Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2012 Proceedings

European Conference on Information Systems (ECIS)

5-15-2012

USE OF PARTIAL LEAST SQUARES AS A THEORY TESTING TOOL – AN ANALYSIS OF INFORMATION SYSTEMS PAPERS

Mikko Rönkkö
Aalto University School of Science

Kaisa Parkkila
Aalto University School of Science

Jukka Ylitalo
Aalto University School of Science

USE OF PARTIAL LEAST SQUARES AS A THEORY TESTING TOOL – AN ANALYSIS OF INFORMATION SYSTEMS PAPERS

Mikko Rönkkö, Aalto University School of Science, PO Box 15500, 00076 Aalto, Finland, mikko.ronkko@aalto.fi

Kaisa Parkkila, Aalto University School of Science, PO Box 15500, 00076 Aalto, Finland, kaisa.parkkila@aalto.fi

Jukka Ylitalo, Aalto University School of Science, PO Box 15500, 00076 Aalto, Finland, jukka.ylitalo@gmail.com

Abstract

Motivated by recent critique toward partial least squares path modeling (PLS), we present a research question if the PLS method, as used currently, is at all an appropriate tool for theory testing. We briefly summarize some of the recent critique of the use of PLS in IS as a theory testing tool. Then we analyze the results of 12 PLS analyzes published in leading IS journals testing if these models would have been rejected in the case that the data used for model testing had very little correspondence with the theorized models. Our Monte Carlo simulation shows that PLS will often provide results that support the tested hypotheses even if the model was not appropriate for the data. We conclude that the current practices of PLS studies have likely resulted in publishing research where the results are likely false and suggest that more attention should be paid on the assumptions of the PLS model or that alternative approached like summed scales and regression or structural equation modeling with estimators that have known statistical properties should be used instead.

Keywords: Partial least squares, theory testing, Monte Carlo simulation

1 Introduction

The use of partial least squares path modeling (PLS) as a tool for theory testing has been increasing in the IS community since the late 90's and PLS is currently one of the most common quantitative data analysis methods in the top IS journals (Gerow et al. 2010). The prevalence of the use of the PLS method in IS is unparalleled by any other discipline, except perhaps marketing (Hair, Sarstedt, et al. 2011). Most notably, PLS analyzes are virtually absent in psychology and econometrics, in which most of the statistical techniques used in IS (including PLS) were originally developed. In fact, researchers publishing in these disciplines have not only abandoned the method, but recently voiced concerns about its use in IS. For example, McDonald (cited in Goodhue et al. forthcoming, Appendix B2) recently stated that "PLS is a collection of algorithms that were casually and very foolishly conceived, and cannot be recommended" and Hardin and Marcoulides (2011, p.762) after their discussion on formative measurement point out that "Such benefits and even supposed immunities to fundamental statistical principles (e.g., distributional characteristics, sample size, magnitudes of standard errors, etc.) have also been attached to the PLS method itself, despite the overwhelming evidence to the contrary".

The prevalence of the PLS method in IS can be explained on one hand by the misleading idea that it is an estimator for SEM models and on the other hand by the fact that the way that the method capitalizes on error correlations (Rönkkö & Ylitalo 2010) and fails to reject incorrect models (Evermann & Tate 2010) can fool a researcher concluding that the method would have more statistical power than alternative methods. Thus reliance on the PLS method has possibly resulted in producing and publishing a large number of studies, whose results are actually invalid. In this paper we consider this possibility by analyzing 12 studies published in leading IS journals and retesting their models with various data that should not provide support for the tested models.

The paper provides further evidence that PLS does not reject a model even if it was incorrect. The paper contributes to the existing simulation studies using synthetic models (Evermann & Tate 2010; Rönkkö & Ylitalo 2010; Goodhue et al. forthcoming) by extending these studies to models from real, published papers. Moreover, more studies highlighting the weaknesses of PLS are needed to raise the awareness of these issues in the IS community. Although evidence against using PLS exists in the literature, the articles discussing the method and providing guidelines for the use of the method (e.g., Gefen et al. 2011) generally do not discuss the critique the method has received. Moreover, some researchers have suggested that IS journals are biased against publishing results that challenge the currently held assumptions about PLS (Goodhue et al. forthcoming).

Our results suggest that the risk of publishing substantively biased results is very real and cannot be ignored. We start the paper by presenting a short overview and a small extension to the existing research criticizing the PLS method. Then we will describe our simulation settings followed by the simulation results. Our paper is concluded by presenting guidelines for the users of PLS.

2 Overview and extension of the recent critique of PLS

The PLS method is used to estimate path models where the construct variables of the path diagram are measured with several indicator variables. These models are most commonly analyzed by expressing them as a set of simultaneous structural equations, whose parameters are estimated jointly, or by estimating scores for the constructs and then using these construct scores in separate regression analyzes. The first approach, known as structural equation modeling (SEM), has an advantage over the second approach, known as composite variable analysis (McDonald 1996), in that it can to some extent control for measurement error in the indicator variables (Gefen et al. 2011). A key problem that the IS research community has with PLS is classified as a structural equation modeling technique while what the method actually does is composite variable analysis.

Classifying PLS as a SEM technique is technically correct, but can be hugely misleading to researchers who are not specifically trained in statistics. SEM analysis, like many other statistical analyzes, is a combination of a statistical model and an estimator: A statistical model is a set of equations with one or more free parameters that are to be estimated. An estimator is any algorithm that can be used to estimate the model parameters, but the definition of estimator does not embed any assumptions about these estimates being correct (Lehmann & Casella 1998, p.4). Because of this very broad definition, all composite variable techniques – including using unweighted summed scales as construct scores – are in fact also SEM estimators. The two main qualities of estimators are consistency and unbiasedness. Consistency is based on asymptotic properties of the estimator, that is, the performance of the estimator when sample size approaches infinity. An estimator is consistent if the estimates that it provides converge to the population value as the sample size approaches infinity. Bias is a property associated with the performance of an estimator with finite or small samples. An estimator is unbiased if estimates over repeated samples of the same population are evenly distributed around the population value. The qualities of consistency and unbiasedness are the first things that a new estimator must show (e.g., Bollen 1996). Composite variable analysis has been shown to be both biased and inconsistent estimator for SEM models because measurement error of indicators is included in the composite scores (Dijkstra 1983; Bollen 1989; McDonald 1996) and because of this they are not typically discussed as viable alternatives in SEM text books.

Because arguing that PLS is a SEM estimator is equally correct as arguing that summed scales is a SEM estimator and because both of these methods work the same way by first estimating construct scores and then using separate regression analyzes to estimate the directional paths in the SEM model (Rönkkö & Ylitalo 2010; Goodhue et al. forthcoming). Because of their similarity, comparing the two algorithms is needed to understand any potential advantage that PLS might have. When considering the point estimates, the only difference between estimating a model with summed scales and regression analysis and estimating it with PLS is that PLS calculates the construct scores as differentially weighted sums of the indicators instead of using equal weights. This is commonly argued to minimizes the effect of measurement error (Chin et al. 2003; Gefen et al. 2011), but to our understanding no evidence to date support the assertion that the construct scores calculated with PLS would be any closer to the true scores[1] than scores calculated by using equal weights. However, there is at least one study that shows evidence to the contrary (Rönkkö & Ylitalo 2010). While the model-dependent indicator weighting scheme can be argued to produce superior results when the measurement errors are truly random (Gefen et al. 2011), this is never the case in applied research because indicators are subject to sampling error (Gefen et al. 2011): Instead of being exactly at zero, a sample correlation of two unrelated variables is known to follow student's t distribution. While the assumption of uncorrelated errors is important for PLS (Lohmöller 1989), we are not aware of any PLS studies that test for the correlated error terms, although this would be relatively straightforward to do by applying the standardized root mean square residual (SRMR) fit index that is also commonly used for evaluating SEM estimation results (Kline 2005, p.141; Lohmöller 1989, pp.52–53). Calculating this index from PLS results would involve first calculating a correlation matrix from the outer estimation residuals and raising each element below the diagonal to the power of two and taking a mean of these values.

The presence of correlated errors in the data results in inflated estimates of the path coefficients (Zimmerman & Williams 1977). The problem with PLS is that the algorithm considers this as relevant variance that should be explained and can thus amplify the bias (Rönkkö & Ylitalo 2010). The effect is particularly strong when a tested path does not in fact exist in the population. However, our analysis presented in the Appendix 1 indicates that the paper by Rönkkö and Ylitalo in fact underestimated the problem with error correlations and in reality the indicator weights produced by PLS can depend more on the strength of the correlations between the error terms than on the actual indicator reliability.

[1] The term true scores is used in the meaning given to the term in classic test theory (Nunnally 1978).

Because sampling error increases with decreasing sample size, the PLS estimates of path coefficients tend to get larger when the sample size gets smaller. This feature is present in several existing simulation studies (Aguirre-Urreta et al. 2008; Chin & Newsted 1999; Goodhue et al. forthcoming), but the real cause – bias caused by correlated errors – has gone unnoticed. Instead, e.g. Chin and Newsted (1999) interpreted this effect in their result so that PLS would result in more accurate estimates when sample size gets smaller. However, a closer examination of their detailed results presented as an online appendix for the paper revealed that only those path coefficients that were originally underestimated became more accurate while the path estimates that were originally close to correct or overestimated became more biased.

There is also a belied that PLS can be used as a model test (Evermann & Tate 2010). This belief can be attributed to the idea that PLS is a SEM method and many SEM estimators indeed have test statistics for the overall model test. The problem with the heuristics that are currently used as a de-facto model test is that they are generally not based on any statistical theory and have been presented without evidence of their capability to detect model misspecification. For example, the AVE statistic was presented by Fornell and Bookstein (1982) based solely on the assumption that because it can be used with SEM models, it would work with PLS. Similarly, Tenenhaus, Amato, and Vinzi (2004) presented a set of Goodness of Fit indices under the assumption that because these can assess the predictive power of the model, they can be used as model test. This latter argument has been recently challenged by Henseler and Sarstedt (Forthcoming), who showed that the GoF indices cannot detect model misspecification. Evermann and Tate (2010) have gone even further and tested the currently used set of model quality indices and concluded that these cannot be at all used to test if the model fits the data. In fact, under some conditions in their simulation, some of the tested model quality statistics actually improved when the model was misspecified. Considering that PLS, as currently used, is unable to reject incorrect models, and the fact that error correlations bias the results away from zero, we argue that PLS will often produce support for hypotheses even if the tested model is incorrect. While it is likely that most of the models in published research are in fact correct, a statistical analysis cannot be taken as evidence for this if the particular model would fail to reject incorrect models.

3 Simulation Study of PLS Method as a Hypothesis Testing Tool

To test if data that should not support the tested model indeed would provide positive results when estimated with PLS and if also differently specified models would be supported, we conducted a Monte Carlo simulation using models published in top IS journals as the population models. Because PLS results are completely determined by the raw data covariance matrix (Lohmöller 1989), it is possible to generate simulated data using a published indicator covariance matrix and reproduce the results of a published paper.

The first step in our study was selecting the studies that are used in the analysis. Since we wanted to include only the highest quality articles, we chose to search for PLS papers in the top journals only. Searching for "PLS" or "partial least squares" in Business Source® Complete database resulted in 115 hits in MIS Quarterly and 38 articles in Information System Research. Of these papers, 90 were empirical papers using PLS, while others were methodological paper or just contained the search terms in the reference list. To simplify the data generation and ensure that the models in the papers are comparable, we restricted our analysis to papers using only reflective indicators. Additionally, we excluded papers if they tested interaction or multiple group models or contained hierarchical constructs. Ideally, an indicator-level covariance matrix would be used in the data generating process, but this was reported in only one of the remaining papers. Due to this, we included also papers presenting item-construct cross-loading matrix, from which an approximation of the population covariance matrix can be generated. Since PLS was designed primarily as a prediction model, the construct scores should be valid predictors for indicators. The residual (unique) variance of each indicator in the prediction can be calculated based on the fact that the construct scores and indicators are standardized by default in the popular PLS packages. Only 11 of the remaining articles reported

either of these matrices. One of the articles (Komiak & Benbasat 2006) contained two different models and another article (McLure Wasko & Faraj 2005) two different data for the same model, so altogether thirteen models were examined. Analysis of one paper produced unexpected results and after contacting the authors, we concluded with them that the paper contained an error and dropped it from the analysis. The list of the selected articles is presented in Table 1.

The shortcoming in using the item-construct cross-loading matrices for estimating an indicator covariance matrix is that we lose the part of the correlations between the indicators that is not explained by the constructs. If the model is correct, these error correlations should be very close to zero. However, none of the papers reported any statistics on the unmodeled correlations between items belonging to different constructs, although these are important in assessing if the model is correctly specified (Lohmöller 1989). Because of this, we worked with the assumption that the constructs explained the covariances in the data perfectly. If this were not the case in the original paper, our simulated data would fit the original model better than the real data used in the actual paper thus artificially penalizing any alternative model when compared with the original model. If the original data contained strongly correlated errors, it would be possible that an alternative model that was not supported with our simulated data would have received support from the original empirical data.

After collecting the papers, we generated 500 datasets from each of the reported indicator covariance matrix or item-construct cross-loading matrix using Monte Carlo feature of Mplus 6.0 software (cf., Marcoulides & Saunders 2006). The indicator variables were set to be centered, and variances of all latent and indicator variables were set to one. All indicators and latent variables were assumed to be normally distributed. Although this is not often the case with empirical papers, none of the included papers included skewness or kurtosis statistics and hence we could not make any informed guesses about the actual distributions. However, the exact distribution of the variables is not important because PLS estimates are completely determined by the sample covariance matrix (Lohmöller 1989), which can in any case be approximated using the item-construct cross-loading matrix. The number of observations was set to equal the number of observations presented in the original paper except for the paper by Majchrzak, Beath, Lim, and Chin (2005) for which we used 50 observations instead of the 17 in the original paper so that the chosen data generating process could be used. These data sets were labeled *original data*. Additionally, we used Stata 11 to generate four series of datasets with increasing deviation from the properties of the *original data*. The second datasets were labeled as *mixed variables data*. These data sets were created by randomly choosing 10% (rounded up) of the variables and swapping these with randomly chosen variables reflecting minor measurement model misspecification. The third data labeled *completely mixed variables data* was similar to *mixed variables data* except that all variables were shuffled. The fourth data labeled *equal covariances data* was generated by drawing a sample from normally distributed population with correlations between all variables set to the mean correlation between items in the *original* data reflecting data that was caused by a single factor (i.e. data that are only method variance). The final data was labeled *random data* and was drawn from a population with zero correlations between the items.

After generating 500 replications of each of the five types of data for each of the 12 models, we used these 30 000 data to test three different types of PLS models: The *original model*, a *misspecified model* where two paths were altered and a *random model*. The modified models were generated by first writing the model paths as a lower triangular matrix where one indicated a path and zero that a path was not present between the constructs. The *misspecified model* was created by choosing two random paths and then changing these so that if a path did not exist in the *original model*, it existed in the *misspecified model* and if a path did exist in the *original model*, it was removed in the *misspecified model*. The *random model* was generated by setting the paths to one and zero randomly with equal probability of both values. The modified models were regenerated for each PLS analysis. For PLS estimation, we chose the plspm-package version 0.1-6 of the R statistical software environment. Since some PLS model – data combinations resulted in non-converging solutions, the exact number of successful replications for each modeling condition was slightly below 500. In total estimation process was started for 90 000 PLS models each containing 100 bootstrap samples.

Before analyzing if the distorted data would support the models presented in the selected papers and if the generated data would also support the modified models, we checked if the generated data actually conformed to the data presented in the original papers using two different approaches. First, we choose a small sample of models and estimated these with SmartPLS to see if the results obtained with the *original data* and *original model* conformed to the results presented in the papers. This ad hoc test did not indicate any major problems in the generated data. Second, we did a more systematic test by calculating the mean of absolute difference between the parameter estimates obtained with the *original data* and *original model* to the model coefficients presented in the published papers. This index is reported in Table 1 and indicates that the parameter estimates vary from very close to mediocre fit with the original. The lack of fit can be a sign that the original data were not close to multivariate normality, or that there are correlations that are not capture with the item-construct cross-loading matrix, or as in the case of the dropped paper, that the original item-construct cross-loading matrix was calculated incorrectly. While a large value of the index shows that the generated data was not very close to the original data, it does not directly invalidate the further tests since we can still compare if PLS would give support for hypotheses even if the used data did not come from a population with the hypothesized structure.

Typically a paper presenting PLS analysis first starts by establishing that the measurement model and overall model fit are adequate and then proceeds to interpreting the path coefficients between the constructs (Evermann & Tate 2010). We did this in a three step process: First, we evaluated the quality of the measurement model by looking at factor loadings and if these exceeded the 0.7 threshold that is commonly used. To save space in the paper, we did not inspect AVE or CR indices, since these can be derived from factor loadings and provide no additional information if the factor loadings are known (Fornell & Larcker 1981). Also, we did not analyze discriminant validity of the measurement due to space constraints. However, since the currently used AVE and cross-loading based tests are insensitive to model misspecification (Evermann & Tate 2010), these tests would have been unlikely to reject false models. Second, we estimated the overall model fit by using the overall goodness of fit (GoF) index presented by Tenenhaus et al. (2004). Third, we examined the share of path coefficients that were significant at $p<0.05$ level and compared this figure across models and data types.

Paper	Success index	Constructs	N
Komiak and Benbasat B (2006)	0.027	7	100
Komiak and Benbasat A (2006)	0.037	6	100
Majchrzak, Beath, Lim, and Chin (2005)	0.046	6	17
McLure Wasko and Faraj A (2005)	0.072	7	173
McLure Wasko and Faraj B (2005)	0.073	7	173
Thatcher and Perrewe (2002)	0.119	5	211
Karahanna, Agarwal, and Angst (2006)	0.137	18	278
Lewis, Agarwal, and Sambamurthy (2003)	0.150	5	161
Wixom and Todd (2005)	0.176	19	465
Jiang and Benbasat (2007)	0.188	13	176
Enns, Huff, and Higgins (2003)	0.218	6	69
Compeau and Higgins (1995)	0.253	20	1020

Table 1 List of analyzed studies

4 Results

The results of each of the three analyzes are reported in Table 2. The three groups of columns contain the results for each test: share of factor loadings over 0.7, mean of global goodness of fit index, and share of path coefficients that are significant at $p<0.05$. We tested the significance of the differences in means of goodness of fit indices comparing the combination of *original data* and *original model* with all other combinations using t tests and did a similar comparison with the share of significant paths and factor loadings over 0.7 using proportions tests. Most differences in the table were significant at $p<0.01$. However, because a researcher makes her judgement typically based a single set of models instead of systematically comparing a large set of models and data, we focus on the question would the results from the different models be interpreted differently assuming that they are interpreted in isolation rather than testing which combination of data and model result in the best result.

4.1 Column group 1: Factor loadings over 0.7

Starting from the first group of columns in Table 2, there are no large differences in the number of factor loadings over the .7 limit between the models. The reason for this was that only the structural part of the misspecified models were altered while measurement model misspecification was accomplished by mixing the variables in the data. The *original data* and *equal covariances data* seem to provide the best results for the factor loadings for all models in all papers. The result for *original data* is expected, since all tested models included the correct measurement model specification that was also used as the model from which the data were generated. The fact that *equal covariances data* provides good results is problematic; for this data to provide support for the models, the correlations between the original indicators must be quite high with also all other constructs indicating that there might be lack of discriminant validity. For all papers, the *random data* provides the worst results for the measurement model, which is quite natural considering that there is no factor structure in these data. However, in the McLure Wasko and Faraj (2005) paper half of the factor loadings in the *random data* were over the 0.7 limit. If this study were implemented with newly developed scales, a researcher might drop some of the poorly performing items resulting in gain acceptable results for measurement model even if there was no real structure in the data.

Several of the paper provide anomalous results The paper by Thatcher and Perrewe (2002) provide substantially worse results for *equal covariances data* and *completely mixed variables model* compared to other papers. The reason for this is that this particular paper had two out of six hypotheses into negative direction and hence negative correlations were present in the estimated indicator covariance matrix. The paper by Enns, Huff, and Higgins (2003) had the worst results for the factor loadings because the paper contained a lot of negatively worded items that, like in the Thatcher and Perrewe paper, resulted in negative correlations in the estimated indicator covariance matrix.

For 8 out of the 12 papers the *mixed variables data* produced acceptable factor loadings for the 90% of indicators and for 3 out of 12 papers the results for factor loadings were very good unless the data were completely random. Since it is not completely unheard of that researchers report results where some factor loadings are only close to acceptable (Evermann & Tate 2010), considering the fact that and in the earlier stages of research, lower factor loadings of 0.5 and 0.6 are sometimes accepted (Chin 1998), it is quite likely that studies with misspecified measurement models get published even in the top journals. When we consider that three papers produced acceptable factor loadings for over 90% of the indicators even with *completely mixed variables data*, this is almost certainly the case. Moreover, there are three additional things to consider: First, we could not include the correlations not between the indicators not explained by the constructs because for most of the paper data generating relied on item-construct cross-loading matrices thus artificially penalizing the modified datasets in our analysis. Second, when faced with one or two poor items, it is typical that a researcher just drops these from the analysis. Third, according to Bollen and Lennox (1991) it is possible that when faced with poor

measurement results, some of the scales are switched to formative mode, where low inter-item correlations are accepted or even desirable.

4.2 Column group two: Goodness of Fit

The second group of columns in Table 2 shows the mean of absolute goodness of fit index for each paper, model type, and data combination. The fit indexes for the *random data* are considerably worse than other models. This is not surprising, since *random data* does not have any factor structure except by chance. What is surprising is that for some models the GoF index is close to 0.25 indicating that a model that should not receive any support from the data still on average accounts for a quarter of the variance in each regression of the model.

The fit indexes are over 0.80 for three papers regardless of model misspecification or data used as long as the data are not completely random: Majchrzak et al. (2005), Wixom and Todd (2005), and Jiang and Benbasat (2007). For these articles, it is clear that the results from *random data* are never acceptable regardless of the model but e.g. the choice over *random model* and *original model* is totally arbitrary on the grounds of fit indexes when the *original data* are used.

While for most of the papers the fit indexes from the *original model* are generally the highest, the differences among different models and data are generally so small that the researcher, most likely, would also accept other combinations of data and model than the *original data* and the *original model*. In the case of *original data*, for seven models of the examined twelve the fit index is same whether the *original model* or *misspecified model* was used. More over, the results obtained using *original data* and *original model* models are not always the best ones. In the models A and B by Komiak and Benbasat (2006), the fit indexes of the *completely mixed variables data* and *random model* are the highest. The *completely mixed variables data* also gives the greatest values to fit index for the models A and B by McLure Wasko and Faraj (2005), and the model by Thatcher and Perrewe (2002). Overall, some other data give as good as or better results than *original data* for all models in nine cases out of twelve.

Based on the analysis of the goodness of fit indices we can conclude that if the level of covariances in the data is approximately equal to the *original data*, any model fits approximately as well as the *original model* and any data gives approximately the same level of support for the model as the *original data*. Considering the recent conclusion by Evermann and Tate (2010) that the indices and heuristics used as a model test do not really work and even more recent criticism of the goodness of fit index by Henseler and Sarstedt (forthcoming), this is not all that surprising.

4.3 Column group 3: Share of significant paths

The third group of columns in Table 2 reports the proportion of path coefficients that are significant at $p<.05$. For eight reported models out of twelve, some other model is better or as good as the *original model*. This underlines the conclusion by Evermann and Tate (2010) that the path coefficients are meaningless unless the model correctness has been shown first.

The *original model* is usually the best for the *original data* but these are not always the highest proportions of all. For instance in the paper of Thatcher and Perrewe (2002), 98.9% of path coefficients are significant with the combination of *random model* and *equal covariances data*, but for the combination of the *original data* and *original model* only 73.5% of path coefficients are significant. The corresponding percentages are 92.5% and 77.4% for Karahanna et al. (2006). The data generating succeeded best for the papers of Komiak and Benbasat (2006) and Majchrzak et al. (2005), but also for these paper some other combination of data and model than the *original data* and *original model* results in a larger number of significant path estimates.

Since the between-data and between-model differences are generally relatively small, it is quite likely that a researcher analyzing the data would conclude that the research hypotheses tested using PLS would be supported even if the model was misspecified or the data were flawed.

Table 2 Results for the three tests by model and data

		Share of factorloadings over 0.7					Mean goodness of fit					Share of significant paths				
Paper	Model	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Compeau and Higgins (1995)	original	0.93	**0.98**	0.88	0.73	0.11	**0.74**	0.67	**0.74**	0.71	0.06	0.94	**1.00**	0.91	0.92	0.01
	misspecified	0.93	**0.98**	0.88	0.73	0.12	**0.72**	0.67	**0.72**	0.71	0.06	0.93	**1.00**	0.91	0.92	0.01
	random	0.94	**0.98**	0.89	0.73	0.11	0.66	0.65	0.67	**0.70**	0.05	0.85	**1.00**	0.85	0.94	0.01
Enns. Huff. and Higgins (2003)	original	**0.83**	0.70	0.74	0.62	0.40	0.61	0.61	0.64	**0.65**	0.36	**0.39**	0.35	0.21	0.19	0.03
	misspecified	**0.85**	0.69	0.74	0.62	0.41	0.48	0.49	0.51	**0.52**	0.26	0.46	**0.47**	0.30	0.24	0.02
	random	**0.84**	0.70	0.75	0.62	0.41	0.44	0.46	0.48	**0.50**	0.19	0.58	**0.66**	0.43	0.36	0.02
Karahanna. Agarwal. and Angst (2006)	original	0.93	**1.00**	0.82	0.71	0.24	**0.75**	0.72	**0.75**	**0.75**	0.10	**0.77**	**0.77**	0.72	0.74	0.01
	misspecified	0.93	**1.00**	0.82	0.71	0.23	**0.75**	0.72	**0.75**	**0.75**	0.11	0.75	**0.79**	0.73	0.75	0.01
	random	0.93	**1.00**	0.82	0.71	0.24	0.72	0.71	**0.74**	**0.74**	0.10	0.79	**0.93**	0.84	0.87	0.02
Komiak and Benbasat A (2006)	original	**1.00**	0.92	0.93	0.85	0.40	0.54	0.58	0.58	**0.60**	0.15	0.86	**0.92**	0.85	0.87	0.03
	misspecified	**1.00**	0.92	0.93	0.85	0.40	0.53	0.57	0.57	**0.59**	0.15	0.89	**0.94**	0.88	0.89	0.03
	random	**1.00**	0.92	0.94	0.85	0.40	0.50	0.57	0.55	**0.59**	0.16	0.70	**0.82**	0.72	0.76	0.02
Komiak and Benbasat B (2006)	original	**1.00**	0.92	0.94	0.85	0.41	0.55	0.60	0.60	**0.62**	0.16	0.84	**0.90**	0.80	0.79	0.03
	misspecified	**1.00**	0.92	0.94	0.85	0.40	0.55	0.59	0.59	**0.61**	0.16	0.86	**0.92**	0.83	0.82	0.03
	random	**1.00**	0.92	0.94	0.85	0.40	0.50	0.57	0.55	**0.60**	0.16	0.70	**0.82**	0.70	0.75	0.03
Lewis. Agarwal. and Sambamurthy (2003)	original	**1.00**	0.97	0.95	0.76	0.12	**0.80**	0.70	0.79	0.72	0.16	**1.00**	0.91	0.99	0.85	0.02
	misspecified	**1.00**	0.97	0.95	0.76	0.12	**0.74**	0.69	**0.74**	0.71	0.14	**0.98**	0.92	0.97	0.87	0.01
	random	**1.00**	0.97	0.95	0.76	0.12	0.70	0.69	**0.71**	**0.71**	0.13	0.90	**0.93**	0.88	0.87	0.01
Majchrzak. Beath. Lim. and Chin (2005)	original	0.99	**1.00**	0.98	0.94	0.20	**0.88**	0.84	0.87	0.85	0.25	**1.00**	0.81	0.97	0.80	0.03
	misspecified	0.99	**1.00**	0.98	0.94	0.20	**0.88**	0.83	0.87	0.85	0.24	**1.00**	0.82	0.97	0.82	0.03
	random	0.99	**1.00**	0.98	0.94	0.20	0.83	0.83	**0.84**	**0.84**	0.22	**0.81**	0.75	0.80	0.74	0.02
McLure Wasko and Faraj A (2005)	original	**1.00**	0.98	0.94	0.82	0.52	0.58	0.61	0.62	**0.67**	0.23	**0.48**	0.26	0.43	0.36	0.07
	misspecified	**1.00**	0.98	0.94	0.82	0.52	0.58	0.61	0.62	**0.67**	0.23	**0.48**	0.26	0.43	0.36	0.08
	random	**1.00**	0.99	0.94	0.85	0.52	0.56	0.59	0.60	**0.65**	0.13	0.69	**0.79**	0.70	0.74	0.04
McLure Wasko and Faraj B (2005)	original	**0.98**	0.97	0.91	0.82	0.52	0.63	0.60	0.64	**0.68**	0.23	**0.55**	0.26	0.47	0.35	0.07
	misspecified	**0.98**	0.97	0.91	0.82	0.52	0.63	0.60	0.64	**0.68**	0.23	**0.55**	0.27	0.47	0.35	0.07
	random	**1.00**	0.98	0.94	0.84	0.52	0.58	0.58	0.60	**0.64**	0.13	0.73	**0.81**	0.71	0.73	0.04
Thatcher and Perrewe (2002)	original	**0.89**	0.45	0.79	0.54	0.10	0.57	0.60	0.59	**0.61**	0.15	0.74	**0.95**	0.69	0.42	0.01
	misspecified	**0.89**	0.45	0.79	0.54	0.10	0.54	0.58	0.56	**0.59**	0.13	0.78	**0.97**	0.75	0.45	0.01
	random	**0.89**	0.45	0.79	0.54	0.10	0.50	0.56	0.52	**0.57**	0.12	0.84	**0.99**	0.81	0.49	0.01
Wixom and Todd (2005)	original	**1.00**	**1.00**	**1.00**	**1.00**	0.32	**0.99**	**0.99**	**0.99**	**0.99**	0.08	0.99	**1.00**	0.98	0.94	0.03
	misspecified	**1.00**	**1.00**	**1.00**	**1.00**	0.32	**0.99**	**0.99**	**0.99**	**0.99**	0.08	0.99	**1.00**	0.98	0.94	0.03
	random	**1.00**	**1.00**	**1.00**	**1.00**	0.32	**0.99**	**0.99**	**0.99**	**0.99**	0.09	0.80	**0.86**	0.79	0.77	0.01
Zhenhui Jiang and Benbasat (2007)	original	**1.00**	**1.00**	0.99	0.98	0.28	0.91	**0.94**	0.91	0.91	0.12	0.97	**1.00**	0.95	0.91	0.03
	misspecified	**1.00**	**1.00**	0.99	0.98	0.28	0.91	**0.94**	0.91	0.90	0.11	0.97	**1.00**	0.96	0.92	0.02
	random	**1.00**	**1.00**	0.99	0.98	0.28	0.91	**0.94**	0.91	0.91	0.12	0.78	**0.91**	0.80	0.82	0.02

Data by columns: (1) original data, (2) equal covariances,(3) mixed variables, (4) completely mixed variables,(5) random data.
for each data type are bolded and for each model type they are underlined.

5 Discussion and Conclusions

Recently both the use of PLS algorithm for theory testing and the general level of rigor used when conducting PLS analyzes have been under attack on several fronts: First, the ability of PLS method to generate valid measurement has been questioned and a commonly held belief that the method would control for measurement error has been shown to be incorrect (Rönkkö & Ylitalo 2010; Goodhue et al. forthcoming). Second, most of the currently used goodness of fit criterion have shown to be either problematic or inappropriate for detecting model misspecification (Evermann & Tate 2010; Henseler & Sarstedt forthcoming). Third, the use of too small samples and the general lack of attention to the assumptions of PLS model have been highlighted (Marcoulides et al. 2009; Marcoulides & Saunders 2006; Goodhue et al. forthcoming).

Since several models received support from all data but the *random data*, this paper provides evidence that PLS analysis will generally provide support for hypotheses regardless of the covariance structure in the data as long as the level of covariances is sufficiently high. Since PLS seems to provide support also for models clearly not related to the data, this means that the false-positive rate can be high and it is likely that many of the papers included in our analysis are indeed reporting results that are not valid. Our conclusion then is that either the method is altogether flawed for theory testing or the current practice of ignoring the requirements of the PLS analysis (large number of indicators, large sample size, model must be correct including the assumption that residuals must not correlate, all assumptions of each OLS regression are met) should be followed.

The large sample size and large number of indicators would be relatively easy to follow, and several test for model fit are available in the Lohmöller's (1989) PLS book. Particularly the analysis of residual correlations and using the SRMR index can be useful in detecting model misspecification and should be investigated in further research test the ability of these techniques to detect model misspecification and potentially to establish guidelines on recommended cut-off criteria. Testing the assumptions of each regression model can be done by exporting the construct scores form the PLS software and then re-estimating the regressions in a statistical package and using the diagnostics described e.g. in the regression analysis book by Cohen et al. (2003). Also testing the same model with different methods (e.g. regression with summed scales, or SEM with the maximum likelihood estimator) can increase the confidence in the PLS results.

Another alternative is to avoid the use of PLS altogether as a hypothesis testing tool. This is possible because any model that can be estimated with PLS can also be estimated with summed scales and regression analysis; the only difference is that instead of empirically determining the weights for the indicators, each indicator is weighted equally. A recent study by Goodhue et al. (forthcoming) showed that the results from summed scales and regression analysis are generally close to PLS estimates. The difference is that the summed scales are not as sensitive to correlated errors and well established and tested procedures exist for assessing measurement reliability and validity.

References

Aguirre-Urreta, M. et al., 2008. A Monte Carlo Investigation of Partial Least Squares, With Implications for Both Structural and Measurement Models. *AMCIS 2008 Proceedings*.

Bollen, K.A., 1996. An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, 61(1), pp.109–121.

Bollen, K.A., 1989. *Structural Equations with Latent Variables*, New York, NY: John Wiley & Son Inc.

Bollen, K.A. & Lennox, R., 1991. Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin*, 110(2), pp.305–314.

Chin, W.W., 1998. The partial least squares approach to structural equation modeling. In G. A. Marcoulides, ed. *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, pp. 295–336.

Chin, W.W., Marcolin, B.L. & 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. & Newsted, P.R., 1999. Structural equation modeling analysis with small samples using partial least squares. In R. H. Hoyle, ed. *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage Publications, pp. 307–342.

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S., 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, London: Lawrence Erlbaum Associates.

Compeau, D.R. & Higgins, C.A., 1995. Computer Self-Efficacy: Development of a Measure and Initial Test. *MIS Quarterly*, 19(2), pp.189–211.

Dijkstra, T.K., 1983. Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics*, 22(1-2), pp.67–90.

Enns, H.G., Huff, S.L. & Higgins, C.A., 2003. CIO Lateral Influence Behaviors: Gaining Peers' Commitment to Strategic Information Systems. *MIS Quarterly*, 27(1), pp.155–176.

Evermann, J. & Tate, M., 2010. Testing Models or Fitting Models? Identifying Model Misspecification in PLS. In *ICIS 2010 Proceedings*.

Fornell, C. & Bookstein, F.L., 1982. Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research*, 19(4), pp.440–452.

Fornell, C. & Larcker, D.F., 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18(1), pp.39–50.

Gefen, D., Rigdon, E.E. & Straub, D.W., 2011. An Update and Extension to SEM Guidelines for Administrative and Social Science Research. *MIS Quarterly*, 35(2), p.iii–xiv.

Gerow, J.E. et al., 2010. The Diffusion of Second Generation Statistical Techniques in Information Systems Research from 1990-2008. *Journal of Information Technology Theory and Application (JITTA)*, 11(4), pp.5–28.

Goodhue, D.L., Lewis, W. & Thompson, R., forthcoming. Comparing PLS to Regression and LISREL: A Response to Marcoulides, Chin, and Saunders. *MIS Quarterly*.

Goodhue, D.L., Lewis, W. & Thompson, R., forthcoming. Does PLS Have Advantages for Small Sample Size or Non-Normal Data. *MIS Quarterly*.

Hair, J.F., Sarstedt, M., et al., 2011. An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*.

Hair, J.F., Ringle, C.M. & Sarstedt, M., 2011. PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory & Practice*, 19(2), pp.139–152.

Hardin, A. & Marcoulides, G.A., 2011. A Commentary on the Use of Formative Measurement. *Educational and Psychological Measurement*, 71, pp.753–764.

Henseler, J. & Sarstedt, M., Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics*, forthcoming, pp.1–16.

Karahanna, E., Agarwal, R. & Angst, C.M., 2006. Reconceptualizing Compatibility Beliefs in Technology Acceptance Research. *MIS Quarterly*, 30(4), pp.781–804.

Kline, R.B., 2005. *Principles and practice of structural equation modeling* 2nd ed., New York, NY: The Guilford Press.

Komiak, S.Y.X. & Benbasat, I., 2006. The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly*, 30(4), pp.941–960.

Lehmann, E. & Casella, G., 1998. *Theory of point estimation.* 2nd ed., New York: Springer.

Lewis, W., Agarwal, R. & Sambamurthy, V., 2003. Sources of Influence on Beliefs about Information Technology Use: An Empirical Study of Knowledge Workers. *MIS Quarterly*, 27(4), pp.657–678.

Lohmöller, J.B., 1989. *Latent variable path modeling with partial least squares*, Heidelberg: Physica-Verlag.

Majchrzak, A. et al., 2005. Managing Client Dialogues During Information System Design to Facilitate Client Learning. *MIS Quarterly*, 29(4), pp.653–672.

Marcoulides, G.A., Chin, W.W. & Saunders, C., 2009. A Critical Look at Partial Least Squares Modeling. *MIS Quarterly*, 33(1), pp.171–175.

Marcoulides, G.A. & Saunders, C., 2006. PLS: A Silver Bullet? *MIS Quarterly*, 30(2), p.iii–ix.
McDonald, R.P., 1996. Path Analysis with Composite Variables. *Multivariate Behavioral Research*, 31(2), pp.239–270.
McLure Wasko, M. & Faraj, S., 2005. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, 29(1), pp.35–57.
Nunnally, J., 1978. *Psychometric Theory*, New York: McGraw-Hill.
Rönkkö, M. & Ylitalo, J., 2010. Construct Validity in Partial Least Squares Path Modeling. In *ICIS 2010 Proceedings*.
Tenenhaus, M., Amato, S. & Esposito Vinzi, V., 2004. A global goodness-of-fit index for PLS structural equation modeling. In *Proceedings of the XLII SIS Scientific Meeting*. pp. 739–742.
Thatcher, J.B. & Perrewe, P.L., 2002. An Empirical Examination of Individual Traits as Antecedents to Computer Anxiety and Computer Self-Efficacy. *MIS Quarterly*, 26(4), pp.381–396.
Wixom, B.H. & Todd, P.A., 2005. A Theoretical Integration of User Satisfaction and Technology Acceptance. *Information Systems Research*, 16(1), pp.85–102.
Zhenhui Jiang & Benbasat, I., 2007. Investigating the Influence of the Functional Mechanisms of Online Product Presentations. *Information Systems Research*, 18(4), pp.454–470.
Zimmerman, D.W. & Williams, R.H., 1977. The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology*, 16(2), pp.135–152.

Appendix 1: Effect of correlated errors

Due to space constraints we have written this appendix with the assumption that the reader is familiar with the PLS algorithm. If this is not the case, a good example can be found in e.g the widely cited book chapter by Chin (1998). When calculating the construct scores, PLS iteratively estimates the indicator weights based on how strongly the indicators correlate with a weighted sum of the indicators of constructs that are linked with a regression path (Hair, Ringle, et al. 2011). For a model with two constructs (A predicts B) with three reflective indicators each (a1, a2, a3, b1, b2, b3) this means that the indicators of A are weighted by how strongly they correlate with a weighted sum of indicators b1, b2, and b3. We can apply the formula for correlation of sums and write the correlation between indicator a1 and the composite variable estimate for *B* ($\hat{B}$) as a function of the current values of the indicator weights (*w*) and the correlations between the indicators (*r*):

$$r_{a1\hat{B}} = \frac{w_{b1}r_{a1b1} + w_{b2}r_{a1b2} + w_{b3}r_{a1b3}}{\sqrt{w_{b1} + w_{b2} + w_{b3} + 2(w_{b1}w_{b2}r_{b1b2} + w_{b1}w_{b3}r_{b1b3} + w_{b2}w_{b3}r_{b2b3})}} \quad (1)$$

The correlation between any two indicators is a function of the product of the total effects between the constructs and the indicator reliabilities, the direct error correlation between the items, and error correlations between an item and those constructs that either directly or indirectly cause the other item. The correlations between the indicators can be written as a function of standardized value of the regression coefficient (β), factor loadings (λ), and error correlations (*e*):

$$r_{a1\hat{B}} = \frac{\lambda_{a1}\beta(w_{b1}\lambda_{b1} + w_{b2}\lambda_{b2} + w_{b3}\lambda_{b3}) + e_{a1b1}w_{b1} + e_{a1b2}w_{b2} + e_{a1b3}w_{b3}}{\sqrt{\begin{matrix} w_{b1} + w_{b2} + w_{b3} + 2w_{b1}w_{b2}(\lambda_{b1}\lambda_{b2} + e_{b1b2}) \\ +2w_{b1}w_{b3}(\lambda_{b1}\lambda_{b3} + e_{b1b3}) + 2w_{b2}w_{b3}(\lambda_{b2}\lambda_{b3} + e_{b2b3}) \end{matrix}}} \quad (2)$$

Thee numerator in the equation reveals that the error correlations affect the correlations used in the indicator weighting process so that an indicator with more error correlations will receive a higher weight when forming the composite. Because the factor loading of a1 is multiplied by the product of the population regression coefficient and a weighted sum of the factor loadings of b1, b2, and b3 but the error correlations are multiplied by the indicator weights only, the effect of error correlations is actually higher unless the factor loadings and regression coefficient are very high. Moreover, if there is no effect between A and B in the population, the indicator weighting is completely determined by the error correlations.