Communications of the Association for Information Systems

Volume 38 Article 23

3-2016

Partial Least Squares Structural Equation Modeling Approach for Analyzing a Model with a Binary Indicator as an Endogenous Variable

David Bodoff
University of Haifa, dbodoff@univ.haifa.ac.il

Shuk Ying Ho
The Australian National University

Communications of the Association for Information Systems

Tutorial Paper ISSN: 1529-3181

Partial Least Squares Structural Equation Modeling Approach for Analyzing a Model with a Binary Indicator as an Endogenous Variable

David Bodoff
Faculty of Management
University of Haifa
Haifa, Israel
dbodoff@univ.haifa.ac.il

Shuk Ying Ho
Research School of Accounting
The Australian National University
Canberra, Australia

Abstract:

In this paper, we focus on PLS-SEM's ability to handle models with observable binary outcomes. We examine the different ways in which a binary outcome may appear in a model and distinguish those situations in which a binary outcome is indeed problematic versus those in which one can easily incorporate it into a PLS-SEM analysis. Explicating such details enables IS researchers to distinguish different situations rather than avoid PLS-SEM altogether whenever a binary indicator presents itself. In certain situations, one can adapt PLS-SEM to analyze structural models with a binary observable variable as the endogenous construct. Specifically, one runs the PLS-SEM first stage as is. Subsequently, one uses the output for the binary variable and latent variable antecedents from this analysis in a separate logistic regression or discriminant analysis to estimate path coefficients for just that part of the structural model. We also describe a method—regularized generalized canonical correlation analysis (RGCCA)—from statistics, which is similar to PLS-SEM but unequivocally allows binary outcomes.

Keywords: Partial Least Squares, PLS, Structural Equation Modeling, Binary Endogenous Variables.

This manuscript underwent editorial review. It was received 02/02/2015 and was with the authors for 5 months for 3 revisions. Nancy K. Lankton served as Associate Editor.

1 Introduction

Structural equation modeling (SEM) is a data analysis method that has quickly gained popularity as one of the primary methods for analyzing behavioral data in information systems (IS). In contrast to first-generation techniques such as ANOVA and multiple regression, SEM simultaneously considers relationships among multiple independent and dependent constructs. SEM expresses these complicated variable relationships in a causal model with hierarchical or non-hierarchical, recursive or non-recursive structural equations (Bullock, Harlow & Mulaik, 1994; Hanushek & Jackson, 1977). Additionally, SEM supports latent variables—variables that are not measured directly but are estimated in the model from manifest variables (or indicators). There are two main approaches to SEM: component-based and covariance-based. Component-based SEM aims to explain variance. One example of component-based SEM is the partial least square (PLS) method. We follow the convention in Gefen, Straub and Boudreau (2000) in referring to component-based SEM as PLS-SEM. The other main approach to SEM is covariance-based SEM techniques, which emphasize the overall fit of the observed covariance matrix with the hypothesized covariance model. LISREL, AMOS, EQS, and MPlus are software packages supporting covariance-based SEM.

In this paper, we focus on PLS-SEM. PLS-SEM's increasing popularity in recent years has resulted in many papers that review PLS-SEM's use in various fields of business research, such as accounting (Lee, Petter, Fayard, & Robinson, 2011), IS (Ringle, Sarstedt, & Straub, 2012), strategic management (Hair, Sarstedt, Pieper, & Ringle, 2012a) and marketing (Hair, Sarstedt, Ringle, & Mena, 2012b). Many papers have promoted PLS-SEM as having several advantages over covariance-based SEM, such as its being able to accommodate data-analysis challenges (e.g., small sample sizes, non-normal data and models with formative constructs) and its being more appropriate than other methods for exploratory research (Hair, Ringle, & Sarstedt, 2011; Sosik, Kahai, & Piovoso, 2009). A later wave of papers that take a second look question whether scholars have overclaimed PLS-SEM's perceived advantages (Aguirre-Urreta & Marakas, 2014; Goodhue, Lewis, & Thompson, 2012; Rönkkö & Evermann, 2013). Still other recent work either rebuts those criticisms (e.g., Henseler et al., 2014; Rigdon et al., 2014) or attempts to create a nuanced perspective on PLS-SEM's strengths and appropriate use (e.g., Rigdon, 2012; Sarstedt, Ringle, Henseler, & Hair, 2014). As a result, we can see that an emerging balanced perspective on the main issues seems to have appeared, and PLS-SEM remains a popular tool in IS research.

Against this background, the literature is less explicit and less clear about PLS-SEM's ability to handle binary outcomes. Several papers that we review below caution using PLS-SEM for binary variables. As a result, IS researchers whose data includes binary outcomes appear to avoid using PLS-SEM. In this paper, we examine the different ways in which a binary outcome may appear in a model and distinguish those situations in which a binary outcome is indeed problematic versus those in which one can easily incorporated it into a PLS-SEM analysis. By explicating the details, we hope to enable IS researchers and reviewers to distinguish the different situations rather than avoid PLS-SEM altogether whenever any binary indicator presents itself. We synthesize and organize known material to make it accessible as a proposed consensus in our research community.

This paper proceeds as follows: in Section 2, we provide examples from two top IS journals (*MIS Quarterly* and *Information Systems Research*) from the past five years to illustrate how IS researchers have handled a model with a binary endogenous variable. In Section 3, we present some of the (conflicting) general statements found in the literature regarding PLS-SEM's suitability for analyzing models with binary endogenous variables. In Section 4, we describe the PLS-SEM algorithm and highlight the places where it uses ordinary least squares (OLS). Since OLS is inappropriate with a binary outcome, those places are those where one might think that a binary variable poses a problem. In Section 5, we identify the four ways in which a binary variable arises as an endogenous construct (i.e., outcome) and propose solutions for how one can properly handle structural models with binary outcomes in PLS-SEM. In particular, we give two recommended solutions for using PLS-SEM (or PLS-SEM like) approaches to handle models with binary outcomes. Finally, in Section 6, we conclude the paper.

2 How Do IS Researchers Handle a Model with a Binary Endogenous Variable?

In IS research, one can measure numerous ultimate outcomes with a binary variable. One important example is system adoption. When one measures it with a surrogate such as intention to adopt, then applying PLS-SEM is straightforward because one can model it as a continuous latent construct that several Likert-based items measure. However, in the ideal case, where actual adoption decisions are available, the situation is paradoxically less straightforward because then the variable is binary. Other examples of binary outcomes are adopting or not adopting a late bidding strategy (Goes, Karuga, & Tripathi, 2012), switching user interfaces (Murray & Häubl, 2011), and taking or not taking a product recommendation (Ho & Bodoff, 2014). Table 1 summarizes papers that include binary indicators as outcome variables from the two pre-eminent IS journals (*MIS Quarterly* and *Information Systems Research*) from 2010 to 2014.

Table 1. IS Publications Having Binary Outcome Indicators

	Description	Dependent variable(s)	Analysis method(s)
Banker, Hu, Pavlou, & Luftman (2011)	Examines how a firm's strategic positioning influences its CIO reporting structure.	CIO reporting to CEO vs. CIO reporting to CFO	A logistic regression
Bapna, Goes, Wei, & Zhang (2011)	Presents a finite mixture logit model to predict the likelihood of electronic payments systems adoption in business-to-business settings.	The likelihood of adopting electronic payments systems	A logit model
Caliendo, Clement, Papies, & Scheel-Kopeinig, (2012)	Examines whether installing a spam filter reduces net costs and whether spam costs influence users' decision to install a spam filter.	Whether users install a spam filter	A logit model
Ceccagnoli, Forman, Huang, & Wu (2012)	Examines whether joining an owner's platform ecosystem is associated with an increase in sales and a greater likelihood of issuing an initial public offering (IPO).	Whether an IPO has been issued	Regression using a linear probability mode
De, Hu, & Rahman (2013)	Examines if Web technologies (e.g., zoom and color swatch) reduce the likelihood of product returns.	Likelihood of product return	Logit and probit models
Gefen & Pavlou (2012)	Proposes that the perceived effectiveness of institutional structures sets the boundaries of trust and risk in online market places and examines how trust and risk influence transaction activity.	Transaction activity (won auctions or not)	A logistic regression
Ghose, Goldfarb, & Han (2013)	Examines how Internet clicking behavior varies between mobile phones and personal computers.	Likelihood of clicking on a post	A binary choice model
Goes et al. (2012)	Develops a model to explain online bidders' adoption of different strategies and change in their strategies over time.	Likelihood of choosing a bidding strategy	A multinomial logistic regression
Goes, Lin, & Au Yeung (2014)	Examines if user interactions help generate product reviews and what kind of reviews do such interactions induce.	Whether a user generates ratings or nonratings reviews	A logit model
Gu, Konana, Raghunathan, & Chen (2014)	Focuses on virtual investment communities and examines investors' propensity to seek interactions with others with similar sentiments in these communities.	Whether a thread in the choice set is chosen by investors	A discrete choice model
Hildebrand, Häubl, Herrmann, & Landwehr (2013)	Examines how receiving others' feedback on initial product configurations affects consumers' satisfaction with these self-designed products.	High vs. low satisfaction level	A logistic regression
Ho & Bodoff (2014)	Integrates the elaboration likelihood model with consumer search theory to examine how Web personalization influences online users' attitude formation and decision making.	Whether a user selects a personalized item as the final choice	Path analysis
Langer, Forman, Kekre, & Sun, (2012)	Examines the factors that drive purchase decisions and how these factors change over time and across buyers.	Buyer purchase decision	A probit model

Table 1. IS Publications Having Binary Outcome Indicators

Li, Shang, & Slaughter (2010)	Considers software firms' capabilities and their competitive actions and examines why firms survive in the volatile software industry.	Likelihood of firm survival	Cox proportional hazard regression
Mani, Barua, & Whinston (2010)	Theorizes that performance heterogeneity across business process outsourcing (BPO) exchanges is a function of the design of information capabilities that fit the unique information requirements of the exchange.	Choice of information capabilities for a given BPO relationship	A probit model
Mani, Barua, & Whinston (2012)	Tests if the use and performance effects of the information structure are greater in time and materials BPO contracts than in fixed-price BPO contracts.	Choice of contractual structure	A probit model
Moreno & Terwiesch (2014)	Examines how both a structured measure (in a numerical reputation score) and an unstructured measure (based on the verbal praise left by previous buyers) influence sellers' likelihood of being selected.	Likelihood of awarding a project to a coder	A logit model
Murray & Häubl (2011)	Examines how users' freedom to choose affects their interface preferences.	Choice of interface	A logistic regression
Oestreicher-Singer & Zalmanson (2013)	Examines how users' engagement in online communities influences their willingness to pay for premium music streaming.	Likelihood of paying for a subscription	A logistic regression
Özpolat, Gao, Jank, & Viswanathan (2013)	Uses a data set of nearly 10,000 shopping sessions at an online retailer's website to empirically test the value and effectiveness of assurance seals on shoppers' likelihood of purchasing.	Likelihood of purchase	A Pearson's chi-squared test
Peng & Dey (2013)	Proposes a dynamic view where the fraction of current adopters in a network positively moderates the impact of network centrality and closure and analyze the adoption of software version control technology by open source software (OSS) projects.	Whether or not an OSS project adopted SVN during the observation period	A discrete choice model
Rice (2012)	Examines how online reputation ratings are assigned and, thus, how electronic reputations are formed in transactions in which buyers and sellers interact anonymously.	Buyers' decision to transact	A logistic regression
Singh & Phelps (2013)	Examines the conditions under which prior adopters of competing OSS licenses socially influence how a new OSS project chooses among such licenses.	Likelihood to adopt a given license type	A binary choice model
Susarla & Barua (2011)	Examines what factors influence service providers' likelihood of survival in the application service providers (ASP) marketplace.	Whether the ASP was still operational or had exited	Hazard models of firm survival
Susarla, Subramanyam, & Karhade (2010)	Examines whether contract extensiveness can alleviate holdup and affect contract renewal.	Whether a contract is renewed beyond the initial contract term	A logistic regression
Tan, Sutanto, Phang, & Gasimov (2014)	Proposes that consumer responses to personal communication technologies (PCT)-disseminated commercial messages are influenced by the PCT that carries general symbolic meanings about its nature and purpose and the context culture in which it is used.	Whether users redeem or forward coupons	A logistic regression
Wattal, Telang, Mukhopadhyay, & Boatwright (2012)	Examines how consumers respond to firms' use of two types of information for personalization: product preferences and name.	Whether consumers open the firm's email; and whether they unsubscribe it	A probit model

Table 1. IS Publications Having Binary Outcome Indicators

Wu (2013)	Considers two intermediate mechanisms by which an information-rich network is theorized to improve work performance—information diversity and social communication—and quantifies their effects on productivity and job security.	Likelihood of being laid off during the financial crisis	A logistic regression
Xue, Ray, & Gu (2011)	Examines the relationship between environmental uncertainty and IT infrastructure governance in a sample of business units from Fortune 1000 companies.	Likelihood of decentralizing IT infrastructure decisions	A logistic regression
Ye, Gao, & Viswanathan (2014)	Examines how auction sellers respond to changes in the design of reputation systems on eBay.	Whether the seller participates in the strike	A logit regression model

3 A Literature Review On PLS With Binary Variables

Table 1 indicates that IS researchers tend not to use PLS-SEM to analyze models with binary outcome indicators. One possible reason is that some of the prior literature on PLS-SEM appears to warn against using PLS-SEM whenever binary variables are present in a model. For instance, Hair and colleagues write that “this practice should be considered with caution” (Hair et al., 2012b, p. 421). Jakobowicz & Derquenne (2007, p. 3668) offer a similar remark: “when working with continuous data..., PLS does not face any problems, but when working with nominal or binary data it is not possible to suppose there is any underlying continuous distribution”. Chin, Marcolin, and Newsted (1996, p. 24) acknowledge that, under certain conditions, it may be proper to use a binary indicator in a multi-item latent construct, but they do not elaborate on using PLS-SEM with a single-indicator binary variable.

In practice, it appears that researchers do use PLS-SEM when the data includes a binary measure as an exogenous variable or one of its indicators but not when it is an endogenous (i.e., outcome) variable. For example, in reviewing 204 papers in marketing, Hair et al. (2012b) identified 43 studies with binary variables in the PLS-SEM analysis. In another review, Hair et al. (2012a) identified 112 PLS-SEM models in strategic management, and, of these, 40 models included binary variables. However, as Table 1 shows, researchers have hesitated to use PLS-SEM to analyze models with binary *endogenous* variables. To handle a model with a binary endogenous variable, some IS researchers have adopted logistic regressions (e.g., Goes et al., 2012, Gopal & Koka, 2012) or used covariance-based SEM to implement a path model (e.g., Ho & Bodoff, 2014), which means forfeiting using PLS-SEM that is best suited for models with latent variables when the data set also includes a (often binary) actual behavioral outcome. This situation is paradoxical. In many common settings, it is also unnecessary.

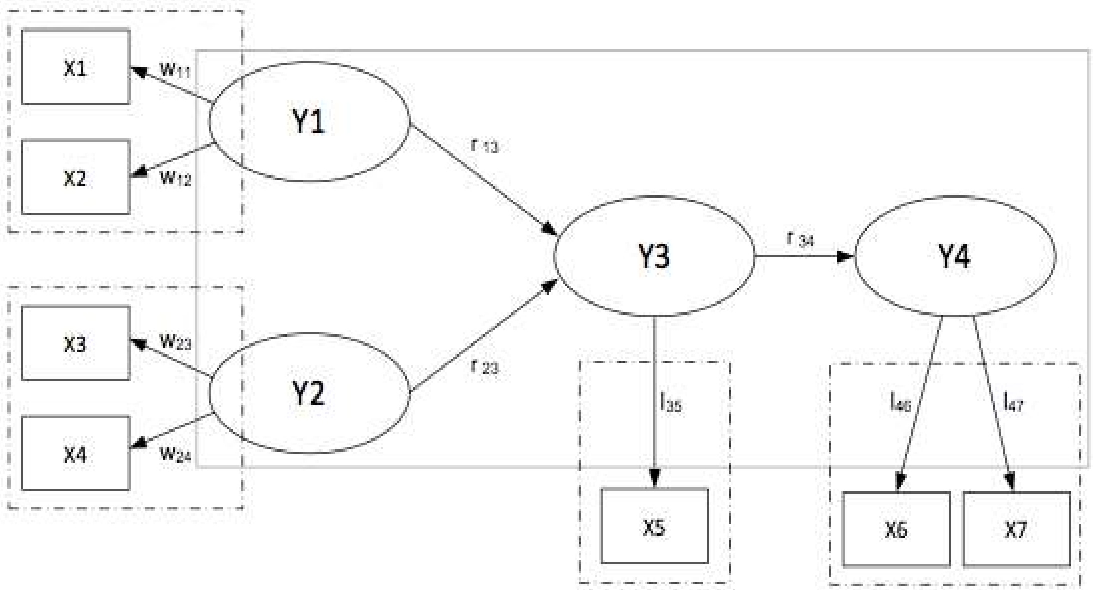
Almost no literature directly addresses whether and how to use PLS-SEM for binary endogenous variables. Many scholars have discussed other PLS variants (especially PLS regression (PLS-R)) for classification problems including binary classification (Barker & Rayens, 2003; Bastien, Vinzi, & Tenenhaus, 2005; Boulesteix, 2004; Rosipal & Kramer, 2006). However, PLS-R’s algorithm is different from PLS-SEM’s. Specifically, PLS-R does not include the PLS-SEM algorithm steps that many scholars view as inappropriate in the presence of binary variables, so we cannot infer from this vast literature on PLS-R that PLS-SEM is appropriate with binary variables. Lohmoller (1989) focuses mostly on PLS-SEM, and he does raise the case of a binary variable. However, the case he addresses (p. 178) concerns a single binary predictor of a categorical outcome that is represented by binary dummy variables. Lohmoller briefly shows how the classical approach of discriminant analysis fits simply in the PLS framework. As with the literature on PLS-R, Lohmoller’s example and brief discussion do not engage with the full spectrum of issues that arise in the case of PLS-SEM and that researchers such as those cited above raise as reasons to disqualify PLS-SEM in the presence of binary endogenous variables.

Russolillo (2012) introduces a variant of PLS-SEM, which he calls “non-metric PLS path modeling”, to handle ordinal and categorical data in the PLS-SEM context. He states the problem (i.e., the reason why one cannot use regular PLS-SEM for ordinal or categorical data), and, importantly for our purposes, the reasons he gives about categorical data all apply to cases with more than two categories. In the case of a binary variable, regular PLS-SEM may be applicable. Indeed, this is Rusolillo’s view (Giorgio Rusolillo, personal communication). To summarize, the statistics literature does not appear to have a single organized stream of research that is explicitly organized around the topic of PLS-SEM with binary endogenous variables. We fill that gap.

4 A Technical Overview of the PLS-SEM Algorithm

In this section, we overview PLS-SEM while focusing specifically on details that may present a problem for some cases of a binary variable. In particular, we highlight the points in the PLS-SEM algorithm and the precise manner in which the algorithm uses OLS regression because scholars often cite such these aspects as problems when using PLS-SEM with a binary variable. For a more extensive technical review on PLS, see Tenenhaus, Vinzi, Chatelin, and Lauro (2005).

Figure 1 depicts an example of PLS-SEM model with two exogenous variables (Y1 and Y2) and one endogenous variable (Y3). As many review papers (such as Haenlein & Kaplan, 2004; Monecke & Leisch, 2012; Tenenhaus et al., 2005; Vinzi, Trinchera, & Amato 2010b) describe, PLS-SEM has two stages of estimation: the first stage estimates the measurement model and the second stage estimates the structural model.



Measurement models
(Indicators x, latent variables Y, and relationships, i.e., outer weights w or l between indicators and latent variables)

	Y_1	Y_2	Y_3	Y_4
x_1	w_{11}			
x_2	w_{12}			
x_3		w_{23}		
x_4		w_{24}		
x_5			l_{35}	
x_6				l_{46}
x_7				l_{47}

Structural models
(Latent variables Y and relationships ρ between latent variables)

	Y_1	Y_2	Y_3	Y_4
Y_1			ρ_{13}	
Y_2			ρ_{23}	
Y_3				ρ_{34}
Y_4				

Figure 1. A Typical PLS-SEM Model

4.1 The First Stage of PLS-SEM's Algorithm

In the first stage, the PLS-SEM algorithm estimates the latent variables. The latent variables are defined as a weighted linear combination of their respective manifest variables (i.e., x1−x7), so we can view the first stage as finding the weights wij—outer weights—for each variable's manifest variables (i.e., x1−x7). The algorithm initially chooses the weights randomly, which then refines them iteratively with an inner and outer step. For reference, Table 2 labels the three major steps in this first stage.

Table 2. IS Publications with Binary Outcome Indicators

Step (1A) An outer step calculates each latent variable as the weighted linear combination (i.e., outer weights) of its manifest variables. As an example, referring to Figure 1, one would calculate Y1's value as $\mathrm{Y1} \propto (\mathrm{X1} * \mathrm{w}_{11} + \mathrm{X2} * \mathrm{w}_{12})$.
Step (1B) An inner step estimates a proxy of each latent variable as a weighted linear combination (i.e., using "inner weights") of its neighboring latent variables that it is connected to in the structural model. Different ways to calculate these inner weights exist, most notably the centroid, factor, and path weighting schemes. Path weighting uses OLS regression. In particular, one calculates a variable's proxy value by regressing its latent value on its predictors. For example, as part of the calculation of a proxy $\widetilde{\mathrm{Y3}}$ for Y3, Y3 is multiple-regressed on its directly connected latent variables (i.e., Y1 and Y2). This process is known in path weighting as the "forward" direction in which the latent values of predictors (Y1 and Y2 in this case) influence the calculation of the proxy of their structurally dependent variable (Y3 in this case).
Step (1C) There are two options. Mode A: each outer weight is estimated as the simple-regression coefficient of that indicator on the proxy latent variable. In our example, the algorithm solves the regression $\mathrm{x}_1 = \beta_0 + \beta_1\widetilde{\mathrm{Y1}}$ and sets $\mathrm{w}_{11} = \beta_1$ and then separately solves regression equation $\mathrm{x}_2 = \beta_0 + \beta_2\widetilde{\mathrm{Y1}}$ and sets $\mathrm{w}_{12} = \beta_2$. Mode B: the outer weights of each variable are estimated as the corresponding multiple-regression coefficient when regressing the latent variable proxy on its own manifest variables. In our example, the algorithm solves multiple regression $\widetilde{\mathrm{Y1}} = \beta_0 + \beta_1\mathrm{x}_1 + \beta_2\mathrm{x}_2$ and then sets $\mathrm{w}_{11} = \beta_1$, $\mathrm{w}_{12} = \beta_2$.

The intuitive idea of steps 1B and 1C is to first find proxy latent values that reflect the model's structural links (1B) and then find weights for the manifest variables to "catch up" (1C) (i.e., weights such that the resulting linear combination of manifest variables will equal that proxy value). Steps 1A-1C repeat until convergence, which is not theoretically guaranteed but is usually obtained.

For our purposes, the most important details are the places where the algorithm uses OLS regression, whose assumptions are violated when the dependent variable is binary. An important detail of step 1B is that, among the schemes for calculating inner weights, path weighting uses regressions for "forward" links in the manner specified in Table 2. At the same time, in the path-weighting scheme, the backward-link from Y3 influences the proxies of Y1 and Y2, but this direction depends on the correlations—not regressions—between Y3 and Y1 and between Y3 and Y2, so the backward direction does not raise any concerns regarding OLS (if there are two directly linked binary variables, then using a Pearson correlation raises a different concern; see Section 5).

Similarly, step 1C assumes that the estimation of the outer weights uses OLS regression. If one uses mode A, the indicator is OLS regressed on the estimated latent variable, whereas, if one uses mode B, the latent variable is OLS regressed on the indicators. Mode A is usually used for reflective models and mode B for formative models. For mode A, the binary indicator raises a concern because mode A OLS regresses the binary indicator on its latent variable; in effect, it treats the binary indicator as the dependent variable in an OLS regression.

4.2 The Second Stage of PLS-SEM's Algorithm

In the second stage, the PLS-SEM algorithm estimates the path coefficients in the structural model ($\rho 13$ and $\rho 23$ in Figure 1). It involves a series of first-generation regressions, usually OLS regressions that use the calculated scores for latent variables (obtained in the first step) as the values of each variable. The algorithm regresses each latent variable using OLS regressions on its direct predecessors (e.g. for Y3, the direct predecessors are Y1 and Y2). There is a separate partial regression model for each endogenous latent variable. That means the algorithm separately considers each endogenous variable and its immediate predictors using first-generation OLS regressions. In Figure 1, latent variable Y4 is regressed on Y3, and, separately from that, Y3 is regressed on Y1 and Y2. As Marcoulides, Chin, and Saunders (2009, p. 172) note:

> *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.).*

4.3 Summarizing the Two Stages of PLS-SEM's Algorithm

To summarize, in the first stage, the PLS-SEM algorithm finds latent values (i.e., values of each latent variable) by assigning weights to its manifest variables while considering the latent values of directly linked variables in the structural model. In the second stage, it performs OLS regressions using the latent values found in the first stage. Taken together, one obtains the estimates for all relationships in the measurement models (i.e., outer weights) and the structural model (i.e., the path coefficients). Table 3 emphasizes some points that are crucial for understanding how PLS-SEM relates to binary single indicators.

Table 3. A Summary of Crucial Points about How PLS-SEM Relates to Binary Single Indicators

1.	Step (1B) in Table 2 depends on the correlations between latent variables under two weighting schemes, but, in the path weighting scheme, it depends on OLS regression for forward-direction connections.
2.	Step (1C) in Table 2 depends on OLS regressions between a latent variable and its own indicators, with the direction of regression depending on mode A versus mode B.
3.	The second stage of PLS-SEM, which calculates path coefficients, uses the result of the first stage, which calculates latent variables, but the two stages do not interact beyond that.
4.	The second stage uses a set of separate OLS regressions, one for each part of the model; there is no connection between the parts. Each endogenous outcome is regressed solely on its direct antecedents (see Vinzi et al., 2010b, p. 56).

With this technical overview, we can specify the following potential problems that may arise when one uses PLS-SEM with a binary indicator.

- "Problem A": if one uses path weighting, step 1B regresses a latent variable on its predictors using OLS, which one could perceive as a problem if the latent variable is binary.
- "Problem B": if one uses mode A, step 1C regresses a latent variable on its indicators using OLS, which one could perceive as a problem if the latent variable is binary.
- "Problem C": in the second stage, the PLS-SEM algorithm uses OLS to estimate path coefficients for the structural model among latent variables, which one could perceive as a problem if a latent variable is binary.

5 Is PLS-SEM Unsuitable for Models with a Binary Endogenous Variable?

As we describe in Section 4, the PLS-SEM algorithm uses OLS at various points, so one may perceive that binary variables pose a problem. To our knowledge, no one has previously systematically organized the possible problems and solutions into a single framework of "how to use PLS-SEM with binary variables". To organize and clarify the issue, we first distinguish between a variety of ways in which binary endogenous variables arise in the PLS-SEM context. Subsequently, we refer back to our technical overview in Section 4 (and, in particular, our discussion on OLS regressions, which Table 3 summarizes). With these building blocks in place, we can be specific about the situations in which PLS-SEM would actually violate OLS assumptions, the precise nature of the violations if any, and ways to remedy them.

5.1 Ways in Which a Binary Variable Arises in an Endogenous Construct

Table 4 presents a two-by-two classification that identifies four ways in which one can use a binary endogenous variable.

Table 4. Role of Binary Variable in Types of Endogenous Variables

	The construct is latent	The construct is observable (and is actually observed)
The construct is represented using a single item	**Quadrant I:** binary variable is sole indicator for a latent variable. des Reis and Soares (2006) use a single binary indicator to capture intention to use an electronic procurement system (EPS). The binary variable is assigned a “1” if the company has an intention or is taking action to implement an EPS; otherwise, the variable is assigned a “0”.	**Quadrant III**: binary item is equivalent to a theoretically meaningful variable. As an example, Ho and Bodoff’s (2014) construct captures whether a consumer chooses a personalized or non-personalized item for their purchase. It contains only one item and this item is binary. The item is “an online user selects a personalized item as the final choice (1 = select a personalized item, 0 = otherwise)”. The researchers conceptualize the construct as fully encompassing the variable’s single-dimensional meaning.
The construct is represented using multiple items	**Quadrant II:** binary variable is one of many indicators for a latent variable. This latent variable may contain only binary indicators or a mixture of indicators with different data types. Alwitt and Pitts (1996) used items: “The manufacturer cares about the environment (1 = “the most important” of five factors, 0 = “otherwise”)” and “the company uses biodegradable plastic (1 = “the most important of five factors”, 0 = “otherwise”) to capture individuals’ perceptions of a company's environmental policy.	**Quadrant IV**: N/A Explanation: for an observable variable, there is no notion of “indicators”. Rather, it is a single variable that has its own meaning and measure. Hence, quadrant IV does not meaningfully exist.

The first dimension distinguishes between latent variables and observable variables. An observable variable cannot be observed or directly measured. An observable variable has no notion of “indicators”. Rather, an observable variable is a single variable that has its own meaning and measure. By contrast, a latent construct—whether reflective or formative—cannot be directly observed but is instead inferred from one or more observable variables, which, when used in this way, are called manifest variables or indicators of the latent variable.

The second dimension refers to whether a given study measures the variable using a single item or multiple items. These two dimensions yield four quadrants: I) latent constructs represented with a single indicator, II) latent constructs represented with multiple indicators, and III) directly observable variables that are measured using a single item that equals the variable itself. Quadrant IV does not exist since, if a variable is not latent (i.e., neither reflective nor formative), one has no need for multiple indicators to try to capture its full scope. Thus, overall, a binary variable may arise in any of three ways: 1) as the only indicator of a latent construct, 2) as one of many indicators of a latent construct, and 3) as a directly observable variable.

Table 5 shows which of the problems enumerated in Section 4 arises in which quadrant. In the rest of this section, we elaborate each of these problems further. We focus especially on showing that quadrants I and III do not violate any PLS-SEM assumptions and quadrant II is a matter of debate. For quadrants I and III, we explain why problems A and B do not exist and offer solutions to address problem C. For Quadrant II, we discuss researchers’ differing opinions.

In quadrant I, a single binary indicator represents a latent construct. The latent variable can be reflective or formative. This setup has two problems, both of which Hair et al. (2012b) raise. The first problem is that this setup “proves problematic for approximations in the PLS-SEM algorithm since path coefficients are estimated by OLS regressions” (Hair et al., 2012b, p. 421). The problem they raise actually includes several distinct issues, which affect both quadrants I and III; we elaborate further when we discuss quadrant III below.

Table 5. Summary of Potential Problems when Using a Binary Variable in Each of the Four Quadrants

	The construct is latent	The construct is observable
The construct is represented using a single item	**Quadrant I:** *Problem A:* OLS may be used when estimating inner weights → problem does not exist here (see Section 5.1.1). *Problem B:* OLS used when estimating outer weights → problem does not exist here (see Section 5.1.1). *Problem C:* OLS used when estimating path coefficients (see Section 5.1.2). Additional modeling problem: single indicator for underlying latent variable → a conceptual, not statistical issue, and not particular to binary variables.	**Quadrant III:** *Problem A:* OLS may be used when estimating inner weights → problem does not exist here (see Section 5.1.1). *Problem B:* OLS used when estimating outer weights → problem does not exist here (see Section 5.1.1). *Problem C:* OLS used when estimating path coefficients (see Section 5.1.2).
The construct is represented using multiple items	**Quadrant II:** *Problem B:* OLS used when estimating outer weights (see Section 5.1.1).	**Quadrant IV:** N/A

The second problem faced by models in this quadrant is the use of a single item—binary or not—to capture a latent variable. As Diamantopoulos, Sarstedt, Fuchs, Wilczynski, and Kaiser (2012) note, researchers would view using single-item measures to capture a latent construct poorly except in some specific conditions they enumerate. Moreover, formative constructs have a more basic problem with a single-item measure since one would normally expect formative measurement models to be more capacious than reflective ones (Ringle et al., 2012, p. vii). Thus, even if we resolve the statistical issues surrounding binary variables (our focus), in practical terms, quadrant I will remain with a conceptual modeling problem and so is unlikely to be commonly used in practice.

Quadrant II includes models that use a binary indicator as part of a multiple-item measure for a latent construct. Some researchers believe that such use raises problems (especially those problems that we term problem A and problem B). Others researchers disagree. We elaborate on this difference of opinion in Section 5.1.1. Quadrant II does not experience problem C because the latent construct is not binary in this case—only one of the indicators is.

In quadrant III, a variable is an observable behavior that a binary single indicator captures, which makes the indicator identical with the whole variable. The difference between quadrants I and III is in how the researcher conceives the variables. Not every variable is latent. Conceptual and theoretical considerations should determine whether to conceive of a variable as latent or observable (see Appendix 1 for an example). In our view, quadrant III represents a common situation in IS research. We can naturally conceptualize all the examples in Table 1 as belonging to this quadrant. In fact, we presume that most if not all of those researchers had this in mind because, otherwise (i.e., if they intended to conceive of the binary outcome as an indicator of a latent variable; that is, quadrant I), they would have needed to somehow justify why they used a single indicator.

In Sections 5.1.1 to 5.1.5, we explain that Problems A and B do not apply to Quadrant III. We also present the debate about whether Problems A and B apply in principle to any of the quadrants, and we show how one can easily solve Problem C.

5.1.1 Problems A and B: OLS Used when Estimating Inner and Outer Weights

Recall that problem A is that, with path weighting, a latent variable is regressed on its immediately linked predictors to estimate a proxy for that variable, which one may see as a problem when the variable is binary. Obviously, one can avoid this problem by using one of the other two schemes: centroid or factor. In light of the finding that the various schemes have "very little" effect on the latent value estimates (Tenenhaus et al., 2005), choosing centroid or factor schemes seems an easy way to solve this problem altogether, but we proceed on the assumption that one has chosen path weighting. Also recall that problem B is that, if one uses mode A, the (binary) indicator is regressed on the latent variable using OLS. This again may appear to be a problem with a binary variable. However, for the reasons we now explain, these issues definitely pose no problem for the cases of quadrants I and III and may not pose a problem for quadrant II.

Problems A and B do not apply in quadrants I and III because, when one measures a variable using a single indicator as in quadrants I and III, the indicator becomes identical with the variable, and none of the steps of PLS-SEM's first stage are used for that variable. PLS-SEM will simply assign an outer weight of 1.0 to that indicator, and the value of the latent variable is just the (possibly normalized) value of the indicator. Therefore, although step 1B (if path weighting is used) hypothetically regresses the variable on its predictors using OLS and although step 1C (in mode A) hypothetically regresses the variable on its indicators using OLS, none of occurs with a single indicator variable.

Second, researchers more generally disagree about whether it is correct to approach any of these questions in terms of violations of PLS-SEM's statistical assumptions. For example, in quadrants I-II, if one uses mode A, then step (1C) actually uses OLS to regress the binary indicator on its latent variable. Hair et al. (2012b), Jakobowicz & Derquenne (2007), and others espouse that this situation is a problem because OLS is used to regress the indicator—binary, in this case—on the (continuous) estimate of the latent variable and OLS is not appropriate for a binary dependent variable. Statisticians such as Tenenhaus, Vinzi, Russolillo, and others (see Russolillo, 2012; Tenenhaus & Hanafi, 2010; Vinzi, Russolillo, & Trinchera, 2010a; Vinzie et al., 2010b) hold an opposing view. In their view, adopting mode A is a statistical decision to use a covariance-based criterion that sets the outer weights to the covariance between the indicator and the estimated latent variable (i.e., the regression coefficient when all scores are standardized) and is not in any way meant as a model of the data (and similarly for mode B, which is a statistical decision to use a correlation-based criterion). Somewhat paradoxically, this latter perspective reflects a more purist mathematical approach, which emphasizes the criteria being employed—covariance, which makes perfect sense for a continuous and binary variable—and not the algorithm to get there (i.e., OLS). In this view, even where one actually applies step (1C) to find a weight for any binary indicators (i.e., quadrant II, unlike quadrant III where the weight is simply set to 1.0), one is not modeling that the data follows the (normal) distribution associated with OLS, and there is no problem even in principle with applying this step to a binary indicator. More generally, this view holds that PLS-SEM does not optimize any established or even clearly defined statistical criterion, and so, by the same token, one cannot say that it operates on the basis of any "statistical assumptions". It is, rather, an algorithm that has many good properties and that one can understand in completely geometric (i.e., not statistical) terms. In this view, any discussion of statistical assumptions is misplaced. In any case, PLS-SEM's appropriateness for quadrants I and III does not depend on this position. As we explain earlier, quadrants I and III do not use OLS in any way that the binary variable is the dependent variable.

5.1.2 Problem C: OLS Used when Estimating Path Coefficients

Problem C occurs in PLS-SEM's second stage, which uses OLS regressions to estimate path coefficients of a structural model. One OLS regression assumption is that the error term of the endogenous variable is normally distributed. The error term for each case is defined as the observed value of the endogenous variable minus the predicted value given x. When the dependent variable is dichotomous, the error term is not normally distributed (because, for any value of x, there are only two possible values that the residuals can take.) Clearly, there is a problem when the endogenous variable in the PLS-SEM is binary, but it is a problem that one can easily overcome. One simply has to use a more appropriate analysis method such as discriminant analysis or logistic regression for only that part of the structural model in only the second stage. In particular, one runs the PLS-SEM first stage as is. This results in estimates of latent variables, which are output and saved. One takes the binary variable and its latent variable antecedents as they were output and saved and uses a logistic regression or discriminant analysis to estimate path coefficients. For the other parts of the structural model, one finds path coefficients using OLS regressions as shown in the usual PLS-SEM output. Recall that there is no iterative connection between the first and second stages of PLS-SEM (see (3) in Table 3), and, in the second stage, there is no connection between the different parts of the model (see (4) in Table 3). This step involves transferring the statistic output from PLS-SEM to another statistics software (such as SPSS) and running an appropriate analysis. The need to save the latent variables in a file arises only because, at the present moment, no software of which we are aware has the option to use other analyses besides OLS in PLS-SEM's second stage. However, no inherent reason for this software limitation exists. (In fact, we suspect that it has not been made available because of the triviality of outputting any latent variables that are found in PLS-SEM's first stage and running any analysis one likes.) There is nothing extraordinary about what has come to be known as PLS-

SEM's second stage in which one usually runs OLS automatically on the structural model. Note again that there is no connection between PLS-SEM's two stages.

To summarize, binary variables present several challenges. Our main practical suggestion is for quadrants I and III and especially for quadrant III, which also does not suffer from any non-statistical question about how a single indicator (binary or otherwise) can fully capture a latent variable. As such, we simply suggest one to run PLS-SEM's first stage as usual for the whole model and to run the second stage (which is just OLS for each part) as usual for the whole model except for estimating the path coefficients for links that predict the binary variable. For that, we suggest a logistic regression. In the example of Figure 1, if x5 is binary, the entire first stage of PLS-SEM is valid, and all the final path coefficients of the second stage are also valid with the exception of the l_{35}, which should be determined separately.

5.1.3 Example

We provide a short example to show the steps of the simple procedure we recommend for quadrants I and III. We took the example from the sample data and model that one can download together with the XLSTAT-PLSPM Excel Add-in[1]. Figure 2 shows the research model. We focus on the complaints construct, which has a single indicator. In the original data, it is continuous, but we mean split it into a binary variable to suit our purpose. Such a measure could indicate whether or not a customer had ever complained. A researcher could conceive this measure as a directly observed non-latent variable and, thereby, justify their using the single indicator (quadrant III). Otherwise, the researcher may conceive the construct as something broader, but only this single indicator is available (Quadrant I). Either way, we explain the practical steps next.

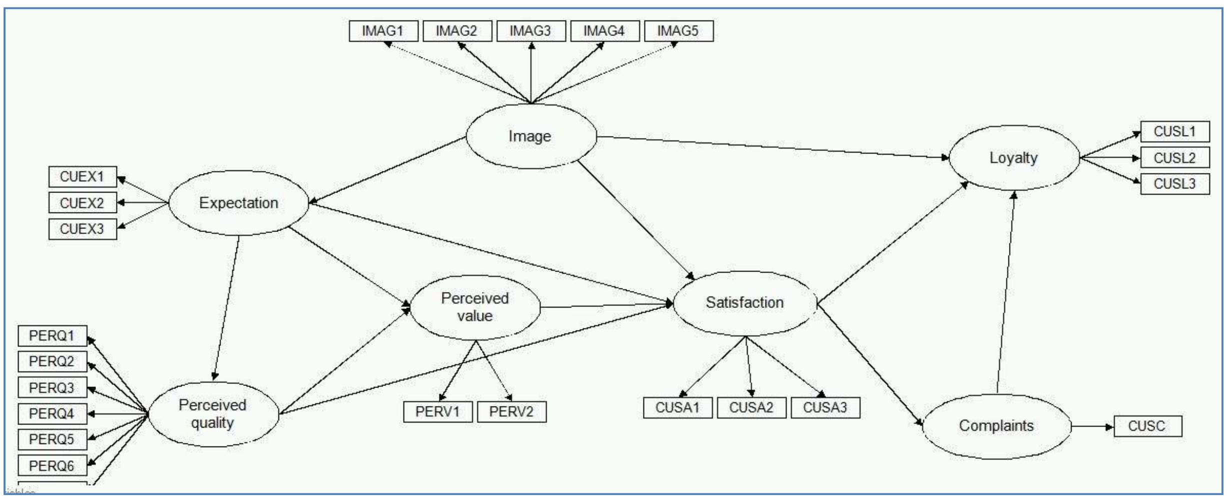


Figure 2. XLSTAT-PLSPM Example

Table 6a shows some records of original data for the satisfaction and dichotomized complaints constructs. We ran the XLSTAT-PLSPM software and opted to output latent values for each variable. The software executed both PLS-SEM stages and generated path coefficients for all of the model's structural links. The final result retains all these path coefficients except for the link from satisfaction to the binary outcome complaints because the OLS regression is not right for predicting the binary complaints. For this one link, we need to take a separate step. Table 6b shows the normalized latent values that the software output for those two constructs (one would ordinarily use standardized values instead of normalized, but then the binary 1's and 0's turn into some other meaningless pair of numbers, so, for clarity, we use normalized values here). Notice that the latent complaints variable is identical with the original data because there is only a single indicator; in Section 5, we refer to this basic fact in explaining why problems A and B do not apply with single-item constructs. In any event, we took the two latent variables and imported them into

[1] Find the software at http://www.xlstat.com/en/products-solutions/plspm.html

SPSS or any statistics software capable of running logistic regressions[2]. Table 6c shows the result on the full dataset.

Table 6. Additional Information for XLSTAT-PLSPM Example

A) Sample of original data			
CUSA1	**CUSA2**	**CUSA3**	**CUSC0**
55.56	33.33	66.67	0
100	100	77.78	1
77.78	66.67	66.67	0
100	100	100	0
100	77.78	77.78	0
77.78	66.67	66.67	1
B) Normalized latent values			
Satisfaction		**Complaints**	
52.55		0	
91.91		1	
70.21		0	
100		0	
84.87		0	
70.21		1	
C) Logistic regression result			
		B	**Sig.**
Satisfaction		.079	.000
Constant		-5.982	.000

5.1.4 Re-visiting the Cautions from Hair and Colleagues

Researchers sometimes cite Hair et al. (2012b) as a basis for not using PLS-SEM whenever a structural model has a binary variable. However, carefully reading their analysis shows that their cautions are both correct and limited and in no way contradicted by anything we have proposed. Hair et al. point to two problems about the appropriateness of using PLS-SEM to handle structural models with a binary single indicator for an endogenous variable. The first problem is that "researchers may decide to use a binary single indicator to measure an endogenous construct…, which proves problematic for approximations in the PLS-SEM algorithm since path coefficients are estimated by OLS regressions" (p. 421). We understand this problem to refer to what we call problem C, which we resolve by simply replacing OLS for that section of the path with an appropriate data-analysis method such as discriminant analysis. The second problem they raise is "using binary indicators in reflective models violates this OLS assumption, because reflective indicators are regressed on the latent variable scores when estimating outer weights" (pp. 421). This problem appears to refer to what we call problem B in which latent variables are regressed on their manifest indicators. Recall that this problem does not apply at all in the case of a single indicator (quadrants I and III). In the case of multiple indicators (quadrant II), scholars differ in opinion. (Note that careful attention to Hair's use of words indicates that they never said that this is a problem in the case of a single indicator. They say the first problem arises when one uses a "single binary indicator" and the latter when one uses "binary indicators in reflective models" but not when it is a single indicator). To our understanding, Hair et al. completely agree that the only problem that arises in quadrant III is the easily rectifiable issue of finding path model coefficients.

We found two items in the literature that seem related to our approach, but, on closer inspection, they do not use PLS-SEM for a binary variable. Henseler (2003) suggests saving the latent values and running

[2] If one uses SmartPLS version 3.1.9 or older as their PLS-SEM software, one should use the standardized latent value scores. There is a bug in the way these older versions output the unstandardized latent values (see http://www.smartpls.de/release_notes).

separate analyses with the binary variable, but his idea is to first remove the binary variable from the model when running the PLS-SEM first stage, which we argue is unnecessary. Festge & Schwaiger (2007) appear to talk about running PLS-SEM, but the case they consider is actually when the outcome variable was originally continuous. Only after running the PLS-SEM do they median split the outcome variable in preparation for a second data analysis using logistic regression. We did not find any examples of researchers who run PLS-SEM on a model with a binary outcome.

5.1.5 Solution 2: Regularized Generalized Canonical Correlation Analysis (RGCCA)

In this section, we add an additional perspective that lends support to using PLS-SEM with binary endogenous outcomes. A stream of research in statistics aims to establish PLS on a more sure statistical footing. One prominent and widely cited work in this stream is regularized generalized canonical correlation analysis (RGCCA) (Tenenhaus & Tenenhaus, 2011) and its corresponding software implementation in the XLSTAT-PLSPM add-in to Excel and RGCCA package for R. RGCCA has a close relationship with PLS-SEM. It has a tau parameter. If the tau parameter is set to 0, then RGCCA is identical to PLS (first stage) mode B. If the tau parameter is set to 1, RGCCA and PLS (first stage) mode A are very close. When the tau parameter varies between 0 and 1, the latent variable mode stands in between mode A and mode B.

The key point for our purposes is that, unlike PLS-SEM, whose use with binary indicators has raised doubts, RGCCA unequivocally allows binary outcomes because, unlike PLS-SEM, RGCCA optimizes an explicit criterion, and this criterion is perfectly valid for binary variables. For example, Figure 7 in Tenenhaus and Tenenhaus (2011) has a categorical, single-indicator outcome (the two variables “stable demo” and “dictatorship” are dummy variables to encode the three-category outcome). RGCCA provides researchers with an alternative to the first stage of PLS-SEM. Researchers may obtain the scores of the latent variables from the first stage and then choose the appropriate regressions or other analyses to analyze the path coefficients for the relationships between the latent variables.

In two ways, RGCCA is relevant to our discussion of structural equation modeling with binary outcomes. First, one can use the RGCCA algorithm itself to calculate latent variables, and it is valid for binary outcomes. Second, for those who wish to use PLS-SEM, RGCCA is closely related to PLS-SEM, so the fact that RGCCA is known to be valid may evidence PLS-SEM’s validity to the extent one believes it is meaningful to even speak about the statistical “validity” and violations of statistical assumptions for a method such as PLS-SEM that has no explicit global criterion.

6 Conclusion

Some of the prior literature on PLS-SEM appears to warn scholars against using PLS-SEM for analyzing models with binary endogenous variables. For that reason, many IS researchers may have pre-emptively avoided using PLS-SEM in such cases and instead used other methods such as covariance-based methods or path modeling. In this paper, we distinguish between various ways in which one may use a binary variable and the different potential problems that may arise in each case. In the particular case of a binary item that is an observable variable (quadrant III; see Table 4), one faces no problem using PLS-SEM’s first stage, which finds latent variables. Today’s software runs OLS regressions as a second stage, but, since this is not appropriate for a binary outcome, one should use the latent predictors of the binary variable(s) as input in a separate analysis. A separate conceptual problem arises when using a binary (or any other single) variable as the sole indicator for a broader latent construct (quadrant I), and yet another problem may arise (although it is a matter of debate) when it is one of many indicators (quadrant II). However, authors better handle these issues, too, when their specific nature is clear. We contribute to that clarity. Since the solution is trivial when one conceives the variable as being directly observed and more open to debate when one conceives the variable as latent, this paper may rekindle debate about the validity of conceiving certain outcomes such as consumer choice or technology adoption as directly observed (i.e., as wholly encompassed by a single indicator). We would also welcome such a result.

References

Aguirre-Urreta, M. I., & Marakas G. M. (2014). Partial least squares and models with formatively specified endogenous constructs: A cautionary note. *Information Systems Research*, *25*(4), 761-778.

Alwitt, L. F., & Pitts, R. E. (1996). Predicting purchase intentions for an environmentally sensitive product. *Journal of Consumer Psychology*, *5*(1), 49–64.

Banker, R., Hu, N., Pavlou, P. A., & Luftman, J. (2011). CIO reporting structure, strategic positioning, and firm performance. *MIS Quarterly*, *35*(2), 487-504.

Bapna, R., Goes, P., Wei, K. K., & Zhang, Z. (2011). A finite mixture logit model to segment and predict electronic payments system adoption. *Information Systems Research*, *22*(1), 118-133.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166-173.

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalized linear regression. *Computation Statistics and Data Analysis*, *48*, 17-46.

Boulesteix, A. L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, *3*(1). Retrieved from http://www.degruyter.com/dg/viewarticle/j$002fsagmb.2004.3.1$002fsagmb.2004.3.1.1075$002fsagmb.2004.3.1.1075.xml;jsessionid=E2A0A1146FCAFFB86B3DF0D8C4C087C5

Bullock, H. E., Harlow, L. L., & Mulaik, S. A. (1994). Causation issues in structural equation modeling research. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(3), 253-267.

Caliendo, M., Clement, M., Papies, D., & Scheel-Kopeinig, S. (2012). The cost impact of spam filters: Measuring the effect of information system technologies in organizations. *Information Systems Research*, *23*(3), 1068-1080.

Ceccagnoli, M., Forman, C., Huang, P., & Wu, D. J. (2012). Cocreation of value in a platform ecosystem: The case of enterprise software. *MIS Quarterly*, *36*(1), 263-290.

Chin, W., Marcolin, B., & Newsted, P. (1996). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and voice mail emotion/adoption study. In *Proceedings of International Conference of Information Systems*.

De, P., Hu, Y., & Rahman, M. S. (2013). Product-oriented Web technologies and product returns: An exploratory study. *Information Systems Research*, *24*(4), 998-1010.

des Reis, A., & Soares, A. (2006). Electronic procurement systems: identifying factors that foster their adoption. In *Proceedings of European Conference on Information Systems*.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434-449.

Festge, F., & Schwaiger, M. (2007). The drivers of customer satisfaction with industrial goods: An international study. *Advances in International Marketing*, *18*, 183-212.

Gefen, D., & Pavlou, P. A. (2012). The boundaries of trust and risk: The quadratic moderating role of institutional structures. *Information Systems Research*, *23*(3), 940-959.

Gefen, D., Straub, D. W., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, *4*, 1-30.

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How is the mobile Internet different? Search costs and local activities. *Information Systems Research*, *24*(3), 613-631.

Goes, P. B., Karuga, G. G., & Tripathi, A. K. (2012). Bidding behavior evolution in sequential auctions: Characterization and analysis. *MIS Quarterly*, *36*(4), 1021-1042.

Goes, P. B., Lin, M., & Au Yeung, C. (2014). Popularity effect in user-generated content: Evidence from online product reviews. *Information Systems Research*, *25*(2), 222-238.

Gopal, A., & Koka, B. R. (2012). The asymmetric benefits of relational flexibility: Evidence from software development outsourcing. *MIS Quarterly*, *36*(2), 553-576.

Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *MIS Quarterly*, *36*(3), 981-1001.

Gu, B., Konana, P., Raghunathan, R., & Chen, H. M. (2014). The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, *25*(3), 604-617.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics*, *3*(4), 283-297.

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

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012a). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5-6), 320-340.

Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012b). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, *40*(3), 414-433.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. NY: Academic Press.

Henseler, J. (2003). From switching intention to switching behavior. Towards an integration of PLS and logistic regression. In *Proceedings of the PLS'03 International Symposium* (pp. 143-145).

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M., & Calantone, R. J. (2014). Common beliefs and reality about partial least squares: Comments on Rönkkö & Evermann (2013). *Organizational Research Methods*, *17*(2), 182-209.

Hildebrand, C., Häubl, G., Herrmann, A., & Landwehr, J. R. (2013). When social media can be bad for you: Community feedback stifles consumer creativity and reduces satisfaction with self-designed products. *Information Systems Research*, *24*(1), 14-29.

Ho, S. Y., & Bodoff, D. (2014). The effects of web personalization on user attitude and behavior: An integration of the elaboration likelihood model and consumer search theory. *MIS Quarterly*, *38*(2), 497-520.

Jakobowicz, E., & Derquenne, C. (2007). A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy. *Computational Statistics and Data Analysis*, *51*(8), 3666-3678.

Langer, N., Forman, C., Kekre, S., & Sun, B. (2012). Ushering buyers into electronic channels: An empirical analysis. *Information Systems Research*, *23*(4), 1212-1231.

Lee, L., Petter, S., Fayard, D., & Robinson, S. (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, *12*(4), 305-328.

Li, S., Shang, J., & Slaughter, S. A. (2010). Why do software firms fail? Capabilities, competitive actions, and firm survival in the software industry from 1995 to 2007. *Information Systems Research*, *21*(3), 631-654.

Lohmoller, J. B. (1989). *Latent variable path modeling with partial least squares*. New York, NY: Springer.

Mani, D., Barua, A., & Whinston, A. B. (2010). An empirical analysis of the impact of information capabilities design on business process outsourcing performance. *MIS Quarterly*, *34*(1), 39-62.

Mani, D., Barua, A., & Whinston, A. B. (2012). An empirical analysis of the contractual and information structures of business process outsourcing relationships. *Information Systems Research*, *23*(3), 618-634.

Marcoulides, G. A., Chin, W., & Saunders, C. (2009). Foreword: A critical look at partial least squares modeling. *MIS Quarterly*, *33*(1), 171-175.

Monecke, A., & Leisch, F. (2012). semPLS: Structural equation modeling using partial least squares. *Journal of Statistical Software*, *48*(3), 1-32.

Moreno, A., & Terwiesch, C. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research*, *25*(4), 865-886.

Murray, K. B., & Häubl, G. (2011). Freedom of choice, ease of use, and the formation of interface preferences. *MIS Quarterly*, *35*(4), 955-976.

Oestreicher-Singer, G., & Zalmanson, L. (2013). Content or community? A digital business strategy for content providers in the social age. *MIS Quarterly*, *37*(2), 591-616.

Özpolat, K., Gao, G., Jank, W., & Viswanathan, S. (2013). The value of third-party assurance seals in online retailing: An empirical investigation. *Information Systems Research*, *24*(4), 1100-1111.

Peng, G., & Dey, D. (2013). A dynamic view of the impact of network structure on technology adoption: The case of OSS development. *Information Systems Research*, *24*(4), 1087-1099.

Rice, S. C. (2012). Reputation and uncertainty in online markets: An experimental study. *Information Systems Research*, *23*(2), 436-452.

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5-6), 341-358.

Rigdon, E. E., Becker, J. M., Rai, A., Ringle, C. M., Diamantopoulos, A., Karahanna, E., Straub, D., & Dijkstra, T. (2014). Conflating antecedents and formative indicators: A comment on Aguirre-Urreta and Marakas. *Information Systems Research*, *25*(4), 780-784.

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: A critical look at the use of PLS-SEM in *MIS Quarterly*. *MIS Quarterly*, *36*(1), iii-xiv.

Rönkkö M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. *Organizational Research Methods*, *16*(3), 425-448.

Rosipal, R., & Kramer, N. (2006). Overview and recent advances in partial least squares. In Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (Eds.), *Subspace, latent structure and feature selection techniques* (pp. 34-51). New York, NY: Springer.

Russolillo, G. (2012). Non-metric partial least squares. *Electronic Journal of Statistics*, *6*, 1641-1669.

Sarstedt, M., Ringle, C. M., Henseler, J., & Hair, J. F. (2014). On the emancipation of PLS-SEM: A commentary on Rigdon (2012). *Long Range Planning*, *47*(3), 154-160.

Singh, P. V., & Phelps, C. (2013). Networks, social influence, and the choice among competing innovations: Insights from open source software licenses. *Information Systems Research*, *24*(3), 539-560.

Sosik, J. J., Kahai, S. S., & Piovoso, M. J. (2009). Silver bullet or voodoo statistics? A primer for using the partial least squares data analytic technique in group and organization research. *Group Organization Management*, *34*(1), 5-36.

Susarla, A., & Barua, A. (2011). Contracting efficiency and new firm survival in markets enabled by information technology. *Information Systems Research*, *22*(2), 306-324.

Susarla, A., Subramanyam, R., & Karhade, P. (2010). Contractual provisions to mitigate holdup: Evidence from information technology outsourcing. *Information Systems Research*, *21*(1), 37-55.

Tan, C. H., Sutanto, J., Phang, C. W., & Gasimov, A. (2014). Using personal communication technologies for commercial communications: A cross-country investigation of email and SMS. *Information Systems Research*, *25*(2), 307-327.

Tenenhaus, M., & Hanafi, M. (2010). A bridge between PLS path modeling and multi-block data analysis. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 99-123). New York, NY: Springer.

Tenenhaus, A., & Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, *76*(2), 257-284.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis*, *48*(1), 159-205.

Vinzi, V. E., Russolillo, G., & Trinchera, L. (2010a). *A joint use of PLS regression and PLS path modelling for a data analysis approach to latent variable modeling.* Retrieved from www.researchgate.net/publication/228524574_A_Joint_Use_of_PLS_Regression_and_PLS_Path_Modelling_for_a_Data_Analysis_Approach_to_Latent_Variable_Modelling/file/9fcfd50925a0424d95.pdf

Vinzi, V. E., Trinchera, L., & Amato, S. (2010b). PLS path modeling: From foundations to recent developments and open issues for model assessment and improvement. In V. E. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares* (pp. 47-82). New York, NY: Springer.

Wattal, S., Telang, R., Mukhopadhyay, T., & Boatwright, P. (2012). What's in a “name”? Impact of use of customer information in email advertisements. *Information Systems Research*, *23*(3), 679-697.

Wu, L. (2013). Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Information Systems Research*, *24*(1), 30-51.

Xue, L., Ray, G., & Gu, B. (2011). Environmental uncertainty and IT infrastructure governance: A curvilinear relationship. *Information Systems Research*, *22*(2), 389-399.

Ye, S., Gao, G., & Viswanathan, S. (2014). Strategic behavior in online reputation systems: Evidence from revoking on eBay. *MIS Quarterly*, *38*(4), 1033-1056.

Appendix A: Choice between Latent and Observable Binary Variables

Choosing whether to conceive of a variable as latent or observed requires that researchers grapple with interesting conceptual questions regarding the meaning of their model. System adoption is an excellent example. A researcher may choose to conceptualize that, in the context of a given research model, whether a user or firm adopts a particular system is a directly observed behavior that fully captures the theoretically predicted behavior. In this case, one conceives it as a directly observed variable, not a latent variable, and it falls into quadrant III. Alternatively, the researcher may conceive that the observed adoption of a particular system is just one indicator of some broader latent variable. For example, the theoretically predicted outcome may be called “system adoption” but may really refer to the firm’s use of technology in a broader sense, with the adoption of systems (or of a specific system) being only one indicator of that more general usage or tendency. In such a case, the observed system adoption is just one of many possible indicators of an underlying latent construct. In this case, the example would fall into quadrant I if the binary outcome were the only indicator used in the study. Apparently along these lines, in the marketing context, Hair et al. (2012a) list a “choice situation” as an example of using a single indicator for a latent construct. Typically, though not necessarily, one will conceive survey items as indicating a latent variable rather than as variables that are measured in more objective ways, such as a financial or behavioral outcome. In the IS field, one could conceive actual system adoption as one of many possible indicators of a latent intention to adopt that is viewed as the outcome of ultimate interest. Yet, this approach would turn on its head the historical conceptualization in IS that has treated actual adoption decisions as the ultimate variable of interest and intention-to-adopt as a second-best surrogate for it. In addition, many studies that have system adoption as an outcome have only a single indicator for that outcome, so, if we conceive system adoption as an indicator for a broader latent construct, all those studies would suffer from the serious drawback of having used a single indicator to capture a latent construct.

About the Authors

David Bodoff is Senior Lecturer in the Department of Business Administration at the University of Haifa. His research is centered on search, including technical and behavior aspects of search engines, as well as consumer search. His research draws on computer science, economics, and game theory, and has been published in *ACM Transactions on Information Systems*, *Experimental Economics*, *MIS Quarterly*, *Information Systems Research*, HCI, and in conferences such as *ACM SIGIR* and *NIPS*.

Shuk Ying Ho is currently a professor at the Australian National University. She completed her PhD in the Hong Kong University of Science and Technology in 2004. Her doctoral dissertation examined how web personalization influences the behavior of online users and her current research portfolio reflects a continuing interest in this area. Her research focuses on the area of human computer interaction, electronic commerce, technology adoption, and electronic government. Her publications have appeared in *MIS Quarterly*, *Information Systems Research*, *European Journal of Operational Research*, *The Journal of Organizational Computing and Electronic Commerce*, *The Journal of American Society Information Science and Technology*, *The Journal of E-Government*, *Electronic Markets* and *International Journal of Human-Computer Interaction*.