Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Research Note—Statistical Power in Analyzing Interaction Effects: Questioning the Advantage of PLS with Product Indicators

Dale Goodhue, William Lewis, Ronald Thompson,

To cite this article:
Dale Goodhue, William Lewis, Ronald Thompson, (2007) Research Note—Statistical Power in Analyzing Interaction Effects: Questioning the Advantage of PLS with Product Indicators. Information Systems Research 18(2):211-227. https://doi.org/10.1287/isre.1070.0123

Information Systems Research
Vol. 18, No. 2, June 2007, pp. 211–227
ISSN 1047-7047 | EISSN 1526-5536 | 07 | 1802 | 0211

informs

DOI 10.1287/isre.1070.0123

Research Note

Statistical Power in Analyzing Interaction Effects: Questioning the Advantage of PLS with Product Indicators

Dale Goodhue
MIS Department, Terry College of Business, University of Georgia, Athens, Georgia 30606,
dgoodhue@terry.uga.edu

William Lewis
College of Administration and Business, Louisiana Tech University, P.O. Box 10318,
Ruston, Louisiana 71272, william.w.lewis@gmail.com

Ronald Thompson
Babcock Graduate School of Management, Wake Forest University, Winston-Salem, North Carolina 27109,
ron.thompson@mba.wfu.edu

A significant amount of information systems (IS) research involves hypothesizing and testing for interaction effects. Chin et al. (2003) completed an extensive experiment using Monte Carlo simulation that compared two different techniques for detecting and estimating such interaction effects: partial least squares (PLS) with a product indicator approach versus multiple regression with summated indicators. By varying the number of indicators for each construct and the sample size, they concluded that PLS using product indicators was better (at providing higher and presumably more accurate path estimates) than multiple regression using summated indicators. Although we view the Chin et al. (2003) study as an important step in using Monte Carlo analysis to investigate such issues, we believe their results give a misleading picture of the efficacy of the product indicator approach with PLS. By expanding the scope of the investigation to include statistical power, and by replicating and then extending their work, we reach a different conclusion—that although PLS with the product indicator approach provides higher point estimates of interaction paths, it also produces wider confidence intervals, and thus *provides less statistical power* than multiple regression. This disadvantage increases with the number of indicators and (up to a point) with sample size. We explore the possibility that these surprising results can be explained by capitalization on chance. Regardless of the explanation, our analysis leads us to recommend that if sample size or statistical significance is a concern, *regression or PLS with product of the sums should be used instead of PLS with product indicators* for testing interaction effects.

Key words: interaction effects; moderator effects; regression; PLS; product indicators; statistical power; statistical accuracy; Monte Carlo simulation
History: Robert Zmud, Senior Editor; George Marakas, Associate Editor. This paper was received on June 4, 2004, and was with the authors 22 months for 4 revisions.

1. Introduction

Information systems (IS) researchers frequently need to test for the existence and strength of interaction effects between constructs measured with multiple items. Often these types of interactions are modeled as moderating constructs (see, for example, Venkatraman 1989) and tested using multiple regression. In this case, the interaction variable has typically been calculated as (the sum[1] of the items of the first con-

[1] Sometimes the average is used rather than the sum. This difference is not consequential for this work.

struct) multiplied by (the sum of the items for the second construct). This might be called a *product of the sums* approach. Unfortunately, in general the product of two measures is much less reliable than either of the two original measures. This may be part of the explanation for why so many IS studies testing interaction effects have failed to detect a moderating influence (Weill and Olson 1989).

In an article published in *Information Systems Research*, Chin et al. (2003) argued that a "product-indicator" approach for measuring interaction effects[2] could be used with partial least squares (PLS), and would be more effective in detecting interaction effects than the more common product of the sums approach previously employed (with both regression and PLS). Chin et al. (2003) tested their assertions using an extensive Monte Carlo simulation analysis that included analyzing 500 data sets in each of 36 different conditions of sample size and number of indicators. Their interpretation of their results was that PLS with the product indicator approach (hereafter referred to as PLS-PI) was superior to regression with a product of the sums approach (hereafter referred to simply as regression), in that the average PLS-PI interaction path coefficient was larger and closer to the true parameter value. In addition, they suggested that the ability to accurately estimate an interaction effect was strengthened by increasing the number of indicators, but was not strengthened, or was even weakened, by increasing the sample size above some minimum level (Chin et al. 2003, pp. 204–205).

Because this sample size result was counter to our understanding of the role of sample size in detecting effects, and because of our interest in the relative efficacy of different ways of handling interaction effects, we decided to replicate and extend their results. After considerable time working with this challenge, we came to the conclusion that the Chin et al. (2003) article, although an important first step in using Monte Carlo simulation to study this type of issue, did not take full advantage of the information available from the simulation. In fact, we came to believe that Chin et al. (2003) could be inadvertently misleading readers as to the efficacy of the PLS-PI approach, because they focused solely on accuracy of point estimates. With this study, we investigate the *statistical power* of the techniques, as well as the accuracy of point estimates.

1.1. Framing the Question

To keep this paper as brief as possible, we do not repeat material that is well described in the Chin et al. (2003) article. To frame our work, we focus on *three potentially interesting questions* that a researcher designing a study involving an interaction might consider. These are:

1. For a given number of indicators and a given sample size, what is the *expected value* of the point estimate for the interaction effect that the researcher will see, and how close will it be to the true value? (This first question addresses point estimate accuracy.)

2. What is the likelihood that the researcher will see a *positive*[3] interaction point estimate (regardless of statistical significance)?

3. What is the likelihood that the researcher will see a *statistically significant* positive interaction estimate? (This last question addresses the issue of statistical power.)

To answer these questions we use Monte Carlo simulation to focus on the differences in the efficacy of regression and PLS-PI. Using seven data sets provided to us by Chin et al. (2003) from the analyses reported in their *ISR* paper,[4] we replicate and extend their analysis. By comparing our findings with the published findings of Chin et al., we show that their analysis of the exact same data sets answered the first two of the above questions, but left the third unanswered; specifically, the statistical power of the technique. When we do answer the third question, we conclude that PLS-PI does produce larger point estimates of the interaction effect, but that fewer of

[2] Others have used a product indicator approach with LISREL (Kenny and Judd 1984, Joreskog and Yang 1996), but, to the best of our knowledge, Chin et al. (1996) were the first to suggest using it with PLS in their ICIS conference paper that served as a precursor to their 2003 paper in *Information Systems Research* (*ISR*).

[3] We recognize that a researcher could instead be hypothesizing a negative interaction effect; to simplify the presentation of our arguments we make reference to positive interaction effects only.

[4] We would like to express our sincere appreciation to Professors Chin, Marcolin, and Newsted for sharing a subset of their data with us.

these are statistically significant compared to regression. In other words, PLS-PI seems to consistently have less statistical power than regression.

We then focus on two additional questions that surface from these findings:

4. What could explain these seemingly contradictory results of greater point estimate accuracy but less statistical power?

5. Are there weaknesses in our research approach that might bring our findings into question?

In the course of answering these questions, we replicate a version of the central analysis done in Chin et al. (2003), looking at point estimate accuracy and statistical power for interaction effects across 36 different conditions of sample size and number of indicators, and using both PLS-PI and regression. We also conduct a series of ancillary analyses to get a sense of the robustness of our results.

Our additional analyses confirm the results of our earlier analysis of the seven data sets from Chin et al. (PLS-PI produces higher point estimates but has less statistical power), but the pattern of results also suggests that the power disadvantage of PLS-PI (versus regression) increases with the number of indicators and (up to a point) with the sample size. Further analyses suggest that the different ways in which PLS-PI and regression "capitalize on chance" (i.e., how they handle random variation in the data) lead to these results.

The primary contributions of our work are: (1) based on our results, we conclude that if having a sufficient sample size to achieve statistical significance for an interaction path is a concern, regression or PLS with a product of sums approach (as we will explain later) would be preferred to PLS-PI under most of the conditions (sample size, effect size, and number of and reliability of indicators) that management information systems (MIS) researchers typically encounter; and (2) the way that PLS handles random variations in data may lead (especially when many indicators are used for single constructs) to both overestimation of the strength of relationships and underestimation of their statistical significance. We provide suggestions for additional work to further investigate these and related issues.

2. Using Monte Carlo Simulation to Assess Point Estimate Accuracy and Statistical Power

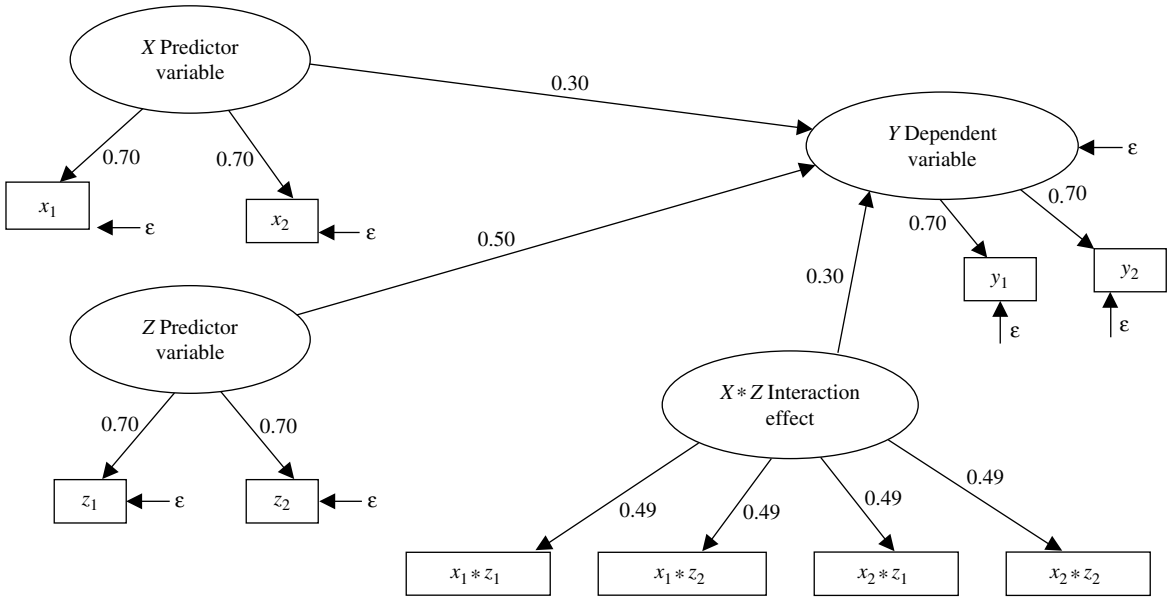
2.1. Monte Carlo Simulation

The Monte Carlo simulation technique has been used by numerous researchers to investigate questions such as how the size of the estimation bias in PLS compares to that in LISREL (Cassel et al. 1999), or the impact of different correlation structures on the chi-square goodness-of-fit tests for structural equation modeling (Fornell and Larcker 1981). The general approach is to define an underlying "true" model (such as the one shown in Figure 1) of relationships between constructs (including any random error), and then to use random number generators to generate simulated data, based on the model (or variations thereof).

Let's take a moment to examine the model in Figure 1 more closely. Figure 1 is adapted from the Chin et al. paper (2003, p. 198, Figure 2), and is the general model for all of the Chin et al. and our analyses. The model has two predictor variables (X and Z), an interaction variable ($X * Z$), and a dependent variable (Y). X and Z are related to Y as follows. The path coefficient for the first predictor variable (X) is set at 0.30, for the second predictor (Z) it is set at 0.50, and for the interaction effect[5] ($X * Z$) it is 0.30. In Figure 1, the loadings for all indicators (for X, Z, and Y) are set at 0.70. Random error terms are added to each indicator to give it a standard normal distribution with a mean of zero and a variance of one. For the Chin et al. (2003) study, and for this study, this basic model is modified in a number of ways to allow us to see how the statistical techniques perform under different circumstances. Specifically, the number of indicators for each of the X, Z, and Y variables varies across six

[5] There might be concern that multicollinearity is present in the analysis of such a model, because it would appear that $X * Z$ is correlated with X and Z. However, if X and Z both have $\mu = 0$, then X, Z, and $X * Z$ are all independent, as can be shown mathematically: $\mathrm{Cov}(X, X * Z) = \mathrm{E}[X * (X * Z)] - \mu_X * \mu_{X*Z}$ (Larsen and Marx 1981, Theorem 10.1). However, $\mu_X = 0$, and $\mathrm{E}[X * (X * Z)] = \mathrm{E}[X^2 * Z] = \mathrm{E}[X^2] * \mathrm{E}[Z] = 0$ because X, Z are independent and $\mathrm{E}[Z] = 0$. This result would not hold if there were some skew to the data, but is true as long as X and Z are centered, bivariate normal, and independent (Aiken and West 1991).

Figure 1 **Model with Two Indicators per Main Construct and Four Product Indicators for the Interaction Construct**



different values, and the sample size varies across six different values. In addition, there are different patterns of indicator loadings that are tested.

To conduct the tests using Monte Carlo simulation, we used two complementary approaches: (1) We used a subset of data from Chin et al. (2003), and (2) similar to Chin et al., we started with the model in Figure 1 and used an SAS program to generate additional simulated questionnaires that exhibit the desired properties. Appendix A[6] shows an example SAS program. Using this approach, it is possible to create simulated data with whatever sample sizes and properties are desired.

2.2. Power Analysis

Before discussing our replication of the analysis in Chin et al. (2003), we need to step back and clarify the relationship between statistical power and statistical significance. Statistical significance is generally well understood by researchers, with the standard being that unless there is less than a 5% chance of being mistaken (or 1%, depending on the desired protection against a type I error), relationships between constructs should not be considered supported. Power is, arguably, less well understood and less carefully attended to in published research (Baroudi and Orlikowski 1989, Sawyer and Ball 1981, Mazen et al. 1987). The power of a statistical test is "the probability of rejecting H_0, when H_1 is true" (Larsen and Marx 1981). In other words, power is the probability that the researcher will find a statistically significant relationship, when the relationship is actually there.[7]

When statistical power is low, there is a high likelihood that a researcher will not find statistically significant results, even when the relationship actually exists. Moreover, even the failure to find a relationship is not scientifically valuable, because low power means that researchers cannot have confidence that the relationship does not exist.

Exceptions aside, the general convention is that the power of a statistical test should be at least 0.80 (Cohen 1988, p. 56). In IS (Baroudi and Orlikowski 1989), in marketing (Sawyer and Ball 2001), and in management (Mazen et al. 1987), behavioral studies should generally be designed so that there is at least an 80% chance of finding relationships that exist. For a

[6] Appendixes are available online at http://isr.pubs.informs.org/ecompanion.html.

[7] Power is one minus the probability of a type II error, or one minus the probability of concluding there is no relationship when there is one. Statistical significance is the probability of a type I error—that is, the probability that a researcher will conclude that a relationship is there, when it really is not there.

more detailed discussion of how various factors affect statistical power (e.g., measurement error, reliability, the number of indicators for a construct, the effect size, and the sample size), please see Appendix B.

2.3. Clarifying the Difference Between the Two Interaction Approaches

Again consider Figure 1. The standard regression approach for testing an interaction would be to center all indicators, and then (for each data point or questionnaire) take the sum (or average) of the X indicators to obtain a value for X, do the same for Z, then multiply the value of X by the value of Z to give a value for I, the interaction term. X, Z, and I would then be regressed against Y. In contrast, the product indicator approach suggested by Chin et al. (2003) involves (after centering) taking the product of each possible combination of one indicator from the X construct and one indicator from the Z construct and using each of these "product indicator" terms as an indicator of the interaction. For example, in Figure 1, the interaction term would have four indicators: $x_1 * z_1$, $x_1 * z_2$, $x_2 * z_1$, and $x_2 * z_2$. PLS would be used to test the impact on Y of X and Z (each with two indicators) and the interaction I (with four indicators).

3. A Replication and Extension

To give a more concrete example, we will replicate and expand the analysis of a data set used by Chin et al. in their 2003 paper, and provided to us by Professors Chin, Marcolin, and Newsted. This data set is a variation of Figure 1. Chin et al. (2003) labeled this variation the A1 model (Chin et al. 2003, p. 208, Table 10), so we too will use that label. By using the exact same data as used by Chin et al., we can create a direct comparison of our results with theirs.

In the A1 model, the strength of the paths between X, Z, I (the interaction), and Y are all the same as those shown in Figure 1, but there are six indicators (not two) for each primary construct (36 for the interaction construct), and there are unequal loadings in the following pattern for X and Z: The first two indicators have a loading of 0.8; the second two have 0.7, and the last two have 0.6 loadings. The six indicators for Y all have loadings of 0.7, (see Chin et al. 2003, p. 207, Table 10). This is probably a moderately realistic example of the kinds of loadings MIS researchers typically encounter.

Using this A1 variation of Figure 1, Chin et al. generated 500 sets of 100 questionnaires (cases) each, for a total of 50,000 questionnaires. Each questionnaire contained values for the six indicators of latent construct X, the six indicators of latent construct Z, and the six indicators of latent construct Y. From this information, the 36 product indicators of $X * Z$ were calculated for use in PLS-PI, and the product of the sums was calculated for use in regression.

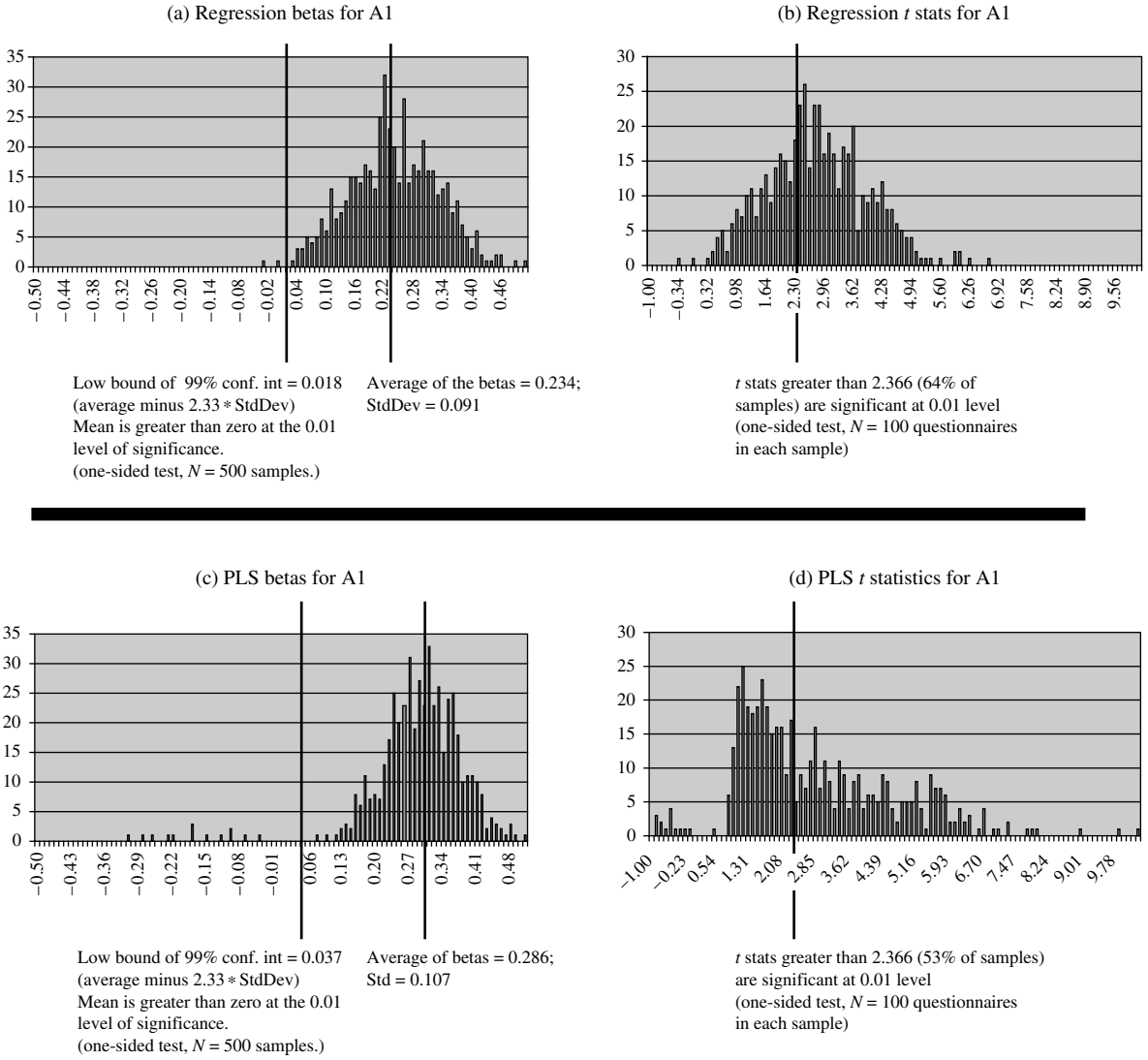
Both Chin et al. and we have analyzed this data as if it were the data from 500 different researchers, each of whom collected 100 questionnaires from the same underlying population. To begin with, we look at the results that would be obtained if all those researchers used regression to analyze their data. If the underlying truth is the six-indicator A1 version of Figure 1, we wish to know what a new researcher (say the 501st researcher) should expect to find when she analyzes her data. More specifically, we would like to answer the three questions stated in the introduction: First, what is the expected value of the interaction effect a new researcher would see and how close would it be to the true value? Second, what is the likelihood of seeing a *positive* interaction effect? Third, what is the likelihood of seeing a *statistically significant* positive interaction effect?

For the regression analyses, Figure 2(a) shows the frequency distribution of the 500 point estimates of the strength of the path from the interaction (I) to the dependent construct (Y). Figure 2(b) shows the frequency distribution for the 500 associated t statistics.

It may be helpful to orient the reader to the histogram in Figure 2(a) before we interpret it. Each vertical bar shows the number of researchers (out of 500) that found a particular beta value. For example, there were 32 researchers that found a beta value of about 0.22 (the longest bar on the graph). The bulk of the beta values found by the 500 researchers falls between about 0.04 and 0.46. (Recall that the true value from Figure 1 is 0.30.) The distribution of beta values is roughly normal, with an average value of 0.234 and a standard error of 0.091.[8]

[8] This dispersion of beta values around the average is due to the fact that Chin et al. (2003) introduced random error into their data generation process, as specified by the A1 variation of Figure 1, and as illustrated in Appendix A (online).

Figure 2 Frequency Distribution for Beta and *t* Stats, for Regression and PLS Using Data Set A1, with 500 Samples of 100 Questionnaires Each



Notes. Data provided by Chin et al. and described in Chin et al. (2003, A1 Condition of Table 10, p. 208). Conclusion: For these conditions, PLS has higher beta. Regression has more power, although neither has the recommended 0.80 power.

3.1. Answering the Three "Questions of Interest" Using Figure 2

Given the information displayed in Figure 2(a), we can answer two of the three key questions stated in the introduction of the paper. To answer the first question, we note that any new researcher using $N = 100$ in this population will likely find an interaction beta near 0.234, as indicated by the right-hand vertical line extending below the histogram, labeled as the "Average of the Betas." Because we know the true value is 0.30, this amounts to an average bias of minus 22% for regression analysis.

To answer the second question, we note that zero is outside the 99% confidence interval[9] around the

[9] For a one-sided test of whether the 500 beta estimates reflect an underlying parameter greater than zero, $n = 500$, $df = 499$, and the 0.01 critical value of the t statistic is 2.33. We note that Chin et al. used 2.36 in their analysis (2003, p. 206, Table 9), which would be the proper value for $df = 99$. This difference does not affect the results.

average (0.234 +/−2.33 ∗ StdDev). Any new researcher can, with a confidence level of 0.01, expect to find an interaction beta greater than zero. This is indicated by the left-hand vertical line extending below the histogram, labeled as "Low Bound of 99% Conf. Int = 0.018."

To answer the third question (what is the likelihood that any new researcher will see a *statistically significant* positive interaction effect—i.e., the power), we need to look at the distribution of the t statistics[10] from the 500 researchers, as shown in Figure 2(b). We note that the t statistics fall mostly in the range from about zero to six.

Now think of this from the point of view of a single researcher with a single sample of size $n = 100$. Note that the single researcher does not have at his or her disposal all the information we have displayed in Figures 2(a) and 2(b). The single researcher has only a sample of $n = 100$, a single beta value for the interaction, and a single t statistic. If she is using a one-sided test of significance at the 0.01 level, she must find a t statistic of at least 2.366 before she can say that the interaction beta is significantly greater than zero.[11] As can be seen in Figure 2(b), based on the regression analysis from each of the 500 researchers in our simulation, only 320, or about 64%, of the researchers found a t statistic that large. In other words, if one additional researcher were to sample the same population using six indicators with these loadings and a sample size of 100, that researcher has only about a 64% chance of finding a significant interaction effect, even though the path is really there. This is equivalent to saying that for these conditions, regression has a power of 0.64. Because it is generally recommended that researchers design their studies with at least a power of 0.80 (Cohen 1988, p. 56; Baroudi and Orlikowski 1989; Sawyer and Ball 1981; Mazen et al. 1987), we would say that this analysis with $N = 100$ has somewhat lower power than desired.

The issue of real interest in this paper is, how would the answers to the three key questions change if we move from regression to PLS-PI. Is PLS-PI actually better than regression at detecting the interaction effects? Consider Figures 2(c) and 2(d), which are the results of using PLS-PI[12] to analyze the exact same data.

Comparing Figures 2(a) and 2(c), we note that the average value of the betas from the 500 researchers is 0.234 for regression, and 0.286 for PLS-PI. At least for this A1 data, the PLS-PI point estimate for the strength of the interaction is higher than regression, and closer to the true value, with an average bias of only minus 5%. The Chin et al. analysis of this same data produced essentially the same results[13] (2003, p. 208, Table 10): 0.234 for regression and 0.285 for PLS-PI. In this light, PLS-PI seems better than regression in relation to the first question of interest.

For the second question, the likelihood that a researcher will encounter a *positive* beta coefficient, both Chin et al. and we looked at the average value of the 500 betas, and the standard deviation of the 500 betas, and calculated a single t value for the combined 500 betas. This t statistic is compared to the hurdle rate of $t = 2.33$, $p \leq 0.01$. For both Chin et al. and for our analysis, the t statistic for PLS-PI is larger than that for regression, indicating a greater confidence that any given researcher will find a *positive* interaction effect. Therefore, for both the first and second questions of interest, both Chin et al. and we would conclude that PLS-PI dominates regression.

The third question asks how likely it is that a single researcher (perhaps the 501st) will find a *statistically significant* positive interaction term. This third question can only be answered by looking at the distribution of the t statistics for the 500 different researchers (i.e., the 500 different samples). This is shown in Figure 2(b) for regression and Figure 2(d) for PLS-PI. To calculate the t statistics for PLS-PI as shown in Figure 2(d), we used bootstrapping with 100 resamples[14] for each of the 500 researchers.

[10] For consistency, we will talk about t values throughout, although above a sample size of about 100, t values and z values are indistinguishable (Larsen and Marx 1981, p. 296).

[11] For a one-sided test of whether the beta is greater than zero in a regression with a constant and three independent variables, $n = 100$, df = 96, and the 0.01 critical value of the t statistic is 2.366.

[12] We used PLS Graph v.3.0 (Chin 2001) for all our PLS analyses.

[13] There are slight but nonsubstantive differences between the results we obtained for PLS and those Chin et al. obtained, because of differences in the way data was rounded in processing.

[14] We deal in §6.1 with the question of whether 100 bootstrapping resamples are sufficient in this context.

What is striking in comparing Figures 2(b) and 2(d) is that although regression has an approximately normal distribution of t statistics, PLS-PI's distribution is highly skewed to the lower end. Whereas 64% of the 500 researchers found significant interaction paths using regression, only 53% of the 500 did when using PLS-PI. Remember that both approaches used the exact same input—the same 500 sets of 100 questionnaires each, as generated by Chin et al. We will return to the question of what causes this unexpected distribution of t statistics, but for now we simply note that for the A1 model and $n = 100$, PLS-PI has an advantage over regression in that it produces a higher point estimate of the beta value, but it has less ability to detect a significant interaction. That is, it has less statistical power—53% versus 64% for regression. Because power is a proportion, we can calculate a 95% confidence interval around the two power values: $53\% + / - 4.4\%$ and $64\% + / - 4.2\%$. Because these confidence intervals do not overlap, it is clear that regression has a statistically significant power advantage over PLS for the A1 data.

Of importance for our work, Chin et al. (2003) did not address this third question of interest. They did not use bootstrapping to calculate 500 t statistics, one for each of the 500 researchers. In fact, no significance testing at the level of the 500 individual researchers was reported, and presumably none was done. Instead, they calculated a single mean and standard error of those 500 betas. By using a single one-sided t test with $p < 0.01$, they were determining whether the mean was high enough above zero to conclude that the 501st researcher would find a positive value[15]—our "Question 2." However, this calculation does not tell the probability that the new beta will be statistically significant—our "Question 3." That is, it does not address statistical power. It fails to answer the question researchers designing a study need to ask: Does this test (sample size, effect size, etc.) give a power of 80%?

[15] Carte and Russell (2003, p. 482) describe this focus on the value of the interaction beta rather than the test of significance of the interaction beta as the first of nine common errors in analyzing interaction effects. They also point to an earlier version of the Chin et al. (2003) paper as an example of this error.

4. A Further Replication and Extension

4.1. Repeating the Analysis with Six More Data Sets from Chin et al.

Do the above results (higher beta estimates, but lower power for PLS-PI) apply only to the A1 data set, or do they apply more generally? In particular, because with regression we typically give equal weights to all indicators, and PLS assigns differentiated weights when there is more than one indicator for a construct, we might expect the balance to shift toward PLS when the true model includes more varied indicator loadings. We were able to test this proposition, again using data sets (labeled A1 through A7 in the top half of Chin et al. 2003, Table 10, p. 208) shared with us by Chin et al. Each of the seven data sets (A1 through A7) included 500 samples (i.e., 500 researchers), each of which used six indicators per main effect construct and $N = 100$. Each of the seven data sets used different indicator loading patterns. These, as well as the point estimate accuracy and power results, are shown in Table 1.

Our results for the point estimate for the interaction path are the same as reported by Chin et al. on their Table 10, verifying that we are using the exact same data they did, and we see that the pattern demonstrated in our Figure 2 is repeated in our Table 1. Generally, as the reliability of the main effect constructs goes up (due to higher average indicator loadings, see Columns 2 and 3), the power goes up for both PLS and regression. However, for any given reliability, PLS-PI yields higher point estimates for beta but less power than regression. Although the advantage of regression over PLS-PI in terms of power is reduced when widely different indicator loadings are used (for example, data sets A4, A5, and A7), even with the most extreme variation, regression still has higher power than PLS-PI for all of these data sets. Repeating the statistical test carried out earlier for the A1 data set, we find that the power difference between PLS-PI and regression is statistically significant at the 0.05 level for all but one (A7) of the seven data sets.

Table 1 PLS with Product Indicators vs. Regression (the Chin et al. (A1–A7 Data)

Data set	Main effect construct indicator loadings	Reliability (main effect constructs)	Power			Beta values		
			PLS	Reg	Regression advantage	PLS	Reg	PLS advantage
A1	2@0.8; 2@0.7; 2@0.6	0.85	0.53	0.64	**0.17**	0.29	0.23	**0.05**
A2	3@0.8; 3@0.7	0.89	0.55	0.69	**0.14**	0.29	0.25	**0.04**
A3	3@0.8; 3@0.6	0.85	0.45	0.62	**0.17**	0.28	0.23	**0.05**
A4	2@0.8; 2@0.6; 2@0.4	0.77	0.47	0.56	**0.09**	0.30	0.22	**0.09**
A5	3@0.8; 3@0.4	0.77	0.42	0.52	**0.10**	0.30	0.21	**0.09**
A6	3@0.7; 3@0.6	0.81	0.46	0.59	**0.13**	0.28	0.22	**0.06**
A7	2@0.7; 2@0.6; 2@0.3	0.70	0.34	0.41	**0.07**	0.30	0.19	**0.11**

4.2. Will the Pattern of Results Change If Integer Data Is Used?

Because we know that most survey research uses questions that have integer values (for example, disagree to agree on a -3 to $+3$ scale), we wondered if this pattern of results would also be apparent if the basic data was rounded to integers rather than being accurate out to four decimal places (as the Chin et al. data was). To test that, we rounded the X, Z, and Y indicator values from the A1 through A7 data sets to integers and repeated the analysis. Our detailed results are shown in Appendix C—online, but we will summarize those here.

Obviously, there is some loss of information as we go from real to integer data—power for both regression and PLS-PI is reduced by about three to five percentage points—but the pattern of results is unchanged. PLS-PI still has an advantage over regression in terms of higher point estimates for beta; regression still has an advantage over PLS-PI in terms of power. Widely different indicator loadings (for example A4, A5, or A7) reduce regression's power advantage, but do not eliminate it. Therefore, the pattern we found in Figure 2 and Table 1 will, in fact, be faced by researchers using questionnaires with integer data as well.

4.3. What Explains the Lower Power of PLS-PI?

Our finding that PLS-PI has higher point estimates and lower power than regression is certainly surprising. To try to understand the reason for our observed pattern of results, we first investigated what factors might reduce or increase the differences between regression and PLS-PI.

Our analysis of the A1 through A7 data (with sample size and number of indicators constant) demonstrated that in general higher reliability of main effect constructs increases the power of both PLS-PI and regression to detect interaction effects, but doesn't remove the advantage of regression. Therefore, we decided to follow the Chin et al. (2003) general approach in their main analysis and investigate the impact of sample size and number of indicators. Chin et al. held indicator loadings equal at 0.7 for each indicator, and investigated the impact of sample size ($N =$ 20, 50, 100, 150, 200, and 500) and number of indicators (2, 4, 6, 8, 10, and 12). We followed their lead, looking at the impact of sample size and number of indicators using the same set of six sample sizes and six numbers of indicators.

Although Chin et al. used real numbers out to four decimal places for their indicator values, we decided to use integer indicator values in our analysis. Our rationale was: (1) we have already shown with our analysis of the A1 through A7 data (in Appendix C—online) that moving from real to integer values does not affect the pattern of results, and (2) we believe that it is more appropriate to use integer values—almost all IS researchers use surveys with integer values and it is quite rare for them to use real values out to four decimal places. Therefore, our results will more closely resemble what most IS researchers will actually experience.

Our data for the 36 condition combinations (six different sample sizes and six different numbers of indicators) were generated using SAS and the SAS random number generator RANNOR.[16] All data

[16] The SAS RANNOR function is a random number generator returning random values distributed normally with mean of zero

were generated from the basic model shown in Figure 1, with enough random error variance added to each indicator to give it a total variance of one. Appendix A (online) shows the actual SAS program that was used to generate data for the condition of two indicators per main effect construct, and four indicators per interaction construct. As can be seen in the code in the appendix, we rounded all indicator values generated by the Monte Carlo process to integer values, which were generally in the range of −4 to +4. Each of the 36 condition combinations contained 500 data sets with the specified characteristics. We analyzed each data set using both PLS-PI and multiple regression (using the product of the sums approach). The results are shown in Tables 2 and 3. Table 2(a) (for PLS-PI) and Table 2(b) (for regression) show the average beta values for the 500 data sets in each condition, as well as the t statistic values (the estimates are on the top of each cell, with the t-statistics listed below).

Comparing our Table 2(a) and 2(b), we see the same pattern here as was revealed in our replication of the analysis of A1 though A7. *In every cell* PLS-PI has a higher point estimate for beta (averaged across all 500 samples) (Question 1). In addition, relative to whether a researcher will see a *positive* point estimate for the interaction beta (Question 2), we observe the same general pattern of cells as Chin et al.—cells where the t statistic is high enough that researchers can have 99% confidence they will find a positive beta are shown as bold in Tables 2(a) and 2(b). PLS has one more such cell than does regression. From these results one would conclude that PLS-PI has an advantage over regression in terms of higher point estimates for the interaction beta, and is as likely or more likely to find a *positive* interaction beta coefficient, the same results as reported by Chin et al. (2003).

To provide an indication of the size of the estimation error (i.e., how close the estimate came to the

and variance of one. RANNOR is used to generate the score for X. When the true score loading of x_1 from X is 0.7, adding a 0.714 loading for random error variance is just enough random variance to give x_1 a total variance of 1.0. Therefore, we set x_1 equal to $0.7*X+0.714*$RANNOR. This is the approach we used throughout to decide the amount of random variance to add to each indicator. See Appendix A (online) for more details.

Table 2 Comparing Interaction Point Estimates of Regression vs. PLS-PI (Equal Loadings, 0.70)

(a) PLS-PI interaction beta values

	Number of indicators for main effect constructs					
Sample size	2i $\alpha=0.66$	4i $\alpha=0.79$	6i $\alpha=0.85$	8i $\alpha=0.88$	10i $\alpha=0.91$	12i $\alpha=0.92$
20	0.14	0.24	0.25	0.30	0.30	0.30
	0.41	0.63	0.62	0.69	0.72	0.63
50	0.16	0.24	0.3	0.33	0.34	0.34
	0.85	1.05	1.38	1.54	1.54	1.28
100	0.17	0.25	**0.29**	**0.30**	**0.32**	**0.32**
	1.36	1.93	**3.05**	**2.59**	**3.01**	**2.83**
150	0.16	**0.24**	**0.27**	**0.29**	**0.30**	**0.31**
	1.86	**2.74**	**3.25**	**3.99**	**3.93**	**4.68**
200	**0.17**	**0.23**	**0.26**	**0.28**	**0.29**	**0.30**
	2.39	**3.39**	**4.36**	**4.86**	**5.23**	**5.46**
500	**0.16**	**0.21**	**0.24**	**0.26**	**0.27**	**0.27**
	3.71	**5.43**	**6.25**	**6.99**	**7.23**	**8.12**

Notes. Top is beta; bottom is t across 500 samples. Bold = researcher expect positive beta, $p < 0.01$.

(b) Regression interaction beta values

20	0.12	0.19	0.22	0.24	0.25	0.26
	0.46	0.77	0.89	1.01	1.11	1.08
50	0.15	0.19	0.22	0.24	0.25	0.26
	1.02	1.30	1.72	1.96	1.98	1.99
100	0.14	0.2	**0.23**	**0.24**	**0.25**	**0.25**
	1.46	2.10	**2.60**	**2.91**	**2.92**	**2.83**
150	0.14	**0.2**	**0.22**	**0.24**	**0.25**	**0.25**
	1.80	**2.83**	**3.10**	**3.27**	**3.46**	**3.66**
200	0.15	**0.2**	**0.23**	**0.24**	**0.25**	**0.25**
	2.21	**3.05**	**3.42**	**3.90**	**4.02**	**4.12**
500	**0.15**	**0.2**	**0.23**	**0.24**	**0.25**	**0.25**
	3.38	**4.77**	**5.51**	**6.14**	**6.57**	**7.10**

Note. Bold = researcher expect positive beta, $p < 0.01$.

(c) PLS-PI mean relative bias (%)

20	−0.53	−0.20	−0.17	**0.01**	**0.01**	0.00
50	−0.47	−0.20	0.00	**0.10**	**0.14**	**0.12**
100	−0.43	−0.17	−0.03	**0.01**	**0.07**	**0.07**
150	−0.47	−0.2	−0.10	−0.03	**0.01**	**0.03**
200	−0.43	−0.23	−0.13	−0.07	−0.03	−0.02
500	−0.47	−0.30	−0.2	−0.15	−0.11	−0.11

Notes. RMB; using equal loadings at 0.70. Bold if overestimate.

(d) Regression mean relative bias (percentages)

20	−0.60	−0.37	−0.27	−0.20	−0.17	−0.13
50	−0.50	−0.37	−0.27	−0.18	−0.18	−0.12
100	−0.53	−0.33	−0.23	−0.21	−0.16	−0.17
150	−0.53	−0.33	−0.27	−0.20	−0.17	−0.15
200	−0.50	−0.33	−0.23	−0.20	−0.16	−0.15
500	−0.50	−0.33	−0.23	−0.20	−0.15	−0.13

true value of 0.30 for the interaction term), we calculated the mean relative bias (Chin et al. 2003, p. 208). These values are shown in Table 2(c) (PLS-PI), and Table 2(d) (regression). PLS-PI has a lower absolute mean relative bias than regression in every case.

One unexpected result of this analysis is that there appears to be a region (shown in bold on Table 2(c)) where PLS-PI systematically overestimates the value of the beta for the interaction term, most prominently for $N = 50$ with 8, 10, or 12 indicators and for $N = 100$ with 10 or 12 indicators. If these cells were randomly spaced throughout Table 2(c), we might suppose this is due to random variation. However, because they are grouped together in a distinct region, it raises the question of whether in this region PLS-PI tends to overestimate. We note that Chin et al. show the same pattern, in the same region (2003, p. 204, Table 7, for 10 and 12 indicators, $N = 50$ and $n = 100$). Recall that in this analysis (as opposed to the analysis for A1 through A7 data sets) Chin et al. and we used different data sets, so the overestimation pattern is not a function of a particular set of data. We will briefly return to this issue in the discussion section of the paper.

When we address the third "interesting question" (statistical power, or whether the researcher will find a statistically significant beta—the question not addressed by Chin et al.), we also find the same pattern as we found with our A1 through A7 analysis. This is shown in Table 3(a) (PLS-PI power), Table 3(b) (regression power), and Table 3(c) (the regression power advantage). In 35 of the 36 cells in Tables 3(a), 3(b), and 3(c), regression has higher power (more of the 500 researchers found a statistically significant interaction beta) than does PLS-PI. In the one cell where it appears that PLS-PI and regression have equal power, if the results are shown to three decimal places, regression still has a slight advantage.

Is the power advantage of regression statistically significant? Repeating the statistical test used in §3.2 to determine whether the difference between the regression power and the PLS-PI power is statistically significant ($p < 0.05$) in any given cell, we find that it is in 16 of 36 cells of Table 3(c) (shown in bold). Once again, these statistically significant cells are in a particular region, suggesting a systematic effect rather than a random one.

Table 3 Comparing Power of Regression vs. PLS with Product Indicators

(a) PLS-PI, power

	Number of indicators for main effect constructs					
Interaction reliability	2i $\alpha = 0.66$	4i $\alpha = 0.79$	6i $\alpha = 0.85$	8i $\alpha = 0.88$	10i $\alpha = 0.91$	12i $\alpha = 0.92$
Sample size						
20	0.01	0.03	0.04	0.06	0.05	0.06
50	0.08	0.15	0.21	0.24	0.26	0.26
100	0.21	0.41	0.45	0.46	0.52	0.48
150	0.27	0.55	0.66	0.67	0.66	0.68
200	0.44	0.67	0.75	**0.81**	**0.82**	**0.84**
500	**0.83**	**0.97**	**0.99**	**1.00**	**1.00**	**1.00**

Note. Equal loadings at 0.70; bold = power > 0.80.

(b) Regression, power

20	0.03	0.08	0.09	0.10	0.10	0.15
50	0.10	0.17	0.30	0.32	0.33	0.43
100	0.21	0.50	0.58	0.63	0.72	0.71
150	0.32	0.67	0.78	**0.83**	**0.89**	**0.91**
200	0.48	0.77	**0.89**	**0.94**	**0.95**	**0.97**
500	**0.88**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**

Note. Equal loadings at 0.70; bold = power > 0.80.

(c) Regression power advantage over PLS-PI

Interaction reliability	2i $\alpha = 0.66$	4i $\alpha = 0.79$	6i $\alpha = 0.85$	8i $\alpha = 0.88$	10i $\alpha = 0.91$	12i $\alpha = 0.92$	Row average
Sample size							
20	0.02	0.05	0.05	0.04	0.05	0.09	0.05
50	0.02	0.02	0.09	0.08	0.07	**0.17**	0.08
100	0.00	0.09	**0.13**	**0.17**	**0.20**	**0.24**	0.14
150	0.05	**0.12**	**0.12**	**0.16**	**0.23**	**0.23**	0.15
200	0.04	**0.10**	**0.14**	**0.13**	**0.14**	**0.13**	0.11
500	0.05	0.02	0.01	0.00	0.00	0.00	0.01
Column average	0.03	0.07	0.09	0.10	0.12	0.14	

Note. Equal loadings at 0.70; bold = significant ($p < 0.05$).

Further, we can use a nonparametric approach to test a null hypothesis that overall PLS-PI and regression are equally likely to have larger power in any particular one of the 36 cells. If the truth is that they are equally likely to have higher power in any cell, how likely is it to find regression higher 35 of 36 times? This is equivalent to asking the probability of tossing a coin 36 times and getting heads 35 times, when the probability of getting heads is 0.50. Given a binomial distribution with a probability of 0.50, the probability of getting more than 34 successes out of 36 tries is 5.48E−10. In other words, it is *extremely* unlikely that we would see the results in Figure 3(c) if regression and PLS-PI had the same power. It is very

clear that regression has a statistically significant and persistent advantage over PLS-PI in terms of power. Whether the average advantage of 9% is substantively significant is a judgment call, but in our opinion it clearly is.

What is most interesting, however, are the marginal averages along the right side and bottom of Table 3(c) (the power advantage of regression over PLS-PI). Along the right side are averages for each sample size (across all numbers of indicators). Along the bottom are averages for each number of indicators (across all sample sizes). As one reads down the right side averages, it appears that the power advantage of regression goes up with sample size until about $N = 200$ (where regression power begins to top out at nearly 1.00, at which point the regression advantage begins to lessen as PLS power also approaches 1.00). As one reads across the bottom, the power advantage of regression clearly increases consistently with increases in number of indicators. We will speculate on what these patterns might mean next.

5. Discussion

Our results appear somewhat contradictory. On the one hand, we confirm what Chin et al. (2003) found—that PLS-PI produces a point estimate for the interaction beta that is consistently higher than regression. In fact, PLS-PI seems to systematically overestimate the beta path under certain circumstances. On the other hand, when we assess the ability to detect a statistically significant interaction coefficient, we find that regression has a definite advantage over PLS-PI. The regression power advantage seems to increase with sample size—to a point—and with the number of indicators.

After much contemplation on the overall results, we offer the following possible explanation. Our contention is that PLS-PI "capitalizes on chance" by taking advantage of PLS's ability to weight interaction indicators based on their correlation with the dependent variable Y. This has the effect of inflating the point estimates for the interaction path. Bootstrapping then compensates for this capitalization on chance, by exposing (rightly so) that the high point estimates for the interaction path are unstable. The result is higher standard errors, lower t statistics, and lower power for PLS-PI when beta values are raised by capitalization on chance. While not very apparent when there are only three or four indicators per construct, this phenomenon is amplified when there are many indicators for a particular construct. PLS-PI normally results in a large number of indicators for the interaction term, creating the condition for this phenomenon to occur. Our results shown in Table 3(c) support this interpretation.

Appendix D shows in much more detail an ancillary analysis of the possibility that PLS-PI may capitalize on change. In Appendix D, we start with a description of the three-stage estimation iteration process used by PLS, and show where in the process we might expect to see capitalization on chance. We then consider how such capitalization on chance would affect the results if there were a single outlier interaction indicator with a very high correlation with the Y construct, and describe why with PLS-PI we might expect to see both higher beta estimates than regression, and lower power. We then move away from the scenario of a single outlier data point and consider how random variation (which exists in all questionnaire data) could also lead to overestimated betas and lower power. The more indicators there are for the interaction construct, the greater the possibility of capitalizing on chance. Finally, we examine one particular PLS-PI run that exhibited strong indications of capitalization on chance. There we see that the correlations between the 36 product indicators and the dependent construct Y seem to drive the 36 indicator weights assigned by PLS. Further, the bootstrapping results for the statistical significance of the 36 indicator weights are all consistent with the suggested capitalization on chance mechanisms discussed in the first part of Appendix D.

Two other results in our analysis from the main paper are also consistent with capitalization on chance. First, both Chin et al. and we found that PLS-PI systematically overestimates the interaction beta in a certain region (that is, a certain set of sample sizes and numbers of indicators), as mentioned in §4.3 and shown in Table 2(a). If these overestimates were spread randomly throughout the cells, it would be explainable as random variation. The fact that these overestimates are consistently found in a defined region suggests there may be something in

this region that promotes capitalization on chance in PLS-PI. Second, as can be seen by comparing Tables 2(a) and 2(b), the point estimate accuracy advantage of PLS *declines* markedly as sample size increases. This is consistent with the idea that with larger sample size, it becomes more difficult to capitalize on random differences to increase beta point estimates. Although alternative explanations are possible, we believe that the evidence from our analysis does suggest that PLS capitalizes on chance.

5.1. Can We Curtail PLS' Ability to Capitalize on Chance?

The evidence suggests it is the large number of interaction indicators that gives PLS-PI so much scope to capitalize on chance. To further test this possibility, we decided to see the effect of using PLS with the more traditional "product of the sums" approach rather than "product indicators." In other words, we calculated a value for the product of the sums for PLS, exactly the same way we did for regression—weighting all indicators equally, determining scores for X and Z, and multiplying those scores together to give the product of the sums. Then we used PLS with 6 (or 8, 10, etc.) indicators for X, Z, and Y, and a single indicator for the interaction. This approach eliminated the possibility of capitalization on chance in the interaction path, because PLS could not give different weights to the 36 (or 64, 100, etc.) interaction indicators, but still allowed us to use the PLS method for all other aspects of the analysis.

We will refer to this approach as PLS-PS (where the "PS" stands for "product of the sums"). How would the PLS-PS approach affect our results in terms of power? The answer is shown in Table 4.[17] For the same six sample sizes and same numbers of indicators for each main effect construct, we show the power resulting from PLS-PS analysis (Table 4(b)), the comparable results for regression (Table 4(a), repeated from Table 3(b)), and the regression advantage (Table 4(c)). In fact, as can be seen from Table 4(c), any clearcut power advantage of regression disappears when PLS-PS is used.

[17] We also used PLS-PS to produce the path estimates for the interaction term, which were essentially equivalent to those produced by regression. For space reasons, these are not displayed here, but are available from the authors.

Table 4 Comparing Power of Regression vs. PLS with the Product of the Sums

(a) Regression, power

Sample size	Number of indicators for main effect constructs					
	2i	4i	6i	8i	10i	12i
20	0.03	0.08	0.09	0.14	0.15	0.18
50	0.10	0.17	0.30	0.34	0.34	0.46
100	0.21	0.50	0.58	0.63	0.72	0.72
150	0.32	0.67	0.78	**0.83**	**0.89**	**0.91**
200	0.48	0.77	**0.89**	**0.94**	**0.95**	**0.97**
500	**0.88**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**

(b) PLS-PS, power

20	0.06	0.04	0.05	0.10	0.09	0.11
50	0.11	0.18	0.29	0.34	0.38	0.45
100	0.22	0.51	0.60	0.66	0.72	0.71
150	0.33	0.66	0.76	**0.84**	**0.88**	**0.90**
200	0.47	0.77	**0.87**	**0.93**	**0.95**	**0.97**
500	**0.85**	**0.99**	**1.00**	**0.98**	**0.99**	**1.00**

(c) Regression power advantage over PLS-PS

Sample size	2i	4i	6i	8i	10i	12i	Row average
20	−0.03	0.04	0.04	0.04	0.05	0.07	0.03
50	−0.01	−0.01	0.01	0.00	−0.03	0.01	0.00
100	−0.01	−0.01	−0.02	−0.03	0.01	0.01	−0.01
150	−0.01	0.01	0.02	−0.01	0.01	0.02	0.00
200	0.01	0.00	0.02	0.01	0.01	0.00	0.01
500	0.03	0.00	0.00	0.02	0.01	0.00	0.01
Column average	0.00	0.00	0.01	0.01	0.01	0.02	

Note. Equal loadings at 0.70; bold = power > 0.80.

We have presented quite a large amount of evidence that supports the notion that PLS-PI does capitalize on chance, but that bootstrapping compensates (or overcompensates). Therefore, PLS-PI results in larger point estimates for beta, but lower power. This phenomenon seems most severe in the $n = 100$ to $n = 150$ range, and definitely increases with the use of more indicators. However, switching from the PLS with the product indicator approach (PLS-PI) to PLS with the product of the sums approach (PLS-PS) eliminates this power disadvantage.

6. Limitations, Additional Analyses, Opportunities for Future Research

Our results are surprising, because they lead to different recommendations from those obtained from

research previously published in this journal. In such cases, it is appropriate to be somewhat skeptical and to ask whether there is some flaw in our analysis that, if recognized, would affect and perhaps invalidate our interpretation. We have tried to ask those questions, and have understandably been assisted by our reviewers.[18] Below we address possible threats to our conclusions and present additional sensitivity analysis to determine how much our results might change under different circumstances. Our conclusion is that the general pattern of results is quite robust. We end the section with opportunities for future research.

6.1. Number of Bootstrapping Resamples

It might be suggested that we should use bootstrapping with 500 resamples (rather than 100). Five hundred resamples is the usual recommendation when using bootstrapping to estimate a parameter using a single sample (Chin 1998). However, we draw 500 samples (500 researchers) from the same population for each cell in our analysis, and use bootstrapping with 100 resamples on each of those. This amounts to 50,000 resamples for each cell, and hence we expect that moving from 50,000 to 250,000 resamples in each cell would not affect the outcome.

As a test of this assumption, we selected two cells ($N = 20$ and $N = 50$, with four indicators) for a spot check. We analyzed the same data using 100 resamples and 500 resamples for each of the 500 cases in each cell, and compared the results. The power of the interaction beta changed by 0.002 at most, not nearly enough to affect any of our results. Therefore, given the computational complexity burden already faced, we used 100 resamples per sample (or, effectively, 50,000 resamples in each cell). See Appendix E (online) for more details.

6.2. Bootstrapping vs. Normal Theory Testing

A second issue that might be raised is that comparing statistical significance tests based on normal theory testing with regression, versus bootstrapping with PLS, is like comparing apples and oranges—the two are not equivalent. Following this logic, we should, for example, use normal theory testing for both PLS and regression. We could do this by using indicator loadings generated by a preliminary PLS run to calculate construct scores, and then running regression with normal theory testing on these construct scores, giving us a normal theory test for PLS. These results could then be compared with straight regression (equal indicator loadings). We would counter that in published research, both PLS with bootstrapping and regression with normal theory testing are used to answer the same question (can we have 99% (or 95%) confidence that the hypothesized relationship exists?) and that the results of these two approaches are interpreted as if they are equivalent by researchers.

Nevertheless, to address this concern more directly, we conducted additional analyses to compare regression with normal theory testing to PLS with normal theory testing, along the lines suggested above. Consistent with Goodhue et al. (2006), who conducted similar comparisons in a noninteraction setting, we included data sets where the actual interaction effect was zero, to test for type I error. We observed that PLS with normal theory testing did achieve slightly higher power than PLS with bootstrapping under some conditions. However, we also observed that when the true effect was zero, PLS with normal theory testing resulted in over 30% false positives for the three cells we tested. That is, about a third of the time (in about 500 of 1,500 runs) PLS with normal theory testing found a statistically significant result where there was no actual effect. This is clearly an unacceptably high type I error rate, because 5% is the usual acceptable upper limit of false positives. Because we have never seen the use of PLS with normal theory testing advocated in published research, obviously we (or others) would need to explore this issue much more thoroughly and systematically before offering any definitive guidelines. However, given our initial results and absent any other careful research, we would strongly discourage its use. See Appendix F (online) for more details.

6.3. Nonnormality

It is possible that the nonnormality of the interaction term may have influenced our results. Recall that regression requires an assumption of normally distributed error terms. Regression is thought to be

[18] We would like to thank the anonymous reviewers for a number of these suggestions, each of which made the study stronger.

relatively robust to violations of this assumption, although kurtosis (more or less peaked) is thought to be more problematic than skewness (biases toward one side or the other) (Neter and Wasserman 1974, p. 513). Although our data for the X, Z, and Y indicators were normally distributed (by design), the interaction data that were computed as the product of two normally distributed variables were highly kurtotic (in both PLS-PI and regression analyses). To see if this had any impact on the results, we selected three cells ($N = 20$, 50, and 100 for the four-indicator data). We then transformed the interaction data in those cells (by taking the square root) to reduce the level of kurtosis to be within the normally accepted range. We reran both PLS-PI and regression with the transformed data, and compared the results to our original results. The differences were negligible (at most a change of 0.02 in power), suggesting that the nonnormality of the interaction terms for both PLS-PI and regression did not materially affect our results. See Appendix G (online) for more details.

6.4. Future Research Directions

It would certainly be valuable to have others confirm our findings that PLS-PI produces higher interaction beta estimates, but lower statistical power. Assuming that is confirmed (as we are confident it will be), there should be further exploration into the reasons for these results, including further investigation of our suggested mechanism for how the PLS algorithm creates these outcomes when there are many indicators. Such investigations would provide the basis for greater clarity about the research conditions under which PLS use is most appropriate.

On another front, it would certainly be useful to have additional comparison analyses using actual data sets from practice, as opposed to data sets generated from simulations. Although no single data set could be expected to demonstrate the overall pattern, if a collection of data sets were analyzed we would expect the same pattern to emerge. However, we note that the earlier work advocating PLS-PI (Chin et al. 2003) also used simulated data, so our work is not unique in that respect.

A third opportunity for future research would be to investigate the impact of adjusting regression beta values to account for unreliability. For interaction betas that were shown to be significantly different from zero, this might be one way to get closer to the true values. A fourth opportunity would be to address the issue of the use of formative, rather than reflective, indicators. In this work (as with Chin et al. 2003), only reflective indicators were specified.

Finally, Chin et al. (2003) primarily used the significance target of 0.01 instead of the more common 0.05 as the focus of their discussion. To be consistent with their work, we have followed their lead in all our comparisons above. However, most researchers would probably be satisfied with an expectation that they would find paths statistically significant at the 0.05 level. With that in mind, we suggest that future research in this domain use the target of 80% power for 0.05 levels of statistical significance. This would not change the general pattern apparent in our results, but would reduce the sample sizes needed to achieve 80% power. See Appendix H (online) for the results of our main analysis, recast in terms of $\alpha = 0.05$ significance hurdles.

7. Conclusion

In June 2006, the editor-in-chief of *MIS Quarterly* called for a careful reexamination of beliefs about strengths and weaknesses of PLS (Marcoulides and Saunders 2006). Our study fits within the general scope of that call. PLS with product indicators has been proposed as a more effective way of detecting interaction effects (Chin et al. 2003). By carefully considering the statistical power implications of the PLS-PI approach for estimating interaction effects, and replicating and extending the work of Chin et al., our work provides evidence that suggests that a recalibration of our beliefs about PLS-PI is necessary. It indicates quite persuasively that PLS with product indicators, although consistently producing higher average point estimates of the interaction path than regression, also has a persistent disadvantage in terms of statistical power relative to regression or PLS with product of the sums.

Our proposed explanation is that PLS-PI's lower power (and perhaps even its higher average point estimates) may be the result of capitalizing on chance when there are many indicators for the interaction term, in which case the researcher will pay a price in terms of lower t statistics. PLS-PI always creates

many interaction indicators—16 interaction indicators when there are 4 indicators for each main effect, 36 indicators with 6 indicators per main effect, etc. Thus, PLS-PI creates the exact situation that seems to lead to capitalization on chance, and thus produces a power disadvantage that grows with the number of indicators.

Regardless of the explanation, the implications of these findings differ for two different categories of researchers. For researchers designing a study that includes interaction effects, the reduced power of PLS-PI drives up the sample size needed for acceptable power. For example, assume we wished to test a model such as shown in Figure 1 (a 0.30 interaction path and four indicators per main effect construct). Using interpolation from Tables H-1 and H-2 in Appendix H (online), we can determine the sample size needed for a power of 0.80 with a 0.05 level of significance for this analysis: 200 for PLS-PI and 126 for regression. (Tables 3(a), 3(b), and 4(b) could be used, but this is based on a very high significance hurdle of 0.01). Therefore, for this analysis, PLS-PI requires either using a larger sample size or accepting a loss in statistical power.

Our analysis strongly suggests that if sample size is an issue, or if there is doubt about whether there is sufficient power, PLS-PI should not be the preferred approach for studying interaction effects. The choice between regression and PLS-PS might depend upon the complexity of the model, because PLS-PS can more easily handle much more complex models.

For researchers who have already completed analyses using PLS-PI, or those who are reading and interpreting published studies employing PLS-PI, there is a different set of insights. First, if PLS-PI was used and statistically significant results were found, these significant results are in no way suspect. There is nothing in our findings to suggest that statistically significant results found using PLS-PI are any less valid than significant results found using other statistical tests. However, suppose the PLS-PI analysis resulted in paths that did not achieve statistical significance. Our results suggest it would be a mistake to interpret this as evidence that the path does not exist, unless it was clear that the sample size was sufficient to give reasonable power. If the PLS-PI results were close to significant, we would even suggest rerunning the analysis using a technique with greater statistical power—either regression or PLS-PS.

What do our results say about PLS in general? They certainly raise the (disturbing and intriguing) possibility that PLS, under certain circumstances, might capitalize on chance and suffer a loss of statistical significance because of it. As our study suggests, this problem seems most clearly apparent when 16 or more indicators are used for a single construct. Under more ordinary conditions with four or fewer indicators per construct, PLS remains a powerful and easy to use tool for statistical analysis. Unfortunately, testing for interactions using PLS-PI creates the exact conditions favoring this capitalization on chance.

References

Aiken, L. S., S. G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, Beverly Hills, CA.

Baroudi, J., W. Orlikowski. 1989. The problem of statistical power in MIS research. *MIS Quart.* **13**(1) 87–106.

Carte, T., C. Russell. 2003. In pursuit of moderation: Nine common problems and their solutions. *MIS Quart.* **27**(3) 479–501.

Cassel, C., P. Hackl, A. Westlund. 1999. Robustness of partial least-squares method of estimating latent variable quality structures. *J. Appl. Statist.* **26**(4) 435–446.

Chin, W. W. 1998. The partial least squares approach to structural equation modeling. G. A. Marcoulides, ed. *Modern Methods for Business Research*. London, UK, 295–336.

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

Chin, W. W., B. Marcolin, P. Newsted. 1996. 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. *Proc. 17th Internat. Conf. Inform. Systems*, Cleveland, OH, 21–41.

Chin, W. W., B. Marcolin, P. Newsted. 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. *Inform. Systems Res.* **14**(2) 189–217.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillside, NJ.

Fornell, C., D. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *J. Marketing Res.* **18** 39–50.

Goodhue, D., W. Lewis, R. Thompson. 2006. Small sample size and statistical power in MIS research. R. Sprague, ed. *Proc. 39th Hawaii Internat. Conf. Systems Sci.*, (*CD*), IEEE Computer Society Press, Los Alamitos, CA, 1–10.

Joreskog, K. G., F. Yang. 1996. Nonlinear structural models: The Kenny-Judd model with interaction effect. G. A. Marcoulides, R. E. Schumacker, eds. *Advanced Structural Equation Modeling, Issues and Techniques*. Lawrence Erlbaum Assoc., Mahway, NJ, 57–88.

Kenny, D. A., C. M. Judd. 1984. Estimating the nonlinear and interactive effects of latent variables. *Psych. Bull.* **96** 201–210.

Larsen, R. J., M. L. Marx. 1981. *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Marcoulides, G. A., C. Saunders. 2006. PLS: A silver bullet? *MIS Quart.* **30**(2) iii–ix.

Mazen, A., M. Magid, M. Hemmasi, M. F. Lewis. 1987. Statistical power in contemporary management research. *Acad. Management J.* **30**(2) 369–380.

Neter, J., W. Wasserman. 1974. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* Richard D. Irwin, Inc. Homewood, IL.

Sawyer, A. G., A. D. Ball. 1981. Statistical power and effect size in marketing research. *J. Marketing Res.* **18**(3) 275–290.

Venkatraman. 1989. The concept of fit in strategy research: Toward verbal and statistical correspondence. *Acad. Management Rev.* **14**(3) 423–444.

Weill, P., M. H. Olson. 1989. Managing investment in information technology: Mini case examples and implications. *MIS Quart.* **13**(1) 3–17.