

PLS: Caveat Emptor

On the Adoption of Partial Least Squares in Psychological Research: Caveat Emptor

Mikko Rönkkö¹

Aalto University, School of Science
Department of Industrial Management, Institute of Strategy and Venturing
PO Box 15500, FI-00076 Aalto, Finland
phone: + 358 50 387 8155
email: mikko.ronkko@aalto.fi

Cameron N. McIntosh

Public Safety Canada
269 Laurier Avenue
Ottawa, Ontario, Canada
K1A 0P8

John Antonakis

Faculty of Business and Economics
University of Lausanne
Internef #618
CH-1015 Lausanne-Dorigny, Switzerland

Accepted for publication in *Personality and Individual Differences*

¹ Corresponding author

Abstract

The partial least squares technique (PLS) has been touted as a viable alternative to latent variable structural equation modeling (SEM) for evaluating theoretical models in the differential psychology domain. We bring some balance to the discussion by reviewing the broader methodological literature to highlight: (1) the misleading characterization of PLS as an SEM method; (2) limitations of PLS for global model testing; (3) problems in testing the significance of path coefficients; (4) extremely high false positive rates when using empirical confidence intervals in conjunction with a new “sign change correction” for path coefficients; (5) misconceptions surrounding the supposedly superior ability of PLS to handle small sample sizes and non-normality; and (6) conceptual and statistical problems with formative measurement and the application of PLS to such models. Additionally, we also reanalyze the dataset provided by Willaby et al. (2015; doi:10.1016/j.paid.2014.09.008) to highlight the limitations of PLS. Our broader review and analysis of the available evidence makes it clear that PLS is not useful for statistical estimation and testing.

Keywords: Partial least squares, structural equation modeling, capitalization on chance, significance testing, model fit.

On the Adoption of Partial Least Squares in Psychological Research:

Caveat Emptor

1. Introduction

Researchers have recently suggested that Partial Least Squares (PLS) should be considered as a viable estimator for testing theoretical models in psychological research (Willaby, Costa, Burns, MacCann, & Roberts, 2015). These authors have largely followed the arguments of PLS advocates in the marketing, information systems research, and strategic management domains – where the technique has become a methodological mainstay – making several claims about the superiority of PLS relative to the more commonly-used technique in psychology, latent variable structural modeling using (SEM). However, the veracity of such claims has been challenged by a number of recent works highlighting serious problems with the PLS method (e.g., Antonakis, Bendahan, Jacquart, & Lalive, 2010; Goodhue, Lewis, & Thompson, 2012; McIntosh, Edwards, & Antonakis, 2014; Rönkkö, 2014; Rönkkö & Evermann, 2013; Rönkkö & Ylitalo, 2010); these critiques were not considered in the Willaby et al. (2015) paper and have received little uptake in the PLS literature more broadly.

In light of these issues, the purpose of this paper is to provide a “sober second thought” on the purported advantages of PLS discussed by Willaby et al. (2015) and others, by reviewing the existing methodological literature and providing novel empirical demonstrations. We start by first explaining why labeling PLS as a SEM method is highly misleading; it is much more informative to consider PLS only as an indicator weighting system for creating composite variables. Second, we discuss the limitations of PLS regarding its ability to test how well a given theoretical model represents the observed data (i.e., the lack of well-established

overidentification tests). Third, we examine problems surrounding statistical inference on path coefficients, as well as the undesirable effects on confidence intervals stemming from a “sign change correction” built into some popular PLS software packages. Fourth, we critically discuss a set of inaccuracies that have long been perpetuated in the PLS literature, namely that PLS is better able to deal with small sample sizes and non-normal variables than can SEM, and that PLS somehow provides a “natural” statistical approach to validating formative models. Overall, our goal is raise awareness among applied researchers in the psychology domain concerning the shortcomings of the PLS method that the article by Willaby et al. (2015) and the majority of other current writings about PLS fail to address.

2. PLS is Simply an Indicator Weighting System

Whereas PLS is currently described as a SEM method (Hair, Hult, Ringle, & Sarstedt, 2014), or as an alternative to SEM (Willaby et al., 2015), this characterization is rather misleading¹. The main problem with defining PLS as a SEM method is that it does not have a coherent theoretical foundation for estimation and inference. Unlike classical maximum likelihood-based SEM, which rests on a unified statistical theory for simultaneous estimation of all model parameters (i.e., measurement and structural effects and the associated overidentification tests; Jöreskog, 1978), the PLS approach consists of an ad hoc collection of statistical procedures that have not been formally analyzed (McDonald, 1996). Thus, labeling PLS as an SEM estimator essentially gives it an ascribed rather than achieved status.

What does PLS actually do? Willaby et al. (2015) describe the algorithm as three stages, of which the first two are required for parameter estimation. In Stage 1, the indicators are combined as weighted composites (i.e., weighted sums), which are of course *not* latent variables

¹ The question of whether PLS is a SEM estimator can be recast as a question of whether we can find a definition for the term “estimator” that matches what PLS does. The answer is that we can (e.g., Greene, 2012, p. 155; Lehmann & Casella, 1998, p. 4), but these definitions are so broad that even a pseudo-random number generator would qualify as an “estimator.”

but merely noisy proxies or substitutes for them. Next, in Stage 2, the composites are used in OLS regression analyses to estimate the path coefficients. Therefore, the PLS method is no different from using summed scales, factor scores, principal components, or any other type of weighted composites in an OLS regression. This simple fact is sometimes obscured in the PLS literature. For example, Lu, Kwan, Thomas, and Cedzynski (2011) explain that “[compared to regression with composites] PLS is more mysterious because estimates are produced by complicated algorithms embedded in specialized statistical software” (p. 259). To remove the mystery and show that the simple PLS algorithm can be implemented with little programming in a general purpose statistical package, we provide a replication of Willaby et al.’s final model as a Stata input file in Supplementary Data 1. That PLS is simply regression with composites logically leads to the conclusion that any purported advantages that PLS might provide over regression with summed scales or other composites by virtue of being a SEM method are simply incorrect (e.g., Gefen, Rigdon, & Straub, 2011; see Rönkkö & Evermann, 2013).

Unfortunately, the widespread mischaracterization of PLS as a SEM technique continues to confuse applied researchers. We discuss first the special case where each composite consists of a single indicator (e.g., a total score or scale mean), such as the model shown in Figure 1 by Willaby et al. (2015). Because there is no need for the multi-indicator PLS weighting system in this situation, the model reduces to conventional OLS regression with observed variables, and so reporting the analysis as PLS is actually a misnomer. Furthermore, the comparison of ML and PLS estimates for the model shown in Willaby et al.’s Figure 1 is uninformative because OLS and ML will yield identical estimates for the saturated and recursive case (i.e., no feedback loops or correlated errors). The comparison is therefore essentially between two maximum likelihood

estimators (ML and OLS), which should provide identical estimates and also do so in our replication included in Supplementary Data 2².

The question of whether PLS is a useful way to estimate multiple-indicator latent variable models is more complex to answer. Given that PLS is best considered as simply an indicator weighting system, it is important to pause and consider the two following questions: Is it a good idea to estimate a latent variable SEM by approximating the latent variables with composites? Second, what is the best way to calculate the composites? The answer to the first question is a resounding no because composites inherit the measurement error in their constituent indicators (Bollen, 1989, pp. 305–306), which renders PLS both inconsistent³ (i.e., estimates do not asymptotically converge to the true population values) and biased (i.e., it over or underestimates the population parameters in finite samples) when applied to latent variable models (Dijkstra, 1983; Hui & Wold, 1982). The issue of consistency is critical if estimates can be of any use in informing policy and practice. As mentioned by Wooldridge in his highly-regarded econometrics textbook (2009), “Although not all useful estimators are unbiased, virtually all economists agree that **consistency** is a minimal requirement for an estimator. The Nobel Prize–winning econometrician Clive W. J. Granger once remarked, ‘If you can’t get it right as n goes to infinity, you shouldn’t be in this business.’ The implication is that, if your estimator of a particular population parameter is not consistent, then you are wasting your time.” (p. 168).

² Note that the direct and indirect effects in the mediation model that Willaby et al. (2015) tested are not causally identified, due to violation of the “order condition” (i.e., at least as many exogenous variables, which are excluded from the y equation, are needed to predict the mediators). Unfortunately, we cannot deal with this important issue here, given that the accompanying dataset does not allow sufficient augmentation of the mediation model to satisfy this condition. However, interested readers can refer to a recent chapter by Antonakis, Bendahan, Jacquart, and Lalive (2014) for a user-friendly introduction to causal identification.

³ Dijkstra and his coauthors (Dijkstra & Henseler, 2015; Dijkstra & Schermelleh-Engel, 2014) have proposed using the correction for measurement error attenuation (cf., Muchinsky, 1996) to make PLS consistent (PLSc). However, this correction is generally considered obsolete with the wide availability of modern SEM techniques (e.g., Moosbrugger, Schermelleh-Engel, & Klein, 1997, p. 97). Furthermore, PLSc addresses only one of the two main sources of bias in PLS: aside from attenuation due to measurement error, the PLS estimates are also biased because of capitalization on chance (Rönkkö, 2014), so finite-sample bias is still inevitable (Rönkkö, McIntosh, & Antonakis, 2015).

The second question has attracted much less attention in the literature. A key problem here is that although PLS weights are often referred to as being “optimal”, the literature is not clear for which specific purpose the weights are optimal. Some authors argue that the weights are optimal for maximizing prediction (e.g., Fornell & Bookstein, 1982; Sosik, Kahai, & Piovosio, 2009), whereas others suggest that they maximize reliability (e.g., Chin, Marcolin, & Newsted, 2003); these are of course two very different criteria (Penev & Raykov, 2006; Rönkkö, 2014). Henseler and Sarstedt (2013) attempt to reassure users by stating that the PLS scores are “*optimal in some sense*” (p. 566), but this obviously provides no resolution to the issue. To get some clarification on this issue, we recently asked a leading PLS proponent on a fairly large private research methods- focused email list to comment on the purported advantages of PLS weights over unit weights, and he admitted that he did not know of any proven, concrete advantage that PLS weights would provide⁴.

Another major problem in claiming that the PLS weights are “optimal” is that the PLS literature largely ignores the fact that a number of alternative, more well-established indicator weighting schemes have existed for decades, each of which was developed to solve a specific problem and mathematically proven to be the optimal solution. For example, (a) regression factor score weights maximize the expected reliability of the composite by minimizing the error in the constituent items (Grice, 2001); (b) correlation-preserving factor score weights generate composites that maintain the estimated factor correlations (Grice, 2001; ten Berge, Krijnen, Wansbeek, & Shapiro, 1999); and (c) principal component weights maximize the indicator variance explained by the composite variable (Dunteman, 1989, p. 15). Unfortunately, there is a paucity of studies attempting to systematically determine whether PLS actually yields relatively better composites than the above approaches with respect to the particular criterion that the

⁴ The PLS proponent declined our request for permission to cite or quote the email discussion.

composites were designed to optimize. Considering that PLS weights converge asymptotically to much more well-known factor-analytic weights (Dijkstra, 2010, p. 210), it is not clear why PLS would provide an advantage over this or any other more well-established technique.

The most common approach for constructing composite variables for use in OLS regression analysis is to sum the indicators using equal or unit weights (cf., Bobko, Roth, & Buster, 2007), and therefore it is natural that articles promoting PLS argue that the PLS weights would provide an advantage specifically over this popular indicator weighting system by improving composite reliability (e.g., Gefen et al., 2011). However, the current evidence suggests that PLS rarely enhances reliability over unit weights, and even in those cases where it does, the difference is only marginal (Henseler et al., 2014; McIntosh, Edwards, & Antonakis, 2014; Rönkkö & Evermann, 2013; Rönkkö & Ylitalo, 2010; Rönkkö, 2014). Moreover, decades of research addressing a wide variety of situations shows that the advantage of even ideal weights that maximize composite reliability in a sample over unit weights is non-existent or trivial (e.g., Bobko et al., 2007; Cohen, Cohen, West, & Aiken, 2003, pp. 97–98; Cohen, 1990; McDonald, 1996; Raju, Bilgic, Edwards, & Fler, 1999).

We now use the data provided by Willaby et al. (2015, Supplementary Material 5) to show that the PLS composites are so highly correlated with unit-weighted composites that one system cannot have an advantage over another in terms of composite reliability. Table 1 shows that the unit-weighted composites were nearly perfectly correlated (i.e., 0.976-0.998) with the PLS composites for the model in Figure 2, meaning that in terms of reliability, the differences between the two sets of composites are marginal. More interestingly, Table 1 reveals that the within-method bivariate correlations, which correspond to regression paths between the multi-item composites in the model, are always higher for PLS composites and that this effect is

stronger for weaker correlations. For example, the correlations between Organization and Mechanics with Conscientious are 24% and 19% higher, respectively, in the PLS vs. unit-weighted composite case. If the only difference between the two sets of composites were their reliabilities – as often suggested by the PLS literature – the differences between these correlations should be less than 2.6%⁵, but the observed differences are much larger.

----- Insert Table 1 about here -----

The real reason why the within-method correlation matrices differ is that the PLS weights capitalize on chance to make the correlations larger (Goodhue, Thompson, & Lewis, 2013; Rönkkö, 2014). It is widely known that in most conventional statistical procedures (e.g., correlation, regression, SEM), random sampling error increases the variance of the estimates but not their bias, given that random over and underestimation cancel each other out across repeated sampling so that estimates are correct on average (i.e. unbiased). Even in the case where sampling error and unreliability interact to produce “chance correlations” among the measurement errors in the observed variables (cf., Charles, 2005; Rönkkö, 2014; Zimmerman, 2007), bias is not a concern for conventional correlation and regression methods (Rönkkö, 2014). Unfortunately, in PLSs’ model-dependent weighting scheme, the effects of chance correlations among measurement errors are much more pervasive, leading to bias as well as inefficiency (Goodhue, Lewis, & Thompson, 2007; Goodhue et al., 2013; Rönkkö & Evermann, 2013; Rönkkö, McIntosh, & Antonakis, 2015; Rönkkö, 2014). In particular, Rönkkö’s (2014) simulation studies, which compared regression coefficients across scenarios where chance correlations among measurement errors were present versus artificially removed, showed

⁵ If we know that reliability is the only difference between the composites, then we can apply the well-known attenuation formula (e.g., Cohen, Cohen, West, & Aiken, 2003, sec. 2.10.2) to calculate an upper limit for the difference between the correlations calculated between the different sets of composites. The maximum difference in attenuation between Conscientiousness and Organization can be expressed as $\sqrt{.998^2 + .976^2} \cdot .974$, or 2.6% attenuation.

unequivocally that capitalization on chance fallaciously inflates the correlations between PLS composites; this effect was also clearly visible in all other simulation studies reviewed in the article. Removal of chance correlations, however, had only negligible effects on the estimates obtained from SEM or regression using unit-weighted composites.

Given this overwhelming evidence, the PLS proponents have also recognized that the technique capitalizes on chance to inflate regression estimates. Nevertheless, they argue that referring to this phenomenon as “capitalization on chance” carries an unfair “negative connotation” in the PLS context (Sarstedt, Ringle, & Hair, 2014, p. 133). Yet, the definitions for this term provided in the statistical literature accurately capture the phenomenon that we described in the previous paragraph; that is, it is all about taking advantage of idiosyncratic variance in a finite sample, which therefore means that the population parameter is not accurately estimated. For example, Stevens (2012) defines capitalization on chance as “Seizing on the properties of the sample” (p. xii); and Corsini (2002) refers to it as “The process of basing a conclusion on data biased in a particular direction by chance factors. Purely random factors often seem to show interpretable patterns, and capitalization on chance involves mistaken inferences from these patterns.” (p. 139). The term is used in a similar way in the SEM context (e.g., Bentler, 2010, p. 218; Kline, 2005, p. 33). However, the perspective of Sarstedt and colleagues (2014, p. 133) is that PLSs’ tendency to amplify the effects of chance factors in the data is apparently an “advantage,” given that it offsets the attenuating effects of the measurement error inherent in the composite variables. The problem is that the two sources of bias are not guaranteed to exactly cancel out the impact of attenuation (except by chance) and, as shown by Rönkkö (2014), will often lead to positively biased and inefficient estimates. Considering that we have more than a hundred years of research showing how the effects of measurement error can

be adjusted in regression analysis with composites through the well-known correction for attenuation (cf., Charles, 2005; Muchinsky, 1996), or using errors-in-variables regression (Fuller, 1987), relying on a capitalization on chance in small samples is hardly the optimal approach for dealing with measurement error attenuation (Rönkkö, 2014, pp. 176–177).

3. Model Testing

Does PLS provide any advantage in model testing? Unfortunately, PLS is simply a series of separate regressions that each are just-identified; as such, it cannot provide an overidentification test to examine if the constraints placed on the model parameters (e.g., pathways hypothesized to be zero) are consistent with the observed data. According to Djikstra (2014), Herman Wold, the PLS originator, never actually intended a restrictive model for imposing constraints on the data, making PLS as an analysis system incompatible with a model-testing mindset. Thus, Willaby et al.'s (2015) suggestion that PLS can be used for model testing is untenable.

Instead of relying on inferential statistics derived from statistical theory, current PLS practice focuses on a set of heuristics for assessing model fit. These approaches traditionally include comparing the composite reliability statistics (CR) and average variance extracted (AVE) statistics against rule-of-thumb cutoffs (i.e., the Fornell-Larcker criterion, Fornell & Larcker, 1981), as well as less commonly used Goodness of Fit (GoF) indices based on the model R^2 statistics (Evermann & Tate, 2013; Rönkkö & Evermann, 2013). The problem with the first set of indices is that in a PLS analysis, the factor loadings – estimated as bivariate correlations between the indicators and the composites that they comprise – are severely positively biased (Aguirre-Urreta, Marakas, & Ellis, 2013; Evermann & Tate, 2013; McDonald, 1996; Rönkkö & Evermann, 2013), in turn rendering the CR and AVE statistics (calculated based on the loadings)

positively biased with respect to the true model characteristics that they are intended to reflect. On the other hand, the main shortcoming with the GoF indices is that the predictive power of a model (i.e., R^2) is a poor yardstick for assessing model fit, because not all well-fitting models are strongly predictive and not all strongly predictive models are well-fitting (Henseler & Sarstedt, 2013; McIntosh et al., 2014); additionally, inconsistent estimators can produce models with high R^2 s. Although there are only a limited number of studies addressing these PLS model quality heuristics, they all provide strong evidence that these measures cannot reliably detect model misspecification (Evermann & Tate, 2013; Henseler & Sarstedt, 2013; Rönkkö & Evermann, 2013).

Recognizing that the current practices of model quality assessment cannot reliably detect model misspecification, Henseler, Ringle, and Sarstedt (2015) recently proposed a new statistic that is calculated solely from the indicator correlation matrix without any link to the PLS algorithm. We fully agree with the idea of disconnecting model evaluation from PLS estimates and weights. However, the advantages of Henseler et al.'s new coefficient over inferential procedures for testing discriminant validity in modern SEM frameworks (Raykov, 2011) remain to be determined.

We demonstrate the inability of PLS model quality heuristics to detect model misspecification by replicating Willaby et al.'s (2015) both "expanded" model and the final, reduced version of this model (displayed in their Figure 2), using techniques that have been proven to detect misspecified models (with the exception of equivalent models, which can never be fully known in real applications; Raykov & Marcoulides, 2007). The initial, larger model contained 7 latent variables (1 exogenous, 6 mediators), whereas the final model contained only 4 latent variables (1 exogenous, 3 mediators); each model also included the ultimate observed

outcome of GPA. We estimated a confirmatory factor analysis (CFA) models for both the large and small model using the Lavaan software for SEM (Rosseel, 2012). Neither model fit the data according to the chi-square test of exact fit: $\chi^2(368) = 605.865, p < .001$; and

$\chi^2(1007) = 1703.524, p < .001$. To assure that the significant lack of fit cannot be explained by a having small sample size relative to the model size (i.e., the number of observed variables), we also calculated the Swain correction to the chi-square, which compensates for the model size effect (Herzog & Boomsma, 2009). This alternative test still shows that both models are significantly misspecified: $\chi^2_{Swain}(368) = 560.58, p < .001$; and

$\chi^2_{Swain}(1007) = 1503.92, p < .001$. Approximate fit indices (based on the original ML chi-square), which we report for descriptive purposes further indicate substantial misfit of the model, because none meet the commonly recommended cut-off values (e.g., Hu & Bentler, 1999), 1999): RMSEA = 0.066, CFI = 0.840, and TLI = 0.823 for the smaller model; and RMSEA = 0.069, CFI = 0.738, and TLI = 0.719 for the larger model. Because PLS does not offer any tests or indices of how well the theoretical model reproduces the observed data, Willaby et al. were left unaware that their model of the drivers of GPA was in fact misspecified, which in turn implies that the parameter estimates are untrustworthy (Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; Kolenikov, 2011; McIntosh, 2007; Yuan, Marshall, & Bentler, 2003).

4. Statistical Inference and Sign Change Corrections

The current strategy used to test the significance of individual PLS parameter estimates no less problematic than the approach to assessing overall model quality. The dominant

framework for statistical inference in PLS-based research is null hypothesis significance testing (NHST). Under this approach, an observed test statistic is referred to a theoretical distribution where the null hypothesis of no effect holds, in order to determine the probability of obtaining a test statistic at least as extreme as the observed one if the null were true; this probability is known as the p value. If this probability falls below a pre-defined threshold (typically, 0.05), we can then infer that there is strong evidence against the null hypothesis (Cashen & Geiger, 2004;). In PLS-based studies, p -values are generally obtained by comparing the ratios of the parameter estimates and their bootstrap SEs against the t -distribution (cf., Chin, 1998), a procedure known as the parametric one-sample t -test in the broader statistical literature. This procedure is very problematic because the sampling distribution of PLS estimates is non-normal under the null and thus the ratio of an estimate to its standard error cannot follow the t -distribution, making such comparisons meaningless (Rönkkö & Evermann, 2013; see also McIntosh et al., 2014).

A recent study by Henseler et al. (2014) suggests that confidence intervals could be used to overcome the issue noted above, asserting in particular that users of PLS-Graph need not be concerned, despite the evidence that the confidence intervals produced by the software are severely biased (Lu, Kwan, Thomas, & Cedzynski, 2011; cited in Henseler et al., 2014)⁶. Because non-normality of the estimates makes normal approximation confidence intervals inappropriate, we focus here on empirical confidence intervals, which are formed by using values selected from the bootstrapped estimates as the upper and lower confidence limits. Different selection rules yield different confidence intervals, such as percentile intervals, bias-corrected percentile intervals, or bias-corrected and accelerated intervals (Davison & Hinkley, 1997, Chapter 5). Using empirical confidence intervals with PLS estimation has not been

⁶ We refer to a particular PLS implementation because neither the study by Lu et al. (2011) nor the user manual of the PLS-Graph software (Chin, 2001) report what technique was used to calculate the confidence intervals.

thoroughly studied, and there is at least one serious issue that has been overlooked in the PLS literature.

Statistical inferences with PLS is further complicated by the so-called “sign change corrections”, which involve selective sign reversals on all quantities (i.e., weights, loadings, and regression coefficients) across the individual bootstrap replications prior to calculating summary statistics from the bootstrap replications. This procedure ensures that the signs of these values in the bootstrap replicates are consistent with those from the analysis of the original data set. However, the more general bootstrap literature does not recognize any similar corrections (e.g., Davison & Hinkley, 1997; Efron & Tibshirani, 1993), and in fact, the sign change corrections violate the basic principle of bootstrapping, which states that the replicated statistic should be the same (uncorrected) statistic that was used in the original estimation (Davison & Hinkley, 1997, Chapter 1).

These corrections are often motivated by citing an article by Wold (1985; see e.g., Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005; Henseler, Ringle, & Sinkovics, 2009), in which he explained that the signs of the indicator weights are indeterminate; the PLS composites are standardized by multiplying the raw composites by a scalar f which can have either a positive or negative sign (Wold, 1985, Eq. 16b). To resolve this ambiguity, he suggested choosing the sign so that majority of the indicators are positively correlated with the composites. Unfortunately this decision rule is not implemented in the PLSX software (Lohmöller, 1988), which PLS-Graph relies on for the PLS algorithm. To compensate for this omission, the PLS-Graph software included two post-processing options called “Individual Sign Correction” (ISC) and the “Construct-level Sign Correction” (CSC) that are applied to bootstrap replications. The ISC is applied individually to each parameter estimate so that all replications have the same sign

as the original estimate. The CSC is less radical, comparing the signs of weights one block at a time and basing the reversal decisions on these comparisons. At least two different versions of this latter correction have been presented in the literature: Hair et al. (2014, p. 135) explain that the sign reversals are based on the signs of indicator weights, whereas Tenenhaus et al. (2005) describe a decision criterion that also takes the magnitude of the weights into account. The PLS proponents argue that sign change corrections are needed because standard errors “increase[...] dramatically without any real meaning if the sign changes are not properly taken into account (Tenenhaus et al., 2005)” (Henseler et al., 2009, p. 307). Rather than focusing on the magnitude, the correct criterion would be whether the bootstrap SEs are actually unbiased estimates of the SD of the estimates over repeated samples, but none of the more recent papers on the correction address the bias of SEs.

Instead of making inference more robust, applying the sign change corrections makes matters worse. Particularly, when the more aggressive ISC is used, all bootstrap replications are constrained to have the same sign, effectively truncating the bootstrap distribution at zero. We demonstrate this in Figure 1, which shows the distribution of the bootstrap replications of one parameter estimate after ISC, as reported in Willaby et al.’s (2015) Supplementary Data 4. There is no scenario where this type of zero-truncated bootstrap distribution would be desirable, because it ensures that the empirical confidence intervals will never contain zero, resulting in a *100% false positive rate*. This effect can also be seen in the confidence intervals reported in Supplementary Data 4 provided by Willaby et al., who seemingly did not notice that two methods of inference – p value-based NHST and empirical confidence intervals – led to different conclusions for the same data set. If they had relied on their empirical confidence intervals to evaluate the model pathways, they would have retained their original model in its entirety.

---- Insert Figure 1 about here ----

We further illustrate the behavior of different types of confidence intervals, the parametric one-sample t -test and the various sign-correction alternatives using the simple population structural equation model studied by both Rönkkö and Evermann (2013) and Henseler, Ringle, and Sarstedt (2015). The population model has two uncorrelated latent variables, each measured with three indicators loading at 0.6, 0.7, and 0.8. Note that because the latent variables are uncorrelated, estimating the regression between the composites representing the latent variables allows us to directly determine the null distribution (and Type I error rates or “false positives”) of the associated t -statistic under the various sign change options. We examined the behavior of the t -statistic using a Monte Carlo simulation with 1000 replications of samples of 100. The standard errors required for calculating the t -statistic were calculated by bootstrapping each Monte Carlo replication 500 times. The analyses were conducted using the *matrixpls* package for R (Rönkkö, 2015). The initial simulation resulted in 13 non-convergent results, which we reran to obtain 1000 convergent replications. The R code for the simulation is available in Supplementary Data 3.

Figure 2 shows the sampling distribution of the calculated t -statistics compared to the t -distribution, which is typically used as the reference distribution for NHST with PLS. For this particular model, the CSC was almost always applied to both composites if it was triggered, and thus the sampling distributions of the t -statistics with this correction are indistinguishable with the case where no correction was applied and are omitted for clarity. The effect was the same for the decision rule presented by Hair et al. (2014, p. 135) and Tenenhaus et al. (2005). Because the relationship between the two composites in our example model is a standardized simple regression slope (i.e., bivariate correlation), both the CSC version presented by Tenenhaus et al.

(2005) and Wold's (1985) original sign indeterminacy solution lead to identical results.

However, for models containing multiple regressions with correlated predictors, the techniques will not be equivalent.

The ISC correction results in much larger t -values, which is to be expected considering that: (1) the purpose of the sign-change corrections is to produce smaller standard deviations; and (2) the ISC is more aggressive than the CSC. Based on Figure 2, it is clear that the test statistics do not follow the reference distribution, and consequently, any inferences made by comparing these to a t -distribution would likely be incorrect. These results are further analyzed in Table 2, which shows the number of times that the null hypothesis of no effect is rejected for a given significance level. Because the null hypothesis was true in our population model, these statistics unambiguously showcase false positive rates. With both the CSC and uncorrected results, the false positive level is nominal for the weakest significance level ($p < 0.10$), but starts to increase when moving to smaller p -values. The results of using the ISC show false positive rates up to two orders of magnitude larger than the nominal rates. Based on these findings, it is clear that none of these techniques should be used for statistical inference.

----- Insert Figure 2 and Table 2 about here -----

To be sure, PLS proponents have noted that the ISC frequently rejects the null hypothesis in cases where results are non-significant with no sign change correction (Hair et al., 2014, Chapter 5; Hair, Sarstedt, Pieper, & Ringle, 2012). However, they seem unaware of the extreme statistical phenomenon that we have revealed here, namely a 100% false positive rate. Rather, works providing guidance in the use of PLS typically present the CSC as a middle ground or compromise between the two extremes of no sign correction and the ISC, and advise that the CSC be used as a check on whether significant results obtained under the ISC can be considered

valid (Hair et al., 2014, Chapter 5, 2012). However, we suspect that the pervasive and well-known problem of *confirmation bias* (cf., Nickerson, 1998) will have the unfortunate effect of motivating many researchers to, perhaps unwittingly, build a case for justifying the acceptance of significant results under the ISC; indeed, such justification is not in short supply, as authoritative papers on PLS are available that directly recommend the use of this procedure (cf., Henseler et al., 2009, p. 307).

5. On Other Purported “Advantages” of PLS Over SEM

In addition to the issues raised thus far, Willaby et al. (2015) repeat some additional purported advantages over ML SEM that are frequently discussed in the PLS literature: lower sample size requirements, less restrictive distributional assumptions, more enhanced capabilities for exploratory modeling, and a more natural approach to formative measurement. In this section, we briefly reflect on these four claims in light of both current methodological developments and empirical evidence.

The issue of sample size and non-normal data is poorly understood (Dijkstra, 2010, sec. 3; Goodhue et al., 2012; Marcoulides & Chin, 2013; McIntosh et al., 2014; Rönkkö & Evermann, 2013), and PLS proponents have not addressed why the technique would be expected to work well in these scenarios. Instead, the choice of PLS over ML SEM is often justified by correctly stating that the latter may be biased in small samples and with non-normal data, and then claiming that PLS must therefore be used instead. However, this argument is illogical because it suggests that we replace a potentially biased estimator with one that is known to be both biased and inconsistent. Arguing that an estimator simultaneously makes fewer assumptions about the data and works well with smaller samples also violates a basic principle in statistical estimation: the parameters are estimated from a combination of sample information

(i.e., the observed data) and nonsample information (i.e., parameter constraints, assumed distributions). The problem with small sample size is that there is insufficient information in the available data for accurate estimation, which can only be addressed by introducing more information to the estimator, either by collecting more data or adding more nonsample information in the form of assumptions and parameter constraints. It should therefore be obvious that the idea that a method works better with small samples is fundamentally incompatible with the assertion that it requires less restrictive assumptions. Indeed, in a recent simulation study on PLS with both small samples and non-normal data, Westland (2015, Chapter 3) found extensive bias and high false positive rates, and concluded that: "Responsible design of software would stop calculation when the information in the data is insufficient to generate meaningful results, thus limiting the potential for publication of false conclusions. Unfortunately, much of the methodological literature associated with PLS software has conflated its ability to generate coefficients without abnormally terminating as equivalent to extracting information" (p. 42).

Moreover, numerous innovations have been developed over the last couple of decades in the SEM domain for dealing with small samples and non-normality. For example, corrections to the chi-square statistic will help adjust for the "model size effect" (i.e., low $N:p$ ratio; Herzog & Boomsma, 2009; Yuan, Tian, & Yanagihara, 2013) and non-normal distributions can be addressed with numerous techniques including multivariate transformations (e.g., Yuan, Chan, & Bentler, 2000), modified test statistics (e.g., Tong, Zhang, & Yuan, 2014), and robust estimators (e.g., Satorra & Bentler, 1994). Given the existence of these well-supported corrective and analytically-derived procedures, there is simply no need to resort to regression with composites; . Some PLS methodologists have explicitly recognized the evolving capabilities of SEM to deal with violations of distributional assumptions, and state that non-normality is an "obsolete"

reason to choose PLS over SEM (cf. Gefen et al., 2011, p. vii). However, such acknowledgments are still rare in the PLS literature.

Willaby et al. (2015) also suggest that PLS is appropriate for exploratory research. An overarching problem with this assertion is that in terms of model specification, PLS is in fact no different from SEM, because both techniques require precise *a priori* decisions on the number of latent variables and their respective indicators (the measurement model), as well as on the system of model pathways (the structural model) (McIntosh et al., 2014; Rönkkö & Evermann, 2013). Moreover, using the PLS indicator weighting is altogether questionable unless supported by a very strong theory (Rönkkö & Evermann, 2013). Given that the weights are determined by the relationships between adjacent composites, the PLS indicator weighting system requires a strong nomological network (i.e., high correlations among path-connected constructs) for accurate weight calibration (e.g., Dijkstra & Henseler, 2015). This requirement is essentially irreconcilable with the notion of exploratory research, which has been defined by leading PLS supporters as the “search for latent patterns in the data in case there is no or only little prior knowledge on how the variables are related” (Hair et al., 2014, p. 3). If a researcher has a paucity of information on the strength of model pathways, unit weights provide a much more appealing way of constructing composite variables in exploratory scenarios, given that they assume nothing about correlations between the latent variables and have been shown to be robust across a wide range of situations (e.g., Bobko et al., 2007; Cohen et al., 2003, pp. 97–98; Cohen, 1990; McDonald, 1996; Raju et al., 1999). Additionally, SEM offers a wide array of well-developed exploratory modeling tools. For example, there are modification indices that help locate misspecified parameter constraints if the initial model fails to fit the data (Hancock, 1999), as

well as several automated model search procedures for situations where a clear motivating theory is absent (Marcoulides & Ing, 2012).

The final area where Willaby et al. (2015) claim an advantage of PLS over SEM is formative measurement. Unfortunately, they fail to point out that formative measurement is a highly controversial subject involving multiple recent debates (e.g., Aguirre-Urreta & Marakas, 2014a; Bainter & Bollen, 2015; Howell, Breivik, & Wilcox, 2007; Lee, Cadogan, & Chamberlain, 2014). The numerous statistical issues involved with estimating and testing such models, as well as a lack of measurement theory that would be compatible with the formative approach, have led to calls for abandoning formative measurement altogether (Edwards, 2011; Hardin, Chang, Fuller, & Torkzadeh, 2011) or at least changing the term to break the connection with formative models and measurement (Markus & Borsboom, 2013, p. 172; Rhemtulla, Bork, & Borsboom, 2015). A major statistical issue with formative measurement with survey data is that “all variance in the items is treated as meaningful” (Willaby et al., 2015, pp. 2–3). In other words, the use of formative indicators necessitates rejecting the notion of measurement error, which is implausible. Considering that alternative modeling approaches can be used to accomplish the same goals as formative measurement but without the unrealistic assumption of error-free variables, the use of formative indicators is not justified (Edwards, 2011).

Even if we make the unrealistic assumption that formative model is correct, it is unclear why we should expect the PLS indicator weighting system would work well with this type of model (Rönkkö, Evermann, & Aguirre-Urreta, 2015). The idea that PLS would be a useful tool for estimating formative models cannot be found in the seminal papers (e.g., Wold, 1982, 1985), nor in the extensions presented by Lohmöller (1989). Instead, the idea appears to originate from Fornell and Bookstein (1982), who seemed to have confused two techniques for constructing

PLS composites (i.e., Mode A and Mode B) with the structure of the statistical model that is estimated, after which the subsequent literature started to build on this basic misunderstanding (Chin, 1998; Fornell & Cha, 1994; Tenenhaus et al., 2005). As Rigdon (2012) aptly notes, the parallels drawn in the PLS literature between Mode A and reflective measurement “may be described as an article of faith among regular PLS path modeling users and among those who write about the method” (p. 344). Therefore, that PLS is naturally well-suited for estimating formative models appears to be a fundamental misconception, which became an accepted fact after being repeated by applied PLS researchers and methodologists alike. Fortunately, recent research has started to seriously address the veracity of these claims (Aguirre-Urreta & Marakas, 2014b; Rigdon et al., 2014; Aguirre-Urreta & Marakas, 2014a). Lastly, even if one adopts the emerging view that formative indicators do not have any causal properties, and that formative variables are only convenient aggregates rather than conceptually meaningful entities (Bollen, 2011; Cadogan, Lee, & Chamberlain, 2013), PLS will still likely be a suboptimal approach for building composites, given the intrinsic limitation of capitalization on chance. Fixed weights are an entirely reasonable alternative (Howell, 2013), which is further supported by recent simulation evidence (Rönkkö, Evermann, et al., 2015).

6. Conclusions

In their article, Willaby et al. (2015) set out to “seed interest in PLS and quell any concerns regarding its viability” (p. 1). As with the general literature on PLS, Willaby et al. (2015) present it as an undisputed technique that has “undergone extensive development, testing, and application” (p. 5). Although PLS is already extensively used in the information systems domain (Ringle, Sarstedt, & Straub, 2012), and is gradually and worryingly receiving greater uptake in many management disciplines (Antonakis, Bastardo, Liu, & Schriesheim, 2014; Hair

et al., 2012; Rönkkö, 2014), it has not undergone formal testing. Herman Wold, the PLS originator, was more interested in developing various algorithms than assessing their properties (Dijkstra, 1983), and the more recent literature on PLS is similarly lacking in terms of any serious analysis of its performance (Rönkkö & Evermann, 2013). Looking beyond the misleading characterization of PLS as a SEM technique and considering it instead as an indicator weighting system provides a sobering view on the method. Considering that the literature on PLS has not yet agreed on what the purpose of the indicator weighting is, let alone proven that the weights are optimal for that purpose, it is clear that there is very little evidence to support the use of PLS weights when forming composite variables.

To conclude, given the intractable problems demonstrated with regard to capitalization on chance, compromised statistical significance testing, and lack of model testing capabilities, PLS should not be adopted as a tool for psychological research. Instead, applied researchers should focus their efforts on familiarizing themselves with the growing body of innovations in the SEM area, so that they are best-equipped to deal with the various specific modeling scenarios they encounter. In the event that composite-based approximations to latent variable models are actually needed, there is very little reason to use anything else than unit-weighted scales (e.g., Bobko et al., 2007; Cohen et al., 2003, pp. 97–98; Cohen, 1990; McDonald, 1996; Raju et al., 1999).

7. References

- Aguirre-Urreta, M. I., & Marakas, G. M. (2014a). A Rejoinder to Rigdon et al. (2014). *Information Systems Research*, 25(4), 785–788. <http://doi.org/10.1287/isre.2014.0545>

- Aguirre-Urreta, M. I., & Marakas, G. M. (2014b). Partial Least Squares and Models with Formatively Specified Endogenous Constructs: A Cautionary Note. *Information Systems Research*, 25(4), 761–778. <http://doi.org/10.1287/isre.2013.0493>
- Aguirre-Urreta, M. I., Marakas, G. M., & Ellis, M. E. (2013). Measurement of Composite Reliability in Research Using Partial Least Squares: Some Issues and an Alternative Approach. *The DATA BASE for Advances in Information Systems*, 44(4), 11–43. <http://doi.org/10.1145/2544415.2544417>
- Antonakis, J., Bastardo, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly*, 25(1), 152–179. <http://doi.org/10.1016/j.leaqua.2013.10.014>
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(6), 1086–1120. <http://doi.org/10.1016/j.leaqua.2010.10.010>
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. Day (Ed.), *The Oxford Handbook of Leadership and Organizations*. Oxford University Press, Oxford (pp. 93–117). New York: Oxford University Press.
- Bainter, S. A., & Bollen, K. A. (2015). Moving Forward in the Debate on Causal Indicators: Rejoinder to Comments. *Measurement: Interdisciplinary Research and Perspectives*, 13(1), 63–74. <http://doi.org/10.1080/15366367.2015.1016349>
- Bentler, P. M. (2010). SEM with simplicity and accuracy. *Journal of Consumer Psychology*, 20(2), 215–220. <http://doi.org/10.1016/j.jcps.2010.03.002>

- Bobko, P., Roth, P. L., & Buster, M. A. (2007). The Usefulness of Unit Weights in Creating Composite Scores A Literature Review, Application to Content Validity, and Meta-Analysis. *Organizational Research Methods, 10*(4), 689–709.
<http://doi.org/10.1177/1094428106294734>
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Son Inc.
- Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly, 35*(2), 359–372.
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265–284.
- Cadogan, J. W., Lee, N., & Chamberlain, L. (2013). Formative variables are unreal variables: why the formative MIMIC model is invalid. *AMS Review, 3*(1), 38–49.
<http://doi.org/10.1007/s13162-013-0038-9>
- Cashen, L. H., & Geiger, S. W. (2004). Statistical Power and the Testing of Null Hypotheses: A Review of Contemporary Management Research and Recommendations for Future Studies. *Organizational Research Methods, 7*(2), 151–167.
<http://doi.org/10.1177/1094428104263676>
- Charles, E. P. (2005). The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets. *Psychological Methods, 10*(2), 206–226.
<http://doi.org/10.1037/1082-989X.10.2.206>
- Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

- Chin, W. W. (2001). *PLS-graph user's guide*. Houston, TX: Soft Modeling Inc.
- Chin, W. W., Marcolin, B. L., & Newsted, P. R. (2003). A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and an Electronic-Mail Emotion/Adoption Study. *Information Systems Research, 14*(2), 189–217.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. London: Lawrence Erlbaum Associates.
- Corsini, R. J. (2002). *The Dictionary of Psychology*. New York, N.Y.: Brunner/Routledge.
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.
- Dijkstra, T. K. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics, 22*(1-2), 67–90.
- Dijkstra, T. K. (2010). Latent variables and indices: Herman Wold's basic design and partial least squares. In E. V. Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of Partial Least Squares* (pp. 23–46). Berlin; New York: Springer.
- Dijkstra, T. K. (2014). PLS' Janus Face – Response to Professor Rigdon's "Rethinking Partial Least Squares Modeling: In Praise of Simple Methods." *Long Range Planning, 47*(3), 146–153. <http://doi.org/10.1016/j.lrp.2014.02.004>
- Dijkstra, T. K., & Henseler, J. (2015). Consistent Partial Least Squares Path Modeling. *MIS Quarterly, 39*(2), 297–316.

- Dijkstra, T. K., & Schermelleh-Engel, K. (2014). Consistent Partial Least Squares for Nonlinear Structural Equation Models. *Psychometrika*, *79*(4), 585–604.
<http://doi.org/10.1007/s11336-013-9370-0>
- Dunteman, G. H. (1989). *Principal components analysis*. Newbury Park, CA: Sage.
- Edwards, J. R. (2011). The Fallacy of Formative Measurement. *Organizational Research Methods*, *14*(2), 370–388. <http://doi.org/10.1177/1094428110378369>
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.
- Evermann, J., & Tate, M. (2013, October 8). Is My Model Right? Model Quality and Model Misspecification in PLS -- Recommendations for IS Research. Social Science Research Network. Retrieved from <http://papers.ssrn.com/abstract=2337699>
- Fornell, C., & Bookstein, F. L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research*, *19*(4), 440–452.
- Fornell, C., & Cha, J. (1994). Partial least squares. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp. 52–78). Cambridge, MA: Blackwell.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.
- Fuller, W. A. (1987). Measurement error models. Wiley series in probability and statistics. *Measurement Error Models. Wiley Series in Probability and Statistics*.
- Gefen, D., Rigdon, E. E., & Straub, D. W. (2011). An Update and Extension to SEM Guidelines for Administrative and Social Science Research. *MIS Quarterly*, *35*(2), iii–xiv.

- Goodhue, D. L., Lewis, W., & Thompson, R. (2007). Statistical power in analyzing interaction effects: questioning the advantage of PLS with product indicators. *Information Systems Research, 18*(2), 211–227.
- 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.
- Goodhue, D. L., Thompson, R., & Lewis, W. (2013). Why You Shouldn't Use PLS: Four Reasons to Be Uneasy about Using PLS in Analyzing Path Models. In *2013 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 4739–4748).
<http://doi.org/10.1109/HICSS.2013.612>
- Greene, W. H. (2012). *Econometric analysis*. Boston: Prentice Hall.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A primer on partial least squares structural equations modeling (PLS-SEM)*. Los Angeles: SAGE.
- Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012). 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. <http://doi.org/10.1016/j.lrp.2012.09.008>
- Hancock, G. R. (1999). A sequential Scheffé-type respecification procedure for controlling type I error in exploratory structural equation model modification. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(2), 158–168.
<http://doi.org/10.1080/10705519909540126>

- Hardin, A. M., Chang, J. C.-J., Fuller, M. A., & Torkzadeh, G. (2011). Formative Measurement and Academic Research: In Search of Measurement Theory. *Educational and Psychological Measurement, 71*(2), 281–305. <http://doi.org/10.1177/0013164410370208>
- Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three – Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841–850. <http://doi.org/10.1016/j.paid.2006.10.001>
- Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., ... Calantone, R. J. (2014). Common Beliefs and Reality About PLS Comments on Rönkkö and Evermann (2013). *Organizational Research Methods, 17*(2), 182–209. <http://doi.org/10.1177/1094428114526928>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135.
- Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in International Marketing, 20*, 277–319.
- Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics, 28*(2), 565–580. <http://doi.org/10.1007/s00180-012-0317-1>
- Herzog, W., & Boomsma, A. (2009). Small-Sample Robust Estimators of Noncentrality-Based and Incremental Model Fit. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 1–27. <http://doi.org/10.1080/10705510802561279>

- Howell, R. D. (2013). Conceptual clarity in measurement—Constructs, composites, and causes: a commentary on Lee, Cadogan and Chamberlain. *AMS Review*, 3(1), 18–23.
<http://doi.org/10.1007/s13162-013-0036-y>
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Is formative measurement really measurement? Reply to Bollen (2007) and Bagozzi (2007). *Psychological Methods*, 12(2), 238–245.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Hui, B. S., & Wold, H. (1982). Consistency and Consistency at Large of Partial Least Squares Estimates. In K. Jöreskog & S. Wold (Eds.), *Systems under indirect observation : causality, structure, prediction*. Amsterdam: North-Holland.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Jöreskog, K. G., Andersen, E. B., Laake, P., Cox, D. R., & Schweder, T. (1981). Analysis of Covariance Structures [with Discussion and Reply]. *Scandinavian Journal of Statistics*, 8(2), 65–92.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.
- Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology*, 41(1), 119–157.

- Lee, N., Cadogan, J. W., & Chamberlain, L. (2014). Material and efficient cause interpretations of the formative model: resolving misunderstandings and clarifying conceptual language. *AMS Review*, 4(1-2), 32–43. <http://doi.org/10.1007/s13162-013-0058-5>
- Lehmann, E., & Casella, G. (1998). *Theory of point estimation*. New York: Springer.
- Lohmöller, J.-B. (1988). The PLS Program System: Latent Variables Path Analysis with Partial Least Squares Estimation. *Multivariate Behavioral Research*, 23(1), 125–127.
- Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg: Physica-Verlag.
- Lu, I. R. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, 28(3), 258–268. <http://doi.org/10.1016/j.ijresmar.2011.03.006>
- Marcoulides, G. A., & Chin, W. W. (2013). You Write, but Others Read: Common Methodological Misunderstandings in PLS and Related Methods. In H. Abdi, W. W. Chin, V. E. Vinzi, G. Russolillo, & L. Trinchera (Eds.), *New Perspectives in Partial Least Squares and Related Methods* (pp. 31–64). New York, NY: Springer.
- Marcoulides, G. A., & Ing, M. (2012). Automated structural equation modeling strategies. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 690–704). New York, NY: Guilford Press.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers in test validity theory: measurement, causation and meaning*. New York, N.Y.: Psychology Press.
- McDonald, R. P. (1996). Path Analysis with Composite Variables. *Multivariate Behavioral Research*, 31(2), 239–270.

- McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences*, 42(5), 859–867. <http://doi.org/10.1016/j.paid.2006.09.020>
- McIntosh, C. N., Edwards, J. R., & Antonakis, J. (2014). Reflections on Partial Least Squares Path Modeling. *Organizational Research Methods*, 17(2), 210–251. <http://doi.org/10.1177/1094428114529165>
- Moosbrugger, H., Schermelleh-Engel, K., & Klein, A. (1997). Methodological problems of estimating latent interaction effects. *Methods of Psychological Research Online*, 2(2), 95–111.
- Muchinsky, P. M. (1996). The Correction for Attenuation. *Educational and Psychological Measurement*, 56(1), 63–75. <http://doi.org/10.1177/0013164496056001004>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
- Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, 5(2), 241–301. <http://doi.org/10.1037//1082-989X.5.2.241>
- Penev, S., & Raykov, T. (2006). On the Relationship Between Maximal Reliability and Maximal Validity of Linear Composites. *Multivariate Behavioral Research*, 41(2), 105–126. http://doi.org/10.1207/s15327906mbr4102_1
- Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, 23(2), 99–115.

- Raykov, T. (2011). Evaluation of convergent and discriminant validity with multitrait–multimethod correlations. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 38–52.
- Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling*, *8*(1), 142–149.
- Raykov, T., & Marcoulides, G. A. (2007). Equivalent Structural Equation Models: A Challenge and Responsibility. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 695–700. <http://doi.org/10.1080/10705510701303798>
- Rhemtulla, M., Bork, R. van, & Borsboom, D. (2015). Calling Models With Causal Indicators “Measurement Models” Implies More Than They Can Deliver. *Measurement: Interdisciplinary Research and Perspectives*, *13*(1), 59–62.
<http://doi.org/10.1080/15366367.2015.1016343>
- Rigdon, E. E. (2012). Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods. *Long Range Planning*, *45*(5–6), 341–358.
<http://doi.org/10.1016/j.lrp.2012.09.010>
- Rigdon, E. E., Becker, J.-M., Rai, A., Ringle, C. M., Diamantopoulos, A., Karahanna, E., ... Dijkstra, T. K. (2014). Conflating Antecedents and Formative Indicators: A Comment on Aguirre-Urreta and Marakas. *Information Systems Research*, *25*(4), 780–784.
<http://doi.org/10.1287/isre.2014.0543>
- Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). A Critical Look at the Use of PLS-SEM in MIS Quarterly. *MIS Quarterly*, *36*(1), iiv–8.

- Rönkkö, M. (2014). The Effects of Chance Correlations on Partial Least Squares Path Modeling. *Organizational Research Methods, 17*(2), 164–181.
<http://doi.org/10.1177/1094428114525667>.
- Rönkkö, M. (2015). matrixpls: Matrix-based Partial Least Squares Estimation (Version 0.6.0). Retrieved from <http://cran.r-project.org/web/packages/matrixpls>
- Rönkkö, M., & Evermann, J. (2013). A Critical Examination of Common Beliefs About Partial Least Squares Path Modeling. *Organizational Research Methods, 16*(3), 425–448.
<http://doi.org/10.1177/1094428112474693>
- Rönkkö, M., Evermann, J., & Aguirre-Urreta, M. I. (2015). Estimating Formative Measurement Models in IS Research – Analysis of the Past and Recommendations for the Future. *MIS Quarterly, submitted*.
- Rönkkö, M., McIntosh, C., & Antonakis, J. (2015). On the Small-Sample Properties of “Consistent Partial Least Squares for Nonlinear Structural Equation Models.” *Psychometrika, submitted*.
- Rönkkö, M., & Ylitalo, J. (2010). Construct Validity in Partial Least Squares Path Modeling. In *ICIS 2010 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2010_submissions/155
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.
- Sarstedt, M., Ringle, C. M., & Hair, J. F. (2014). PLS-SEM: Looking Back and Moving Forward. *Long Range Planning, 47*(3), 132–137. <http://doi.org/10.1016/j.lrp.2014.02.008>
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables*

analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA, US: Sage Publications, Inc.

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.

<http://doi.org/10.1177/1059601108329198>

Stevens, J. P. (2012). *Applied Multivariate Statistics for the Social Sciences, Fifth Edition* (5 edition). New York: Routledge.

Ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and Its Applications*, 289(1–3), 311–318. [http://doi.org/10.1016/S0024-3795\(97\)10007-6](http://doi.org/10.1016/S0024-3795(97)10007-6)

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). PLS Path Modeling. *Computational Statistics & Data Analysis*, 48(1), 159–205.

<http://doi.org/10.1016/j.csda.2004.03.005>

Tong, X., Zhang, Z., & Yuan, K.-H. (2014). Evaluation of Test Statistics for Robust Structural Equation Modeling With Nonnormal Missing Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 553–565.

<http://doi.org/10.1080/10705511.2014.919820>

Van der Sluis, S., Dolan, C. V., & Stoel, R. D. (2005). A Note on Testing Perfect Correlations in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(4), 551–577.

http://doi.org/10.1207/s15328007sem1204_3

- Westland, J. C. (2015). Partial Least Squares Path Analysis. In *Structural Equation Models* (pp. 23–46). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-16507-3_3
- Willaby, H. W., Costa, D. S. J., Burns, B. D., MacCann, C., & Roberts, R. D. (2015). Testing complex models with small sample sizes: A historical overview and empirical demonstration of what Partial Least Squares (PLS) can offer differential psychology. *Personality and Individual Differences, 84*, 73–78.
<http://doi.org/10.1016/j.paid.2014.09.008>
- Wold, H. (1982). Soft modeling - The Basic Design And Some Extensions. In K. G. Jöreskog & S. Wold (Eds.), *Systems under indirect observation : causality, structure, prediction* (pp. 1–54). Amsterdam: North-Holland.
- Wold, H. (1985). Partial Least Squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 581–591). New York: Wiley.
- Wooldridge, J. M. (2009). *Introductory econometrics: a modern approach* (4th ed). Mason, OH: South Western, Cengage Learning.
- Yuan, K.-H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 53*(1), 31–50.
- Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology, 33*(1), 241–265.

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2013). Empirical Correction to the Likelihood Ratio Statistic for Structural Equation Modeling with Many Variables. *Psychometrika*, 1–27.

<http://doi.org/10.1007/s11336-013-9386-5>

Zimmerman, D. W. (2007). Correction for Attenuation With Biased Reliability Estimates and Correlated Errors in Populations and Samples. *Educational and Psychological*

Measurement, 67(6), 920–939. <http://doi.org/10.1177/0013164406299132>

8. Tables and Figures

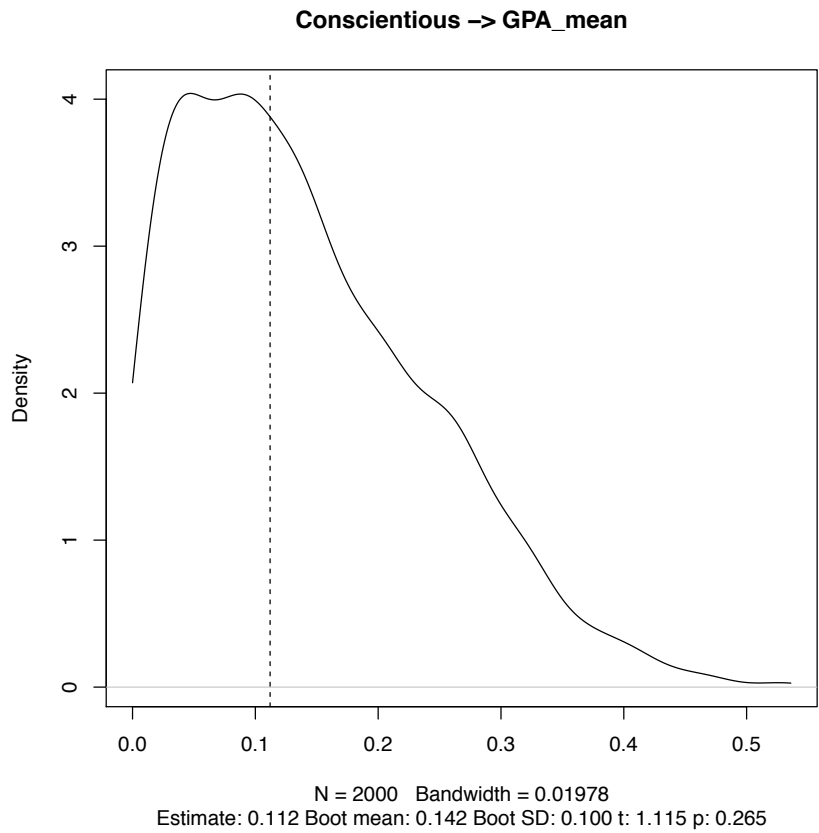


Figure 1 Kernel density plot of a bootstrap distribution from Willaby et al.

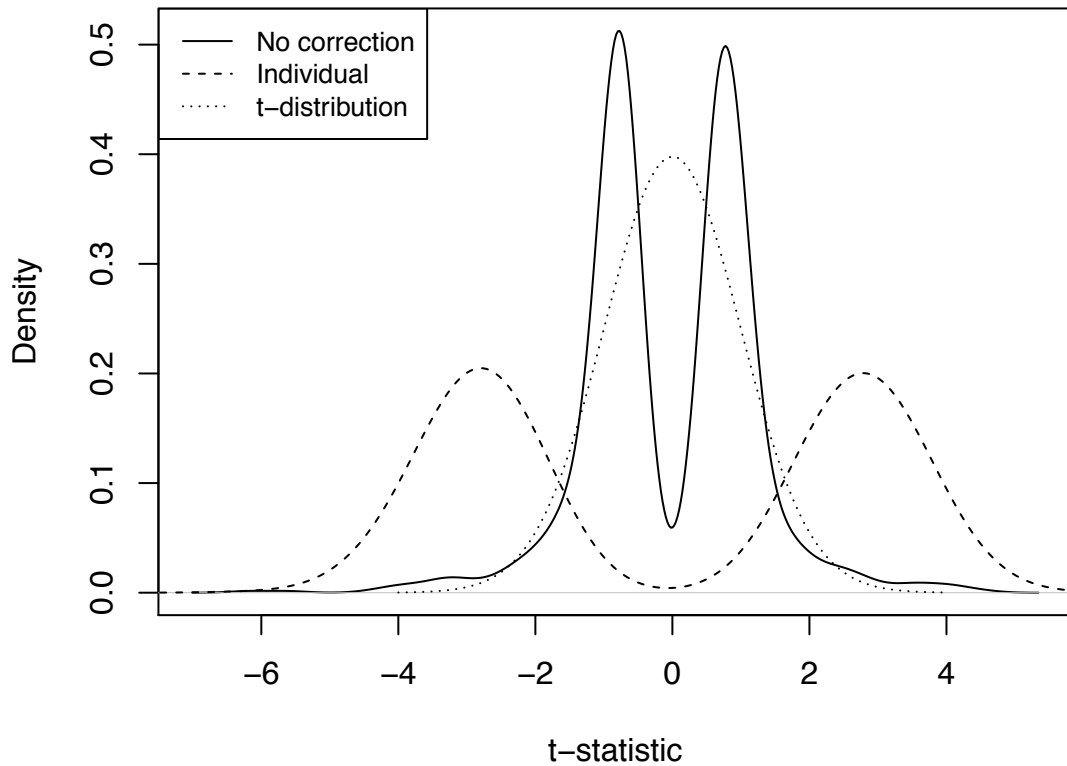


Figure 2 Kernel density plot of PLS t-statistics with and without sign change corrections compared with the t-distribution over 1000 replications.

Table 1 Correlations between PLS composites and equal weights composites

		1	2	3	4	5	6	7	8	9
PLS	1	Conscientious								
	2	GPA	.182							
	3	Organisation	.269	.261						
	4	Temporal flow	.699	-.014	.177					
	5	Mechanics	.252	.351	.143	.063				
Equal weights	6	Conscientious	.998	.182	.28	.687	.242			
	7	GPA	.182	1.000	.261	-.014	.351	.182		
	8	Organisation	.205	.223	.976	.145	.096	.217	.223	
	9	Temporal flow	.694	-.01	.169	.997	.059	.683	-.01	.135
	10	Mechanics	.223	.352	.136	.048	.988	.212	.352	.094

Table 2 Nominal false positive rates and realized false positive rates for three different sign-change corrections over four p-values and confidence intervals (CIs) in 1000 replications

Criterion	Nominal number of false positives	No sign-change correction	Construct level sign-change correction (Hair)	Construct level sign-change correction (Tenenhaus)	Individual sign-change correction
$p < 0.10$	100	101	112	101	957
$p < 0.05$	50	73	72	73	880
$p < 0.01$	10	37	41	37	597
$p < 0.001$	1	20	18	20	215
95% Normal CI	50	61	69	61	435
95% Basic CI	50	444	452	444	385
95% Percentile CI	50	60	61	60	1000
95% BCa CI	50	12	11	12	993*

Note: * = The CI could not be calculated for 7 replications.