PLS, Small Sample Size, and Statistical Power in MIS Research

Dale Goodhue
University of Georgia
dgoodhue@terry.uga.edu

William Lewis
Louisiana Tech University
wlewis@cab.latech.edu

Ron Thompson
Wake Forest University
ron.thompson@mba.wfu.edu

Abstract

There is a pervasive belief in the Management Information Systems (MIS) field that Partial Least Squares (PLS) has special abilities that make it more appropriate than other techniques, such as multiple regression and LISREL, when analyzing small sample sizes. We conducted a study using Monte Carlo simulation to compare these three relatively popular techniques for modeling relationships among variables under varying sample sizes (N = 40, 90, 150, and 200) and varying effect sizes (large, medium, small and no effect). The focus of the analysis was on comparing the path estimates and the statistical power for each combination of technique, sample size, and effect size. The results suggest that PLS with bootstrapping does not have special abilities with respect to statistical power at small sample sizes. In fact, for simple models with normally distributed data and relatively reliable measures, none of the three techniques have adequate power to detect small or medium effects at small sample sizes. These findings run counter to extant suggestions in MIS literature.

1. Introduction

This work is motivated by the authors' belief in three key assertions about the Partial Least Squares (PLS) statistical analysis technique. First, the use of PLS in behavioral research is predominantly an MIS phenomenon. Second, there is a pervasive belief among MIS researchers that *when the sample size is small*, PLS has special abilities that make it more appropriate than other techniques such as multiple regression and LISREL. Specifically, researchers often argue that PLS only requires a sample size of 10 (or even 5) times the most complex relationships within the research model. For example, if the most complex relationship involved a construct with four formative indicators, the argument would be made that the minimum sample size would be 40.

Third, research supporting the claim for PLS having greater efficacy at small sample size is inadvertently misleading the MIS research community, since it has, in effect, asked the wrong question. More specifically, that research has focused on accuracy rather than statistical significance. We argue that statistical significance is a primary consideration and accuracy a secondary one. Since MIS researchers have in a very real sense "championed" the use of PLS to the wider research community, if the third assertion above is true, it is important to clarify the issue of PLS and small sample size.

Because we are interested in comparing the output from different techniques rather than examining how they operate, we do not go into depth in describing PLS or the other techniques (multiple regression and LISREL) that we compare it with. Previous work has provided detailed descriptions of the various statistical analysis techniques, and in some cases has provided excellent comparisons between two or more (e.g., [1]; [4]; [5]; [7]; [10]; [11]; [13]). Interested readers wishing details on the various analysis techniques are encouraged to review these sources.

Our findings suggest that PLS does *not* have an advantage in terms of detecting statistical significance at small sample size. Further, the often cited and generally accepted "10 times" rule of thumb for the minimum sample size in PLS analysis ([1]; [5]) can lead to unacceptably low levels of statistical power.

It is important to note that our examination of these issues is done within a narrow research context. We use a relatively simple research model, measured with relatively strong (very reliable) indicators. Future research should extend our examination by including more complex models, less reliable indicators, and so on. The paper closes with implications of the findings and suggestions for future research.

0-7695-2507-5/06/$20.00 (C) 2006 IEEE

2. Background: The Three Assertions

2.1. Assertion 1 -- The use of PLS is predominantly an MIS phenomenon.

By asking knowledgeable researchers, we developed a consensus that three top research journals in the field of Management include: *Journal of Management*, *Academy of Management*, and *Organizational Behavior and Human Decision Processes*. For Marketing, three top research journals are: *Journal of Marketing*, *Journal of Marketing Research*, and *Journal of Consumer Research*. For IS, we selected *MIS Quarterly*, *Information Systems Research* and the *Journal of Management Information Systems*. We examined all articles published in each of the nine journals from 2000 to 2003 (inclusive), and determined which methodology was used for every article where one of the three statistical analysis techniques (multiple regression, PLS or some form of covariance-based structural equation modeling, such as LISREL or AMOS) was employed.

The results were surprising, even to us. In fact, only two articles (one in Management and one in Marketing) used PLS in these six journals during this four year period, and both of those cited MIS authors when they justified using PLS ([18] cited [1]; [19] cited [5]). In contrast, almost a third of the relevant MIS articles used PLS. This is powerful evidence that PLS has been wholeheartedly accepted as an important statistical method in the MIS field, but is still by far the exception in Management and Marketing. While we know that PLS is used in other fields, it is clear that the MIS field has adopted PLS in a way not matched by other behavioral disciplines.

2.2. Assertion 2 -- There is a pervasive belief among MIS researchers that PLS has special abilities at small sample size.

Below are quotes from two often-cited sources justifying the use of PLS with small sample sizes:

- Barclay et al. [1]: "It is this segmenting of complex models that allows PLS to work with small sample sizes. . . . Sample size requirements, using the 'rule of thumb' of ten cases per indicator, become ten times the number of predictors [in the most complex relationship]."
- Chin [5]: "If one were to use a regression heuristic of 10 cases per predictor, the sample size requirement would be 10 times [the most complex regression relationship in the model]."

These statements do not address the issue of statistical power directly. Chin and Newsted [7] also did not focus on the issue of statistical power, but they did address it as a secondary component of their work. More specifically:

- Chin and Newsted [7, pg. 327] repeat the statement above, and follow it by a comment that, for more accurate assessment, one should use the power tables provided by Cohen [8]. A few pages later (page 335): "Earlier it was noted that a regression involving four independent variables and medium effect sizes would need a minimum sample size of 80 [according to Cohen's tables]. Interestingly, the [Monte Carlo analysis of PLS] at sample size 50 still generated significant results."

MIS researchers appear to have interpreted these and similar statements to imply that PLS has advantages over techniques such as regression and covariance-based structural equation modeling (CB-SEM) for small sample sizes, including increased power. Below are typical quotes from papers justifying the use of PLS. These comments are by no means unique, and appear to reflect a prevailing perception among many MIS researchers that PLS has greater power for small sample sizes than regression or CB- SEM techniques such as LISREL.

- Kahai and Cooper [14, page 277], using sample size of 31: "One important benefit is the ability of PLS to be employed with less data than other structural modeling packages. As [9] indicate, PLS can be used in situations where there are at least five data points for each path leading to the construct that has the most incoming paths. The minimum amount of data for our analyses is 25, since there are four hypothesized and one control path leading to decision quality."
- Yoo and Alavi [20], using a sample size of 45: "We chose PLS among several structural equation modeling tools, including EQS, AMOS, and LISREL because, unlike other tools, PLS does not require a large sample size [1, 11]."

These comments suggest that the "10-times" heuristic ([1]; [5]), or even the "5-times" heuristic [9], is being used as the guide to the minimum sample size necessary to give sufficient power to detect relationships in PLS analysis. This heuristic is based at least in part on Monte Carlo analyses such as those reported in [7]. If MIS researchers who employ PLS are not using the "10-times" heuristic as a guide to power, they must be ignoring the issue, since they

typically don't explicitly address statistical power in any other way.

2.3. Assertion 3 -- The research on which the claim for adequate power for PLS at small sample size is based has, in effect, asked the wrong question.

There are really three important issues that are often not clearly distinguished in comments about PLS and sample size. The first issue is the question of whether the statistical technique will converge on a solution and avoid inadmissible results. The second issue is how close parameter estimates will be to the true parameter values. The third issue is the question of the power of the statistical approach – how likely a researcher is to find a statistically significant path estimate when a relationship really does exist in the underlying population. We will discuss each of these in turn.

2.3.1. Issue 1: Inadmissable Solutions. It is certainly true that non-convergence and inadmissible solutions (such as negative variances) are a potential problem for LISREL at small sample size ([5]; [13]). Neither PLS nor regression suffer from this problem, unless the sample size (N) is smaller than the number of incoming paths in the most complex relationship. For example, a regression with three independent constructs predicting a fourth construct will not produce results unless there are at least three data points. The same is true for PLS.

2.3.2. Issue 2: Accuracy of the parameter estimates. Much of the MIS Monte Carlo simulation research on PLS and small sample size has focused on the second issue – how close the estimated parameters come to the true value. Chin and Newsted [7], for example, looked at the performance of PLS and regression under a variety of conditions, in two separate studies. They concluded that "the PLS approach can provide information about the appropriateness of indicators at sample size as low as 20. Furthermore, it performed better than the simple summed regression with four or eight indicators [7, page 335]." By "better" they meant it produced parameter estimates that came closer to the true value.

2.3.3. Issue 3: Statistical Significance and Power of a Statistical Test. Statistical significance is well understood by researchers, with the standard being that unless there is less than a 5% chance of being mistaken, relationships between constructs should not be considered supported. Power is, arguably, less well understood and less carefully attended to in published research ([2]; [16]; [17]). The power of a statistical test is "the probability of rejecting H_0, when H_1 is true"[15]. In more basic terms, power is the probability that the researcher will find a statistically significant relationship, when the relationship is actually there. The power of a statistical test is reduced by (among other things) small sample size, a weak underlying relationship, or measures that are clouded by error (low reliability).

Given that in general it is harder to detect statistically significant relationships when sample size is small, and given that PLS is touted as having special abilities at small sample size, it is appropriate to ask whether PLS has more or less power at small sample sizes than regression or LISREL.

Obviously both accuracy and statistical significance are important to researchers. However, we contend that a perfectly accurate but "not statistically significant" estimate of a parameter cannot be assumed to carry scientific knowledge. The fact that the estimate is not statistically significant tells us that although the estimate may be positive, there is also a good chance the true value is zero. On the other hand, a somewhat inaccurate but statistically significant estimate of a parameter does carry scientific knowledge – specifically that there is little chance that the true value is zero. Since we cannot legitimately interpret the value of a parameter that is not statistically significant, we believe that statistical significance and power have to be primary considerations, and accuracy secondary.

To close this introduction, we summarize as follows. We agree with existing literature that suggests PLS and regression are less likely to produce inadmissible solutions than LISREL at small sample size, and that PLS appears to generate larger parameter estimates than regression. However we can find little evidence in the literature that PLS actually has greater power at small sample size (that is, greater ability to detect a path relationship as statistically significant). We argue that it is this issue of the power of the technique that is most important to MIS researchers. We conclude that the frequently implied "10 times" rule for sample size and the assertion that PLS has more power than other techniques at small sample size is in need of being tested more thoroughly.

We now introduce the research model used for our study, and describe the Monte Carlo technique that we employed.

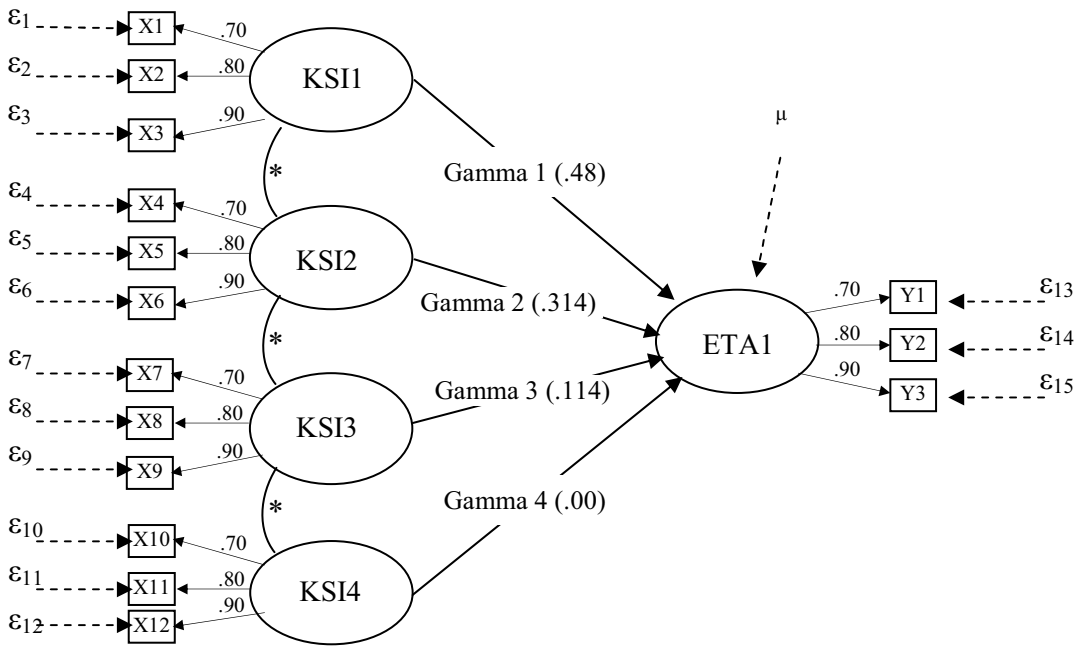

Figure 1. Our Basic Model

3. Methods

3.1. Monte Carlo simulation

The best way to address this issue is to use a Monte Carlo simulation approach, which has been used by numerous researchers to investigate questions such as how the size of the biases in PLS estimates compares to the size of the biases in LISREL estimates [3], or the impact of different correlation structures on the chi-square goodness of fit tests for structural equation modeling [12].

The Monte Carlo simulation approach requires that we start with a pre-specified "true" model that includes both the strength of the paths and the amount of random variance in the linkages (both between constructs, and between constructs and their indicators), such as is shown in Figure 1. Using random number generators and the relationships from this true model, we then generate multiple simulated data sets for each condition we wish to study. For example, we could generate 500 datasets with 40 cases each (i.e., the equivalent of 20,000 questionnaire responses in total), based on the model in Figure 1.

Here a "case" is simply a set of responses generated (using a random number generator) for the indicators in the model; each case corresponds to the questionnaire responses for one respondent. Suppose we wanted to start by looking at the performance of regression on this collection of datasets. We would analyze each of the 500 datasets (each having N=40) using regression. We would then have 500 different regression estimates of the value of a particular causal link, and 500 different regression t-statistic values for that link. Looking at the distribution of the resulting parameter estimates and the t-statistics, we could draw conclusions about the strengths of regression, at this sample size, for this model.

Again looking at Figure 1, suppose that after an analysis such as described above, we found that out of 500 datasets, regression found a statistically significant estimate for the Gamma1 link 253 times, or 50.6% of the time. We would conclude that regression had power of about .50 for this link and this sample size. It would be safe to assume that if we took one more sample of 40 from the same population, we would only have about a 50% chance of finding a significant result for that link. Further, if a researcher tests an equivalent model in the field using a sample size of 40, he or she has only about a 50% change of detecting that relationship, even if it is truly there. The whole process could be repeated for PLS and then for LISREL, and the results compared.

3.2. Chin and Newsted's (1999) study

Chin and Newsted [7] report the results from two different Monte Carlo simulation analyses. Though they used slightly different underlying models for the generation of the data, in both they used approximately the same approach, and in both they used only a strong effect size. In the first study they used 125 "cells" or sets of conditions for sample size, number of indicators, and number of latent variables; in the second study they used 36 sets of conditions. For each cell they did the following. First, they generated 100 distinct datasets for that cell. Then they ran PLS (or regression) 100 times, once for each of the 100 datasets. In this way they found 100 path estimates for each relationship in their model. They then compared the average path coefficients from PLS with those from regression, and with the true value of .40 (see [7, Appendix Tables 7 through 12, as referred to in 7, p331]). From this they concluded that PLS was generally closer to the true value.

3.3. The issue of statistical power

We would argue that the most important thing a researcher designing a study wants to know from such an analysis is, what is the *power* for a particular sample size and number of indicators. For example, if one single additional sample of size 50 with 4 indicators were analyzed, what is the likelihood of finding a statistically significant path. Although [7] analyzed 100 datasets in each cell, they did not report what proportion of the 100 resulted in statistically significant parameters; that is, they did not report the power for each cell.

Chin and Newsted [7] did calculate a statistical significance for the *average* path coefficient in each cell. To do this, for each cell they took the 100 estimated path coefficients, and determined a single average, a single standard error, and a single t-statistic value that summarized all 100 estimated path coefficients for that cell (again, see their appendix). They did not report or address in any way the 100 t-statistics of the path coefficients in each cell. Their single t-statistic for each cell of 100 datasets tells us something very different from power. It tells us, with a given significance level (say .05), whether a new researcher who gathered a new sample from this population could expect a *positive* path coefficient. Knowing that this new researcher could expect a positive path coefficient does not tell us whether he or she could expect a *statistically significant* path coefficient. In other words, it does not tell us the power of the test.

The studies conducted by Chin and Newsted [7]) provided considerable insight into the use of PLS under varying conditions, and a good comparison of PLS with MR, but they did not tell us whether PLS has greater statistical power than other techniques at small sample size. To address the issue of statistical power at small sample size, we designed an experiment using Monte Carlo simulation to extend the Chin and Newsted [7] research in three ways. First, we compared the three common analysis techniques of multiple regression, PLS and LISREL. Second, we created a research model that includes large, medium, small and zero effect sizes. Third, we examined the statistical significance of each path estimate individually, and counted the proportion in each condition that were significant at the .05 level. This gave us an accurate assessment of the power of each of the three techniques in each condition.

3.4. Designing the Monte Carlo model

The model in Figure 1 forms the basis for our analyses. The model has four predictor variables (KSI1, KSI2, KSI3 and KSI4) and a dependent variable (ETA1). The dependent variable and each predictor variable have three indicators. In addition to the path coefficients and the indicator loadings, random error is added to the indicator scores and to the value for ETA1, as shown in the diagram.

We selected the values for the links between the KSI constructs and ETA1 (the Gammas in the diagram), the random variance added to ETA1, and the random variance added to each indicator, based on the following requirements:

- KSI1, KSI2, KSI3, and KSI4 should be independent and normally distributed with unit variance.
- The effect size (measured without error) in the population for our four KSI constructs predicting ETA1 should be strong (.35), medium (.15), weak (.02) and zero (following Cohen's (1988) suggestions for these effect sizes).
- ETA1 should be normally distributed with a unit variance. This required the following values for Gamma1 through Gamma4: 0.48, 0.314, 0.114, and 0.00.
- Each of the five latent constructs should be measured by three indicators, and specifically with indicator loadings of .70, .80, and .90.
- The error variance added to each indicator should be set so that the variance of each indicator (sum of its true score component and its error) would be normally distributed with variance of one.

We selected the unequal values for the indicator loadings for two reasons. First, we wanted to use values that would be similar to what one would expect to see in MIS research studies, and especially studies employing PLS. Many researchers using PLS seem to follow the guideline for minimum loadings recommended by [1] of .707. Second, we wanted to move away from equal indicator loadings, both because real studies typically exhibit some differences and also because equal loadings might favor regression over PLS or LISREL. The result was that each construct had an underlying construct measurement reliability (Cronbach's alpha) of .84.

In other words, by construction (i.e. because we have defined it that way in our model), our constructs have good reliability. Similarly, because we have defined the model in that way, our constructs all have unidimensional measures (no systematic cross loadings.) The advantage of this approach is that it provides a common base model for testing each of the analysis techniques. The disadvantage is that the simplicity of the model could favor regression. We address this possible limitation later.

We chose the sample sizes to use in the study by the following logic. We selected 40 as the minimum reasonable size for PLS by using the "10-times" rule (10 times the number of antecedent constructs). We selected 200 as a conservative estimate of the minimum reasonable size for LISREL. We then partitioned the difference between the upper and lower bounds into three parts, giving us the following four sample sizes: 40, 90, 150, and 200.

We then used our model to generate 500 data sets of simulated questionnaire data for each of the four sample sizes. The data was generated using SAS and the SAS random number generator RANNOR. Each of the resulting 2000 data sets (500 data sets for each of four sample sizes) was then analyzed using multiple regression, PLS, and LISREL.

3.5. Data analysis and results

Chin and Newsted [7] observed that PLS provided estimates for path coefficients that were closer to the true values than those provided by regression. Our results (shown in Table 1) were similar. In addition, the estimates produced by LISREL were consistently larger than those produced by PLS, and generally were the closest to the true value. (Ignore the line labeled 'PLS-R' for now; we discuss this later).

Table 1. Results of Tests of Path Coefficients, Parameter Estimates

Gamma 1 --Strong Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*0.480*	*0.480*	*0.480*	*0.480*
MR	0.398	0.391	0.393	0.393
PLS	0.405	0.399	0.400	0.399
LISREL	0.486	0.485	0.484	0.484
PLS-R	0.403	0.398	0.399	0.399
Gamma 2 – Medium Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*0.314*	*0.314*	*0.314*	*0.314*
MR	0.255	0.258	0.254	0.256
PLS	0.273	0.270	0.263	0.262
LISREL	0.317	0.320	0.315	0.314
PLS-R	0.273	0.269	0262	0.262
Gamma 3 – Weak Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*0.114*	*0.114*	*0.114*	*0.114*
MR	0.099	0.099	0.096	0.089
PLS	0.104	0.114	0.107	0.099
LISREL	0.116	0.120	0.119	0.110
PLS-R	0.105	0.114	0.106	0.099
Gamma 4 – No Effect				
n=	**40**	**90**	**150**	**200**
Predicted	*0.000*	*0.000*	*0.000*	*0.000*
MR	-0.009	-0.001	-0.002	-0.003
PLS	-0.016	0.002	-0.002	-0.003
LISREL	-0.010	0.000	-0.004	-0.003
PLS-R	-0.017	0.003	-0.002	-0.003

Table 2 displays the results of our test for statistical power. We used t-statistics provided by regression and by LISREL, and bootstrapping with 500 resamples for PLS, to determine the statistical significance of each link in our model from Figure 1.

This gave us a separate t-statistic value for each of the 500 datasets within each cell. We then counted the number of datasets that were statistically significant at the .05 level, and divided by 500, giving us the proportion of times the true relationship for that link was detected; in other words, the power of the test.

Table 2. Results of Tests of Path Coefficients, Proportion Significant

Gamma 1 --Strong Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*0.79*	*0.99*	*0.99*	*0.99*
95% C.I.	(.75, .83)	(.98, 1.0)	(.98, 1.0)	(.98, 1.0)
MR	0.75	0.99	1.00	1.00
PLS	0.78	0.98	1.00	1.00
LISREL	0.69*	0.98	1.00	1.00
PLS-R	0.81	0.99	1.00	1.00
Gamma 2 – Medium Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*0.45*	*0.81*	*0.96*	*0.99*
95% C.I.	(.41, .49)	(.77, .85)	(.94, .98)	(.98, 1.0)
MR	0.40	0.76	0.92	0.97
PLS	0.41	0.76	0.93	0.98
LISREL	0.37*	0.79	0.95	0.98
PLS-R	0.50	0.83	0.95	0.98
Gamma 3 – Weak Effect Size				
n=	**40**	**90**	**150**	**200**
Predicted	*<25*	*<25*	*0.29*	*0.35*
95% C.I.	N/A	N/A	(.25, .33)	(.31, .39)
MR	0.09	0.16	0.27	0.29
PLS	0.09	0.16	0.25	0.29
LISREL	0.11*	0.19	0.28	0.31
PLS-R	0.21	0.28	0.39	0.40
Gamma 4 – No Effect				
n=	**40**	**90**	**150**	**200**
Allowable	*0.05*	*0.05*	*0.05*	*0.05*
95% C.I.	(.03, .07)	(.03, .07)	(.03, .07)	(.03, .07)
MR	0.06	0.05	0.04	0.03
PLS	0.07	0.05	0.04	0.03
LISREL	0.08*	0.05	0.04	0.04
PLS-R	0.14	0.12	0.11	0.09

(Note that N=40 is well below any recommended minimum size for LISREL analysis. In fact, 11 of the 500 datasets did not converge at this sample size, even though the underlying model has very well behaved data.)*

To understand the presentation of our findings, consider the entries under Gamma1 (strong effect size), with n=40. The expected power (the row labeled Predicted) is .79. That is, according to Cohen's power analysis for regression, we expect that 79% of the 500 datasets will have statistically significant links in a regression analysis. A standard equation for the confidence interval of a proportion suggests that for n=500 (there are 500 different samples and, therefore, 500 t-statistic values for Gamma1 in this cell), the 95% confidence interval around .79 is about .75 to .83.

Our analysis shows that with regression, 75% of the 500 runs had significant paths for Gamma1. For PLS it was 78%, and for LISREL it was 69%. (Ignore the line labeled 'PLS-R' for the moment; we return to it shortly). The values for regression and PLS are close to the predicted power level, but the value for LISREL was below the 95% confidence range. Note that the sample size of 40 is well below the minimum recommended for LISREL, and hence it is not surprising to see that some runs did not converge, and the overall power was below that desired. Since the values for regression and PLS are near the target for power of 80%, we would conclude that these two techniques have almost sufficient power to detect *strong* effect sizes with N equal or greater than 40. Since both are within the 95% confidence interval, we cannot conclude that one is better than the other.

Continuing with the strong effect size (Gamma1) and looking now at n=90, 150 and 200, we see that all three approaches had very strong power (98% to 100%). Here all three were within the 95% confidence interval for predicted power of regression based on Cohen's (1988) power analysis.

Moving to Gamma2 (medium effect size) at n = 40, we see that Cohen's (1988) power analysis predicts a power of 45% with a confidence interval of .41 to .49. PLS was at the bottom of this limit, but regression and LISREL were slightly below it. However, LISREL results at such small sample sizes must be viewed with some skepticism. In addition, it should be clear that a power of less than 50% is unacceptable, since there is only a fifty-fifty chance that true relationships will be detected. None of the techniques had even close to an acceptable level of power for a medium effect size at N=40.

For medium effect sizes with sample sizes of 90 and above, power is close to acceptable for all three techniques -- about 76-79% for N= 90, and above 90% for larger sample sizes. Again at N=90 and above, both regression and PLS had a level of power slightly below the 95% confidence interval.

For Gamma3 (weak effect size), note that *none of the techniques* had sufficient power (i.e., 80%), even at n = 200. Given the typical sample sizes used in IS research, we suspect most IS research does not have enough power to detect weak effects.

For Gamma4 (no effect), we note that all results fall within the 95% confidence interval around .05,

with the exception of LISREL at N=40. None of the techniques falsely identified significant paths that did not exist for sample sizes of 90 or above, over the normally acceptable limit of .05. The exception for LISREL at a size of 40 should be discounted, for the reasons discussed previously.

One possible criticism of our work to this point is that we are using two different methods for testing for statistical significance; normal theory testing for MR and LISREL, and bootstrapping for PLS. Our rationale for doing so is that we wished to compare the three techniques within the context of how they are normally used by MIS researchers.

To address this possible concern, we conducted additional analyses to see what impact (if any) there would be if we were to test PLS using normal theory testing. More specifically, we ran the PLS analysis for all 500 datasets in each sample size condition, then stripped off the indicator weights for each construct and used those and the raw data to determine construct scores for each data point (each data point can be thought of as a "questionnaire"). These construct scores were then fed into a regression analysis which estimated the betas and t-statistics for each of the 500 datasets, using normal theory testing. The path estimates are displayed in Table 1, and the proportion of t-statistics that are significant (i.e. the power) for each effect size and N are displayed in Table 2, under the label 'PLS-R'.

Three things are worthy of note in the results. First, the path estimates derived from PLS are almost identical, whether bootstrapping or normal theory testing is employed. Second, the power of PLS with regression significance testing (labeled PLS-R in Table 2) dominates all three other approaches for strong, medium and weak effect sizes at N=40, and dominates or is very close for those effect sizes at N=90. At sample sizes of 150 and above, this advantage seems to have disappeared and the power of PLS-R is generally similar to the other techniques and is within the confidence interval of expected power. On the face of it, this is strong evidence that PLS with normal theory significance testing is a more efficacious technique (has more power) than the other techniques at small sample sizes.

However, PLS-R also finds far more significant betas for Gamma 4, for which there is no actual effect. The other techniques all find between 3 and 7% of these false positives, within a 95% confidence interval around the allowable amount of 5%. PLS-R finds between 12 and 14% of these false positives for small and medium sample size, and .11 and .09 at N=150 and N=200, respectively. This is strong evidence that PLS-R detects an unacceptably high number of 'false positives'.

Our interpretation of why this occurs is the following. PLS has more "levers" available to it to capitalize on chance than regression. Regression can only vary the beta coefficients, while PLS can vary both the beta coefficients and the indicator weights. This gives PLS a stronger ability to capitalize on any chance high correlations of a particular indicator and the dependent construct. Especially with small sample size, often these chance high correlations come about through one or a few outlier data points.

Bootstrapping, because of the way it determines the standard error for significance testing, will react to such outliers with a larger standard error, since in the resamples sometimes the outlier data point will be included and sometimes it will not. However, PLS-R allows the PLS algorithm to capitalize on chance, and does not correct for this using bootstrapping. The result is an unacceptably high percent of false positives with PLS-R.

This suggests that the higher power of PLS-R for strong, medium and weak effect sizes may also be due to capitalization on chance. It further suggests that the approach of using PLS to determine indicator weightings and then using those weightings and indicators scores as input to a regression analysis is probably not appropriate.

No published work that we are aware of has advocated this approach. Obviously we need further investigation of this issue before offering any definitive statements, but, based on our results, we would not recommend employing it.

4. Limitations and Opportunities for Future Research

Certain limitations inherent in the study need to be acknowledged. First, the model that we used is quite simple, with four independent and one dependent variable, which may well favor regression. PLS may have more relative advantage when employed with more complicated models. Future research could expand the range of models to include those with mediating and moderating effects, as well as those with a larger number of antecedent constructs.

In addition, the data generated for use in this study were designed to be normally distributed and have relatively high factor loadings with little cross-loadings. While use of such well-behaved data created a level playing field for the three statistical techniques tested, actual field data often exhibits more challenging characteristics. Future studies could be designed to test the three techniques across a variety of data conditions, including low reliability of measures and indicators that cross-load (i.e., load on constructs

other than the one they are intended to measure). Furthermore, our study employed only measures that were reflective in nature. It would be useful to examine the impact of formative indicators [1]. PLS might have an advantage under those conditions.

Finally, future research could investigate the impact of using bootstrapping approaches for all three techniques, or provide a more thorough testing of standard normal theory testing for all three under more diverse conditions (e.g., complexity of the model, weaker reliability of measures, etc.).

5. Conclusions

5.1 Accuracy of estimates

Similar to other researchers [7] we observed that the average path coefficient estimates obtained from PLS were slightly closer to the true values than those obtained by regression. The LISREL estimates were the closest to the true values. However, as argued earlier, without statistical significance, accuracy contributes no scientific knowledge.

5.2 Statistical significance and power

Does PLS with bootstrapping provide more statistical power than other techniques at small sample size? Champions of PLS may take some comfort in the fact that for N = 40, PLS had 3% and 1% higher power than regression for strong and medium effect sizes, but on the other hand PLS had the same power as regression at weak effect size, and found 1% more false positives.

However, since the 95% confidence interval around these power values spans some 4 to 8 percentage points, none of these differences are statistically significant. In that sense we would have to say that there is no evidence that either PLS or regression has an advantage in terms of power at small sample size.

5.3 The "rule of 10"versus Cohen

Using commonly cited sources [1, 5] and the “10 times” rule of thumb, N = 40 is the recommended minimum sample size for this model with PLS. At N=40 with a weak or medium effect size, none of the techniques had even close to the 80% recommended power – all were less than 50%. Only for a strong effect size (and high reliability) did the "10 times" rule lead to acceptable power. On the other hand, Cohen's calculations for the power of regression analysis correctly predicted power in virtually all cases, for all sample sizes, all effect sizes and all techniques. The one exception was LISREL at small sample size, consistent with unanimous suggestions in the literature that LISREL requires greater sample size. At larger sample size, if any technique had the edge in terms of power, it was LISREL, though that edge generally did not put LISREL outside the 95% confidence interval.

This is strong evidence that, general beliefs in the MIS research community to the contrary, the "10 times" rule for sample size should not be used as a guideline when employing either PLS or regression for anything but a strong effect size with high reliability. Stated another way, the "10 times" rule does not take into account effect size, reliability, number of indicators, or other factors which are in one fashion or another are known to affect power. It is therefore a misleading guide for acceptable sample size. We suggest that the MIS research community should move quickly away from the claim that PLS has special abilities at small sample size, and away from the "10 times" rule.

We need to consider what this might mean in terms of existing published research that found significant results using PLS with small sample size. Since the small increase in false positives for PLS versus regression is not statistically significant, there is nothing in our findings to suggest that any previously reported statistically significant results found with PLS are false positives.

Now consider cases where the "10 times" guide to sample size were used, and no statistically significant results were found (in either published or unpublished studies). Our results clearly suggest that it would be incorrect to assume that the relationships tested do not exist. Because power was likely too low, these low power non-statistically significant results do not convey any scientific knowledge. Further, the interpretation of results using regression at small sample sizes would be exactly the same -- significant results are probably there; non-significant results convey no scientific knowledge.

We recognize that our conclusions are based on a quite simple model, with normally distributed data. However, this is not different from earlier Monte Carlo work on PLS and small sample size [7]. Additional research should explore more complex models, different levels of reliability, and non-normal data. However, we submit that if PLS has special abilities at small sample size, we should have seen evidence of those under the conditions used in this study.

Finally, we want to stress that PLS did not perform worse in terms of statistical power than the other techniques for normally distributed data, even though it seems to have no special abilities at small

sample size. It is still a convenient and powerful technique that is appropriate for many research situations, such as complex research models with sample sizes that would be too small for covariance-based SEM techniques. Unfortunately, however, PLS does not provide researchers with a magic bullet for achieving adequate statistical power at small sample sizes.

6. References

[1] Barclay, D., Higgins, C. and R. Thompson (1995). "The Partial Least Squares (PLS) Approach to Causal Modeling: Personal Computer Adoption and Use as an Illustration," *Technology Studies*, 2, 2, pp. 285-309.

[2] Baroudi, J. and W. Orlikowski. (1989). "The Problem of Statistical Power in MIS Research," *MIS Quarterly*, 13, 1, 87-106.

[3] Cassel, C., Hackl, P. and A. Westlund (1999). "Robustness of Partial Least-Squares Method of Estimating Latent Variable Quality Structures," *Journal of Applied Statistics*, 26, 4, May , pp. 435-446.

[4] Chatelin, Y.M., Vinzi, V.E. and M. Tenenhaus (2002). "State-of-art on PLS Path Modeling through the available software," unpublished working paper, IdE France.

[5] Chin, W.W. (1998). "The Partial Least Squares Approach to Structural Equation Modeling," in G.A. Marcoulides (Ed.) *Modern Methods for Business Research*, London, pp. 295-336.

[6] Chin, W.W. *PLS Graph User's Guide*, Version 3.0, Houston, TX: Soft Modeling, Inc., 2001.

[7] Chin, W. W., and Newsted, P. R. (1999). Structural Equation Modeling analysis with Small Samples Using Partial Least Squares**.** In Rick Hoyle (Ed.), *Statistical Strategies for Small Sample Research*, Sage Publications, pp. 307-341.

[8] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, L. Erlbaum Associates, Hillside, NJ.

[9] Falk, R.F. and Miller, N.B. (1992). *A Primer for Soft Modeling*. Akron, OH: University of Akron Press.

[10] Fornell, C. (1984). "A Second Generation of Multivariate Analysis: Classification of Methods and Implications for Marketing Research," Working Paper, University of Michigan, April.

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

[12] Fornell, C. and Larcker, D. (1981). "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," Journal of Marketing Research, 18, 39-50.

[13] Gefen, D., Straub, D. and M.C. Boudreau (2000). "Structural Equation Modeling and Regression: Guidelines for Research Practice," *Communications of the Association for Information Systems*, 4, article 7, October.

[14] Kahai, S.S. and Cooper, R.B. (2003). "Exploring the Core Concepts of Media Richness Theory: The Impact of Cue Multiplicity and Feedback Immediacy on Decision Quality," Journal of Management Information Systems, 20, 1, 263-299.

[15] Larzen, R.J, and Marx, M.L. (1981). *An Introduction to Mathematical Statistics and It's Applications*, Englewood Cliffs, N.J., Prentice-Hall, Inc.

[16] Mazen, A.M., Graf, L.A., Kellog, C.E. and M. Hemmasi. (1987). "Statistical Power in Contemporary Management Research," *Academy of Management Journal*, 30, 2, 369-380.

[17] Sawyer, A. and Ball, D. (1981). "Statistical Power and Effect Size in Marketing Research," *Journal of Marketing Research*, 18, 3, 275-290.

[18] Shea, C.M. and J.M. Howell (2000). "Efficacy-Performance Spirals: An Empirical Test," *Journal of Management*, 26, 4, 791-

[19] White, J.C., Varadarajan, J.P. and P.A. Dacin. (2003). "Market Situation Interpretation and Response: The Role of Cognitive Style, Organizational Culture, and Information Use," *Journal of Marketing*, 67, 3, 63-

[20] Yoo,Y. and M. Alavi. (2001). "Media and Group Cohesion: Relative Influences on Social Presence, Task Participation, and Group Consensus," *MIS Quarterly*, 25, 3, 371-390.