MIS Quarterly

SPECIAL ISSUE

A CRITICAL LOOK AT PARTIAL LEAST SQUARES MODELING

By: **George A. Marcoulides**
University of California, Riverside
george.marcoulides@ucr.edu

Wynne W. Chin
University of Houston
wchin@uh.edu

Carol Saunders
University of Central Florida
carol.saunders@bus.ucf.edu

The first seed for this special issue on partial least squares (PLS) modeling was sown when George A. Marcoulides, editor-in-chief of *Structural Equation Modeling* (*SEM*), submitted an Issues and Opinions article to *MIS Quarterly* (*MISQ*) about the presumed immunity of the method to distributional assumptions and that urged caution when using PLS with small sample sizes. As editor-in-chief of *MISQ*, Carol Saunders started a review process for the paper that turned out to be more problematic than any she had ever seen. Several reviewers turned down the invitation to review. Several reviewers who accepted the invitation never completed their reviews either because of personal problems that had arisen in their families or for unstated reasons. The few reviewers completing the assignment found the paper either too advanced for the MIS target audience or, concomitantly, conveying a message that was so obvious that the paper did not warrant publication. More so than other disciplines, the Information Systems discipline relies heavily on PLS for testing path models. Goodhue, Lewis, and Thompson (2006) found that PLS was used almost one third of the time in three top MIS journals between 2000 and 2003 (inclusive) when testing such models. Carol, believing the cautionary message was an important one, offered to work with George to issue a joint editorial to share that message with the IS community (Marcoulides and Saunders 2006).

Both Carol and George saw much merit in the collaborative effort and thought it would be a good idea to get the *SEM* and *MISQ* audiences talking with one another. The idea of a joint call for papers took root with the goal of providing a forum for exploring PLS within the broad context of Management Information Systems research and to create greater dialogue across the two communities. The stated hope was to advance the exploration of theoretical issues for effectively applying PLS and interpreting its results. This was the first joint call for papers in either journal's history. Wynne Chin, who developed the first graphical-based PLS software and who is recognized for his expertise in PLS, was invited to come aboard as a coeditor of the special issue.

A total of 23 papers were submitted for review. Three were routed to *SEM* and the remaining papers were routed to *MISQ*. A list of researchers who agreed to serve as associate editors and reviewers is provided in Appendix A. Within 18 months of the special issue submission deadline, two papers were accepted and one paper was conditionally accepted for publication in *MISQ*. In this introduction we summarize the two articles included in this Special Issue section. As stated in the call for papers, their abstracts are published in *SEM*. Because the sticking point for the third paper (that was subsequently withdrawn), and for that matter a good number of other rejected papers, relates to attempts at comparing the results of PLS, SEM (generically used to refer to path analysis with latent variables or covariance-based models), and multiple regression, we also briefly comment on issues that must be considered when conducting such comparisons. We conclude with some final thoughts about directions for future research.

An Overview of PLS Articles

We believe that the two articles published in this issue can expand the appropriate use of PLS. The article by Wetzels, Odekerken-Schröder, and van Oppen, entitled "Using PLS Path Modeling for Assessing Hierarchical Construct Models: Guidelines and Empirical Illustration," demonstrates how PLS can be used to create hierarchical constructs (Note: we return to comment on the use of this terminology in a section below). In terms of higher-order construct models, the literature typically models only second-order hierarchical structures. So this article presents a contribution by describing an approach for constructing even higher-order constructs. For example, it illustrates the application of the approach on a fourth-order, reflective, hierarchical construct model. The authors' guidelines describe four steps which are nicely summarized in Table 1 to help facilitate their application. Higher-order constructs are argued to allow greater theoretical parsimony and reduce model complexity. They also make it possible to further measure specificity, which refers to matching the level of abstraction for predictor and criterion variables. Empirically, higher-order constructs may exhibit a higher degree of criterion-related validity.

The second article, "Assessing Between-Group Differences in Information Systems Research: A Comparison of Covariance- and Component-based SEM," by Qureshi and Compeau studies the conditions that are most appropriate for applying alternative approaches to determine the presence or absence of between-group differences. In particular, the authors apply Monte Carlo simulation to examine the strength of moderating effects when using the PLS pooled significance test for multi-groups and SEM multigroup analysis. They find that the component-based PLS approach is more likely to detect differences between-groups than the covariance-based SEM approach when data are normally distributed, sample size is small, and exogenous variables are correlated. Both approaches consistently detect differences under conditions of normality with large sample sizes, and neither technique consistently detects differences across the groups for all paths with non-normally distributed data. However, for some paths under conditions of non-normality, moderate effect sizes, and smaller samples, the PLS approach appears preferable. The paper presents a helpful decision tree when selecting an approach to study between –group differences.

Comparison Across Approaches

Both articles on PLS in this issue make different types of comparisons between PLS and SEM methods of analysis. In the Wetzel et al. piece, the comparison is based solely on theoretical characteristics of the modeling approaches. The Qureshi and Compeau piece creates a baseline case with fixed path coefficients in each of the approaches and then generates comparison models to produce differences in path coefficients at five different effect sizes. That is, for both the PLS and SEM approaches the baseline model (with predetermined values for each of the path coefficients) is judged against five comparison models. Using Monte Carlo simulation, the authors generated 500 replications. They mainly tested differences in the structural model and took steps to assure measurement model invariance.

The results reported in both studies are consistent with previous research investigations, particularly those in which model errors are explicitly taken into account and handled (e.g., Dijkstra 1983; Mathes 1993; McDonald 1996; Schneeweiss 1993). We note that in cases where the model errors are not explicitly taken into account for the estimation of endogenous latent variables, a new approach proposed by Vittadini et al. (2007) would need to be used to appropriately determine the PLS model estimates. This is because in PLS the reflective scheme assumed for the endogenous latent variables is inverted (i.e., the model errors are not taken into account). Savalei and Bentler (2007) have also indicated that with this type of model, PLS estimates (for example, those obtained by regression on factor scores) are generally biased as estimators of regression among latent variables.

Based on the above, it is quite clear that a variety of issues must be kept in mind when attempting these types of comparisons between approaches. The first issue is that any comparison of the performance of multiple regression relative to either PLS or SEM is trivial. Specifically, it is well known that an analysis of the same data and model based on a single regression equation using multiple regression, PLS, or SEM approaches will always result in identical estimates (irrespective of the estimation method used, be it maximum likelihood, unweighted least squares, generalized least squares, etc.). This is due to the fact that a single regression equation is a just-identified model and fits the data in the exact same way irrespective of the fit function that is minimized.

Now with respect to comparisons between PLS and SEM approaches, a number of additional side issues must be considered. First, we concur with McDonald (1996) that when discussing PLS models researchers should avoid the quite common yet confusing convention adopted of referring to both latent traits (common factors) and composites as *latent variables*.[1] As indicated by many researchers (e.g., Fornell and Bookstein 1982; Mathes 1993; Noonan and Wold 1982) the latent variables in PLS are estimated as exact linear combinations of their indicators (or manifest block variables). This essentially implies that "latent" PLS variables are not true latent variables as they are defined in SEM, since they were not derived to explain the covariation of their indicators except approximately (Mathes 1993; McDonald 1996). In contrast, the latent variables in SEM are true latent variables (i.e., hypothetically existing entities or constructs). In other words, they cannot be found as weighted sums of the manifest variables; they can only be estimated by such weighted sums (Schneeweiss 1993). This implies that for certain models, the weight vector of the PLS model is proportional to the SEM latent variable (common factor) loading vector (for further details, see equations 9 through 12 in Schneeweiss 1993). Mathematically, the key to governing the closeness of PLS to SEM latent variables for a particular block is the ratio of the largest eigenvalue of the error covariance matrix to the sum of squared loadings. In situations where this ratio, or by the model specified, is made to be small (e.g., path coefficients and loadings), estimates obtained from PLS and SEM will be very close to each other or approximately equal (Schneeweiss 1993).

The instances for which researchers have reported the two modeling approaches as supposedly showing divergence of results generally have more to do with an incorrect comparison of selected mathematical functions and/or model parameterizations (Dijstra 1983; Marcoulides 2003). For example, let us consider the case in which data from composite variables are going to be analyzed in a path model (Note: this was done in a number of the problematic and rejected papers submitted for consideration for this Special Issue). Depending on the proportionality of the generated weight matrix, a researcher could likely get very different coefficient estimates from each approach. Does that imply that the two methods differ in their estimation? Absolutely not! Any observed differences are merely a function of the differentially parameterized models being analyzed.

A simple illustration of such a phenomenon would include the case in which a weighted composite of variables (such as that generated by a principal component analysis) of a correlation matrix is compared to the weighted composite of the covariance matrix for the same variables. Although it might seem that weighted composites of a correlation matrix can be obtained fairly easily by a simple transformation of the covariance matrix, this is not the case (Joliffe 2002; Marcoulides and Hershberger 1997). In fact, because of the sensitivity of composites to units of measurement or scales of variables, weighted composites of correlation and covariance matrices do not give equivalent information nor can they be mathematically derived directly from each other. For example, using both the correlation and covariance matrices computed for a set of eight variables (n = 72) collected in a clinical study, Jolliffe (2002, see page 40 for the observed data matrix) illustrated how the weighted composite $w_1 = 0.2, 0.4, 0.4, 0.4, -0.4, -0.4, -0.2, -0.2$ (explaining 35 percent of the total variation) would be obtained when the correlation matrix is used, compared to the weighted composite $w_2 = 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0$ (explaining 99 percent of the total variation) obtained when using the covariance matrix for the same data. Clearly, the two weighted composites bear little resemblance and their interpretations would lead one to very different conclusions. Does such a result imply that the two methods differ in their estimation? Absolutely not! Any observed differences are merely a function of the differentially parameterized models being analyzed. We note that the original term used by Dijkstra (1983, p. 71) was *correct parametrization*.

Detailed evidence of the above-discussed phenomenon was also provided by McDonald (1996) in his study concerning path analysis with composite variables. Using a conventional path model with true latent variables (i.e., the case in which a path analysis of this model with SEM latent variables provides the correct model parameters and gives perfect model fit), a numerical example was analyzed using various PLS methods. McDonald indicates that it is quite reasonable to consider a path model with weighted composites as approximating the path model with SEM latent variables. Indeed, both theory and numerical results suggest that certain conditions (e.g., increasing the number of indicators in the model) will improve the approximation. For example, when comparing PLS Mode A (which optimizes composite covariances) to PLS Mode B estimation (which optimizes correlations), the obtained composites $w_A = 0.926, 0.654, 0.926, 0.654, 0.926, 0.654$ and $w_B = 0.982, 0.491, 0.982, 0.491, 0.982, 0.491$ clearly show evidence of some differences between values (we note that the true model SEM latent variable loadings in this example case are 0.8, 0.4, 0.8, 0.4, 0.8, 0.4). With a sufficiently large number of indicators in an examined model, however, the choice of composite weights actually ceases to have an influence on the parameters of the path model. Research has shown that an actual value with regard to the effect of the number of indicators on the approximation can be calculated using a so-called biasing factor formula

[1]Actually, the topic of composites versus common factors alone could serve as the focus of a special issue.

(Lohmöller 1989; McDonald 1996). Ultimately it would seem that without adding an inordinate number of indicators to make the weighting issue irrelevant, it boils down to what a researcher is interested in examining. Optimizing the correlation between blocks leads to one method, whereas maximizing the composite covariance would lead to a different choice. But the models must exhibit a so-called "correct parametrization" before there can be any type of comparison. In summary, it should be clear to the IS research community that comparison of PLS to other methods cannot and should not be applied indiscriminately. Ignoring any of the above issues could lead to incorrect conclusions or lead to overstating the importance of outcomes observed in a study.

Sample Size and Non-Normality of Data Revisited

One additional point that we feel a need to reiterate is the reification of the 10 cases per indicator rule of thumb. In spite of the editorial addressing this point nearly 3 years ago, we continue to see a number of papers applying this rule and continuing the inappropriate attribution to Chin and Newsted (1999). As a point of fact, Chin and Newsted listed in their Table 1 comparison of PLS and SEM that sample size for PLS models requires a power analysis based on the portion of the model with the largest number of predictors (p. 314). They later explained it in a paragraph that begins with the conditional statement "**If one were to use** a regression heuristic of 10 cases per predictor" and concludes as follows: "Ideally, for a more accurate assessment, one needs to specify the effect size for each regression analysis and look up the power tables" (p. 327; emphasis added).

Ultimately, the sample size necessary for adequate power requires not only specifying the effect size the researcher is attempting to detect, but the conditions underlying the sample set. Specifically, Marcoulides and Saunders (2006, p. vi) noted that

> When moderately non-normal data are considered, a markedly large sample size is needed despite the inclusion of highly reliable indicators in the model.…Indeed, a researcher must consider the distributional characteristics of the data, potential missing data, the psychometric properties of the variables examined, and the magnitude of the relationships considered before deciding on an appropriate sample size to use or to ensure that a sufficient sample size is actually available to study the phenomena of interest.

Recent simulation on PLS estimates for multigroup comparisons indeed showed the deleterious impact of non-normal data (Chin and Dibbern 2009). Further study concerning such impacts of non-normal distributed data seems warranted.

In conclusion, it is our hope that the joint Call for Papers, the special issue process itself, and the publication of papers in both *SEM* and *MISQ* will encourage the audiences of these two journals to continue to study PLS and its correct application. We can see nothing but good coming from greater dialogue between these two communities.

References

Chin, W. W., and Dibbern, J. 2009. "A Permutation Based Procedure for Multi-Group PLS Analysis: Results of Tests of Differences on Simulated Data and a Cross of Information System Services between Germany and the USA," in *Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields*, V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang (eds), Berlin: Springer.

Chin, W. W., and Newsted, P. R. 1999. "Structural Equation Modeling Analysis with Small Samples Using Partial Least Squares," in *Statistical Strategies for Small Sample Research*, Rick Hoyle (ed.), Thousand Oaks, CA: Sage Publications, 1999, pp. 307-341.

Dijkstra, T. 1983. "Some Comments on Maximum Likelihood and Partial Least Squares Methods," *Journal of Econometrics* (22), pp. 67-90.

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

Goodhue, D., Lewis, W., and Thompson, R. 2006. "PLS, Small Sample Size and Statistical Power in MIS Research," in *Proceedings of the 39th Hawaii International Conference on System Sciences*, R. Sprague Jr. (ed.), Los Alamitos, CA: IEEE Computer Society Press.

Jolliffe, I. T. 2002. *Principal Component Analysis* (2nd ed.), New York: Springer-Verlag.

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

Marcoulides, G. A. 2003. "PLS Model Specification Searches Using Optimization Algorithms," invited paper presented at the Third International Symposium on PLS and Related Methods (PLS'03), Lisbon, Portugal.

Marcoulides, G. A., and Hershberger, S. L. 1997. *Multivariate Statistical Methods: A First Course*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marcoulides, G. A., and Saunders, C. 2006. "PLS: A Silver Bullet?," *MIS Quarterly* (30:2), pp. iv-viii.

Mathes, H. 1993. "Global Optimization Criteria of the PLS-Algorithm in Recursive Path Models with Latent Variables," in *Statistical Modelling and Latent Variables*, K. Haagen, D. J. Bartholomew, and M. Deistler (eds.), Amsterdam, Elsevier, pp. 229-248.

McDonald, R. P. 1996. "Path Analysis with Composite Variables," *Multivariate Behavioral Research* (31), pp. 239-270.

Noonan, R., and Wold, H. 1982. "PLS Path Modeling with Indirectly Observed Variables: A Comparison of Alternative Estimates for the Latent Variable," in *Systems Under Indirect Observation, Part II*, K.G. Jöreskog and H. Wold (eds.), Amsterdam: North Holland, pp.75-94.

Savalei, V., and Bentler, P. M. 2007. "Structural Equation Modeling," in *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*, R. Grover and M. Vriens (eds), Thousand Oaks, CA: Sage Publications, pp. 330-364.

Schneeweiss, H. 1993. "Consistency at Large in Models with Latent Variables," in *Statistical Modelling and Latent Variables*, K. Haagen, D. J. Bartholomew, and M. Deistler (eds.), Amsterdam: Elsevier, pp. 299-320.

Vittadimi, G., Minotti, S. I., Fattore, M., and Lovaglio, P. G. 2007. "On the Relationship Among Latent Variables and Residuals in PLS Path Modeling: The Formative-Reflective Scheme," *Computational Statistics and Data Analysis* (51:12), pp. 5828-5846.

Appendix A

Associate Editors/Reviewers

We thank the following individuals who served as associate editors and reviewers on the PLS Special Issue.

Inigo Arroniz, University of Cincinnati
Tihomir Asparouhov, Muthén & Muthén
Peter M. Bentler, University of California, Los Angeles
Anne Boomsma, University of Groningen
Fred Davis, University of Arkansas
Raj Echambadi, University of Central Florida
David Gefen, Drexel University
Dale Goodhue, University of Georgia
Lisa L. Harlow, University of Rhode Island
Chris Higgins, University of Western Ontario
Geoffrey Hubona, Virginia Commonwealth University and Georgia State University
John Hulland, University of Pittsburgh
Irene R.R. Lu, York University
John J. McArdle, University of Southern California
Tenko Raykov, Michigan State University
Edward Rigdon, Georgia State University
Detlef Schoder, University of Cologne
Thomas Stafford, University of Memphis
Detmar Straub, Georgia State University
Michel Tenenhaus, HEC
Ron Thompson, Wake Forest University
Pierre Valette-Florence, Institut d'Administration des Entreprises de Grenoble (IAE)
Vincenzo Vinzi, ESSEC Business School