous construct

[7] The results across the simulation conditions also parallel those from Scenario 1 and are available from the authors upon request.

under analysis.

5. Empirical illustration

To illustrate the use of the model selection criteria with empirical data, we draw on the corporate reputation model used by Hair et al. (2017b, 2018). The goal of this model is to explain the effects of competence (*COMP*) and likeability (*LIKE*), representing the two dimensions of corporate reputation (e.g., Raithel & Schwaiger, 2015; Sarstedt, Wilczynski, & Melewar, 2013), on customer satisfaction (*CUSA*) and customer loyalty (*CUSL*). Furthermore, the model includes the following four antecedent constructs of corporate reputation that Schwaiger (2004) identified: (1) the quality of a company's products and services as well as its quality of customer orientation (*QUAL*), (2) its economic and managerial performance (*PERF*), (3) a company's corporate social responsibility (*CSOR*), and (4) its attractiveness (*ATTR*).

The measurement models of *COMP*, *LIKE*, and *CUSL* draw on three reflective items each, whereas *CUSA* is measured with a single item. In contrast, the four antecedent constructs (i.e., *ATTR, CSOR, PERF,* and *QUAL*) have formative measurement models with a total of 21 indicators (Schwaiger, 2004). The model estimation draws on data from two major German mobile communications network providers and two smaller competitors.[8] A total of 344 respondents rated each item on a 7-point Likert scale. Observations with missing values were deleted, leaving a total sample size of 336. Our analysis considers five different model configurations (Figure 3). Model 1 is the theoretically well-established original model that has been extensively used in prior illustrations of PLS-PM (e.g., Hair et al., 2017b; Hair, Sarstedt, Ringle, & Gudergan, 2018), and in showcasing the methodological extensions of the method (e.g., Sarstedt & Ringle, 2010; Matthews, Sarstedt, Hair, & Ringle, 2016). Model 2 is equivalent to Eberl's (2010) conceptualization according to which only *LIKE* influences *CUSL* directly. Model 3 is a further simplified version of

[8] The dataset and SmartPLS model files can be downloaded from https://www.pls-sem.net/downloads/.

Models 1 and 2 in which *LIKE* and *COMP* only influence *CUSA* directly. However, both these models disregard that corporate reputation—as conceptualized and operationalized by Schwaiger (2004)— is an attitude-related construct with one affective dimension (i.e., *LIKE*), and one cognitive dimension (*COMP*). As loyalty manifests itself in customers' relative *attitude* towards a firm (i.e., the attitudinal strength and the degree of attitudinal differentiation; Dick and Basu, 1994), it is reasonable to assume that both dimensions impact customer loyalty, albeit to different degrees. Finally, Models 4 and 5 are more complex configurations of the original model in which the antecedent constructs also directly influence *CUSA* and *CUSL*. Both models are theoretically plausible in that they assume that *LIKE* and *COMP* may only partially mediate the relationship between the four antecedent constructs and *CUSA* as well as *CUSL*. Walsh et al.'s (2009) study on the outcomes of corporate reputation provides further support for Model 5. These authors operationalized corporate reputation similar to Schwaiger (2004) but focusing on the antecedent dimensions while disregarding the two attitude-related dimensions *LIKE* and *COMP*. Their study shows that the more concrete denotation of the corporate reputation concepts (e.g., corporate social responsibility) directly impact customer loyalty. As customer satisfaction is the primary direct antecedent of loyalty, and since satisfaction is grounded in cognitive and affective judgments (Oliver, 1993), assuming direct effects on these two constructs is theoretically plausible. That is, a good reputation acts as a signal of sound company behavior toward market transactions and reduces customers' perceived risk (Bartikowski & Walsh, 2011). However, this signal should not be limited to impact customer satisfaction (Model 4) but also customer loyalty (Model 5).

Our analysis considers all criteria evaluated in the Monte Carlo study, focusing on *CUSA* as the immediate consequence of reputation (Eberl, 2010). We first ensured that all the measurement models met the relevant evaluation criteria (for further details on the PLS-PM results, see Hair et al., 2017b). Table 11 shows the results of our analysis for the structural model comparisons. We

observed that most of the criteria show the strongest preference for Model 5 followed by Model 4, which are the most complex models in our set and the least theoretically defensible. This preference for the saturated models is shared by the PLS-PM in-sample criteria (i.e., R^2), the purported prediction-oriented criteria (i.e., Q^2), some of the model selection criteria (e.g., CP and AIC), and by the out-of-sample criteria (e.g., RMSE). Even the Adjusted R^2, which is designed to adjust for parsimony, shows greater preference for Model 5 followed by Model 4.

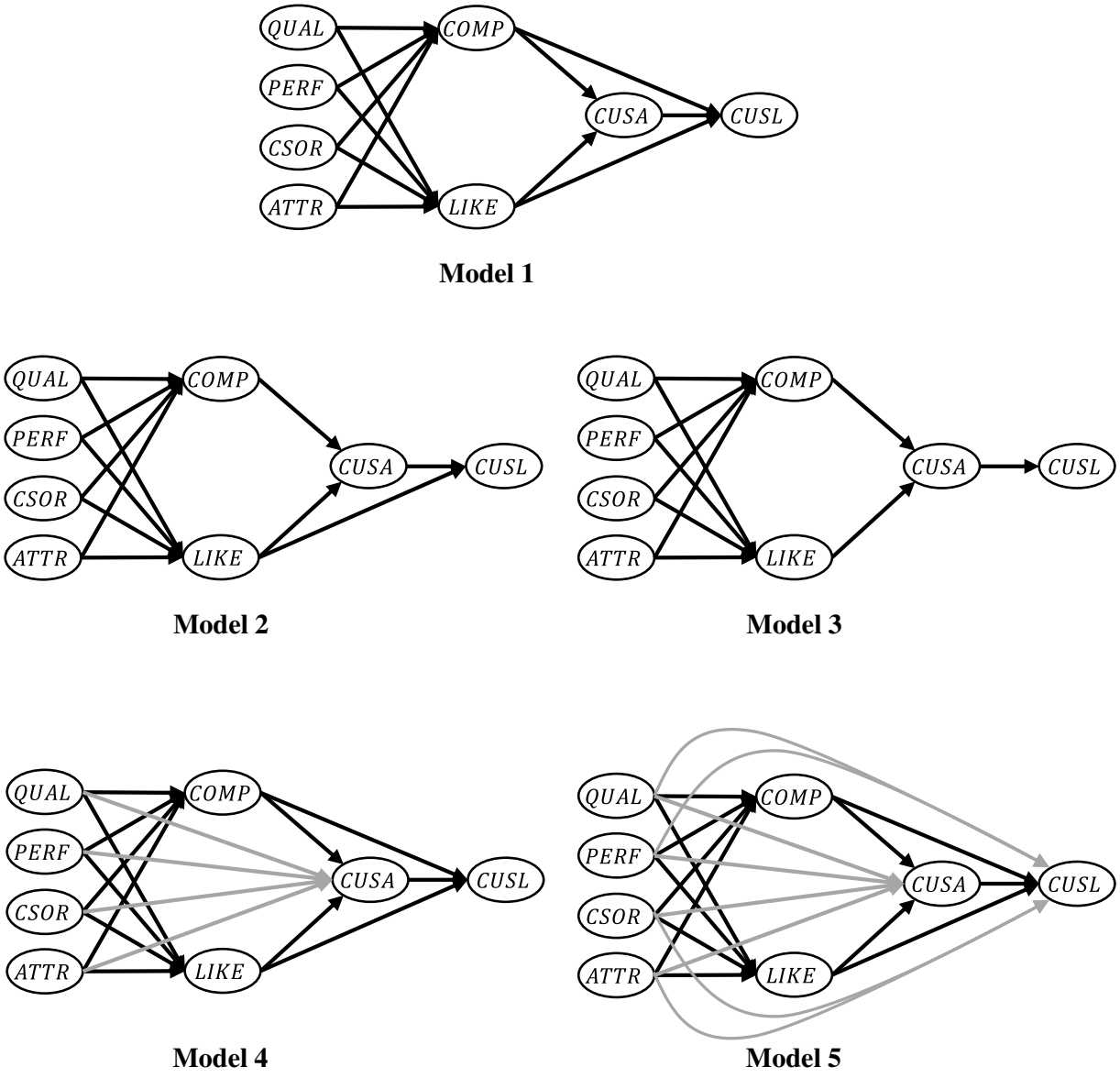


Figure 3: The five alternative models (Corporate Reputation).

In sharp contrast, only two model selection criteria namely, GM and BIC, show the strongest preference for Model 1, which is the most theoretically established model in our set. These findings echo the results of our earlier simulations, where GM and BIC provide the most appropriate balance between predictive performance and correct specification.

Table 11: Criteria values for Alternative Models 1-5 (Corporate Reputation Example)

	Model #	1	2	3	4	5	Selected Model
PLS-PM Criteria	**R^2**	0.2911	0.2906	0.2909	0.3217	0.3285	5
	Adjusted R^2	0.2870	0.2865	0.2867	0.3097	0.3166	5
	Q^2	0.2820	0.2816	0.2817	0.2914	0.2960	5
Model Selection Criteria	**FPE**	0.7192	0.7198	0.7195	0.7044	0.6973	5
	CP	24.1037	24.3810	24.2505	16.4748	13.0000	5
	GM	379.6257	379.9030	379.7725	387.3592	383.8845	1
	AIC	-113.3742	-113.1109	-113.2348	-120.5517	-124.0191	5
	AICu	-110.3610	-110.0977	-110.2216	-113.4795	-116.9469	5
	AICc	232.7438	233.0071	232.8832	225.8781	222.4108	5
	BIC	-101.8523	-101.5889	-101.7128	-93.6672	-97.1346	1
	HQ	-108.7851	-108.5218	-108.6457	-109.8440	-113.3113	5
	HQc	-108.6290	-108.3656	-108.4895	-109.1802	-112.6475	5
Out-of-Sample Criteria	**MAD**	0.6724	0.6726	0.6725	0.6637	0.6612	5
	RMSE	0.8408	0.8409	0.8408	0.8369	0.8338	5
	MAPE	207.9099	207.8683	207.8881	194.6732	193.8904	5
	SMAPE	61.0738	61.0772	61.0701	59.0624	58.8620	5

6. Discussion

Since its inception PLS-PM has avowedly been an exploratory technique for theory building where researchers might want to compare several models (Wold, 1974; 1980). Recent work in the PLS-PM literature has also highlighted its abilities as a predictive technique (Becker et al., 2013; Evermann & Tate, 2016; Shmueli et al., 2016). Because PLS-PM straddles the divide between causal explanation and prediction, researchers using the method need to ensure that the estimated model adequately maps reality while offering sufficient predictive capabilities (Shmueli et al.,

2016). While prior studies have evaluated model selection criteria's efficacy for selecting a specific model among a set of competing models (Sharma & Kim, 2012; Sharma et al., 2018), none have examined their performance from the prediction perspective, where the goal is to select models with high predictive power.

Using a Monte Carlo study, we analyzed the performance of the standard PLS-PM criteria (R^2, Adjusted R^2, and Q^2), and various model selection criteria vis-à-vis the performance of out-of-sample criteria, when selecting the best predictive model among a cohort of competing models. In particular, we explored whether the in-sample criteria can substitute for out-of-sample predictive criteria (most notably RMSE) that require a holdout sample, and under which conditions. Such a substitution is advantageous because splitting datasets into training and holdout samples may cause substantial loss of statistical and predictive power when the overall dataset is not large, as is usually the case with survey-based studies using PLS-PM (Rigdon, 2016). Our study revealed a range of findings, relevant to researchers using PLS-PM.

First, our results show that the model selection criteria, in particular BIC and GM, have significantly higher agreement levels with RMSE over the set of correctly specified models than the PLS-PM criteria (i.e. R^2, Adjusted R^2, and Q^2), thereby achieving a balance between theoretical consistency and high predictive power. This makes BIC and GM ideal candidates for prediction-oriented model selection when the holdout sample is unavailable and the researcher is working under the ***EP*** lens, as is usually the case in PLS-PM studies (e.g., Sarstedt et al., 2017). Among the correctly specified set, model selection criteria showed a stronger preference for the underspecified model compared to RMSE. Another difference of note between the RMSE and model selection criteria is the RMSE's preference for the (incorrectly specified) saturated model in about a quarter of cases, while the model selection criteria tended to avoid it. This difference resulted in the disagreement between model selection criteria and RMSE because the model selection criteria

heavily remain within the correctly specified sphere (i.e., Models 2, 5, and 8). The opposite is true for PLS-PM criteria, which agree over the incorrectly specified saturated model. The only exception is Q^2, which showed a similar yet inferior performance compared to the model selection criteria, in particular to BIC and GM. Because Q^2 has been the only predictive relevance criterion available in PLS-PM so far, researchers have called for its use on a regular basis (Hair et al., 2017b). However, with the introduction of the model selection criteria in the PLS-PM context researchers now have a wider set of criteria to rely on especially when comparing the predictive generalizability of their models. Our results show that BIC and GM are much better candidates for comparing the predictive abilities of models than existing PLS-PM criteria, including the Q^2.[9]

Second, our analysis of the impact of experimental conditions on the performance of the in-sample criteria vis-à-vis the out-of-sample criteria also revealed interesting patterns. With an increase in sample size, all in-sample criteria tended to disagree more with RMSE over the correctly specified set. However, the "sweet spot" for BIC and GM emerges between sample sizes 50 and 200, with a peak at 100, where these criteria heavily agreed with RMSE over correctly specified models. This result is encouraging for researchers using PLS-PM as it suggests that BIC and GM show their best performance precisely at those sample sizes where splitting the dataset into holdout and training samples becomes impractical. With an increase in effect size and item loadings, the model selection criteria show more agreement with RMSE over the correctly specified set, while the PLS-PM evaluation criteria do not. Overall, these results suggest that the best conditions to use BIC and GM appear at the intersection of sample sizes ranging between 50 and 200, and when the item loadings and effect sizes are high. That is, to efficiently utilize the criteria, researchers must work with instruments that exhibit sufficient levels of reliability and validity along with a reasonably

[9] Note that we focus on the model comparison context, because a stand-alone BIC or GM value is not useful for interpretation (unlike the R^2 which can be interpreted in isolation), but rather it is the relative values that make the comparison meaningful (Burnham & Anderson, 2002).

developed theory at the structural level to support higher effect sizes. Finally, we found that our results hold for the scenario where the data generation model has not been included in the competing set. This finding is encouraging because exploratory research, where PLS-PM is often used (e.g., Hair et al., 2017b), may often create conditions where the researcher may not be aware of, or in possession of, all relevant variables that impacted the focal endogenous variable.

Third, it may also be useful in certain cases (e.g., when theory is ill-developed, nonexistent, not of primary concern, or when forecast accuracy is primary concern) to select the "ultimate" predictive model without any regard for theoretical consistency (i.e., utilizing the ***P*** lens). In such cases, a useful strategy is to create all possible models and compute out-of-sample criteria (RMSE) using a holdout sample—as RMSE is often considered a "gold standard" for judging predictive power. This approach shows researchers how much extractable predictive information is contained in the data and help set predictive benchmarks against which the best predictive models within the "theoretically correct sphere" can be judged. The knowledge regarding how much predictive accuracy the data at hand may allow, versus what the theory is able to achieve, can be a useful tool for further theoretical development.

Finally, despite the fact that PLS-PM is a prediction-oriented technique, existing studies have not focused on the performance of out-of-sample criteria but rather on the prediction capabilities of PLS-PM method itself (Becker et al., 2013; Evermann & Tate, 2016). These studies rely on out-of-sample criteria without necessarily comparing their strengths. This is surprising because there is ongoing debate in the forecasting literature regarding the appropriate predictive criteria, where no single measure provides an unambiguous indication of forecasting performance and there is disagreement among researchers (Goodwin & Lawton, 1999; Makridakis, 1993; Davydenko & Fildes, 2013; Tofallis, 2015). There is also no consensus yet regarding which predictive criteria should take precedence over others in the context of PLS-PM and when.

Our study takes an important step in this direction by benchmarking the predictive model selection performance of various out-of-sample criteria in PLS. In essence, we ask: which out-of-sample criteria are more suitable in the context of prediction-oriented model comparisons? Our results suggest that among the out-of-sample criteria, RMSE and MAD behaved the best per expectation (i.e., keeping in mind PLS-PM's theory building aspect), followed by SMAPE. It is worth noting that in about a quarter of the cases, the RMSE selected the saturated model by going outside the theoretically correct sphere. Another important finding in this study is the danger of using MAPE as a predictive model selection criterion. Although MAPE offers the ease of interpretation and is therefore highly popular in practice, its utility for comparing models is limited. In our simulation, the criterion consistently selected incorrect models (Models 3 and 6). This behavior is not unexpected as several studies in other contexts have indicated that MAPE is biased towards models that under-predict and advocate avoiding it as a predictive model selection criterion (Goodwin & Lawton, 1999; Tofallis, 2015). Our study confirms this recommendation in the context of PLS-PM. We therefore call for the use of RMSE or MAD in PLS-PM-based model selection when researchers can afford drawing a holdout set.

While our study offers empirical insights, the actual choice of out-of-sample criteria may also depend on other factors out of the realm of pure statistics. For example, Flores (1986, p. 97) asks, "*Can the decision of which statistic to use be tied with the managerial style?*" He suggests that conservative managers whose goal is to minimize a "*regret criterion*" should rely on RMSE. On the other hand, MAPE and MAD should be preferred if the manager is "*like the baseball manager who will play basing his decisions on the law of averages*." Because our goal in this study was to select the best predictive model, we based our analyses on RMSE to minimize the prediction errors (i.e., regret criterion). Future studies should explore alternative business-driven prediction metrics in more detail including several new forecasting measures that have been recently discussed in the

forecasting literature (e.g. Hyndman & Koehler, 2006; Tofallis, 2015). Furthermore, future studies should extend our simulation design by considering more complex model structures, such as hierarchical component models, interaction terms, mediating effects, and nonlinear effects (e.g., Hair et al., 2017b). While the use of these modeling elements has recently become more en vogue, nothing is known about the out-of-sample criteria's performance when estimating corresponding model types. Furthermore, the predictions derived in our study focused on composite scores rather than on individual item scores. While this approach reflects the common use of models estimated with PLS-PM, generalizability to real-world settings would also benefit from predictions on an item level (e.g., Shmueli et al., 2016).

Our research sheds light on the performance of model selection criteria in the context of PLS-PM, which is by far the most prominent composite-based SEM method in business research and whose use has gained momentum in recent years (e.g., Hair et al., 2017, 2018). However, future research should benchmark PLS-PM's predictive performance against other composite-based SEM methods such as consistent PLS (Dijstra & Henseleer, 2015), generalized structure component analysis (Hwang et al., 2010), and regularized canonical correlation analysis (Tenenhaus & Tenenhaus, 2011). A particularly promising candidate is Universal Structure Modeling (USM) that uses a Bayesian neural network approach to search for interactions, and quadratic and other higher-order effects within path models (Buckler & Hennig-Thurau, 2008). Specifically, USM derives starting values for the model's latent variables through principal component analysis, and then applies the Bayesian neural network approach to discover the optimal system of linear, nonlinear, and interactive pathways among the latent variables, with the aim of maximizing the variance explained (e.g., Rigdon, Ringle & Sarstedt, 2010; McIntosh, Edwards & Antonakis, 2014; Henseler, Hubona & Ray, 2016). As a result, the USM approach typically yields higher explanatory power than PLS-PM-based linear modeling (Albashrawi, Kartal, Oztekin & Motiwalla, 2017). Since USM

has recently received increasing attention in business research (e.g., Garbe & Richter, 2009; Oztekin, Kong & Delen, 2011; Turkyilmaz, Oztekin, Zaim & Demirel, 2013; Turkyilmaz, Temizer & Oztekin, 2018; Al-Ebbini, Oztekin & Chen, 2016), investigating the method's predictive accuracy would be particularly promising.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1), 243-247.

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.

Al-Ebbini, L., Oztekin, A., & Chen, Y. (2016). FLAS: fuzzy lung allocation system for US-based transplantations. *European Journal of Operational Research*, 248(3), 1051-1065.

Albashrawi, M., Kartal, H., Oztekin, A., & Motiwalla, L. (2017). The impact of subjective and objective experience on mobile banking usage: an analytical approach. *Proceedings of the 50th Hawaii International Conference on System Sciences*. Available at: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1159&context=hicss-50

Babin, B. J., Hair, J. F., & Boles, J. S. (2008). Publishing research in marketing journals using structural equation modeling. *Journal of Marketing Theory and Practice*, 16(4), 279-285.

Bartikowski, B., & Walsh, G. (2011). Investigating mediators between corporate reputation and customer citizenship behaviors. *Journal of Business Research*, 64(1), 39-44.

Becker, J. M., Rai, A., & Rigdon, E. (2013). Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance. *34th International Conference on Information Systems*, Milan, Italy.

Bentler, P. M., & Mooijaart, A. B. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin*, 106(2), 315-317.

Berk, R. (2008). *Statistical learning from regression perspective*. Springer, New York.

Buckler, F., & Hennig-Thurau, T. H. (2008). Identifying hidden structures in marketing's structural models through universal structure modeling: an explorative Bayesian Neural Network complement to LISREL and PLS. *Marketing ZfP. Journal of Research and Management*, 4(2), 47-66.

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

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.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295-358). Mahwah: Erlbaum.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, 29(3), 510-522.

Dick, A. S., & Basu, K. (1994).- Customer loyalty: Toward an integrated conceptual framework. *Journal of the Academy of Marketing Science*, 22(2), 99-113.

Dijkstra, T. K., & Henseler, J. (2015). Consistent partial least squares path modeling. *MIS Quarterly*, 39(2), 297-316.

Eberl, M. (2010). An application of PLS in multi-group analysis: the need for differentiated corporate-level marketing in the mobile communications industry. In V. Esposito Vinzi, W. W. Chin, & J. Henseler (Eds.), *Handbook of partial least squares: concepts, methods and applications* (pp. 487-514). Heidelberg et al.: Springer.

Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. *Journal of Business Research*, 69(10), 4565-4582.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data. *Applied statistics*, 47(2), 231-250.

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2), 93-98.
Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1-35.
Garbem J.-N., & Richter, N. F. (2009). Causal analysis of the internationalization and performance relationship based on neural networks – advocating the transnational structure. *Journal of International Management*, 15(4), 413-431.
Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, 30(3), 611-642.
Goodhue, D. L., Lewis, W., & Thompson, R. (2012a). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, 36(3), 981-1001.
Goodwin, P., and Lawton, L. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, 15(4), 405-408.
Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012a). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40(3), 414-433.
Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012b). The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long Range Planning*, 45(5), 320-340.
Hair, J. F., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y L. (2017a). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems*, 117(3), 442-458.
Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017b). *A primer on partial least squares structural equation modeling (PLS-SEM)*, 2nd edition. Thousand Oaks, CA: Sage.
Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017c) 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.
Hair, J.F., Sarstedt, M., Ringle, C. M., & Gudergan, S P. (2018). *Advanced issues in partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, CA: Sage.
Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics*, 28(2), 565-580.
Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M., & Calantone, R. J. (2014). Common beliefs and reality about partial least squares: comments on Rönkkö & Evermann (2013). *Organizational Research Methods*, 17(2), 182–209.
Henseler, J., Hubona, G. S., & Ray, P. A. (2016). Using PLS path modeling in new technology research: updated guidelines. *Industrial Management & Data Systems*, 116(1), 1-19.
Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, 55(1), 1-34.
Hwang, H., Malhotra, N. K., Kim, Y., Tomiuk, M. A., & Hong, S. (2010). A comparative study on parameter recovery of three approaches to structural equation modeling. *Journal of Marketing Research*, 47(4), 699-712.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.
Iyengar, K., Sweeney, J. R., & Montealegre (2015). Information technology use as a learning mechanism: The impact of IT use on knowledge transfer effectiveness, absorptive capacity, and franchisee performance. *MIS Quarterly*, 39(3), 615-641.
Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 255–284). New York, NJ: Seminar Press.
Jöreskog, K.G. and Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In H. Wold & K. Jöreskog (Eds.), *Systems under indirect observation: causality, structure, prediction* (Vol. I), Amsterdam: North-Holland, 263-270.
Kuha, J. (2004). AIC and BIC: comparisons of assumptions and performance. *Sociological Methods &*

Research, 33(2), 188-229.
Lee, L., Petter, S., Fayard, D., & Robinson, S. (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, 12(4), 305-328.
Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting,* 9(4),527-529.
Matthews, L., M., Sarstedt, M., Hair, J. F., & Ringle, C. M. (2016). Identifying and treating unobserved heterogeneity with FIMIX-PLS. Part II – a case study. *European Business Review* 28(2), 208-224.
McIntosh, C. N., Edwards, J. R., & Antonakis, J. (2014). Reflections on partial least squares path modeling. *Organizational Research Methods*, 17(2), 210-251.
McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection* (Vol. 43). Singapore: World Scientific.
Monecke, A. (2012). semPLS: An R package for structural equation models using partial least squares. R Package Version 1.0-08. Available at: https://cran.rproject. org/web/packages/semPLS/index.html
Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44(1), 190-204.
Nau, R. 2016. Statistical forecasting: Notes on regression and time series analysis, in: Durham: Fuqua School of Business, Duke University. Available at: https://people.duke.edu/~rnau/compare.htm.
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*, *39*, 19-35.
Nitzl, C., & Chin, W. W. (2017). The case of partial least squares (PLS) path modeling in managerial accounting research. *Journal of Management Control*, 28(2), 137-156.
Oliver, R. L. (1993). Cognitive, affective, and attribute bases of the satisfaction response. *Journal of Consumer Research*, 20(3), 418-430.
Oztekin, A., Kong, Z. J., & Delen, D. (2011). Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations. *Decision Support Systems*, 51(1), 155-166.
Park, I, Sharman, R., & Rao H. R. (2015). Disaster experience and hospital information systems: An examination of perceived information assurance, risk, resilience, and HIS usefulness. *MIS Quarterly*, 39(2), 317-344.
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.
Peng, D. X., & Lai, F. (2012). Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management*, 30(6), 467-480.
Polites, G. L., & Karahanna, E. (2012). Shackled to the status quo: The inhibiting effects of incumbent system habit, switching costs, and inertia on new system acceptance, *MIS Quarterly*, *36*(1), 21-41.
R Development Core Team. (2014). R: A language and environment for statistical computing. The R foundation for statistical computing, Vienna, Austria.
Raithel, S. & Schwaiger, M. (2015). The effects of corporate reputation perceptions of the general public on shareholder value. *Strategic Management Journal*, 36(6), 945-56.
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.
Richter, N. F., Sinkovics, R. R., Ringle, C. M., & Schlägel, C. (2016). A critical look at the use of SEM in international business research. *International Marketing Review*, 33(3), 376-404.
Rigdon, E. E., Ringle, C. M., & Sarstedt, M. (2010). Structural modeling of heterogeneous data with partial least squares. In N. K. Malhotra (Ed.), *Review of Marketing Research* (Vol. 7), Bingley: Emerald Group Publishing, 255-296.
Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, 34(6), 598-605.
Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-

SEM in MIS quarterly. *MIS Quarterly*, 36(1), iii-xiv.
Ringle, C. M., Sarstedt, M., & Schlittgen, R. (2014). Genetic algorithm segmentation in partial least squares structural equation modeling. *OR Spectrum*, 36(1), 251-276.
Sarstedt, M. & Ringle, M. (2010). Treating unobserved heterogeneity in PLS path modelling: a comparison of FIMIX-PLS with different data analysis strategies. *Journal of Applied Statistics*, 37(8), 1299-318.
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), 329-39.
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.
Sarstedt, M., Ringle, C. M., & Hair, J. F. (2017). Partial least squares structural equation modeling. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research,* Berlin et al.: Springer. Available at: https://link.springer.com/referenceworkentry/10.1007/978-3-319-05542-8_15-1
Schlittgen, R. (2015). SEGIRLS: Clusters regression, pls path and gsca models by iterative reweighting. R package version 0.5. Available at: http://www3.wiso.uni-hamburg.de/fileadmin/bwl/statistikundoekonometrie/Schlittgen/SoftwareUndDaten/SEGIRLS_0.5.tar.gz
Schwaiger, Manfred (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. *Annals of Statistics*, 6(2), 461-464.
Shao J. (1993). Linear model selection by cross-validation. *Journal of American Statistical Association*, 88(422), 486-494.
Sharma, P.N. and Kim, K.H. (2012). Model selection in information systems research using partial least squares-based structural equation modeling, in *Proceedings of the 33rd International Conference on Information Systems*, Orlando, FL.
Sharma, P.N., Sarstedt, M., Shmueli, G., Kim, K.H, and Thiele, K.O. (2018). Model selection in MIS research using PLS-SEM, Working Paper.
Shi, P. and Tsai C.L. (2002). Regression model selection—a residual likelihood approach. *Journal of the Royal Statistical Society Series B*, 64(2), 237-252.
Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289-310.
Shmueli, G. and Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553-572.
Shmueli, G., Ray, S., Estrada, J. M. V., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research*, 69(10), 4552-4564.
Steenkamp, J.-B. E. M., & Baumgartner, H. (2000). On the use of structural equation models for marketing modeling. *International Journal of Research in Marketing*, 17(2/3), 195-202.
Stone M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society Series B*, 39(1), 44-47.
Tenenhaus, M., Amato, S., & Esposito Vinzi, V. (2004). A global goodness-of-fit index for PLS structural equation modelling. In *Proceedings of the XLII SIS scientific meeting*, 739-742.
Tenenhaus, A., & Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2), 257–284.
Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8), 1352-1362.
Turkyilmaz, A., Oztekin, A., Zaim, S., & Demirel, O. F. (2013). Universal structure modeling approach to customer satisfaction index. *Industrial Management & Data Systems*, 113(7), 932-949.
Turkyilmaz, A., Temizer, L., & Oztekin, A. (2018). A causal analytic approach to student satisfaction index modeling. *Annals of Operations Research*, 263(1-2), 565-585.
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.
Venkatesh, V., Brown, S. A., Maruping, L. M., & Bala, H. (2008). Predicting different conceptualizations of system use: the competing roles of behavioral intention, facilitating conditions, and behavioral

expectation. *MIS Quarterly*, 32(3), 483-502.
Vilares, M. J., & Coelho, P. S. (2013). Likelihood and PLS estimators for structural equation modeling: an assessment of sample size, skewness and model misspecification effects. In J. Lita da Silva, F. Caeiro, I. Natário, & C. A. Braumann (Eds.), *Advances in regression, survival analysis, extreme values, Markov processes and other statistical applications* (pp. 11–33). Berlin: Springer.
Vrieze, S. I. (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.
Walsh, G., Beatty, S. E., & Shiu, E. M. K. (2009). The customer-based corporate reputation scale: Replication and short form. *Journal of Business Research*, 62(10), 924-930.
Wold, H. (1974). Causal flows with latent variables: partings of the ways in the light of NIPALS modelling. *European Economic Review*, 5(1), 67-86.
Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 47-74). New York.

Appendix A
Model Selection Criteria

The simplest criteria widely used in PLS-PM is the R^2, which is calculated as:

$$R^2 = 1 - \frac{SS_{error_k}}{SS_{total}}$$

where SS_{error_k} is the sum of squared errors and SS_{total} is total sum of squares. The Adjusted R^2 can be written in terms of p_k, which is the number of predictors plus 1:

$$Adjusted\ R^2 = 1 - \left[\left(\frac{n-1}{n-p_k}\right)\ \left(\frac{SS_{error_k}}{SS_{total}}\right)\right]$$

The efficient and consistent model selection criteria described in this paper can be written as a function of the maximized value of the likelihood function ($\hat{L}$). For example,

$$AIC = -2ln\hat{L} + 2p_k$$

$$BIC = -2ln\hat{L} + p_k ln(n)$$

$$HQ = -2ln\hat{L} + 2p_k ln(ln(n))$$

Under a normal error distribution assumption, these likelihood-based formulas can be written in terms of SS_{error} as shown in Table A1 (Burnham & Anderson, 2002; p.63; McQuarrie & Tsai, 1998). An Excel spreadsheet that illustrates the computation of all model selection criteria considered in this study using the standard output from any PLS software can be downloaded from: https://www.pls-sem.net/downloads/

Table A1: Formulas for model selection criteria based on SS_{error}

Criterion	Formula	Description
Distance-based criteria		
Final Prediction Error (FPE)	$\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}-\mathrm{p_k}}\right)\times\left(1+\frac{\mathrm{p_k}}{\mathrm{n}}\right)$	Selects the best model by minimizing the final prediction error.
Mallow's Cp	$\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{MS_{error}}}\right)-(\mathrm{n}-2\mathrm{p_k})$	Based on mean square error (MSE); $\mathrm{MS_{error}}$ is MSE from the saturated (full) model.
Akaike Information Criterion (AIC)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}}\right)+\frac{2\mathrm{p_k}}{\mathrm{n}}\right]$	Estimates the relative expected KL distance to the unknown true model.
Unbiased AIC (AICu)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}-\mathrm{p_k}}\right)+\frac{2\mathrm{p_k}}{\mathrm{n}}\right]$	Uses the unbiased estimate for population MSE, hence differs from AIC in small samples.
Corrected AIC (AICc)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}}\right)+\frac{\mathrm{n}+\mathrm{p_k}}{\mathrm{n}-\mathrm{p_k}-2}\right]$	Corrects AIC's tendency to overfit (select a complicated model) under small samples.
Consistent criteria		
Bayesian Information Criterion (BIC)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}}\right)+\frac{\mathrm{p_k}\log(\mathrm{n})}{\mathrm{n}}\right]$	Derived using Bayesian argument; adjusts AIC for model complexity by using a stronger penalty for overfitting.
Geweke-Meese Criterion (GM)	$\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{MS_{error}}}\right)+\mathrm{p_k}\log(\mathrm{n})$	Adjusts Mallow's Cp for model complexity by using a stronger penalty for overfitting.
Hannan-Quinn Criterion (HQ)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}}\right)+\frac{2\mathrm{p_k}\log(\log(\mathrm{n}))}{\mathrm{n}}\right]$	Corrects small sample performance of BIC by using a stronger penalty term.
Corrected HQ Criterion (HQc)	$\mathrm{n}\left[\log\left(\frac{\mathrm{SS_{error_k}}}{\mathrm{n}}\right)+\frac{2\mathrm{p_k}\log(\log(\mathrm{n}))}{\mathrm{n}-\mathrm{p_k}-2}\right]$	Corrects small sample performance of HQ and adjusts for model complexity.

Note: $SSerror_{(k)}$ is the sum squares error for the k^{th} model in a set of models; MS_{error} is the mean squared error from the saturated model; SS_{total} is the total sum of squares; p_k is the number of predictors in the k^{th} model plus 1.

Table A2: Formulas for predictive metrics

Criterion	Formula	Description
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	The RMSE represents the sample standard deviation of the differences between predicted values and observed values.
MAPE	$\frac{100}{n}\sum_{i=1}^{n}\left\|\frac{y_i - \hat{y}_i}{y_i}\right\|$	A percentage metric reflecting the mean absolute percentage of predictive error over actual value.
SMAPE	$\frac{100}{n}\sum_{i=1}^{n}\frac{\|y_i - \hat{y}_i\|}{\|y_i\| + \|\hat{y}_i\|}$	A symmetric percentage metric based upon MAPE.
MAD	$\frac{1}{n}\sum_{i=1}^{n}\|y_i - \hat{y}_i\|$	Mean absolute deviation (MAD) of a data set is the mean of the absolute predictive error.
Q^2	$1 - \frac{SSE_q}{SSO_q}$	Stone-Geisser's Q^2 value is a criterion of predictive relevance obtained using the blindfolding procedure. Please refer to Stone (1974) and Geisser (1974) for complete details on the calculation of SSE_q and SSO_q.

Note: y_i represents the actual composite score of the target construct, while $\hat{y}_i$ represents the predicted score.

Appendix B

Figure B1: Comparison of criteria value distributions across the 8 models (Scenario 1)

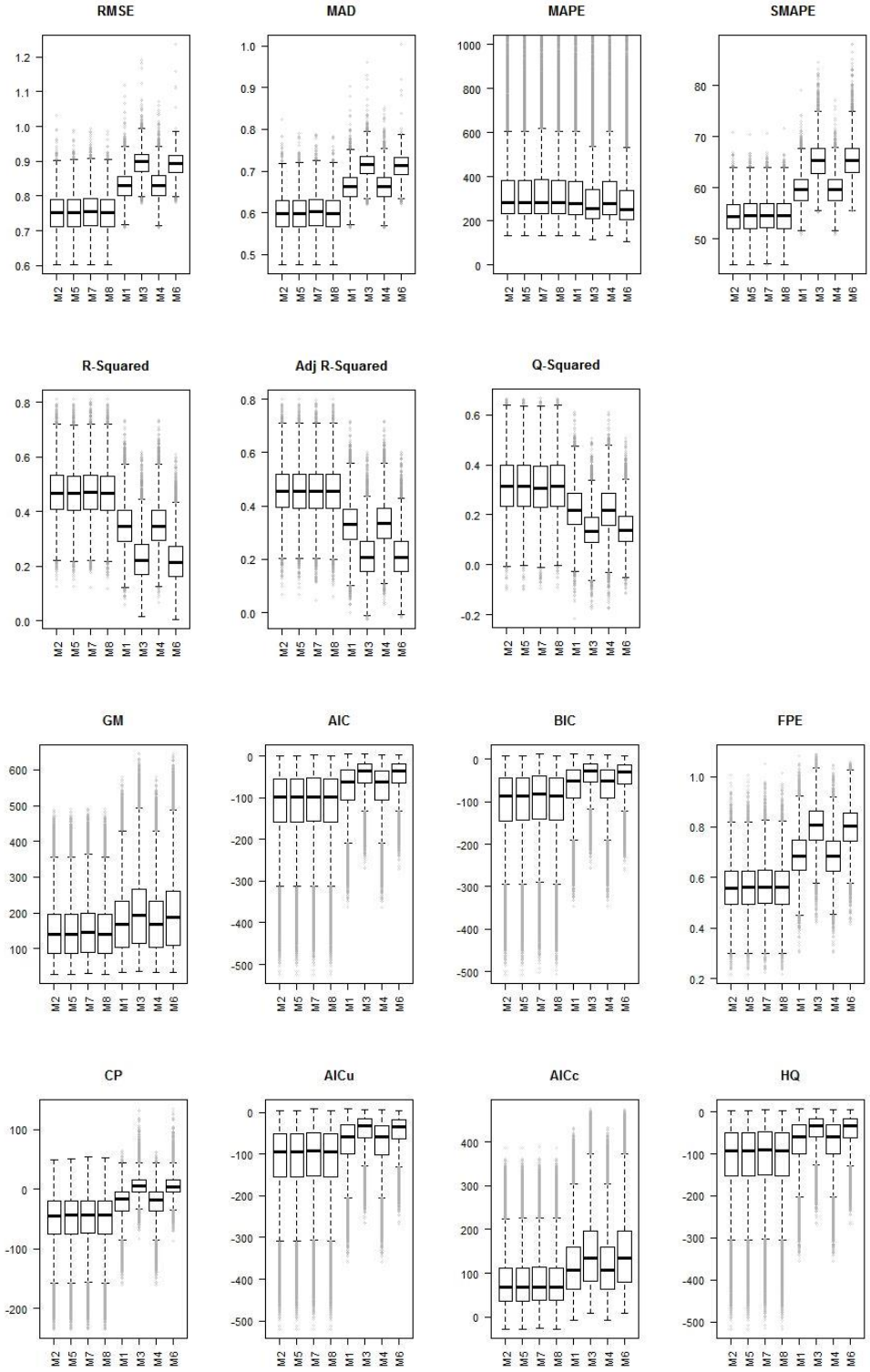


Figure B2: Comparison of criteria value distributions across the 8 models (Scenario 2)

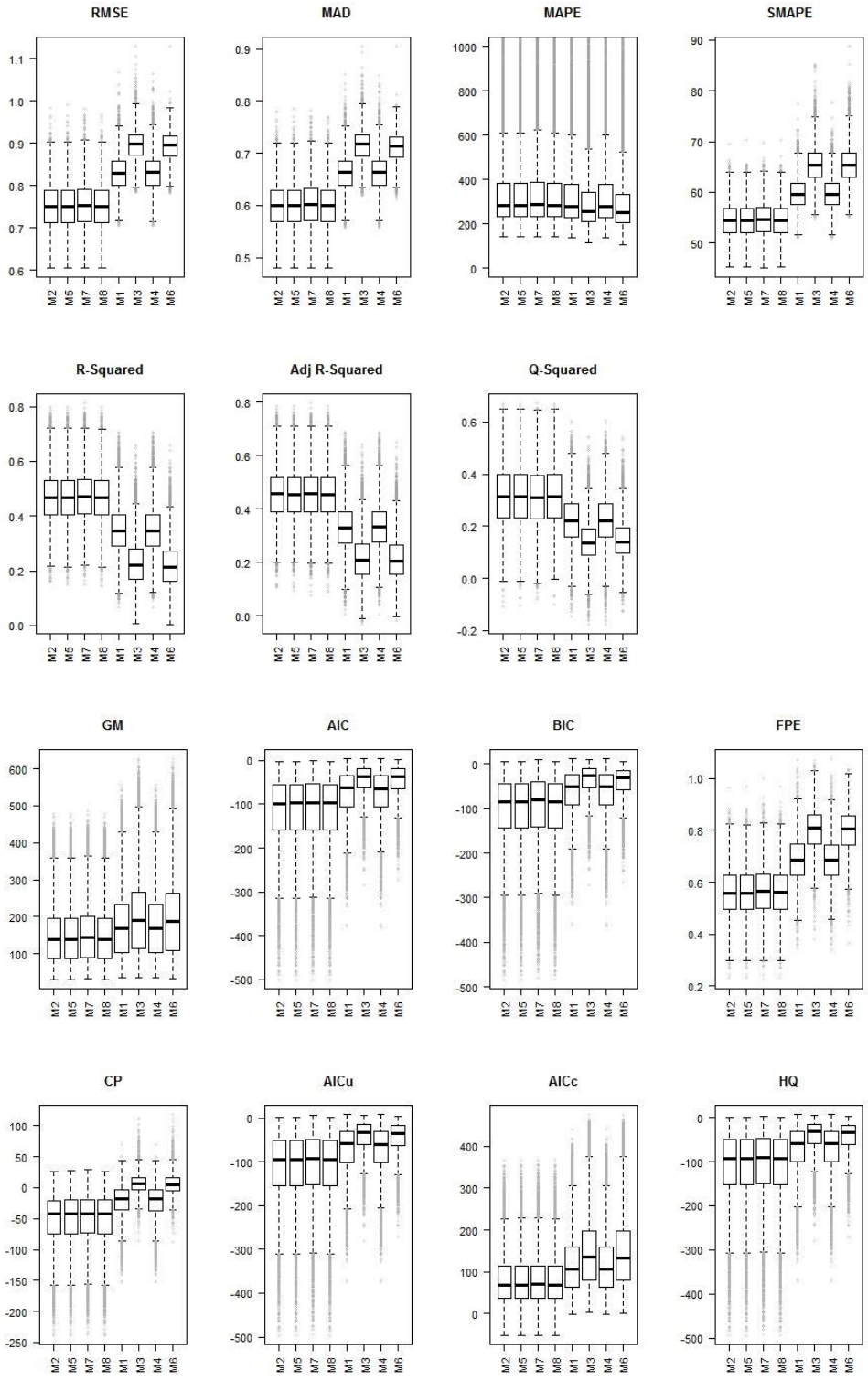


Table B1: Percentage agreement with RMSE by model number (Scenario 2)

Model #		1	2	3	4	5	6	7	8	Total
PLS-PM Criteria	R^2	0.000	0.085	0.000	0.000	0.002	0.000	0.126	0.002	0.214
	Adjusted R^2	0.000	0.166	0.000	0.000	0.010	0.000	0.031	0.012	0.219
	Q^2	0.000	0.103	0.000	0.000	0.038	0.000	0.016	0.057	0.213
Model Selection Criteria	FPE	0.000	0.206	0.000	0.000	0.013	0.000	0.011	0.015	0.245
	CP	0.000	0.225	0.000	0.000	0.015	0.000	0.005	0.019	0.264
	GM	0.000	0.256	0.000	0.000	0.017	0.000	0.000	0.020	0.294
	AIC	0.000	0.206	0.000	0.000	0.013	0.000	0.011	0.015	0.245
	AICu	0.000	0.228	0.000	0.000	0.015	0.000	0.005	0.018	0.266
	AICc	0.000	0.213	0.000	0.000	0.013	0.000	0.010	0.016	0.252
	BIC	0.000	0.248	0.000	0.000	0.017	0.000	0.000	0.020	0.285
	HQ	0.000	0.229	0.000	0.000	0.015	0.000	0.003	0.018	0.266
	HQc	0.000	0.233	0.000	0.000	0.016	0.000	0.003	0.019	0.270

Appendix C

Robustness Checks for Non-Normal Distribution

Table C1: Overall proportion of model choice by each criterion under log-normal distribution (Scenario 1)

	Model #	1	2	3	4	5	6	7	8
PLS-PM Criteria	**R^2**	0.002	0.340	0.001	0.012	0.016	0.000	0.619	0.010
	Adjusted R^2	0.002	0.592	0.003	0.015	0.058	0.003	0.271	0.057
	Q^2	0.005	0.293	0.005	0.004	0.234	0.014	0.162	0.283
Model Selection Criteria	**FPE**	0.002	0.682	0.004	0.016	0.071	0.008	0.147	0.071
	CP	0.002	0.681	0.004	0.016	0.071	0.008	0.147	0.071
	GM	0.003	0.748	0.006	0.013	0.078	0.036	0.035	0.082
	AIC	0.002	0.682	0.004	0.016	0.071	0.008	0.147	0.071
	AICu	0.002	0.724	0.005	0.014	0.075	0.021	0.082	0.077
	AICc	0.002	0.691	0.005	0.015	0.072	0.012	0.130	0.072
	BIC	0.002	0.746	0.006	0.013	0.078	0.040	0.034	0.081
	HQ	0.002	0.730	0.005	0.015	0.076	0.018	0.077	0.078
	HQc	0.002	0.735	0.006	0.014	0.076	0.024	0.065	0.079
Out-of-Sample Criteria	**MAD**	0.002	0.373	0.001	0.001	0.211	0.002	0.162	0.250
	RMSE	0.005	0.361	0.001	0.001	0.214	0.005	0.162	0.252
	MAPE	0.071	0.010	0.255	0.042	0.012	0.570	0.019	0.020
	SMAPE	0.001	0.424	0.000	0.001	0.107	0.000	0.338	0.128

Table C2: Percentage agreement with RMSE by model type under log-normal distribution (Scenario 1)

Model Type		Correctly Specified (Model 2 or 5 or 8)	Incorrectly Specified (Model 1 or 3 or 4 or 6)	Saturated (Model 7)
PLS-PM Criteria	**R^2**	0.279	0.001	0.079
	Adjusted R^2	0.552	0.001	0.014
	Q^2	0.653	0.002	0.011
Model Selection Criteria	**FPE**	0.658	0.001	0.004
	CP	0.658	0.001	0.004
	GM	0.741	0.002	0.000
	AIC	0.658	0.001	0.004
	AICu	0.709	0.002	0.001
	AICc	0.670	0.001	0.003
	BIC	0.738	0.002	0.000
	HQ	0.716	0.002	0.001
	HQc	0.722	0.002	0.001

Table C3: Overall proportion of model choice by each criterion under log-normal distribution (Scenario 2)

	Model #	1	2	3	4	5	6	7	8
PLS-PM Criteria	**R^2**	0.001	0.339	0.001	0.014	0.014	0.000	0.623	0.009
	Adjusted R^2	0.002	0.593	0.003	0.016	0.060	0.002	0.270	0.055
	Q^2	0.006	0.300	0.005	0.006	0.233	0.015	0.153	0.282
Model Selection Criteria	**FPE**	0.002	0.680	0.004	0.016	0.073	0.007	0.152	0.066
	CP	0.002	0.680	0.004	0.016	0.073	0.007	0.152	0.066
	GM	0.002	0.749	0.007	0.014	0.081	0.037	0.034	0.077
	AIC	0.002	0.680	0.004	0.016	0.073	0.007	0.152	0.066
	AICu	0.002	0.724	0.006	0.016	0.078	0.019	0.083	0.072
	AICc	0.002	0.691	0.005	0.016	0.075	0.011	0.133	0.067
	BIC	0.002	0.748	0.007	0.014	0.080	0.040	0.033	0.077
	HQ	0.002	0.733	0.005	0.016	0.079	0.015	0.077	0.073
	HQc	0.002	0.737	0.006	0.015	0.079	0.023	0.064	0.074
Out-of-Sample Criteria	**MAD**	0.002	0.381	0.000	0.000	0.207	0.002	0.160	0.248
	RMSE	0.005	0.362	0.001	0.001	0.208	0.004	0.169	0.251
	MAPE	0.075	0.014	0.249	0.040	0.013	0.567	0.021	0.022
	SMAPE	0.001	0.445	0.000	0.001	0.115	0.000	0.313	0.125

Table C4: Percentage agreement with RMSE by model type under log-normal distribution (Scenario 2)

Model Type		Correctly Specified (Model 2 or 5 or 8)	Incorrectly Specified (Model 1 or 3 or 4 or 6)	Saturated (Model 7)
PLS-PM Criteria	**R^2**	0.270	0.001	0.081
	Adjusted R^2	0.547	0.001	0.013
	Q^2	0.652	0.002	0.012
Model Selection Criteria	**FPE**	0.648	0.001	0.005
	CP	0.648	0.001	0.005
	GM	0.734	0.002	0.000
	AIC	0.648	0.001	0.005
	AICu	0.701	0.002	0.002
	AICc	0.662	0.001	0.004
	BIC	0.733	0.002	0.000
	HQ	0.711	0.002	0.001
	HQc	0.716	0.002	0.001