IS MY MODEL RIGHT? MODEL QUALITY AND MODEL MISSPECIFICATION IN PLS – RECOMMENDATIONS FOR IS RESEARCH

Joerg Evermann
Memorial University of Newfoundland
jevermann@mun.ca

Mary Tate
Victoria University of Wellington
mary.tate@vuw.ac.nz

ABSTRACT

Partial Least Squares (PLS) is a statistical technique that is widely used in the Information Systems discipline to estimate statistical models with structural equations and latent variables. While PLS does not provide a statistical test of model fit to data, PLS users apply a set of heuristics to evaluate the quality of estimated models. In this paper, we investigate to what extent these heuristics are able to identify misspecified models and may be used in place of a statistical test of model fit. The results of our analysis and simulation study indicate that existing heuristics are unable to reliably detect typical model misspecifications. We discuss the implications of this result for theory testing, prediction, and exploratory analysis in Information Systems research.

Keywords: Partial Least Squares, Misspecification, Simulation Study, Structural Equation Modeling, Data analysis, Research methods

INTRODUCTION

Information systems (IS) researchers build theories that are intended, among other things, for explanation and prediction of IS phenomena (Gregor, 2006). While prediction is an important aspect to theorizing, existing IS research has focused primarily on explanation (Shmueli and Koppius, 2011). In the prevalent positivist, quantitative paradigm in IS research (Chen and Hirschheim, 2004), explanatory research takes the form of proposing causal theories which are subsequently tested against observation. The primary goal is for a theory to be correct in the sense that is an isomorphic model of the real-world causal system.

The nature and increasing complexity of IS phenomena leads researchers to represent their theories in models of simultaneous regression equations that include latent and observed variables. Partial Least Squares Path Modeling (PLS) is an increasingly popular technique in IS research to analyze these structural equation models (Ringle et al., 2012). IS research is the primary user of PLS analysis (Rouse and Corbitt, 2008) and the use of PLS is increasing in high-quality IS and marketing journals (Ringle et al., 2012). For example, Information Systems Research published 32 articles using PLS in the ten year span between 2000 and 2009, including important methodological studies on PLS (Chin et al., 2003; Goodhue et al., 2007).

PLS is a parametric statistical technique that estimates values of model parameters from a (training) sample. Just as theories can be used for explanation and prediction (Gregor, 2006), statistical models and techniques can be used for explanation and prediction. For prediction, a statistical technique should approximate observations in a training or test sample using the estimated parameters from the training sample (Hastie et al., 2009). When used for explanation, a statistical technique should allow population parameters of the real-world causal system to be inferred from the sample. It should produce parameter estimates that are consistent and unbiased. Because the parameter estimates depend on the form of the estimated model, that model must accurately represent the real-world causal system. If the two are not isomorphic, the estimated parameters have no real-world counterpart and cannot be interpreted

substantively. Thus, model testing, the determination of whether the model is correct in the sense of being isomorphic to the real-world causal system, plays a critical role in explanatory research.

PLS does not offer a statistical test of model correctness, as covariance-based methods do, and "this issue limits PLS-SEMs's usefulness for theory testing" (Hair et al., 2012, pg. 416). However, a number of model evaluation criteria are proposed in the PLS literature. *In this paper, we investigate whether the existing model evaluation criteria proposed in the PLS literature can, in the absence of a global statistical test, provide reliable indications of model correctness. Given the extent of theory testing research in IS using PLS, it is useful to identify to what degree researchers can rely on PLS results to support the substantive conclusions they can draw from their studies.* This paper contributes an examination of the theoretically expected and actual behavior of PLS model quality heuristics under various model misspecifications and their usefulness for identifying misspecified models.

Using a simulation study, we examine the behavior of a set of widely-used model quality heuristics for a range of models and conditions. Our results show that, individually or jointly, the proposed model quality heuristics do not allow researchers to make a reliable assessment of the correctness of the specified model. This is a weakness of PLS when used in explanatory, theory testing research, especially when compared to covariance-based alternatives that do provide a reliable test (Hair et al., 2012). While PLS may have advantages over other techniques such as increased statistical power for low samples, a better ability to estimate formative models, or a lack of distributional assumptions, the fact that researchers cannot draw reliable conclusions about the correctness of the estimated model in a theory testing context, leads us to recommend against the use of PLS for explanatory, theory testing research.

The next section briefly introduces existing model quality criteria for PLS estimation. We then describe the simulation study followed by a presentation of our results. The paper concludes with a discussion and recommendations for researchers. Eight appendices describe details of the Partial Least Squares Path Modeling technique (Appendix A), provide an overview of prior empirical work on PLS (Appendix B),

show details of the simulation conditions (Appendix C), describe the data generation and model estimation (Appendix D), discuss the effects of cross-loadings and error-variances on reliability estimates (Appendices E, F), present detailed findings (Appendix G) and provide details of our classification and prediction approach (Appendix H).

PLS MODEL QUALITY HEURISTICS

The result of a PLS estimation of a model is evaluated using a number of model quality heuristics (Hair et al., 2012; Ringle et al., 2012). These are not statistical tests, as there are no test statistics with a known probability distributions, no type I and II error rates, and the recommended values for these heuristics are rules of thumb that vary with different authors. In this section, we briefly introduce these model quality criteria. They will be described in more detail as we present and discuss our results. A summary of the heuristics is provided in Table 1, together with representative references to the literature that recommends the use and specific values for these heuristics.

Table 1: Criteria for assessing quality of models using PLS

Criterion		Recommendation	References
1	Indicator loadings	> 0.7	Chin and Gopal, 1995 Gefen et al., 2000 Goetz et al., 2010 Hair et al., 2012 Hulland, 1999 Henseler et al., 2009 Straub *et al.* 2004
2	Cross loadings (relative)	less than indicator loadings (by at least 0.1)	Chin, 2010 Gefen et al., 2000 Gefen and Straub, 2005 Goetz et al., 2010 Hair et al., 2012 Henseler et al., 2009 Hulland, 1999 Straub et al., 2004
3	Cross loadings (absolute)	< 0.4	Straub et al., 2004

Criterion		Recommendation	References
4	Reliability (Cronbach's α)	> 0.7 or > 0.6	Gefen et al., 2000 Goetz et al., 2010 Hair et al., 2012 Hulland, 1999 Straub et al., 2004
5	Reliability (composite reliability ρ)	> 0.7 or > 0.6	Gefen et al., 2000 Hair et al., 2012 Hulland, 1999 Straub et al., 2004
6	AVE	> 0.5	Chin, 1998 Chin, 2010 Gefen and Straub, 2005 Hair et al., 2012 Henseler et al., 2009
7	AVE	> latent correlations	Gefen et al., 2000
8	Root AVE	> latent correlations	Chin, 1998 Chin, 2010 Fornell and Larcker, 1981 Gefen and Straub, 2005 Henseler et al., 2009 Hulland, 1999
9	Coefficient of determination (r^2)	significant (larger is better)	Chin, 1998 Hair et al., 2012 Fornell and Larcker, 1981
10	AVE	> r^2	Fornell and Larcker, 1981
11	Structural Paths	significant at 0 .05	Chin, 2010 Hair et al., 2012
12	GoF (relative)	> 0.9	Chin, 2010 Esposito Vinzi et al., 2010
13	Predictive Relevance (Q^2)	> 0.5 > 0	Chin, 2010 Hair et al., 2012

The evaluation of the measurement model ("outer model") focuses on heuristics for reliability and validity. The assessment of convergent validity focuses on the average variance extracted (AVE) whereas discriminant and indicator validity focuses on the loadings and cross-loadings of the indicators on the intended latent construct. Reliability is typically assessed using either Cronbach's α or composite

reliability ρ. The evaluation of the structural model ("inner model") focuses on the coefficient of determination (r^2) for endogenous latent variables and the statistical significance of the estimated structural relationships.

Recognizing the need for an overall measure of model quality, a goodness-of-fit (GoF) metric has recently been proposed. However, contrary to its name, this GoF metric is a combination of evaluation heuristics for the inner and outer models, and does not indicate the fit of the estimated model to observed data.

Finally, a blindfolding procedure can be used to compute the Q^2 metrics to assess the predictive ability of an estimated PLS model. While this study focuses on identifying misspecified models in an explanatory theory-testing context, this heuristic is typically seen as an important model quality indicator (Chin, 2010; Hair et al., 2012) and should be reported (Ringle et al., 2012).

The inner and outer model evaluation metrics are widely used for model quality assessment and consistently reported in the literature (Hair et al., 2012; Ringle et al., 2012). However, few studies report the goodness-of-fit metric or the test of predictive relevance; the former is a relatively recent development (Tenenhaus et al., 2004) and the latter may not be used by researchers who focus on theory testing instead of prediction.

While other model quality heuristics have been proposed in the PLS literature, such as the effect sizes f^2 or q^2 (Chin, 2010; Ringle et al., 2012; Hair et al., 2012) there are no recommendations of required or desired values for these. Similarly, Lohmöller (1989) recommends additional criteria, such as the unexplained covariance among the indicators (should be "low enough"), the covariances between indicators of different latent variables (should be "near zero"), covariances between residuals of the indicators and the latent variable estimates (should be "near zero"), covariances among the residuals of the indicators and the residuals of the latent variables (should be "low enough"), but without

recommending required or desired values for these criteria. Hence, we have not investigated these criteria and instead focused on criteria that are presented in the form of "hard" constraints, similar to statistical tests, and currently recommended in the Information Systems literature.

Finally, while we focus on PLS model quality criteria, we also estimated each model using CB-SEM and used the p-value of the $\chi 2$ statistical test of model fit for comparison.

SIMULATION STUDY

The iterative and complex nature of the parameter estimation in PLS (Appendix A) and the complexity of realistic models precludes an analytical derivation of the impact of model misspecifications on many of the model quality heuristics. We therefore use a simulation study (Paxton et al., 2001) to investigate the extent to which the PLS model quality criteria are able to identify model misspecification. A simulation study overcomes the inability to know the true model when real world data is estimated. It is a controlled experiment where sample data is repeatedly generated from a known, true model for which the researcher specifies the parameter values. These sample data are then used to estimate parameter values for a range of misspecified models.

Models

Models for simulation studies should be representative of those found in the substantive literature (Paxton et al., 2001). PLS models in highly-ranked IS and marketing journals have a median number of 7 to 9 latent variables with 9 to 11 structural relations (Ringle et al., 2012). Another study on the use of PLS in the marketing literature reported a median of 7 latent variables with a median of 8 structural relations (Hair et al., 2012). Similarly, the PLS studies published in ISR between 2000 and 2009 have a median of 9 latent variables and 12 structural relations. Both Ringle et al. (2012) and Hair et al. (2012) report a median of 3.5 indicators for each reflective construct, which is similar to PLS studies in ISR, which have an average of 3.8 indicators for each reflective construct. Based on these reports, we examine the models

shown in Figures 1 through 3 (for model clarity, indicators are not shown). While models 1 and 2 are relatively simple models, model 3 matches the typical characteristics of PLS models in the literature quite well. Together, they cover model complexity from low to typical.

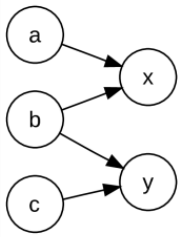


Figure 1: Model 1

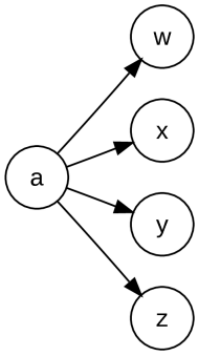


Figure 2: Model 2

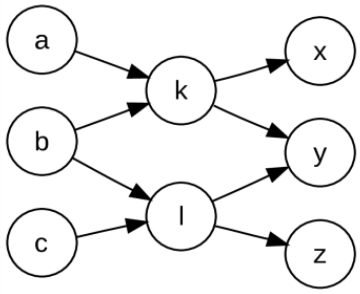


Figure 3: Model 3

Design Factors

Sample size, the number of indicators, indicator loadings, structural effect sizes and measurement specification (formative or reflective) are important design variables for PLS simulation studies (Aguirre-Ureta and Marakas, 2008; Goodhue et al., 2007; Goodhue et al., 2012; Reinartz et al., 2009). This study is intended to be conservative in its evaluation of the PLS model heuristics and assumes ideal samples in the sense of continuous responses, multivariate-normality, and complete data. While there has been much recent attention on the specification of formative measurement, the use of reflective measurement is dominant for PLS models in the IS literature (Ringle et al., 2012). Further, Ringle et al. (2012) report a median sample size of 198 for PLS studies published in MIS Quarterly, and the PLS studies published in ISR have a median sample size of 176. Table 2 shows a summary of the experimental factors in this study and their levels.

Table 2: Study Design Factors

Design factor		**Levels**
s	Sample size	100, 250, 750
i	Number of indicators	3, 5, 7
l	Indicator loadings (non-std.)	.75, 1, 1.25
b	Effect sizes (β,γ)	.25, .75
—	Sampling distribution	Normal
—	Response type	Continuous
—	Missing values	None
—	Measurement model	Reflective

HEURISTICS AND THE IDENTIFATION OF MODEL MISSPECIFICATIONS

We examine four broad types of misspecifications (Paxton et al., 2001). Table 3 presents a summary of our theoretical expectations and empirical findings for each of the model quality heuristics of Table 1 under the four types of misspecifications. Tables C.1, C.2, and C.3 in Appendix C provide detailed lists of all misspecifications. As Table 3 indicates, many of the proposed model quality criteria are not designed to identify typical misspecifications ("0" in row "Expectation" in Table 3), as they were developed to assess factorial validity (convergent and discriminant). This is reflected in our simulation results ("Finding" in Table 3).

Table 3: Theoretical Expectations and Empirical Findings for Model Quality Heuristics

Legend: **"0" expectation or finding of no difference,**
"+" expectation or finding of positive difference
"-" expectation or finding of negative difference
"~" expectation or finding of complex difference
Parentheses indicate expectations or findings only for some conditions or misspecifications

Shading indicates differences between expectations and findings

	Heuristic		Measurement Misspecification	Added Paths	Removed Paths	Reversed Paths
Measurement Related	1 (Loadings)	Expectation	0	0	0	0
		Finding	(-)	0	(+)	0
	2 (Rel. Cross loadings)	Expectation	-	0	0	0
		Finding	(-)	0	(~)	0
	3 (Abs. Cross loadings)	Expectation	-	0	0	0
		Finding	(-)	0	(-)	(-)
	4 (Reliability α)	Expectation	-	0	0	0
		Finding	(~)	0	(+)	0
	5 (Composite reliability ρ)	Expectation	-	0	0	0
		Finding	0	0	0	(-)
	6 (AVE)	Expectation	-	0	0	0
		Finding	(-)	0	(+)	0
Structure Related	7 (AVE-latent correl.)	Expectation	-	0	-	
		Finding	-	-	~	0
	8 (Root AVE-latent correl.)	Expectation	-	0	-	
		Finding	(-)	0	(-)	0
	9 (r^2)	Expectation	0	0	0	
		Finding	0	0	0	0
	10 (AVE - r^2)	Expectation	-	0	+	
		Finding	(-)	0	(+)	(+)
	11 (Structural paths)	Expectation	0	-	0	
		Finding	0	(-)	0	(-)
Global	12 (Relative GoF)	Expectation	-	-	-	-
		Finding	~	~	~	~
	13 (Predictive relevance Q^2)	Expectation	+	0	0	
		Finding	(+)	0	~	~

Cross-Loading Misspecifications

Measurement misspecifications (cross-loadings) can occur when indicators are related to multiple concepts and affect discriminant and convergent validity. Such misspecifications may stem from poorly designed measurement instruments. Many of the model quality heuristics we examined are designed to identify cross-loading misspecifications. This is not surprising as many heuristics express some aspect of factorial, i.e. convergent or discriminant, validity (Straub et al., 2004). While heuristics 2 and 3 (relative and absolute cross-loadings) should be the most immediate indicators of such misspecifications, other heuristics should be affected as well (Appendices E, F). We found that these heuristics do identify misspecifications, but not with the consistency and reliability required to allow researchers to draw firm conclusions about the model correctness. While heuristic 7, comparing the AVE to latent variable correlations, is most reliably affected by measurement misspecifications, we cannot recommend that researchers rely on this heuristic, as it is also affected, by design, by structural misspecifications. On the other hand, heuristic 2, relative cross-loadings, is least affected by structural misspecifications, but a relatively unreliable indictor of measurement misspecifications.

Removed Paths

The second group of misspecification conditions contains latent paths in the true, generating model but missing in the estimated model. This reflects the fact that researchers may underspecify the model, i.e. fail to include structural paths. In these conditions, the measurement model does not change, so that many of our heuristics should not be affected. Only heuristics that use structural parameters (latent variable correlations and coefficient of determination r^2) should be affected. While heuristic 9, the coefficient of determination r^2 should be the most immediate heuristic to identify these misspecifications, the fact that no firm recommendation can be given, other than the very weak requirement of statistical significance, means this heuristic is of little use in identifying misspecified models. We found that the comparison of AVE to r^2 (heuristic 10) and comparison of the AVE to latent variable correlations (heuristic 7) were

able to identify some, but not all, of the misspecifications. On the other hand, these two heuristics are also affected by measurement misspecification, so that they may be misleading.

We have also found that these types of misspecifications affect measurement-related heuristics, often in a positive way. Thus, the under-specification of theory in the structural model may artificially inflate measurement-related heuristics and thereby mask possible measurement misspecifications.

Added Paths

The third group is conditions in which latent paths are estimated that are not in the true model. Such structural misspecifications can occur when the theoretical foundation is not fully understood. These conditions reflect the fact that researchers may over-specify the model, i.e. include additional paths. These conditions are relatively harmless if the path estimates for the additional paths are non-significant. In these conditions, the measurement model estimates should not change, nor should the structural parameters, e.g. coefficient of determination r^2, change as the additional paths have not corresponding paths in the generating model and should therefore not be statistically significant. Only heuristic 11, which is an immediate indicator of this misspecification type, should be affected. Our findings confirm that this heuristic is able to identify this type of misspecification in most, but not all cases, with most of the problems occurring at low sample size where sampling fluctuations and standard errors around parameter estimates are relatively large.

Reversed Paths

The fourth group are misspecifications that reverse the relationship between latent variables in the estimated model. In some cases, this means that exogenous variables are modeled as endogenous and vice versa. Reversal of paths represents inverted causality and can occur when theory or construct definitions are not well defined or ambiguous. In such cases, researchers might conceivable specify a model that includes paths in the opposite direction as the generating model. Ideally, these misspecifications should

have no effect on measurement-related heuristics but the iterative nature of the PLS estimation (Appendix A) means that such effects might occur. Further, given the complexity of the estimation and the estimated models, we had no expectations of what effects a reversal of paths might have on structure-related heuristics. We observed that many heuristics are affected by these types of misspecifications. However, no criterion provides a consistent way of identifying these and only these types of misspecifications.

In comparison, our covariance-based SEM estimation using the $\chi 2$ statistical test of model fit was able to reliably identify all cross-loading misspecifications, as well as all removed and reversed paths (with the exception of covariance equivalent models, which can be identified prior to estimation). Similar to PLS heuristic 11, covariance-based SEM identified added paths by estimating the path coefficients to be zero, not by changes to model fit. Also similar to heuristic 11, this identification was reliable for medium and large sample sizes, but not for small samples when combined with complex models.

CLASSIFYING PLS MODELS

Predicting the category of a model, true or misspecified, from the model quality heuristics is a classification task, and suitable for statistical classification techniques (Hastie et al., 2009). The outcome is a categorical variable, either with two categories (true model, misspecified model) or with multiple categories corresponding to specific misspecifications of the model. The Weka data mining suite (Hall et al., 2011) provides more than 30 classification algorithms for categorical response variables, including parametric models, classification tree-based methods, decision rule-based methods and Bayesian methods. Because classification algorithms have different performance on different problem sets, we subjected our data to all algorithms in the Weka suite, using cross-validation. Appendix H reports details of the application of classification methods to identify misspecified models using the PLS model quality heuristics.

The overall performance of the classification algorithms on our complete PLS data is poor. This suggests that it is not possible to provide model classification guidelines that are generalizable to all three models

for all model- and sample-conditions based on the model quality heuristics we have examined. Moreover, the performance of even the best algorithms is poor when the sample size is low (s=100) or when structural path coefficients are weak (b=0.25). Finally, while it is possible to predict model misspecification with acceptable degrees of accuracy for medium to large sample sizes and strong structural effects, this must be qualified due to the nature of the classifiers that are built by the software. Tree-based classifiers, which are consistently the best type of classifiers on our PLS data, have a tendency to build large trees with very specific branching rules. The best classification trees often had dozens or hundreds of branches. More interpretable classifiers lead to more parsimonious guidelines for identifying model misspecifications but have unacceptably low classification accuracy.

In summary, applying classification techniques to classify PLS models based on their model quality heuristics shows limited success. The relatively low classification success rate for low sample sizes and weak structural effects, coupled with the highly complex classification rules that are generated make it impractical to provide simple, parsimonious and general guidelines to a researcher using PLS to identify misspecified models. Moreover, while we have built classification rules for all three models, given the complexity and specificity of the classification rules, it is unlikely that these generalize even to relatively similar models and/or sample characteristics.

DISCUSSION AND RECOMMENDATIONS

While PLS does not offer a statistical test of model fit, we have examined to what extent the existing model quality guidelines can be used to identify misspecified models. We have theoretically examined the proposed heuristics with respect to a variety of common model misspecifications and analyzed their behavior using a simulation study. We found that many of the model quality heuristics are not able to identify misspecifications, nor have they been specifically designed to do so. In the few cases where a heuristic was designed and able to identify a certain type of misspecification, its response to misspecification was often unreliable, depending strongly on sample-, model- or misspecification

condition. Our attempt to classify PLS models by their model quality heuristics supports the general inability to reliably identify misspecified models. One exception was the statistical significance of path estimates, which was a reliable indicator for additional paths at medium to large sample size. In contrast, we found that covariance-based estimation is able to reliably identify all misspecification conditions for medium and large sample sizes, with the exception of covariance-equivalent models which can be identified a-priori and are therefore unproblematic.

We discuss the implications of this finding for three different aspects of the research cycle: Theory testing, exploratory theory building, and prediction.

PLS and Theory Testing Research

In the broader perspective of the scientific process, theory testing is linked to the Popperian idea of falsifiability of theories which recognizes that theories cannot in principle be proven correct, only proven wrong (if that). Thus, to advance this scientific process, reliable tests of model correctness are required.

Many of the model quality heuristics have a valuable place in assessing the quality of a model. It is clearly important to establish the reliability and validity of measurement, and the existing heuristics based on estimated factor loadings offer appropriate tools to do this. However, the parameter estimates that are used in these quality assessments are subject to a properly specified model. Thus, in a theory-testing paradigm, which is dominant in quantitative IS research, the primary objective must be to establish the overall correctness of the model. Only then can the parameter estimates be used to assess further qualities the model, such as measurement validity and reliability, or the proportion of explained variance.

The fact that researchers using PLS cannot draw reliable conclusions about the correctness of the estimated model in a theory testing context, leads us to recommend against the use of PLS for theory testing research. Instead, we recommend that researchers use covariance-based methods with medium or large samples to establish model correctness. As a corollary, we also recommend that researchers be

cautious when concluding that their model is the true generating model or that their statistical results support their theoretical claims even when all model quality criteria are satisfied.

PLS and Exploratory Research

Theory testing is but one step in the overall scientific process, where exploration and theory building are equally important. PLS is often argued to be suitable for an exploratory approach that is useful for the early stages of theory development (e.g. Hair et al., 2011, Esposito Vinzi et al., 2010, Chin et al., 2010) and 20% of published PLS studies in MIS Quarterly cite this as a reason for choosing PLS over alternative techniques (Ringle et al., 2012), though only 4 of 32 studies in ISR provided this motivation.

One way to understand exploratory analysis is that it should reveal patterns in the data instead of testing a pre-specified model. It is clear that PLS does not have this ability because the model must be completely specified prior to the analysis. Another way to understand exploration is as part of theory building or theory improvement. However, as Evermann and Tate (2011) have shown, theory building and theory improvement using quantitative data relies on strong model tests. We can only build credible theories and improve our theories if we are able to recognize a mismatch between model and reality. While theory building should proceed primarily based on substantive considerations (Evermann and Tate, 2011), a statistical technique can support this process if it is able to point out specific problems with the tested model. However, our results show that measurement misspecifications often affect structural parameters and quality heuristics, and vice versa. This makes the use of these quality heuristics as diagnostic tools problematic. *We recommend against the use of PLS for exploratory or theory building research. Instead, in line with Evermann and Tate (2011) we recommend the use of covariance-based methods with medium or large samples for exploratory theory building.*

Related to this issue, we also note an apparent disconnect between the motivation of PLS studies, and the way they are carried out. Many studies that cite exploration or theory development as a reason for choosing PLS are in fact presented as theory testing studies: A literature review is followed by the

derivation of causal theory and formal hypotheses, subsequently tested by examining the statistical significance of parameters in the estimated model. The majority of studies estimate a single model, frequently using the terms "model test" or "testing the model" to describe their analysis. This reflects the assessment by Ringle et al. (2012) who lament the lack of analyses such as model comparison, i.e. exploration, using the f^2 metric, finite mixture analyses to identify distinct sub-samples (Ringle et al., 2010), or tetrad analysis for data-based model construction (Scheines et al., 1998). Our own review of PLS studies published in ISR shows a similar lack of such exploratory analyses.

PLS and Predictive Research

Two approaches to prediction can be distinguished (Shmueli and Koppius, 2011). The first approach to prediction is based only on the data ("empirical prediction"). The form of the statistical or computational model underlying the prediction is irrelevant in this case, as long as predictive aims are satisfied. The second approach predicts based on explanatory theory ("theoretical prediction") and builds on knowledge of the real-world causal system.

A recent study compares the predictive ability of PLS and covariance-based estimation techniques (Evermann and Tate, 2012) and concludes that PLS is superior to covariance-based prediction, based on the Q^2 predictive ability metric, especially for small sample sizes ($n = 100$). This advantage mostly disappears for medium to large samples ($n = 250, 750$). However, that study also shows that the proportion of explained variance in endogenous latent variables, r^2, was higher for covariance-based estimation than for PLS. This suggests that reporting of Q^2 rather than endogenous r^2 is important when the use of PLS is justified by appeal to predictive ability of the model.

Prediction is reported as a reason for using PLS by more than 15% of studies published in MIS Quarterly (Ringle et al., 2012) and 2 of 32 of studies published in ISR. In empirical prediction, the isomorphism of the estimated to the real-world, generating model, i.e. model correctness, is largely irrelevant (Hastie et

al., 2009; Shmueli and Koppius, 2011). Further, prediction does not require statistical inference and parameter tests. Hence, it allows typical requirements for these, such as multivariate normality and large sample sizes, to be relaxed, making PLS an attractive option. The earlier section on model classification and Appendix H provide an example of empirical prediction and discuss some of the available statistical techniques. While prediction may be an appropriate area for the use of PLS, we caution that existing guidelines for predictive studies (Shmueli & Koppius, 2011) and the goals and techniques for prediction (Hastie et al., 2009) suggest that researchers cannot simply claim their PLS-based studies to be predictive and expect to follow the same process of data analysis as for theory testing.

Shmueli & Koppius (2011) also point out the tension between explanation and prediction. Predictive models may not be good explanatory models, and vice versa. Similarly, McDonald (1996) notes that "a path model … is already 'explanatory', and generally suboptimally predictive … If the object of the analysis were to predict the response variables … we cannot de better than to use the multivariate regression … or the corresponding canonical variate analysis" (pg. 266). The recent study by Evermann and Tate (2012) also compared PLS prediction to atheoretical prediction with the expectation maximization (EM) method and found that the latter dominated PLS prediction in all experimental conditions. Thus, the practice of using a PLS estimated model as for both explanation and predicting is likely to yield models that are not optimal for either.

We therefore recommend that the widespread practice of conflating theory testing and prediction in a single model should be discontinued. Instead, we recommend for "theoretical prediction" that researchers first establish model correctness using covariance-based techniques, and may subsequently use PLS estimation from that model for predictive purposes. For "empirical prediction" we recommend

that researchers also apply techniques such as expectation maximization, canonical regression, PLS regression[1] *or principal components regression, which do not rely on a structural model.*

Again, we note an apparent disconnect between the stated aims of many PLS-based studies and their execution and presentation. For example, Shmueli and Koppius (2011) report that only 1 of 11 studies in MIS Quarterly that claim prediction as their main uses appropriate predictive analytics while the remaining 10 studies do not. Similarly, Ringle et al. (2012) report that none of the PLS studies they surveyed, including those that claimed prediction as their main goal, report statistics such as the Q^2 degree of predictive relevance. Our own review of PLS studies published in ISR found that two studies motivated the use of PLS by appealing to predictive ability, yet none reported predictive relevance statistics.

The predictive ability of PLS path modeling has to our knowledge not been compared to that of competing techniques, such as CB-SEM or OLS. Especially the known under-estimation of structural path estimates by PLS (Lohmöller, 1989), and thus of the r^2 values for endogenous constructs, suggests that researchers should, at least for the time being, exercise caution when applying PLS path modeling for prediction.

CONCLUSION

This methodological commentary should be understood as a call for more rigorous theory testing using PLS, in the same line as recently argued for covariance-based modelling (Evermann and Tate, 2011); not

[1] The distinction between PLS path modeling and PLS regression is important. PLS regression is widely used in predictive analytics research (Hastie et al., 2009) and has been evaluated against other predictive techniques such as neural networks. However, this is not the case for PLS path modeling, discussed in this paper.

to reject models, but to publish, discuss, improve and learn from them. We acknowledge that this may not be a popular position. After all, if PLS cannot test a model, a researcher never finds themselves in a position where a model is decisively rejected by the evidence. Given the publication bias in many fields to produce "positive" results, a cynic might say that the inability to reject a model is a good reason to stay with the status quo. However, as Evermann and Tate (2011) argue, and we have discussed to above, the ability to identify misspecified models is critical for improving our theories and is the basis for exploratory theory building. In their words, "we believe that there is value in publishing and discussing models that do not fit" (pg. 651). The Information Systems area is also not alone in this: A call for more rigorous theory testing has recently been voiced in organizational sciences where Gray and Cooper (2010) suggest that "our field [organizational science] seems to privilege corroborating over disconfirming evidence" (p.622) and argue that "organizational studies ... would benefit from a stronger focus on theory development via the pursuit of failure" (p.621).

The acknowledged lack of overall model test and our findings of an inability to use existing model quality criteria in lieu of such a test, compel us to *recommend against the use of PLS for theory testing work*. In light of the availability of a reliable test of model fit for covariance-based estimation, this should be the preferred for establishing model correctness in a theory testing context.

While much of the PLS-based work in Information Systems claims to have either exploratory, theory building goals, or predictive goals, recent data on PLS application and predictive analytics (Ringle et al., 2012; Shmueli and Koppius, 2011) show those claims to be frequently lip-service. We urge researchers who wish to use PLS to more strongly focus on appropriate predictive techniques and evaluation methods. In support of this, we urge PLS methods researchers to examine and compare the predictive ability of PLS path modeling with other predictive analytics techniques.

While we do not believe that the Information Systems field, as a scientific discipline, should abandon the goal of explanation, and especially not because of the choice of statistical method, we agree with Shmueli

& Koppius (2011) that there is room for more predictive studies in Information Systems. We urge authors, reviewers, and editors in our field to enter into the discussion of what the right balance for our research field, and the right balance for individual studies, should be.

REFERENCES

Aguierre-Urreta, M.I. and Marakas, G.M. 2008. "The use of PLS when analyzing formative constructs: Theoretical analysis and results from simulations," *Proceedings of the 29th International Conference on Information Systems (ICIS)*, Paris, France, December 2008

Areskoug, B. 1982. "The first canonical correlation: Theoretical PLS analysis and simulation experiments," in *Systems under Indirect Observation: Causality, Structure, Prediction*. K. Joreskog and H. Wold (eds.), Amsterdam, North Holland, 1982.

Cassel, C.M., Hackl, P. and Westlund, A.H. 1999. "Robustness of Partial Least Squares for estimating latent variable quality structures," *Journal of Applied Statistics*, (26**:**4), pp. 435-446.

Chen, W.S. and Hirschheim, R. 2004. "A paradigmatic and methodological examination of Information Systems research from 1991 to 2001," *Information Systems Journal*, (14), pp. 197-235.

Chin, W. W. 1998. "Issues and opinions on structural equation modeling," *MIS Quarterly*, (22:1), pp. vii-xvi.

Chin, W. W. 2010. "How to write up and report PLS analyses," *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang (eds.), Springer Verlag, Heidelberg, Germany, pp. 655-690.

Chin, W.W. and Gopal, A. 1995. "Adoption intention in GSS: Relative importance of beliefs," *DATABASE*, (26:2), pp. 42–64.

Chin, W.W., Marcolin, B.L. and Newsted, P.R. 2003 "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/Adoption Study," *Information Systems Research,* (14:2), pp. 189-217.

Chin, W.W. and Newsted, P.R. 1999. "Structural equation modeling analysis with small samples using partial least squares," *Statistical Strategies for Small Sample Research,* Hoyle, R.A. (ed.), Thousand Oaks, CA: SAGE Publications.

Esposito Vinzi, V., Trinchera, L., and Amato, S. 2010. "PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement," *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang, (eds.), Springer Verlag, Heidelberg, Germany, pp. 47–81.

Evermann, J. and Tate, M. 2011. "Fitting covariance models for theory generation," *Journal of the Association for Information Systems* (12:9), pp. 632-661.

Evermann, J. and Tate, M. 2012. "Comparing the predictive ability of PLS and covariance models," *Proceedings of the 33rd International Conference on Information System (ICIS)*, Orlando, FL, December 2012.

Fornell, C. and Larcker, D. F.1981. "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, (18), pp. 39–50.

Gefen, D. and Straub, D. 2005. "A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example," *Communications of the Association for Information Systems*, (16:5).

Gefen, D., Straub, D. W., and Boudreau, M.-C. 2000. "Structural equation modeling and regression: Guidelines for research practice," *Communications of the Association for Information Systems,* (4:7).

Goetz, O., Liehr-Gobbers, K., and Krafft, M.2010. "Evaluation of structural equation models using the partial least squares (PLS) approach," *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang, (eds.), Springer Verlag, Heidelberg, Germany, pp. 661–711.

Goodhue, D., Lewis, W., and Thompson, R. 2006. "PLS, small sample size and statistical power in MIS research." *Proceedings of the 39th Hawaii International Conference on Systems Science*, pp. 1–10.

Goodhue, D., Lewis, W., and Thompson, R. 2007. "Statistical power in analyzing interaction effects: Questioning the advantage of PLS with product indicators," *Information Systems Research*, (18:2), pp. 211–227.

Goodhue, D., Lewis, W., and Thompson, R. 2012. "Does PLS have advantages for small sample size or non-normal data?" *MIS Quarterly*, forthcoming.

Gray, P. H., and Cooper, W.H. 2010. "Pursuing failure," *Organizational Research Methods*, (13:4), pp. 620-643.

Gregor, S. 2006. "The nature of theory in Information Systems," *MIS Quarterly*, (30:3), pp. 611-642.

Hair, J.F., Ringle, C.M. and Sarstedt, M. 2011. "PLS-SEM: Indeed a silver bullet," *Journal of Marketing Theory and Practice*, (19:2), pp. 139-151.

Hair, J.F., Sarstedt, M., Ringle, C.M. and Mena, J.A. 2012. "An assessment of the use of partial least squares structural equation modeling in marketing research," *Journal of the Academy of Marketing Science*, (40), pp. 414-433.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. "The WEKA data mining software: An update," *SIGKDD Explorations*, (11:1).

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning – Data Mining, Inference, and Prediction.* 2nd edition. Springer Verlag, New York, NY.

Henseler, J., Ringle, C.M., and Sinkovics, R.R. 2009. "The use of partial least squares paths modeling in international marketing," *Advances in International Marketing*, (20), pp. 277-319,

Hulland, J. 1999. "Use of partial least squares (PLS) in strategic management research: A review of four recent studies," *Strategic Management Journal*, (20), pp. 195–204.

Lohmöller, J.-B. 1989. *Latent Variable Path Modeling with Partial Least Squares*. Physica-Verlag, Heidelberg, Germany.

Mattson, S. 1997. "How to generate non-normal data for simulation of structural equation models," *Multivariate Behavioral Research*, (32:4), pp. 355-373.

McDonald, R.P. 1996. "Path analysis with composite variables," *Multivariate Behavioral Research*, (31:2), pp. 239-270.

Paxton, P., Curran, P.J., Bollen, K.A., Kirby, J., and Chen, F. 2001. "Monte Carlo experiments: Design and Implementation," *Structural Equation Modeling*, (8:2), pp. 287-312.

Reinartz, W., Haenlein, M., and Henseler, J. 2009. "An empirical comparison of the efficacy of covariance-based and variance-based SEM," *International Journal of Research in Marketing*, (26), pp. 332–344.

Ringle, C.M., Wende, S., and Will, A. 2010. "Finite mixture partial least squares analysis: Methodology and numerical examples," *Handbook on Partial Least Squares*, V. Esposito Vinzi, W. Chin, J. Henseler, and H. Wang, (eds.), Springer Verlag, Heidelberg, Germany, pp. 195-218..

Ringle, C.M., Sarstedt, C., and Straub, D.W. 2012. "A critical look at the use of PLS-SEM in MIS Quarterly," *MIS Quarterly*, (36:1), pp. iii-xiv.

Rönkö, M. and Ylitalo, J. 2010. "Construct validity in partial least squares path modeling," *Proceedings of the 31st International Conference on Information Systems (ICIS), St. Louis, MS*, December 2010.

Rouse, A.C. and Corbitt, B. 2008. "There's SEM and "SEM": A Critique of the Use of PLS Regression in Information Systems Research", *Proceedings of the 19th Australasian Conference on Information Systems (ACIS), Christchurch, New Zealand,* pp. 845-855.

Scheines, R., Spirtes, P., Glymour, C., Meek C., and Richardson, T. 1998. "The TETRAD project: Constraint based aids to causal model specification," *Multivariate Behavioral Research*, (33:1), pp. 65-117.

Shmueli, G. and Koppius, O. 2011. "Predictive analytics in Information Systems research," *MIS Quarterly*, (35:5), pp. 553-572.

Straub, D., Boudreau, M.-C., and Gefen, D. 2004. "Validation guidelines for IS positivist research," *Communications of the Association for Information Systems*, (13:24).

Tenenhaus, M., Amato, S., and Esposito Vinzi, V. 2004. "A global goodness-of-fit index for PLS structural equation modelling." *Atta della XLII Riunione Scientifica, Bari, Italy, 9–11 June*.

IS MY MODEL RIGHT? MODEL QUALITY AND MODEL MISSPECIFICATION IN PLS – RECOMMENDATIONS FOR IS RESEARCH

APPENDICES

APPENDIX A: PARTIAL LEAST SQUARES PATH MODELING

This appendix briefly illustrates the partial least squares path modeling technique, based on the detailed exposition by Lohmöller (1989). The basic statistical model of PLS is given in Equations A.1 and A.2 where η_j designates latent variables, y_{kj} designates manifest indicator k of latent variable j, π and β represent regression coefficients, and η and ϵ represent residuals.

$$\eta_j = \beta_{j0} + \sum_i \beta_{ji}\,\eta_i + \nu_j \qquad \text{(Equation A.1, inner model)}$$

$$y_{k_j} = \pi_{k_{j_0}} + \pi_{k_j}\eta_j + \epsilon_{k_j} \qquad \text{(Equation A.2, outer model)}$$

However, this is not the model that is estimated. Instead, for estimation, the latent variables η_j are replaced by a weighted linear composite of their indicators:

$$Y_j = \sum_k \omega_{j_k} y_{j_k} \qquad \text{(Equation A.3)}$$

The weights ω_{j_k} are estimated in either of two ways. In mode A estimation, the manifest variables y_{j_k} are regressed on an approximation $\tilde{Y}_j$ of the composite Y_j :

$$y_{k_j} = \tilde{\omega}_{k_j}\tilde{Y}_j + \tilde{\epsilon}_{k_j} \qquad \text{(Equation A.4, mode A estimation)}$$

In mode B estimation, the approximation $\tilde{Y}_j$ of the composite Y_j is regressed on the manifest variables y_{j_k}:

$$\tilde{Y}_j = \sum_{kj} \tilde{\omega}_{k_j} y_{k_j} + \tilde{d}_j \qquad \text{(Equation A.5, mode B estimation)}$$

With this model, the basic PLS algorithm then follows the steps in Table A.1.

Table A.1: The basic PLS algorithm (Lohmöller, 1989)

Stage 1:	Iterative estimation of weights and LV scores. Starting at step #4, repeat steps #1 to #4 until convergence is achieved
#1	Inner weights $v_{ji} = \begin{cases} \text{sign cov}(Y_j, Y_i) & \text{if } Y_j \text{ and } Y_i \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$
#2	Inside approximation $\tilde{Y}_j = \sum_i v_{ji} Y_i$
#3	Outer weights; estimate ω_{k_j} in Equation A.4 (mode A) or Equation A.5(mode B) using ordinary least squares regression
#4	Outside approximation $Y_j = f_j \sum_{kj} \widetilde{\omega}_{k_j} y_{k_j}$
Stage 2:	Estimation of path and loading coefficients by ordinary least squares regression from Equation A.1 (path coefficients) and Equation A.2 (loadings coefficients) where the latent variables η_j are replaced by their estimate Y_j
Stage 3:	Estimation of the means of latent and manifest variables as weighted sums.

APPENDIX B: PRIOR EMPIRICAL STUDIES ON PLS

Despite the increasing popularity of PLS in the IS and wider management literature, and despite the large number of papers offering advice on the use of PLS, there are few empirical studies that examine the properties of PLS estimates.

Most simulation studies investigate the behavior of PLS for correct models under various conditions for comparison with other statistical techniques, rather than investigating the behavior of PLS for misspecified models. Recent studies by Reinartz et al. (2009) and Goodhue et al. (2006; 2012), as well as an early study by Areskoug (1982) compare the statistical power of parameter significance tests of PLS and CB-SEM or regression when estimating the true model for different sample sizes, numbers of indicators and other factors. Studies by Goodhue et al. (2007) and Rönkkö and Ylitalo (2010) compare PLS to OLS regression in terms of power and parameter accuracy of PLS estimates for interaction effects (Goodhue et al., 2007) and correlated measurement errors (Rönkkö and Ylitalo, 2010).

Those comparative studies, as well as a study by Chin and Newsted (1999) on PLS only, have examined a variety of conditions, focusing on the sample size (Areskoug, 1982; Chin and Newsted, 1999; Goodhue et al., 2006, 2012; Reinarts et al., 2009), the number of indicators (Areskoug, 1982; Chin and Newsted, 1999; Goodhue et al., 2012; Reinartz et al., 2009), the size of effects between latent variables (Goodhue et al., 2006), the loadings of indicators on latent variables (Goodhue et al., 2006, 2012; Reinartz et al., 2009), and the distribution of observed values (Reinartz et al., 2009).

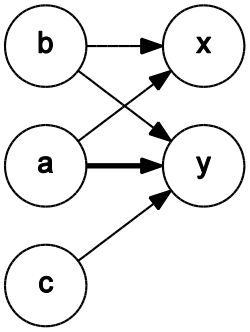
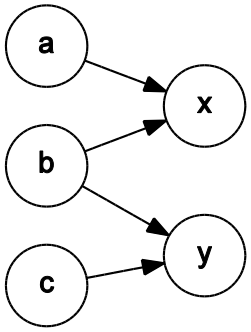
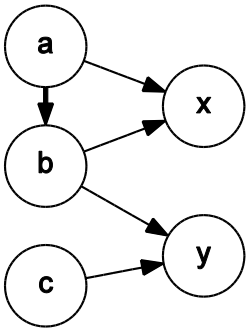
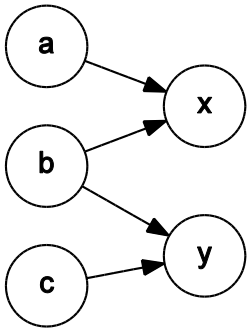
The studies by Areskoug (1982), Goodhue *et al.* (2006, 2007, 2012), Reinarts *et al.*, (2009) and by Chin and Newsted (1999) estimate parameters for the true model from which data was generated. However, in realistic research settings, the true model is unknown and researchers rely on the statistical technique to identify whether their model is a good model, i.e. in close agreement with observations. Only two studies (Cassel et al., 1999; Aguiree-Ureta and Marakas, 2008), examine misspecified models using PLS. Cassel et al. (1999) investigate the robustness of parameter estimates with respect to the skewness of the

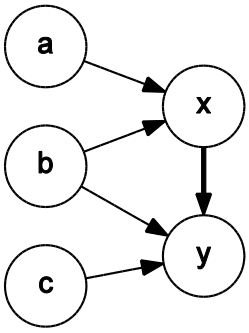
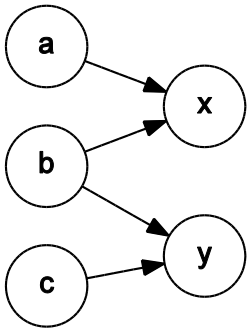
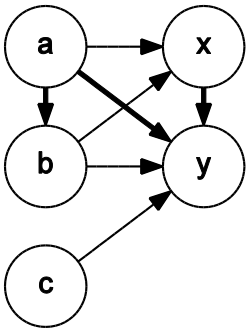
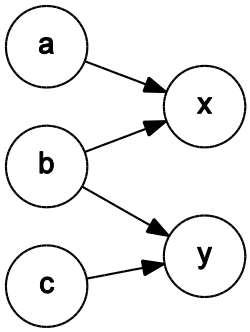
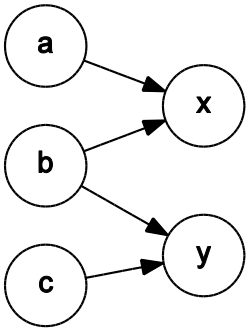
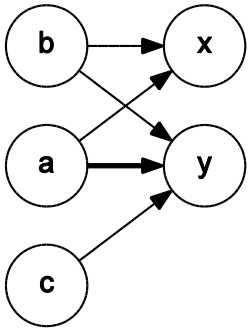
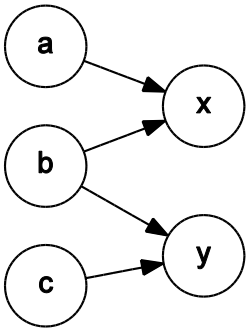
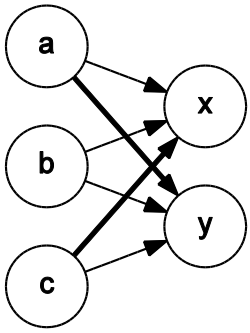
distribution of the observed variables, multi-collinearity among variables, and misspecification of the structural model. They conclude that PLS estimates are quite robust to structural misspecification unless very important regressors are omitted. Their study examined the parameter bias of the misspecified model, but did not examine whether or how misspecifications can be identified in the first place. Their study is also limited to structural misspecification and does not examine measurement model misspecifications such as cross-loadings. Aguierre-Ureta and Marakas (2008) examine misspecification in the PLS context, but limit their study to measurement misspecification of formative or reflective indicators. They do not include structural misspecification or cross-loading misspecification of the measurement indicators.

This brief overview of existing studies on PLS shows that we know very little about the behavior of PLS under misspecification conditions, and the ability of existing model quality heuristics to identify misspecified models.

APPENDIX C: MISSPECIFICATION CONDITIONS

Table C.1: Model Misspecifications for Model 1 (Only the structural model is shown, changes to the base model in bold arrows)

Misspecification Condition		Generating Model	Estimated Model
M1XL1	True model contains one cross-loading from *A* on *C* and from *X* on *Y*	Not shown in the structural model	
M1XL2	True model contains two cross-loadings from *A* on *C* and from *X* on *Y*		
M1XL3	True model contains two cross-loadings from *A* on *X* and from *C* on *Y*		
M1L1	True model contains latent path from *A* to *Y*		
M1L2	True model contains latent path from *A* to *B*		

Misspecification Condition		Generating Model	Estimated Model
M1L3	True model contains latent path from *X* to *Y*		
M1L4	Combination of M1L1 and M1L3		
M1L5	Estimated model contains latent path from *A* to *Y*		
M1L6	Estimated model contains latent path from *A* to *Y* and *C* to *X*		

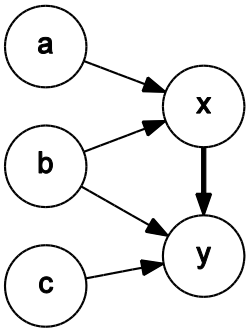
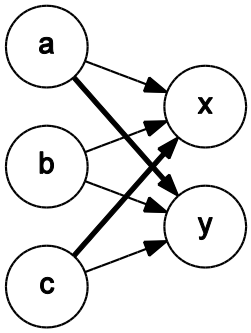
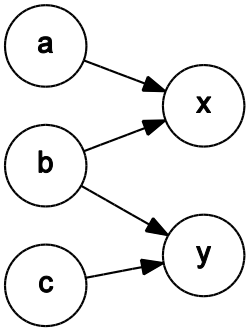
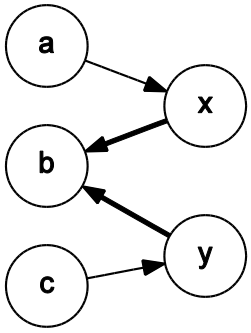
Misspecification Condition		Generating Model	Estimated Model
M1L7	Combination of M1L3 and M1L6		
M1R1	Estimated model reverses paths between *B* and *X* and *B* and *Y*		

Table C.2: Model Misspecifications for Model 2 (Only the structural model is shown, changes to the base model in bold arrows)

Misspecification Condition		Generating Model	Estimated Model
M2XL1	True model contains one cross-loading from *W* on *X* and from *Y* on *Z*	Not shown in the structural model	
M2XL2	True model contains one cross-loading from *A* on *X* and from *A* on *Z*		

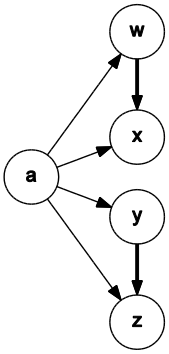
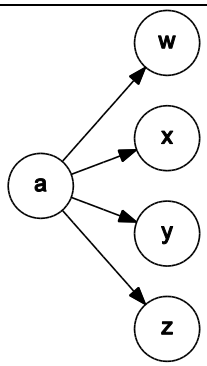
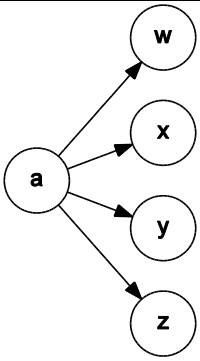
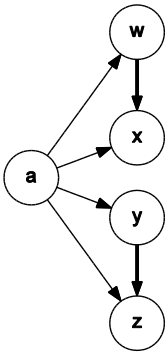
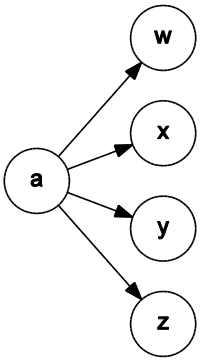
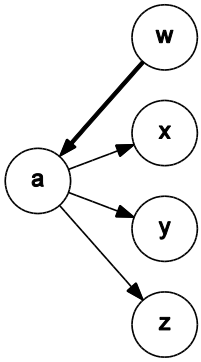
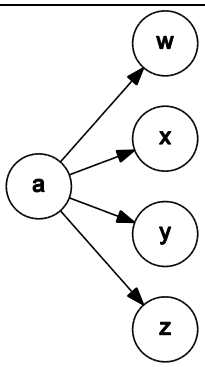
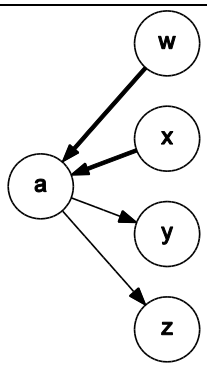
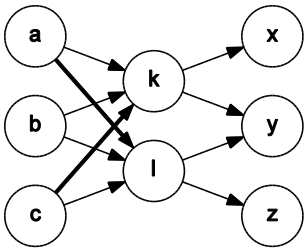
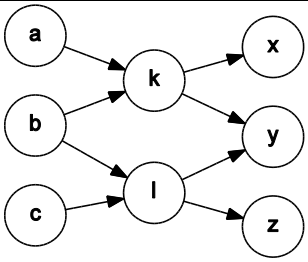
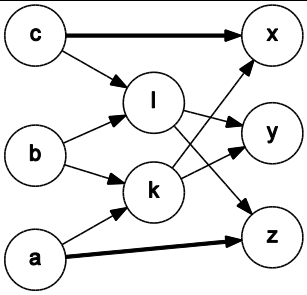
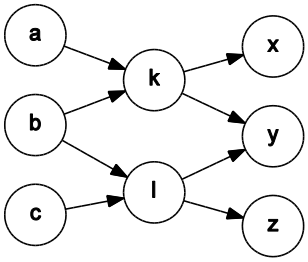
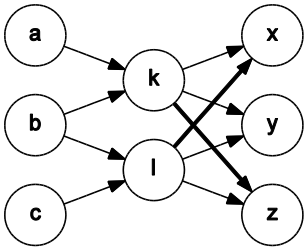
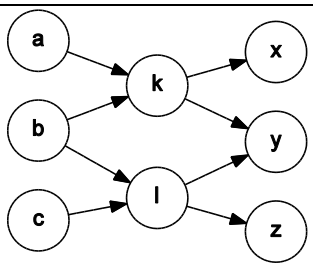
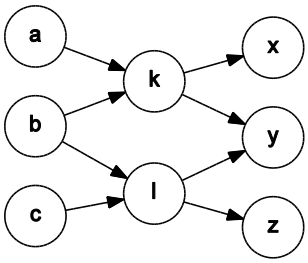
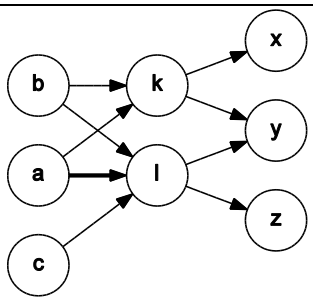
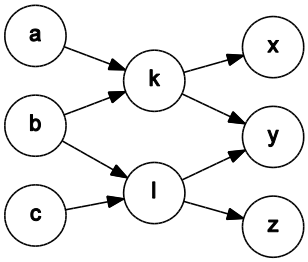
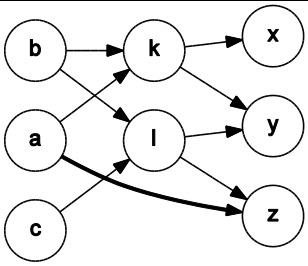
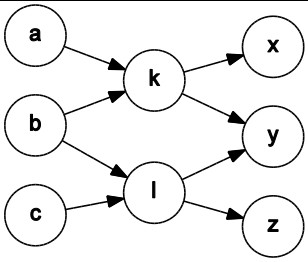
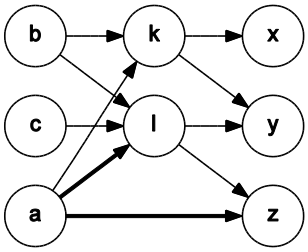
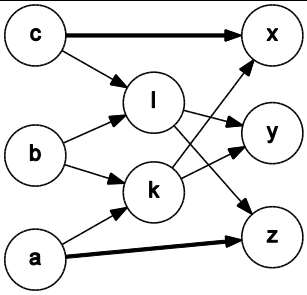
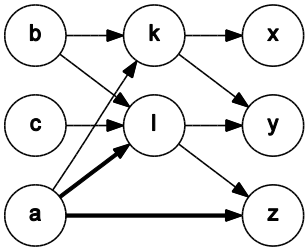
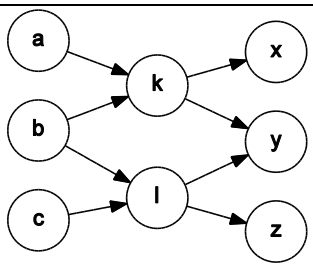
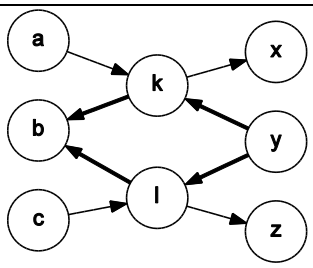
Misspecification Condition		Generating Model	Estimated Model
M2L1	True model contains latent path from *W* to *X* and *Y* to *Z*		
M2L2	Estimated model contains latent path from *W* to *X* and *Y* to *Z*		
M2R1	Estimated model reverses path between *A* and *W*		
M2R2	Estimated model reverses paths between *A* and *W* and *A* and *X*		

Table C.3: Model Misspecifications for Model 3 (Only the structural model is shown, changes to the base model in bold arrows)

Misspecification Condition		Generating Model	Estimated Model
M3XL1	True model contains one cross-loading from *A* on *C* and from *X* on *Y*	Not shown in structural model	
M3XL2	True model contains two cross-loadings from *A* on *C* and from *X* on *Y*		
M3XL3	True model contains two cross-loadings from *A* on *X* and from *C* on *Z*		
M3L1	True model contains latent paths from *A* to *L* and *C* to *K*		
M3L2	True model contains latent paths from *A* to *Z* and *C* to *X*		
M3L3	True model contains latent paths from *K* to *Z* and *L* to *X*		

Misspecification Condition		Generating Model	Estimated Model
M3L4	Estimated model contains latent path from *A* to *L*		
M3L5	Estimated model contains latent path from *A* to *Z*		
M3L6	Combination of M3L4 and M3L5		
M3L7	Combination of M3L3 and M3L6		
M3R1	Estimated model reverses paths from *B* to *K*, *B* to *L*, *K* to *Y*, and *I* to *Y*		

APPENDIX D: DATA GENERATION AND MODEL ESTIMATION

Data generation can be done either from a model-implied covariance matrix or by generating values for exogenous latent variables, and using the structural and measurement relationships with known parameters to create values for endogenous latent and manifest variables (Mattson, 1997). We followed the first approach to generate 200 samples from the true model for each experimental condition.

Using that data, we estimated 3×3×3×2=54 experimental conditions for each of the 12+7+12=31 true model and misspecification conditions, each with 200 bootstrap resamples for parameter significance tests. In total, we performed more than 67,294,800 PLS estimations. We used the R statistical system (version 2.10), with the `plspm` (version 0.1-4) and `lavaan` (version 0.31) packages. Using a cluster of four eight-processor virtual machines on Amazon's EC2 compute "cloud", data generation, model estimation, and collection of results took approximately 2 weeks.

APPENDIX E: RELIABILITY UNDER CROSSLOADINGS

For a model with a latent variable X_1 with two indicators y_1, y_2 we can write:

$$cov(y_1, y_2) = E[\ (\lambda_1 X_1 + \epsilon_1)(\lambda_2 X_1 + \epsilon_2)] = \ \lambda_1 \lambda_2 var(X_1)$$

We have assumed for simplicity that errors do not covary with latent variables or each other.

$$var(y_1) = \ \lambda_1^2 var(X_1) + var(\epsilon_1)$$

$$var(y_2) = \ \lambda_2^2 var(X_1) + var(\epsilon_2)$$

So that the correlation can be written as

$$cor(y_1, y_2) = \frac{\lambda_1 \lambda_2 var(X_1)}{\sqrt{\lambda_1^2 var(X_1) + var(\epsilon_1)}\sqrt{\lambda_2^2 var(X_1) + var(\epsilon_2)}}$$

For the case of equal loadings, equal error variances, and unit latent variance this simplifies to

$$cor(y_1, y_2) = \ \frac{\lambda^2}{\lambda^2 + var(\epsilon)}$$

The variance of the sum can be written as

$$var_{sum} = E[(\lambda_1 X_1 + \ \epsilon_1 + \lambda_2 X_1 + \ \epsilon_2)(\lambda_1 X_1 + \ \epsilon_1 + \ \lambda_2 X_1 + \ \epsilon_2)]$$

After some arithmetic and assumptions that errors do not covary with latent variables and each other, this can be simplified to

$$var_{sum} = \lambda_1^2 var(X_1) + \ \lambda_2^2 var(X_1) + 2\ \lambda_1 \lambda_2 var(X_1) + var(\epsilon_1) + var(\epsilon_2)$$

For the case of equal loadings, equal error variances, and unit latent variance this simplifies to

$$var_{sum} = 4\lambda^2 + 2\ var(\epsilon)$$

So that

$$\alpha = \frac{2}{2-1}\frac{2\,cor(y_1, y_2)}{var_{sum}} = \frac{4\,\lambda^2}{[\lambda^2 + var(\epsilon)][4\lambda^2 + 2var(\epsilon)]}$$

Now consider a cross-loading of latent X_2 on indicator y_1 with strength λ_3. With this cross-loading, the expression for y_1 changes to

$$y_1 = \lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2$$

So that, assuming uncorrelated errors, the variance, covariance, and correlation change to

$$var(y_1) = E[(\lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2)(\lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2)]$$

$$var(y_1) = \lambda_1 var(X_1) + \lambda_3 var(X_2) + 2\,\lambda_1\lambda_3\phi + var(\epsilon_1)$$

$$cov(y_1, y_2) = E[(\lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2)(\lambda_2 X_1 + \epsilon_2)] = \lambda_1\lambda_2 var(X_1) + \lambda_1\lambda_3\phi$$

$$cor(y_1, y_2) = \frac{\lambda_1\lambda_2 var(X_1) + \lambda_1\lambda_3\phi}{\sqrt{\lambda_2^2 var(X_1) + var(\epsilon_2)}\sqrt{\lambda_1 var(X_1) + \lambda_3 var(X_2) + 2\,\lambda_1\lambda_3\phi + var(\epsilon_1)}}$$

Similarly, the expression for the variance of the sum of the scale changes to

$$var_{sum} = E[(\lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2 + \lambda_2 X_1 + \epsilon_2)(\lambda_1 X_1 + \epsilon_1 + \lambda_3 X_2 + \lambda_2 X_1 + \epsilon_2)]$$

After some arithmetic and assumptions that errors do not covary with latent variables and each other, this becomes

$$var_{sum} = \lambda_1^2 var(X_1) + \lambda_2^2 var(X_1) + 2\,\lambda_1\lambda_2 var(X_1) + 2\,\lambda_1\lambda_3\phi + 2\,\lambda_2\lambda_3\phi + \lambda_3^2 var(X_2) + var(\epsilon_1) + var(\epsilon_2)$$

For the case of equal loadings, equal error variances, and unit latent variance these expressions simplify to

$$cor(y_1, y_2) = \frac{\lambda^2 + \lambda^2\phi}{\sqrt{\lambda^2 + var(\epsilon)}\sqrt{2\lambda^2 + 2\lambda^2\phi + var(\epsilon)}}$$

$$var_{sum} = 5\,\lambda^2 + 4\,\lambda^2\phi + 2\,var(\epsilon)$$

So that the reliability becomes

$$\alpha = \frac{2}{2-1}\frac{2\,(\lambda^2 + \lambda^2\phi)}{(5\,\lambda^2 + 4\,\lambda^2\phi + 2\,var(\epsilon))\sqrt{\lambda^2 + var(\epsilon)}\sqrt{2\lambda^2 + 2\lambda^2\phi + var(\epsilon)}}$$

For comparison purposes, Figure E.1 shows a plot of reliability with and without cross-loadings for a range of error variances and unit loadings.

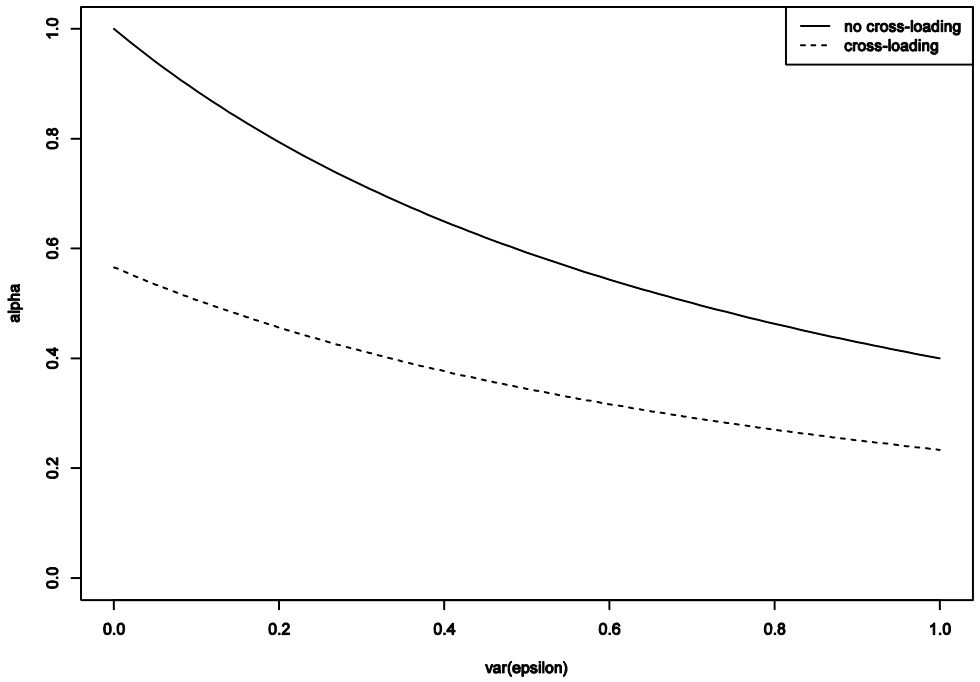


Figure E.1: Reliability under cross-loading conditions

APPENDIX F: CROSS-LOADINGS AND ERROR VARIANCES

The variance of indicator x of a latent variable X can be written as

$$var(x) = \lambda_1^2 \, var(X_1) + var(\epsilon) \qquad \text{(Equation F.1)}$$

Where λ is the regression coefficient and ϵ the measurement error of the indicator. Rewriting this, the measurement error estimate is

$$var(\epsilon) = var(x) - \lambda_1^2 var(X_1) \qquad \text{(Equation F.2)}$$

A cross-loading in the generating model inflates the indicator variance:

$$var(x') = \lambda_1^2 \, var(X_1) + \lambda_2^2 \, var(X_2) + var(\epsilon) \geq var(x) \qquad \text{(Equation F.3)}$$

When the estimated model is misspecified and omits the influence of X_2 on indicator x, the estimation equation is Equation F.1, not Equation F.3. Hence, the two right-most terms in Equation F.3 will comprise the error variance estimate (ϵ') :

$$var(\epsilon') = \lambda_2^2 var(X_2) + var(\epsilon) \geq var(\epsilon) \qquad \text{(Equation F.4)}$$

APPENDIX G: DETAILED DISCUSSION OF HEURISTICS AND RESULTS

For each model quality criterion in Table 1, we compared the mean of the criterion for the 200 samples for the generating model to the mean of the criterion for the 200 samples for the estimated (true or misspecified) model. The following subsections present our expectations and findings for each model quality heuristic.

The complex, iterative nature of the PLS parameter estimation (Appendix A) results in an interdependence of measurement and structural parameter estimates, and also precludes the analytical derivation of many expectations about the effect of misspecifications on various heuristics. Hence, our expectations in Table 3 are partially based on the intent of the heuristic and in that sense represent an "ideal" expectation. For example, the cross-loading heuristic 2 is intended to address measurement model properties, and we would therefore have an "ideal" expectation of no change for structural misspecifications. However, in practice, the interdependence of parameter estimates means that it is possible that cross-loadings are affected by changes to the structural model.

A brief note on terminology: We designate a complete decision rule such as "AVE must be greater than 0.5 for a good model" as a *heuristic*; within a heuristic, we designate the value to be considered, e.g. "AVE" as the *criterion*.

Indicator Loadings (absolute) (Heuristic 1)

Low measurement loadings indicate unreliable indicators, i.e. those with high error variance. This criterion should not be susceptible to the structural misspecifications introduced here. It should also not be susceptible to the cross-loading misspecifications introduced here, as cross-loadings that were added to the generating model had the same strength as the intended loadings. They should not have lowered the intended indicator loadings.

This heuristic is operationalized as the proportion of indicators with loadings greater than 0.7 and should be approximately 1 for all true models. We found that this proportion was as low as 0.80 for true models, especially for the complex model 3 when sample size was low and structural effects were weak (s=100, b=0.25).

Our data showed a decrease in the proportion of loadings satisfying the heuristic for all cross-loading conditions, but only for weak structural relationships (b=0.25) and with a stronger effect for low loadings (l=0.75). The decrease in the proportion of indicators satisfying the heuristic depends on the model, as the models contain different numbers of indicators (Figure G.1).

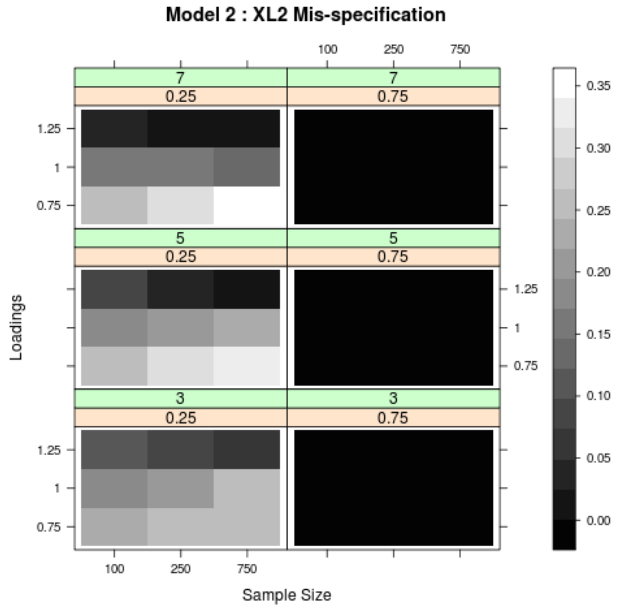


Figure G.1: Indicator loadings (absolute) criterion (criterion 1) decrease for M2XL2

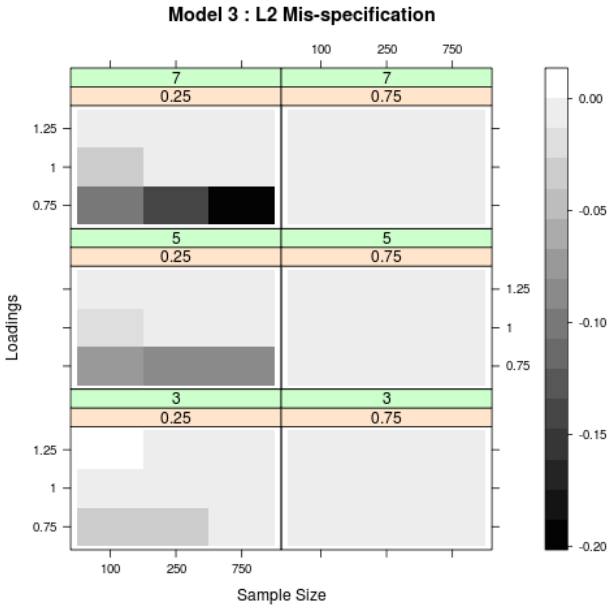


Figure G.2: Indicator loadings (absolute) criterion (criterion 1) decrease for M3L2

(note the negative scale extent)

Contrary to our expectations, the criterion was affected by structural misspecifications that removed paths in the estimated model. Conditions M2L1, M3L1, M3L2, M3L3 and M3L7 showed a significant *increase*

(up to 15%) in the proportion of indicators satisfying the heuristic, again only for weak structural relationship and a stronger effect for low loadings (Figure G.2).

In summary, this heuristic is neither designed nor able to reliably identify common model misspecifications.

Cross-Loadings (relative) (Heuristic 2)

This heuristic was designed to identify the type of cross-loadings introduced here. The heuristic was operationalized as the average proportion of indicator cross-loadings that are more than 0.1 below the loading on the intended construct. For the true models, this proportion should be 1. This was the case for M1, but not for M2 and M3, where the proportion was as low as 0.7 when structural effects were strong (b=0.75), independent of sample size.

We observed that all except two (M1XL1, M1XL2) cross-loading conditions had a significant effect, which was more pronounced when structural coefficients were high (b=0.75) (Figures G.3, G.4). However, this criterion was also susceptible to structural misspecification in conditions M1L4, M2L1, M3L1, M3L3, and M3L7 (decreases) and also in condition M3L2 (improvement) when structural effects were high (b=0.75) (Figure G.5).

In summary, this heuristic correctly identified most conditions it was designed for, but it also incorrectly and misleadingly identified many structural misspecifications as measurement problems.

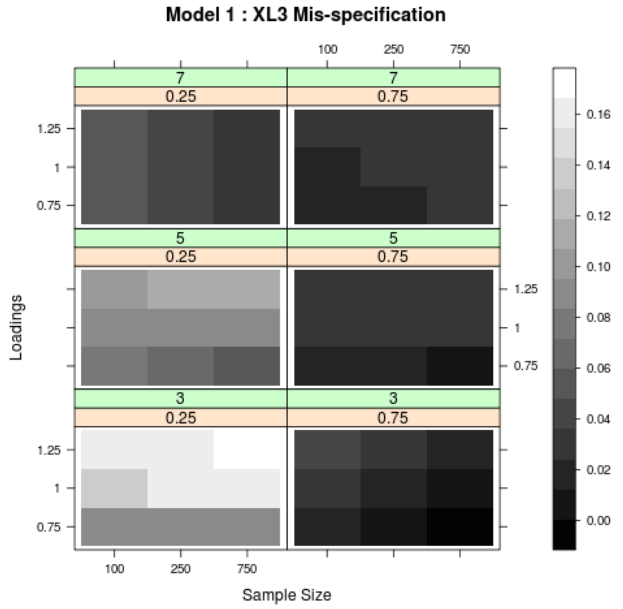


Figure G.3: Cross-loadings (relative) criterion (criterion 2) decrease for M1XL3

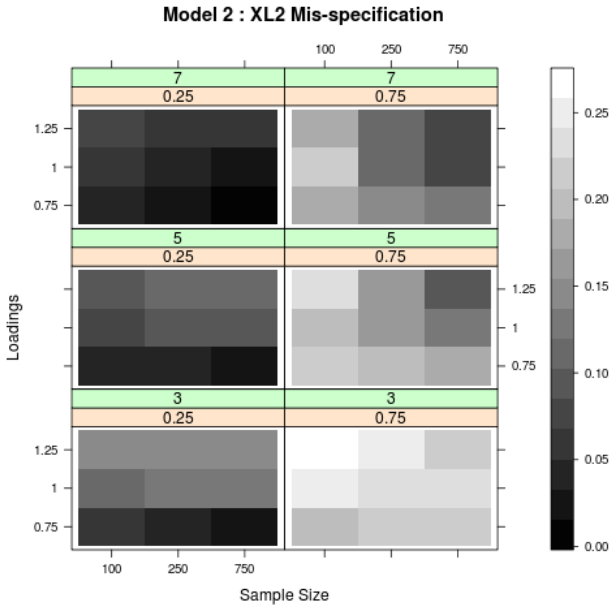


Figure G.4: Cross-loadings (relative) criterion (criterion 2) decrease for M2XL2

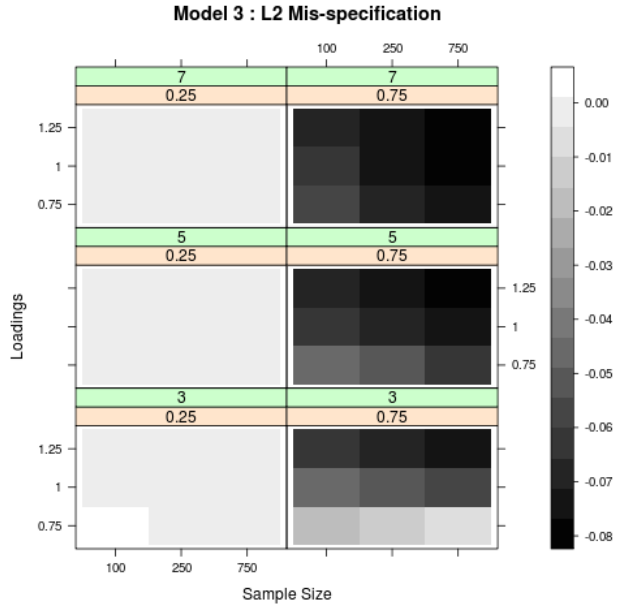
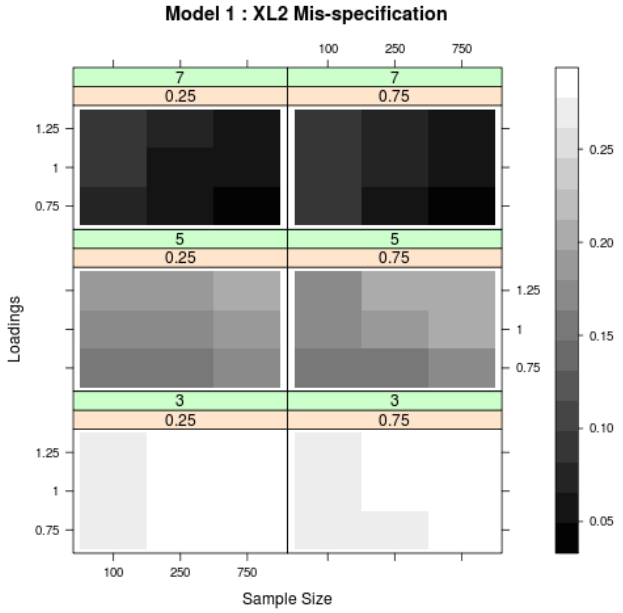


Figure G.5: Cross-loadings (relative) criterion (criterion 2) decrease for M3L2

Cross-Loadings (absolute) (Heuristic 3)

This heuristic was also designed to identify the cross-loading misspecifications in this study. It was operationalized as the average proportion of indicator cross-loadings less than 0.4. For the true models, this proportion should be, and was in fact 1. We expected the same effects on this heuristic as for the previous. The criterion behaved in the expected direction for M1XL1, M1XL2, M3XL1, M3XL2 but conditions M1XL3 and M3XL3 had no effect on this criterion (Figure G.6). However, this heuristic misleadingly affected by the structural misspecifications in conditions M1L4 and M3R1 when structural coefficients were high (b=0.75) and decreasing with sample size (s) (Figures G.7, G.8).



FigureG.6: Cross-loadings (absolute) criterion (criterion 3) decrease for M1XL2

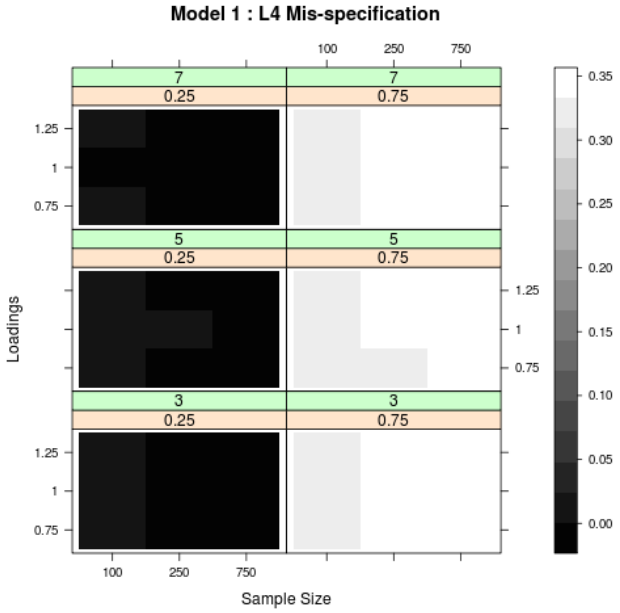


Figure G.7: Cross-loadings (absolute) criterion (criterion 3) decrease for M1L4

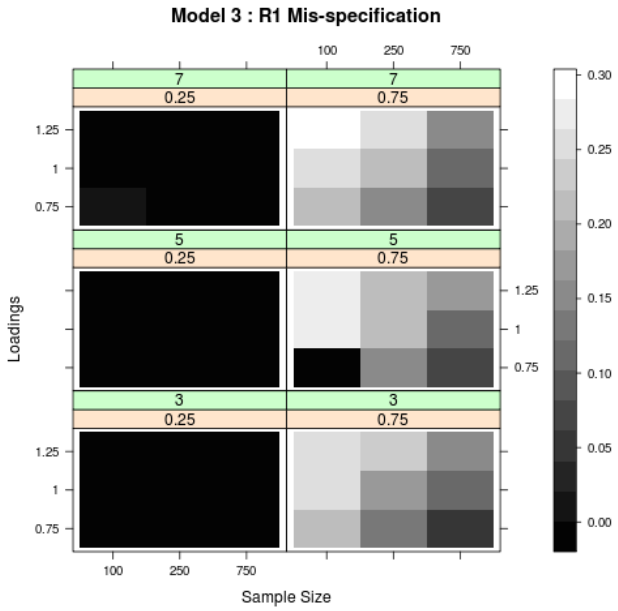


Figure G.8: Cross-loadings (absolute) criterion (criterion 3) decrease for M3R1

Similar to the previous criterion, this cross-loading criterion behaved as expected for some but not all conditions that it was intended to identify, and it was, contrary to its design, also susceptible to misleadingly identifying some structural misspecifications as measurement problems. In contrast to the previous criterion, this criterion appears more robust to variations in sample size, loading and structural effects. It also provided a more reliable baseline for the true models.

Reliability (Cronbach's α) (Heuristic 4)

Reliability is "the extent to which measurements are repeatable" (Nunnally, 1967). Reflective measurement items are intended to be semantically equivalent and can thus be thought of as repeated measures of the same concept. Cronbach's α expresses the idea that repeated measures yield the same value, i.e. are reliable, if they are highly correlated. Consequently, α is defined in terms of the correlations among the indicators for a latent variable (Peter, 1979):

$$\propto = \frac{p}{p-1}\frac{2\sum_{i,j;i<j} r_{ij}}{\frac{n-1}{n}\sigma_{sum}^2}$$

Here, p is the number of indicators for the latent variable, r_{ij} are the correlations among the indicators, n is the sample size and σ_{sum}^2 is the variance of the sum of the indicators. The numerator expresses the covariance among the items, while the denominator expresses the total sample variance of the summed scale.

The definition of this reliability measure shows that at best it is designed to identify measurement misspecifications. In the presence of cross-loadings, this reliability measure is expected to decrease (Appendix E), whereas structural misspecifications should have no effect as they do not affect the correlations among items.

This heuristic was operationalized as the proportion of latent variables with a Cronbach's α reliability greater than or equal to 0.7. We expected this to be 1 for all true models. However, for true models with few indicators and low loadings (l=0.75) this criterion ranged between 0.8 and 0.95, indicating that some latent variables in the true models did not meet the reliability requirement. In these conditions, we observed *improvements* of this criterion for removed path structural misspecification conditions and some measurement misspecifications (Figure G.9).

In contrast to the decision heuristic, the mean α value behaved as theoretically expected. They ranged between 0.77 and 0.90 for the true models and decreased for some measurement misspecifications (M1XL1, M1XL2, M3XL1, and M3XL2), though not to the extent that they dropped below the recommended 0.7 value. In contrast to our expectations, we observed *improvements* to the criterion for many structural misspecification conditions where paths are removed in the estimated model (M1L1, M1L2, M1L3, M1L4, M1L7, M2L1, M2R2, M3L1, M3L2, M3L3, and M3L7) (Figure G.10).

In summary, while Cronbach's α is affected as expected, the threshold of 0.7 for this heuristic is too low to identify measurement misspecifications. The variation of α for true models suggests that this makes a poor baseline and model and condition specific thresholds should be identified. The improvement of α for removed path misspecifications raises the possibility that the positive effect of a structural misspecification can mask the negative effect of a measurement misspecification and suggest a good model when in fact the model contains two types of misspecifications.

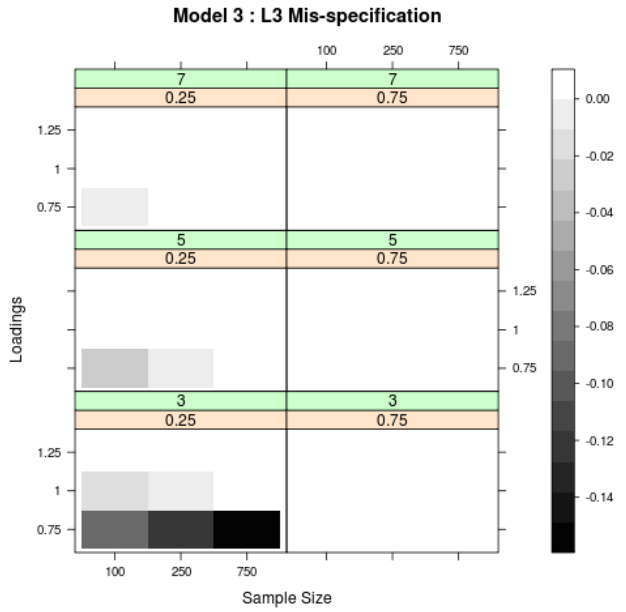


Figure G.9: Reliability alpha criterion (criterion 4) decrease for M3L3

(note the negative scale extent)

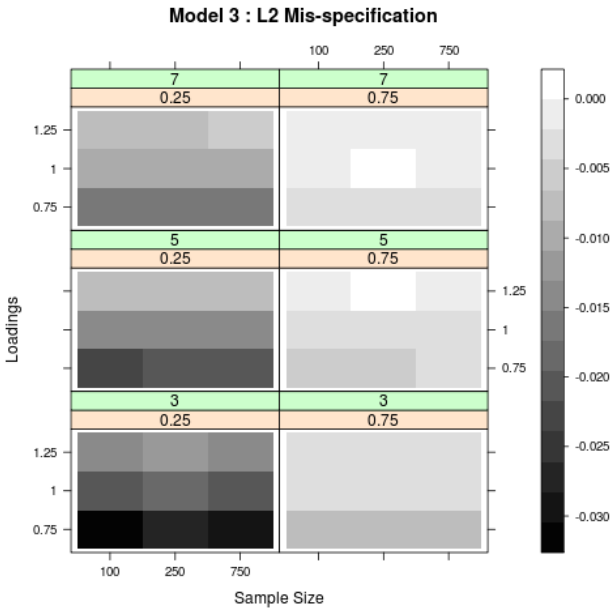


Figure G.10: Mean reliability alpha decrease for M3L2

(note the negative scale extent)

Composite Reliability (ρ) (Heuristic 5)

In contrast to Cronbach's α, the composite reliability ρ is defined not in terms of indicator correlations, but in terms of factor loading estimates:

$$\rho = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2 + \sum_i var(\epsilon_i)} = \frac{\sum_i \sum_{j \neq i} \lambda_i \lambda_j + \sum_i \lambda_i^2}{\sum_i \sum_{j \neq i} \lambda_i \lambda_j + \sum_i \lambda_i^2 + \sum_i var(\epsilon_i)}$$

The denominator depends on the estimated error variance. In cross-loading measurement misspecifications, the estimated error variance is inflated to include that part of the indicator variance that is contributed by the cross-loading factor, but which is not explicitly captured by a regression path in the estimated model (Appendix F). Hence, composite reliability is expected to decrease for measurement misspecifications.

Our findings showed that there were no latent variables whose composite reliability was less than the threshold of 0.7 in any of the misspecified models. The mean ρ was 0.95 or greater for the true models. We observed a decrease of composite reliability for some reversal of path conditions (M2R1, M2R2). Other misspecification conditions had no effect. In summary, this heuristic is unable to identify model misspecifications.

AVE-Absolute (Heuristic 6)

The average variance extracted (AVE) is also defined in terms of factor loading estimates and expresses the average ratio of indicator variance due to the underlying factor over the total indicator variance:

$$AVE = \frac{\sum_i {\lambda_i}^2}{\sum_i \lambda_i^2 + \sum_i var(\epsilon_i)}$$

As with composite reliability, the denominator depends on the estimated error variance. In cross-loading measurement misspecifications, the estimated error variance is inflated to include that part of the indicator variance that is contributed by the cross-loading factor but which is not explicitly captured by a regression path in the estimated model (Appendix F). Hence, AVE is expected to decrease for measurement misspecifications but it should not be affected by any of the other misspecifications in this study.

We have operationalized this heuristic as the average proportion of latent variables that satisfy the 0.5 threshold value. This proportion was 1 for true models, except when loadings and structural paths were weak (l=0.75, b=0.25) where this proportion dropped to approx. 0.8 – 0.95.

As expected, this criterion was affected by only very few misspecification conditions. It decreased for some, but not all measurement misspecifications. Contrary to expectations, the criterion improved for some structural misspecification conditions that remove paths from the estimated model (M2L1, M3L1, M3L2, M3L3, M3L7) (Figure G.11) and it decreased for some, but not all measurement misspecifications (M2XL1, M2XL2, M3XL1, M3XL2, M3XL3 when l=0.75, b=0.25) (Figure G.12). Similar to the reliability heuristic, this is not a reliable heuristic for identifying the intended misspecifications and may misleadingly identify structural model problems as measurement issues, possibly masking measurement problems by structural problems.

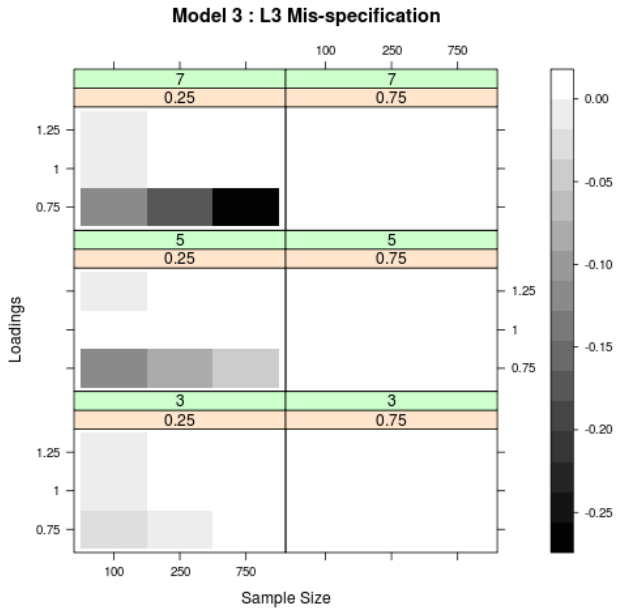


Figure G.11: AVE (absolute) criterion (criterion 6) decrease for M3L3

(note the negative scale extent)

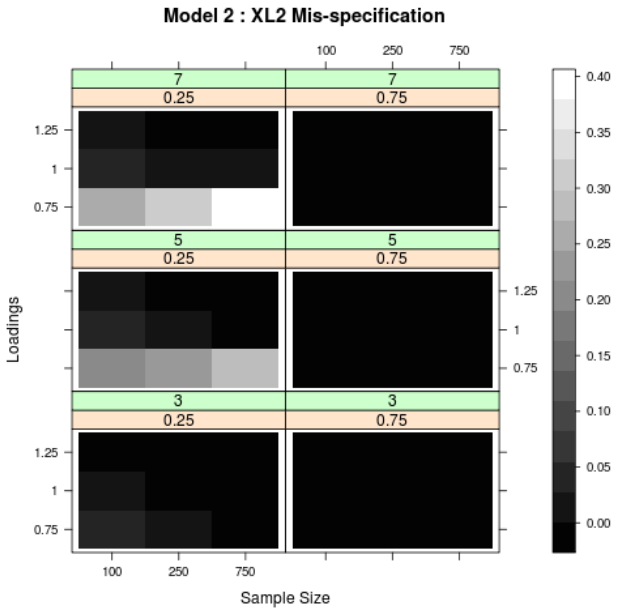


Figure G.12: AVE (absolute) criterion (criterion 6) decrease for M2XL2

(note the negative scale extent)

(Root) AVE-Latent Correlations (Heuristics 7 and 8)

This heuristic encompasses properties of both the structural and the measurement model. It is operationalized as that proportion of latent variables whose AVE exceeds all of its correlations with other

latent variables. This proportion was 1 for the true M1, but was lower for the true M2 and M3 for weak loadings and strong structural effects (as low as 0.3 for M2, as low as 0.95 for M3).

Because the AVE decreases for cross-loading conditions, this criterion should be affected similarly. While the complexity of the models and structural misspecifications precludes a precise prediction, this heuristic may be affected by structural misspecifications as these affect the latent variable correlations.

Our findings confirm these expectations for most misspecification conditions, but with an effect size that strongly depends on the model, type of misspecification, and strength of structural coefficients. We observed decreases with a small effect size (up to 3% change) for M1L4, M1XL1, M1XL2, and M3XL2, and with modest effect size (up to 30% change) for M1XL3, M2L1, M2L2, M3L2, and M3XL3, and with strong effects (up to 90% change) for M2XL1 and M2XL2. The models M3L2, M3L3, and M3L7 showed improvements for this heuristic for some conditions and decreases in others (Figures G.13, G.14).

In summary, this heuristic tends to identify true models with weak measurement and strong structural effects as poor. It is able to identify misspecifications only when structural effects are strong. This heuristic is unsuitable for both weak structural effects (it will not identify misspecified models) and strong structural effects (it will mistakenly reject true models).

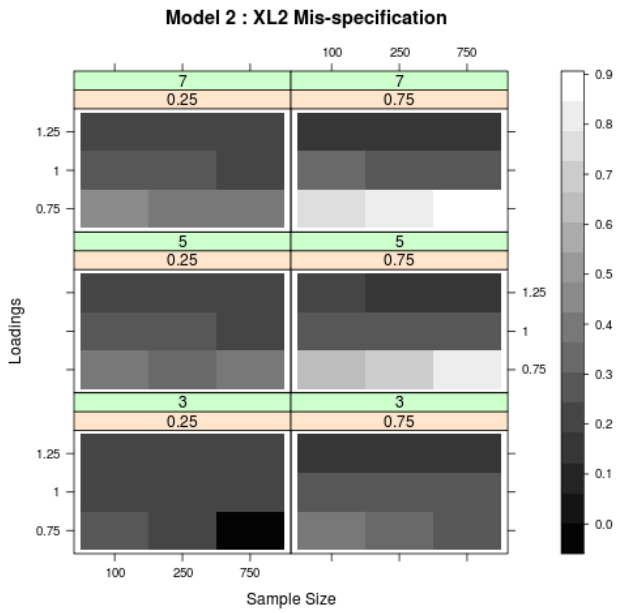


Figure G.13: AVE-Latent correlation criterion (criterion 10) decrease for M2XL2

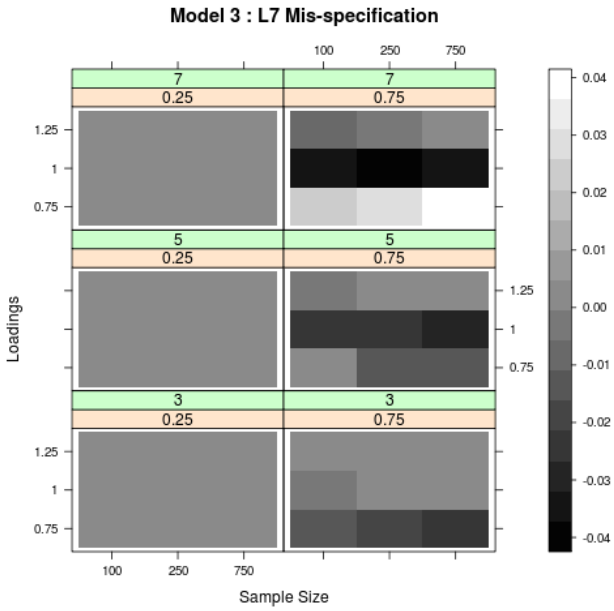


Figure G.13: AVE-Latent correlation criterion (criterion 10) decrease for M3L7

(note the negative scale extent)

Heuristic 8 requires that the square root of the AVE should be greater than the latent variable correlations and is a relaxation of heuristic 7. Thus, we expect the effects to be generally less pronounced than for heuristic 7, and our findings confirm that this heuristic is too weak to be able to reliably identify misspecification conditions.

Significant R2 (Heuristic 9)

As in regression models, the coefficient of determination for endogenous latent variables r^2 should be significant, but "acceptable levels depend on research context" (Hair et al., 2012). However, the mere statistical significance of the coefficient of determination is a low threshold, as none of the misspecifications introduced here, especially the number of removed paths, are sufficient to reduce the r^2 values to insignificant levels. Consequently, our findings indicate no effect of any misspecification

conditions on the statistical significance of the coefficient of determination. We conclude that statistical significance of the coefficient of determination is not a sufficiently strict heuristic.

AVE-R2 (Heuristic 10)

This heuristic combines the two previously discussed ones; it includes aspects of the measurement model, the AVE, and of the structural model, the coefficient of determination r^2. We operationalized this heuristic as the proportion of endogenous latent variables whose AVE is greater than its coefficient of determination r^2.

For all three true models, this proportion was 1, except when loadings were low and structural effects small (l=b=0.75) when the proportion dropped to approx. 0.8.

Given our expectations for the AVE and the coefficient of determination, we expect this heuristic to identify measurement misspecifications, because the AVE should drop while the coefficient of determination should not. Because the AVE should be unaffected by structural misspecifications, we expect this heuristic to be unable to identify added paths, as these do not correspond to paths in the generating model and therefore do not lead to increases in the r^2 value. In contrast, there should be a positive effect for removed paths, as the r^2 value is expected to decrease.

Our findings confirmed the positive effect for almost all conditions where paths are removed, and many conditions where paths are reversed: The criterion improved for conditions M1L1, M1L2, M1L3, M1L4, M1L7, M1R1, M2L1, M2R1, M3L2, M3L3, M3L7, and M3R1 when l=b=0.75 (Figure G.15).

We also found that this criterion decreased for many, but not all measurement misspecifications: The criterion decreased for conditions M2XL1, M2XL2, M3XL1, M3XL2, and M3XL3 when b=0.75 and loadings are low (l=0.75 or l=1.0), for M1XL1, M1XL2 when b=0.75 and for M1XL3 under all conditions (Figure G.14).

In general, this appears to be a useful heuristic but only when loadings were low, relative to structural effects. This lack of reliability across experimental conditions is problematic.

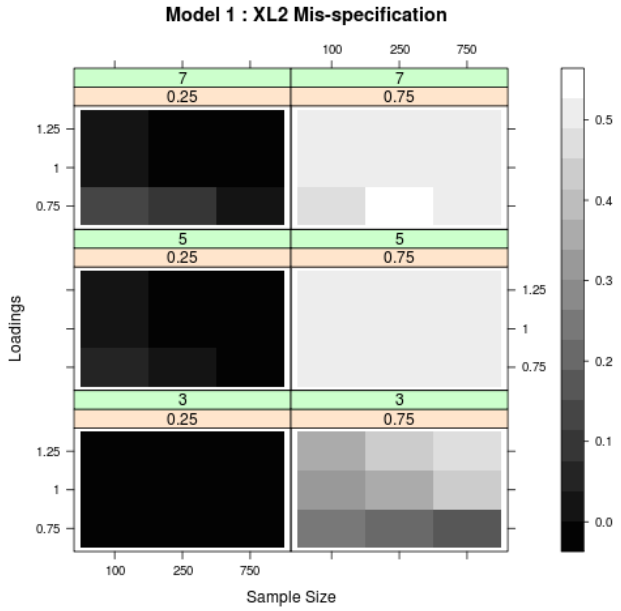


Figure G.14: AVE-R2 criterion (criterion 10) decrease for M1XL2

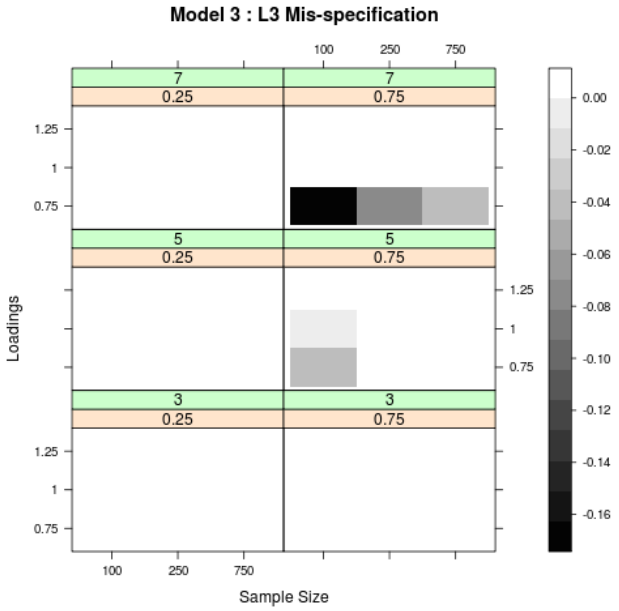


Figure G.15: AVE-R2 criterion (criterion 10) decrease for M3L3

(note the negative scale extent)

Significant structural path estimates (Heuristic 11)

This heuristic concerns only the structural part of the model and requires that all structural path estimates be significantly different from zero. This heuristic is operationalized as the proportion of structural regressions that are statistically significant.

We expect that this heuristic is not affected by the removal of paths, as the remaining paths should be statistically significant. The addition of paths to the estimated model that are not in the generated model should have an effect on this heuristic, as these paths should not be statistically significant. Ideally, reversed paths should also have a negative effect, as the reverse path is not in the generating model, though bivariate regressions are by their nature invariant under reflection.

Our findings confirmed our expectations of no effects for measurement misspecification and removed paths. Our findings also confirmed our expectations for a negative impact of additional paths on this heuristic for many but not all conditions.

Specifically, the proportion of significant paths should drop to .8 for M1L5 (we added one additional path to the four already in M1), which we observed except when indicator loadings were low or when structural effects were strong and the number of indicators was low. In those cases, there was no change, meaning that PLS overestimated the significance of the structural path (Figure G.16).

For M1L6, M1L7 and M2L2 the proportion should drop to 0.75 as two additional paths were added to the four already in M1 or M2, which is what we observed for M1L6 but not for M1L7 or M2L2. For M1L7, we saw a drop to 0.84 when b=0.75 or b=0.25 and s=750, and no change when b=0.25 and s=100. For M2L2, we saw very complex behavior, depending on sample size, strength of structural relationships, and indicator loadings.

We expected the criterion to drop to approx. 0.88 for M3L4 and M3L5 (we added one additional path to the eight already in M3). We observed this for M3L4 only when structural effects were weak and sample size was small, and we observed it for M3L5 only when structural effects were weak, independent of sample size.

The criterion should drop to 0.8 for M3L6 and M3L7 (we added two additional paths to the eight already in M3). This was observed for M3L7, but not M3L6 for which the criterion showed complex behavior depending on strength of structural relationships.

In summary, it appears that PLS estimates models in such a way that frequently even parameters that do not correspond to population parameters are significant. Our data does not show a consistent set of conditions under which this is true, instead differing from model to model. This suggests that this heuristic is unsuitable for reliably identifying misspecified models in practice.

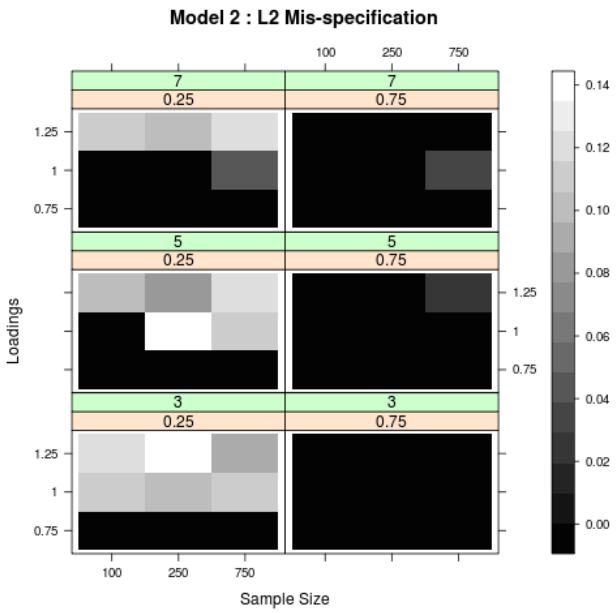


Figure G.16: Significant structural path estimates criterion (criterion 17) decrease for M2L2

Goodness-Of-Fit (relative) (Criterion 12)

The relative goodness-of-fit (GoF) index is "a measure of the global goodness of fit as it is a compromise between communality and redundancy" (Tenenhaus, 2004). It is defined as the geometric mean of the product of mean communality and redundancy of latent variables (Chin, 2010; Esposito Vinzi et al., 2010):

$$GoF = \sqrt{\overline{communality} \times \overline{redundancy}}$$

Based on the absolute GoF defined by Tenenhaus et al. (2004), Esposito Vinzi et al. (2010) propose the relative GoF index that is bound between 0 and 1 as:

$$GoF_{rel} = \sqrt{\left[\frac{1}{J}\sum_{j}\left(\frac{1}{p_j}\sum_{i=1}^{p_j}\frac{r^2(x_{ij}, y_j)}{\lambda_j^{(1)}}\right)\right] \times \left[\frac{1}{J'}\sum_{j:y_j endogenous}\frac{R^2(y_j; y_{j1}, \ldots y_{jk})}{\rho_j^2}\right]}$$

Here J is the total number of latent variables, J' is the number of endogenous latent variables, x_{ij} denotes indicator i of latent variable j, y_i indicates latent variable j, r^2 are correlation coefficients, R^2 are coefficients of determination for endogenous latent variables, λ_j^1 is the first/largest eigenvalue of the principal component decomposition of the indicators for latent variable j and ρ_j is the is the first canonical correlation between the matrix of manifest variables for latent variable y_j and the matrix of manifest variables for all predictors of y_j.

The relative GoF for all true models was consistently above the recommended thereshold of 0.9 when b=0.75 and greater than 0.7 and approaching 0.9 when b=0.25. This suggests that the relative GoF is a good baseline when structural effects are strong.

The complex definition of the relative GoF that includes both structural and measurement model properties precludes a simple analytical prediction. Instead, based on the intent of the GoF to capture both measurement and structural model quality, this heuristic should be negatively affected by all misspecifications in this study.

The relative GoF decreased for many structural and measurement misspecifications (M1L5, M1L6, M1XL1, M2L2, M2XL1, M2XL2, M3L5, M3L6, M3XL1, M3XL2) (Figure G.17), though not sufficiently to yield a value less than the recommended threshold of 0.9. However, contrary to expectations, the rGoF *increased* for some structural misspecifications (M1L2, M1L3, M1L4, M1L7, M2L1, M2R1, M2R2, M3L1, M3L3, M3L7) and showed complex behavior for other misspecification conditions (M1L1, M1R1, M1XL2, M1XL3, M3L2, M3R1, and M3XL3) (Figure G.18). Given this complex behavior, the relative GoF index is not a reliable heuristic to identify model misspecifications.

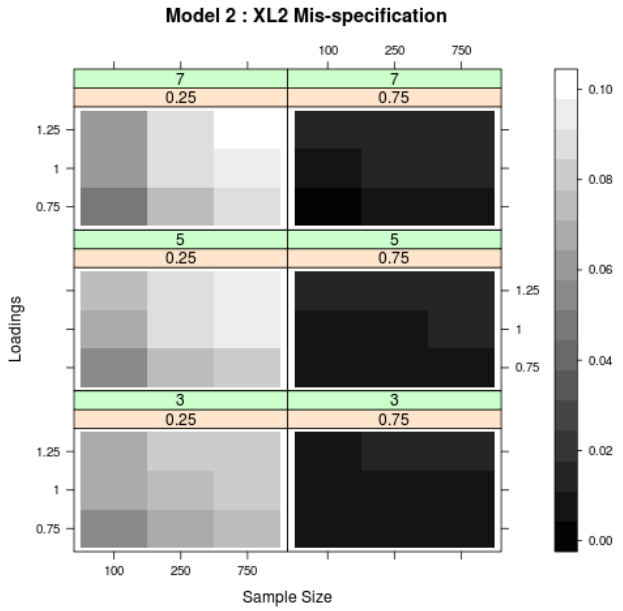


Figure G.17: Relative GoF criterion (criterion 12) decrease for M2XL2

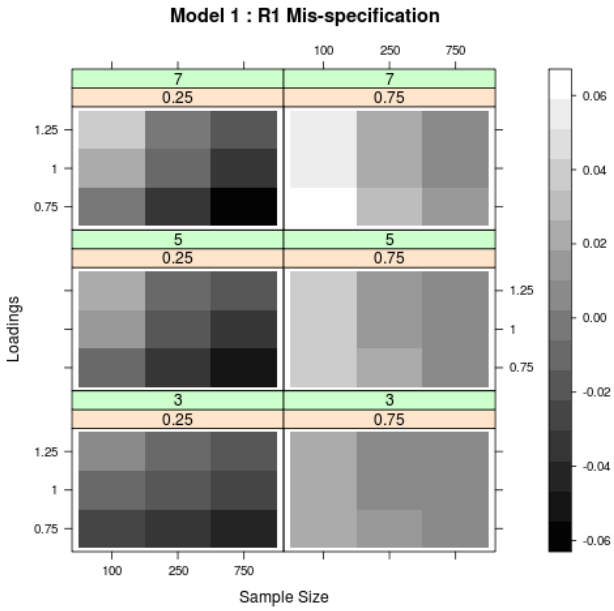


Figure G.18: Relative GoF criterion (criterion 12) decrease for M1R1

(note the negative scale extent)

Predictive Relevance (Heuristic 13)

Predictive relevance is assessed using blindfolding, a procedure where the researcher omits a fraction $\frac{1}{k}$ of observations from the data set, estimates the model parameters, and uses the estimated model to predict the omitted observations. The predictions are compared to observations using the squared difference (E). At the same time, the difference between the variable mean and the observed values are also compared using the squared difference (O). The process is repeated with k different omitted fractions k times and the predictive relevance is then calculated as

$$Q^2 = 1 - \frac{\sum_k E_k}{\sum_k O_k}$$

Chin (2010) suggests that "in general, a cross-validated redundancy Q^2 above 0.5 is indicative of a predictive model." In contrast, the recommendation of 0 by Hair et al. (2012) means that any

improvement in predictive accuracy over the variable mean demonstrates predictive relevance, a very low threshold.

Predictive accuracy is expected to be affected by measurement misspecifications, as the measurement cross-loading introduces an additional predictor to some manifest variables, which should improve predictive accuracy. Redundancy-based predictive accuracy should also be affected by the coefficient of determination r^2 for endogenous latent variables. A larger r^2 implies a smaller residual, in theory yielding improved predictive accuracy. Larger r^2 values might be achieved when structural paths and predictors are added. However, in our case, the additional paths do not have corresponding generating paths and should therefore have no effect.

In line with our expectations, our findings did not show any effect of added paths on this criterion. However, counter to expectations, this criterion was affected in complex ways by other structural misspecification conditions. Our findings also confirmed that predictive ability improved in many, but not all, measurement misspecification conditions. Given this complexity of responses and the limited ability to identify measurement misspecifications, predictive relevance is not a reliable heuristic to identify misspecified models, nor was it designed to be.

CB-SEM χ 2 Test (Criterion 18)

For comparison purposes we also used CB-SEM estimation and the χ2 statistical test of model fit to detect misspecifications. This criterion is operationalized as the mean p-value of the χ2 test. We expected this to be above 0.05 for all true models and below 0.05 for all misspecifications. This was the case for the true M1 and M2 except when s=100 and i=7 and for M3 except when s=100 and i=5 or i=7, confirming the known fact that the χ2 requires reasonably large sample size.

The misspecifications M1L1, M1L3, M1L4, M1L7, M1R1, M1XL1, M1XL2, M1XL3, M2L1, M2XL1, M2XL2, M3L1, M3L2, M3L3, M3R1, M3XL1, M3XL2, and M3XL3 were reliably identified with the

mean p-value dropping to zero. However, for some models (M1L3, M1L4, M1L7, M1R1, M2L1, M2XL2, and M3L1) this required either strong structural effects or a large sample size when structural effects were weak (Figure G.19).

There were no changes in the p-value of the test for models M1L2, M1L5, M1L6, M2L2, M2R1, M2R2, M3L4, M3L5, and M3L6. The added path in M1L2 is already captured in the free exogenous covariance, thus makes no difference to the model fit, i.e. M1L2 is covariance-equivalent to the true M1. The added paths in the models M1L5, M1L6, M2L2, M3L4, M3L5, and M3L6 also make no difference to model fit, as their coefficients are simply estimated to be zero. M2R1 and M2R2 can also be shown to be covariance-equivalent to the true M2.

In summary, the $\chi 2$ test can reliably identify misspecifications for medium and large samples, a result well known from the literature.

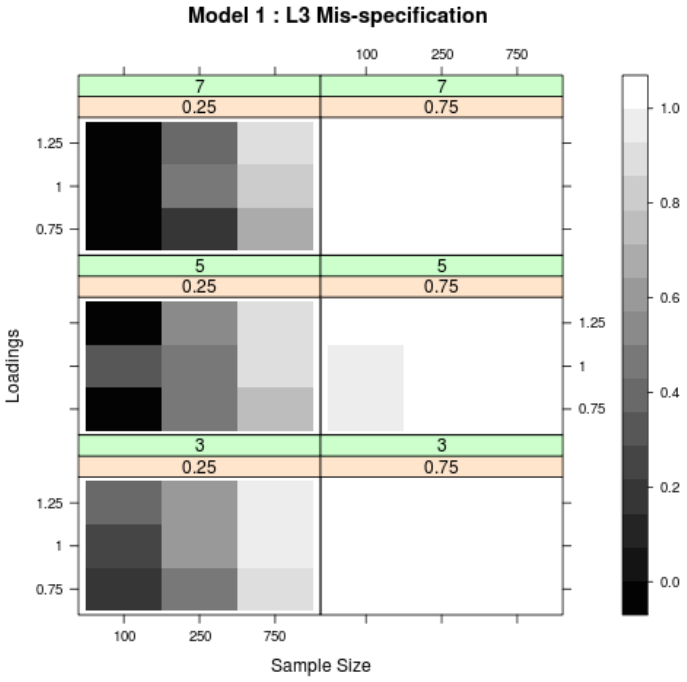


Figure G19: CB-SEM χ^2 test p-value decrease for M1L3

APPENDIX H: PREDICTION MODELS

Predicting the category of a model, true or misspecified, from the model quality heuristics is a classification task, and suitable for predictive statistical classification techniques (Hastie et al., 2010). The outcome is a categorical variable, either with two categories (true model, misspecified model) or with multiple categories corresponding to specific misspecifications of the model. Many classification algorithms and heuristics ("classifiers") have been proposed in the literature.

We used version 3.6 of the Weka data mining suite (Hall et al., 2009) to analyze the PLS data we collected. This software provides more than 30 classifiers for nominal response variables, including parametric models like logistic ridge regression, classification tree-based methods like J48 (an implementation of the C4.5 algorithm), decision rule-based methods such as JRip, and Bayesian methods (Table H.1). Because classifiers have different performance on different problem sets, we subjected our data to all classifiers in the Weka suite. For each classifier, we used the default parameter setting and employed 10-fold cross-validation. Cross-validation is similar to the blindfolding test of predictive relevance used in PLS analyses in that in a k-fold cross-validation, $\frac{1}{k}$ of the data set is treated as test data while the remainder is used as training data for the classifier. The average performance statistics for k repetitions are reported.

Table H.1: Classifiers in the Weka data mining suite

Weka Classifier	**Type**	**Description**
BayesianLogisticRegression	Bayesian	Bayesian logistic regression with normal priors
BayesNet	Bayesian	Bayes network learning classifier
ComplementNaiveBayes	Bayesian	Complement class naïve Bayesian classifier
DMNBtext	Bayesian	Discriminative multinomial naïve Bayesian
NaiveBayes	Bayesian	Naïve Bayes using estimator classes
NaiveBayesSimple	Bayesian	Naïve Bayes using normal attribute distribution
Logistic	Functional	Logistic ridge regression

Weka Classifier	Type	Description
RBFNetwork	Functional	Normalized Gaussian radial basis function network
SimpleLogistic	Functional	Boosted linear logistic regression with automatic attribute selection
SMO	Functional	Support vector machine
SPegasos	Functional	Stochastic primal estimated sub-gradient solver for support vector machines
HyperPipes	Other	Hyper-pipe classifier
IB1	Other	1-Nearest neighbor classifier
IBk	Other	k-nearest neighbor classifier
KStar	Other	Instance-based classifier using entropy-based distance function.
LWL	Other	Locally weighted learning using decision stumps
VFI	Other	Voting feature intervals classifier
ConjunctiveRule	Rule-based	Single conjunctive rule learner
DecisionTable	Rule-based	Simple decision table majority classifier
DTNB	Rule-based	Decision table/ naïve Bayes hybrid classifier
JRip	Rule-based	Repeated incremental pruning to produce error reduction (RIPPER) propositional rule learner
NNge	Rule-based	Nearest-neighbor using generalized exemplars
OneR	Rule-based	1R classifier using the minimum-error attribute
PART	Rule-based	Builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule
Ridor	Rule-based	Ripple-down rule learner
ZeroR	Rule-based	Always predicts mode
ADTree	Tree-based	Alternating decision tree learning
BFTree	Tree-based	Best-first decision tree classifier
DecisionStump	Tree-based	Decision stump based on entropy measure
FT	Tree-based	Functional trees with logistic regression functions at the inner nodes and leaves
J48	Tree-based	C4.5 revision 8 decision tree
J48graft	Tree-based	Grafted C4.5 revision 8 decision tree
LADtree	Tree-based	Multi-class alternating decision tree using LogitBoost
LMT	Tree-based	Logistic model trees

Weka Classifier	Type	Description
NBTree	Tree-based	Decision tree with Naive Bayes classifiers at the leaves
RandomForest	Tree-based	Forest of random trees
RandomTree	Tree-based	Tree that considers K randomly chosen attributes at each node, without pruning
REPTree	Tree-based	Decision tree learner using reduced-error pruning
SimpleCart	Tree-based	Minimal cost-complexity pruning classification and regression tree

As it may be difficult to provide a single classifier for all experimental conditions, we divided our data into sub-sets by sample size, loadings and strength of structural coefficients. Based on the discussion of our results in Appendix G, we expect more difficulties in classifying at low sample size due to a relatively large effect of sampling variations at small sample size. Further, we expect more difficulties in classifying at low structural coefficients, as the effects of the structural relationships on the various model quality heuristics would be weaker.

Each classifier in the Weka suite provides a wealth of performance statistics, such as true positives and false positives for each of the two classes (true model, misspecified model), precision and recall. We focus on the Cohen's Kappa statistic which provides a good summary of the agreement of classifier prediction with the actual class of the model and takes into account the chance agreement. This is important as our data contains only approximately 10% true models. A naïve classifier that by default classifies every model as misspecified would still achieve an impressive average precision and recall over both classes. Using Cohen's Kappa statistics takes this into account.

Table H.2 provides the average F-statistic (precision and recall) over both classes, and Cohen's Kappa statistic of agreement of classification and actual class of model (true, misspecified) for the classifier with the highest Kappa value on each subset and the full data set.

Table H.2.: Performance of best classifier on PLS model misspecification prediction for various experimental conditions and full sample

Sample Size	Loadings	Beta	F (avg)	Kappa	Classifier
100	0.75	0.25	.889	.3216	RandomForest
100	0.75	0.75	.945	.6857	RandomForest
100	1.00	0.25	.893	.3524	RandomForest
100	1.00	0.75	.958	.7630	RandomForest
100	1.25	0.25	.890	.3371	RandomForest
100	1.25	0.75	.962	.7824	RandomForest
250	0.75	0.25	.907	.4445	RandomForest
250	0.75	0.75	.968	.8214	RandomForest
250	1.00	0.25	.924	.5563	RandomForest
250	1.00	0.75	.980	.8867	RandomForest
250	1.25	0.25	.935	.6266	RandomForest
250	1.25	0.75	.983	.9022	RandomForest
750	0.75	0.25	.945	.6800	RandomForest
750	0.75	0.75	.968	.8184	LMT
750	1.00	0.25	.964	.7941	RandomForest
750	1.00	0.75	.975	.8582	RandomForest
750	1.25	0.25	.969	.8253	LMT
750	1.25	0.75	.984	.9091	LMT
Full data set			.941	.6626	RandomForest

As the results in Table H.2 show, the overall performance of the classifiers on the complete PLS data set is poor (Kappa = 0.6626) and suggests that it is difficult to provide model classification guidelines that are generalizable to all three models for all model and sample conditions. As we expected, the performance of even the best classifiers is poor when sample size is low (s=100) or when structural path coefficients are low (b=0.25). Table H.2 indicates that it is possible to predict model misspecification

with acceptable degrees of accuracy (Kappa ≈ 0.8 or better) for medium to large sample sizes and strong structural effects.

However, this result must be qualified due to the nature of the classifiers that are built by the software. Given the large size of our data set (334800 PLS estimations), tree-based classifiers, which are consistently the best type of classifiers (Table H.2), have a tendency to overfit, i.e. to build large trees with very specific branching rules. For example, the tree built by the LMT classifier for ($s = 750, l = 1.25, b = 0.75$) has 44 leaves at the end of each is a logistic regression model, the random forest classifiers for the other conditions each consist of 10 decision trees, each with dozens or hundreds of leaves. While reducing the number of leaves in each decision tree is possible to achieve a more parsimonious set of guidelines, this severely degrades the Kappa value of classification accuracy. More interpretable classifiers, such as ADtree, Ridar, PART, and JRip, lead to more parsimonious guidelines about identifying model misspecifications, but have unacceptably low Kappa values. Table H.3 shows the performance of the full suite of classifiers for ($s = 750, l = 1.25, b = 0.75$). For example, the tree built using the ADtree classifier contains only 21 leaf nodes, but its Kappa value is unacceptably low at 0.7242. Tables H.4 and H.5 compare the confusion matrices for the LMT and ADtree classifiers to show the implications of the different Kappa values on the number of false positives and false negatives.

Table H.3. Classifier performance for $s = 750, l = 1.25, b = 0.75$.

Classifier	**F (avg)**	**Kappa**
Logistic	0.857	0
BayesianLogisticRegression	0.857	0
DMNBtext	0.857	0
RBFNetwork	0.857	0
SimpleLogistic	0.857	0
SMO	0.857	0
SPegasos	0.857	0
LWL	0.857	0

Classifier	F (avg)	Kappa
ConjunctiveRule	0.857	0
ZeroR	0.857	0
DecisionStump	0.857	0
OneR	0.857	0.0070
ComplementNaiveBayes	0.678	0.0280
VFI	0.624	0.1568
NaiveBayes	0.725	0.2297
HyperPipes	0.819	0.3660
DecisionTable	0.912	0.4384
BayesNet	0.932	0.6359
ADTree	0.955	0.7242
DTNB	0.966	0.8079
RandomTree	0.976	0.8613
NNge	0.977	0.8664
NBTree	0.977	0.8683
IB1	0.978	0.8743
IBk	0.978	0.8743
BFTree	0.978	0.8771
KStar	0.979	0.8806
Ridor	0.979	0.8828
FT	0.980	0.8860
JRip	0.980	0.8881
SimpleCart	0.981	0.8955
PART	0.981	0.8963
REPTree	0.982	0.9008
J48graft	0.982	0.9015
J48	0.983	0.9027
RandomForest	0.984	0.9087
LMT	0.984	0.9091

Table H.4. Confusion matrix of LMT classifier for $s = 750, l = 1.25, b = 0.75$

		Classified As	
		True	Misspecified
Actual Class	True	1738	62
	Misspecified	247	16553

Table H.5. Confusion matrix of ADtree classifier for $s = 750, l = 1.25, b = 0.75$

		Classified As	
		True	Misspecified
Actual Class	True	1144	656
	Misspecified	122	16678

In summary, applying predictive modelling techniques to classify models based on their model quality heuristics shows limited success. The relatively low classification success rate for low sample sizes and weak structural effects, coupled with the highly complex classification rules that are generated make it impractical to provide simple, parsimonious and general guidelines to a researcher using PLS to identify misspecified models. Moreover, while we have built classification rules for all three models, given the complexity and specificity of the classification rules, it is unlikely that these generalize even to relatively similar models and/or sample characteristics.