FRACTURES IN THE EDIFICE OF PLS: REJOINDER TO COMMENTARIES ON MARKETING OR METHODOLOGY

Journal:	*European Journal of Marketing*
Manuscript ID	EJM-02-2022-0138
Manuscript Type:	Commentary
Keywords:	PLS, Methodology, Statistical analysis, Measurement

FRACTURES IN THE EDIFICE OF PLS: REJOINDER TO COMMENTARIES ON *MARKETING OR METHODOLOGY*

When we were informed that we were to receive two commentaries on our paper *Marketing or Methodology: Exposing the Fallacies of PLS with Simple Demonstrations* (hereafter MoM), we approached them with some trepidation, given our prior experiences with debates in this area. Indeed, we had a fair idea of what we might expect to see in those commentaries, and who the authors might be. We were surprised then to see a commentary from Ke-Hai Yuan, and intrigued to read its content. We were less surprised to see a commentary from Joe Hair and colleagues, although as we will explain later, we were certainly surprised to see that most of the substantive content in Hair et al's commentary actually *supported* our position that PLS is a poor choice of method for typical marketing research tasks.

We discuss each commentary in turn, but before that we start with a few general points. The single most important point in MoM is to some extent not specifically about PLS at all: **The first decision a researcher must make when selecting an analysis method for multi-item scale data is whether they are going to work with composites[1] or latent variables in their model.** Once this decision is made, one can select between the various estimation methods available for the chosen task. If this process is followed, the false equivalence that is drawn in much pro-PLS literature between PLS and ML-SEM vanishes. If researchers made explicit their decision to use composites or latent variables, and justified that decision clearly, many of the problems that are so evident in existing research practice that we point out in MoM would be far less prevalent.

To be very clear, PLS is *not* a method to directly estimate latent variable models in the way that ML-SEM or factor analysis is. It is a method to *construct composites*, and

[1] To be more precise, we presented this as a choice between latent variables and scale scores that can be either linear composites or non-linear functions of the observed data. But because scale scores are nearly always calculated as linear composites, in practice the choice is between latent variables and composites.

decisions to use PLS should be made amongst the different composite methods, *after* deciding to use composites rather than latent variables in the model. As we make clear in MoM, composites have their uses. However, to have a viable place in a researcher's toolkit, PLS must have *useful advantages over other composite methods*, rather than be compared against latent variable modelling methods like factor analysis.

Unfortunately, it appears that many researchers do not understand that composites and latent variables are not interchangeable in models, and that there are important implications of using one or the other. We do not wish to enter in any depth the debate about conceptualization of constructs here (although see Lee and Cadogan's two papers in this special issue, and associated commentaries, for a more focused treatment). However, it stands to reason that if a theory includes concepts that are characterized as latent (i.e. not directly observable; as are many in marketing and related fields), then latent variable methods such as ML-SEM, or common factor analysis, should be the first choice of operationalization. Such methods are not directly conceptually interchangeable with composite methods, and therefore if one wishes to use a composite method in place of a latent variable method, the choice should be justified (perhaps through the need for computational simplicity in parameter estimation). Once the reason for *this* decision to use composites is established, one must justify the decision to use any given composite method, of which there are many available. Almost no existing literature using PLS provides any kind of justification for using composites, and simply uses PLS in the place of ML-SEM, presuming modeled concepts and estimators are interchangeable. This is simply not the case.

Even if the choice to use composites is well-justified, one must still justify which method is to be used to create those composites. In MoM, we showed clearly that claims of advantages of PLS over other composite methods either a) were not based on any evidence, b) were based on invalid evidence or incorrect interpretation of evidence, or c) were evident only to a trivial degree and/or in highly unrealistic settings. Further, we showed that even if PLS *did* have the claimed advantages, they were heavily outweighed by the clear and well-established drawbacks of PLS. Table 1 summarizes the claims that we made about PLS in MoM, regarding both its advantages and disadvantages, and whether they are supported by evidence.

TABLE 1

To make things crystal clear for readers, there is nothing in the comments of Yuan or Hair et al which convincingly rebuts any of our points (although there are certainly points worthy of discussion). However, particularly in Hair et al's comment, there is a large amount of rhetoric designed to make readers *think* there is important contradictory evidence to MoM. Conversely, Yuan's comment contains copious statistical detail which certainly *looks* impressive. However, none of the points made by Yuan invalidate points made in MoM (or even appear that they are intended to do so), or provide any strong evidence to support the continued use of PLS. Indeed, as we will show, a good part of both Hair et al's comments, and Yuan's results, can be most correctly understood as speaking *against* the use of PLS in typical marketing (and related) research studies.

Although we will try to respond to the commentaries as comprehensively as possible, we do not have space to cover every single point made in both commentaries, and indeed many of the points made in Hair et al's comment do not merit even the most cursory reply; the comments we do not respond to in depth are either irrelevant to our points, or simply obfuscate the core issues rather than being, to use a refrain from PLS aficionados, 'silver bullets' that invalidate MoM.

We will take the comments of Hair et al and Yuan in turn, beginning with Hair et al's attempts to rebut our key points. We will continue with a discussion of Yuan's commentary, which diverges significantly from our key points but still contains a lot of material that needs to be addressed. We finish with a general summary and set of conclusions for how best to move forwards. We show again that there is no reason at all to use PLS, given the numerous superior alternatives already available. In light of existing evidence, we conclude that PLS as it is currently used in marketing and management research should be immediately abandoned.

RESPONSE TO HAIR'S CRITIQUE OF OUR MAIN POINTS

Before addressing Hair et al's (hereafter H21) comments on our specific points, we would like to particularly address H21's argument that researchers should not report our Composite Equivalence Index (CEI). CEI is a simple, method-agnostic tool that researchers can use to provide direct comparisons between different composite methods (see Figure 6 of MoM); in brief, it assesses the degree of correlation between composites constructed by various techniques. It is not intended to privilege one method over another, but simply to show where competing methods of constructing composites make an empirical difference, and where they do not. In cases where the CEI shows that there is no substantive difference between different ways of constructing a composite (i.e., the inter-composite correlations are very close to unity), it is obvious that the simplest method should be preferred, such as unit-weighted composites. However, when the CEI shows there *are* differences, the onus is on the researcher to explain why the method they wish to use is beneficial. For example, researchers could support their choices through careful consideration of the simulation-based evidence of different composite methods or alternatively by explaining why the specific weights for the indicators make sense considering the indicator wording and the underlying theory.

H21 object to the use of CEI, but their reasoning for doing so is unclear. H21 simply state "Unfortunately, the CEI lacks discriminatory power, conceals reliability concerns in reflective measurement models, overlooks the likelihood of differences in relative indicator contributions in formative measurement models, and violates the principles of index construction procedures—see Hair et al. (2021)" (p. 22). Unfortunately, the cited article does not appear to exist. Instead, we found a similarly titled article from the same journal and the same author team (Sharma et al., 2022) but that article does not address CEI. As such, it seems that H21 think it is important to object to the use of CEI, but not important to explain their logic for doing so. Indeed, it appears that the PLS emperor has no clothes.

If H21 were confident that PLS offered strong advantages over other composite methods, surely they should *advocate* for its reporting as a new standard. Using CEI forces researchers to provide both evidence and rationale for their choice of composite method. Thus, advocating against the CEI is essentially tantamount to advocating the suppression of evidence, which seems to be a more ideological than scientific stance. As such, we strongly advocate that editors and reviewers who are interested in improving methodological transparency of research should demand consistent reporting of the CEI.

Point 1: PLS does not maximize explained variance or R^2

The first point we made in MoM was that claims made in the PLS literature that PLS maximizes explained variance or R^2 were based on no evidence whatsoever. Further, we explained that claims regarding PLS's maximization of 'explained variance' could not be generally true, because it was impossible to determine which criteria are meant by this term. For example, for any model there are different R^2s (i.e., for each indicator block vs. global explained variance), and it is never specified which of these PLS is supposed to maximize. Finally, we showed that it was trivial to find a method providing a higher global R^2 than PLS, which at a stroke rendered any general claim that PLS maximized explained variance meaningless.

In their commentary, H21 appear to try to rebut our critique. We find four elements in their argumentation. First, the general theme in H21 is that because MoM cites the 2014 PLS Primer book, it is based on outdated literature. This is hardly the case with the R^2 claim, because the same claim is present in also Hair's more recent work:"The goal [of PLS] is to maximize the explained variance of the endogenous latent variables. " (Hair et al., 2021b, p. 190).

Second, H21 attempt to move the goal posts by writing that PLS "obtain[s] composite scores that minimize the residuals in the relationships between composites and indicators ... *as well as* those between composites" (emphasis added, pg. 5). It is important to point out first that the claim about R^2 maximization that Hair and his coauthors made in the 2014 book and continue to advance to date, and which is repeated in countless empirical articles, makes no reference to indicators. As documented in the earlier literature (Rönkkö, McIntosh, Antonakis, et al., 2016), there are various different unproven and ambiguous claims about the purpose of the weights and we simply chose to address the one that we viewed as the most popular.

We will now address this new claim. The careful reader might note that "as well as" here is not to be understood as meaning "jointly" or "at the same time", but rather to mean *alternatingly* through the iterative algorithm. From that, H21 conclude that "the PLS estimates are *coherent* in the sense that all the residual variances are minimized *jointly* (emphasis added, pg. 5). First, we note the use of "coherent" instead of "optimal". It is not clear what it means for estimates to be coherent, and this is not a term in obvious usage within the relevant statistical literature. Second, there has been no proof in the literature that the iterative, alternating optimization used in PLS will in fact lead to a joint optimum.

Further, the reader might rightfully ask what 'joint' means in this case anyway? For example, is the 'joint optimum' a weighted sum, a mean, or something else?

More empty arguments follow in the next paragraph of H21. Again, "minimizing residuals jointly" (pg. 5) is not true in the sense that most researchers would use the term "jointly" and H21 use the phrase "establish a balance between these two key objectives" (pg. 5) without specifying in detail what this balance might be. Without step-by-step proofs and/or direct empirical evidence from simulations, these claims are meaningless, and we can only assume their purpose is to confuse readers into thinking H21 can provide evidence against our main claim. The paragraph concludes by pointing to the optimality of OLS regression when producing the model parameters during the second stage of PLS. This point distracts from the core problem – the purpose of PLS weights – because by that time, any optimization is too late because the component weights have already been determined. Another distraction is the reference to the results by Hanafi (2007) for PLS mode B, when H21 are well aware from their own research that the overwhelming majority of PLS applications use Mode A, an inconvenient fact they silently omit.

Third, H21 makes a strawman argument by claiming that we "call for a global optimization criterion" (p. 4). Of course, we are well aware that simply having an optimization criterion (which PLS lacks) does not guarantee the optimality properties of an estimator. Indeed, in MoM we clearly point out that "It is also unclear why R^2 maximization would be useful for estimating parameters in a complex model consisting of multiple equations" (p. 6). What we say is simply that if a researcher is interested in a maximization problem, the problem should be clearly defined, after which a suitable maximization technique should be chosen.

Fourth, perhaps the most baffling part of H21's critique of our point is the criticism of the results of our canonical correlation analysis. Specifically, H21 suggest that "weight estimates show bewildering behavior" (pg. 6), point out indicators which "incorrectly" (pg. 6) imply certain substantive relationships and other "clearly flawed" (pg. 7) implications. The purpose of the example was not to draw any substantive conclusions, but simply to show that a specific and often repeated claim about R^2 maximization is incorrect. The negative weights are a result of *sign indeterminacy* of composite methods and can be trivially solved by reversing all signs. The fact that signs vary within indicator block just serves to show again what we have repeatedly stated: it is unclear why maximizing R^2 would be desirable in complex modeling context.

By p. 8, it seems clear that H21 are more interested in promoting a personal agenda than advancing rigorous knowledge development. H21 note that GSCA is a composite technique that *does* provide an optimization criterion. Given that GSCA can *also* adjust for cross-loadings and correlated errors, one must wonder why H21 did not simply recommend the use of GSCA instead of PLS. Indeed, in a recent paper another of the key popularizers of PLS *does* recommend that GSCA be used in component-based models because it performs "equally as well or better than PLSPM [PLS path modelling]" (Cho, Sarstedt, and Hwang, in press). Again though, these authors stop short of drawing the now-obvious conclusion that PLS should simply be abandoned, which seems the only sensible position given the ever-increasing evidence base against its use.

Point 2: PLS Weights Do Not Meaningfully Improve Reliability

Our second point was presumably more subtle than the first, given that either H21 found it impossible to understand, or have intentionally misrepresented it. To be clear, at no point in MoM did we claim that "composite scores produced by PLS-SEM are less reliable than those of unweighted composites" (H21 p8). The original claim by Rönkkö and Evermann (2013) was that PLS does not provide an advantage in terms of reliability. In MoM, we merely pointed out that under realistic conditions there is generally no appreciable or detectable difference between a PLS composite and an unweighted sum-score and that this is simple to demonstrated with the CEI index even without the use of any simulations. We even showed situations where there *were* differences in favour of PLS. However, these differences were a) trivial in real terms, b) disappeared when correct scale development procedures (as recommended by Hair himself) were applied, and c) came at the expense of serious drawbacks of PLS which are elaborated elsewhere in MoM (e.g., capitalization on chance). So to quote ourselves, we *did not state* that PLS is less reliable than unweighted sum-scores, but instead that "improving reliability is certainly not a reason to use PLS" (*this issue*).

To support their claims, H21 cite the proofs by Rigdon (2012), and Cook and Forzani (2020) and simulation studies by Hair et al. (2017) and Yuan et al. (2020). It is important to make clear that the proofs simply show that the difference in reliability between PLS Mode B composites and ML-based factor scores is *nonzero in a population*. They do *not* show that the difference is either a) meaningfully large, or b) evident in finite samples. As such, our own points are entirely consistent with that, as explained above.

Yuan et al. (2020) does not provide any direct evidence on reliability differences between the indicators. What they show is that both PLS and summed scale estimates ($\hat{\beta}_{pls}$ and $\hat{\beta}_s$ in their Table 3 and Table 5) are all negatively biased but PLS estimates less so. But this is not because of any reliability advantage. Instead, this results from how PLS capitalizes on chance, creating a positive bias that to some extent cancels out the effect of measurement error, as H21 are fully aware of because the effect has been documented in a number of articles. The positive bias is evident in two features of the results. First, the positive bias due to capitalization on chance decreases with sample size, but measurement error bias does not and, as a consequence, we see an overall increase in negative bias as sample size increases and converges to the same level that summed scales have (Rönkkö, 2014). If PLS indeed provided a reliability advantage, we should observe that PLS estimates were consistently *less* biased and the level of bias would *not* be affected by sample size. Second, the PLSc estimates ($\hat{\beta}^{c}_{plsc}$ and $\hat{\beta}^{dhc}_{plsc}$) are positively biased in small samples. This happens because the effect of random measurement error has been eliminated by correction for attenuation, but the capitalization on chance effect is not addressed, causing a positive bias, which too has been documented in the prior literature (Rönkkö, McIntosh, & Aguirre-Urreta, 2016). The other simulation study by Hair et al. (2017) is simply irrelevant to the reliability debate because they generated data from models where indicators are assumed to be perfectly reliable (i.e. indicator-level random error was not modelled) and no reliability data are thus reported. Indeed, even the term reliability appears just once on the article.

Much of the rest of this section of H21 is completely irrelevant. For example, H21 argue that summed scale estimates are more sensitive to low quality indicators like cusl2 (pg. 10). We acknowledge this point, which is why we use Hair's own textbook approach to identify the low-quality indicator and remove it prior to the analysis. This is a procedure that *Hair himself recommends as good research practice* (e.g., "Indicators with very low loadings (below 0.40) should, however, always be eliminated from the measurement model (Hair, Hult, Ringle, & Sarstedt, 2022)." Hair et al., 2021b, p. 77), a fact that seems to have been conveniently forgotten in H21's critique. Bizarrely, H21 conclude that "it is difficult to conceive why one would prefer equal weights over differential weights" (pg. 11), when well over 30 years of psychometric research has shown exactly the opposite, and to repeat our quote from Cohen (1990, p. 1306) "as a practical matter, most of the time, we are better off using unit weights: +1 for positively related predictors, -1 for negatively related predictors, and 0, that is, throw away poorly related predictors". Indeed, the *only* time one might prefer a differential weighting algorithm to create composites is either when a) one has a specific reason to weight indicators differently (e.g., a theoretical reason, or knowledge that indicators which must be included in the composite have very low correlations by design), or b) one is engaging in the poor-practice of throwing the indicators into the algorithm without *any* prior assessment, as specifically recommended *against* by Hair himself. Furthermore, *even if one was to prefer a weighting algorithm*, one would not choose PLS, because of the numerous other drawbacks we point out in MoM.

Point 3: PLS Weights Bias Composite Correlations

Our third point in MoM was to show that – untold in the PLS advocacy literature, but well-established in the critical literature – using PLS to create weighted composites predictably leads to bias in the correlations between those composites in a number of very common real-world scenarios. Specifically, bias occurs a) when the scales are weakly correlated, b) where there are cross-loadings or correlated errors between items in different scales, and c) particularly when sample sizes are small.

This section of our article addresses two but related problems, which H21 seem to confound: model misspecification and chance correlations, which occur because of sampling error. For example, in our example data (available in Online Supplement 2 of MoM) the die variables are uncorrelated with all other variables in the population but sample correlations with other variables vary between -0.20 and 0.09. These non-zero sample correlations are what we refer to as chance correlations. Evidence for the problem of chance correlations has been available now for almost a decade, published in the world's leading organizational methodology journal (Rönkkö, 2014). Not only can this issue be seen in nearly every simulation study of PLS that Rönkkö (2014) analyzed, but the effect is also clearly visible in the simulation results by Yuan et al. (2020); there is a clear pattern that on average the PLS estimates get large as sample size increases but no such effect is seen for the equal weight composites .

Yet PLS users continue to ignore this issue, and PLS advocates continue to obfuscate it to the extent that even now, some authors claim PLS should be used specifically to cope with small sample sizes. Alarmingly, even the J-B Steenkamp Award judging committee for the 2021 *International Journal of Research in Marketing* state that the simulations in Reinartz, Haenlein, and Henseler's (2009) winning paper comparing SEM and PLS "show that PLS can be a good methodological choice if sample size is low" (*IJRM*, 2021, pp. A3).

H21 continue this ignoble tradition, with the suggestion that our "claim boils down to the observation that a violation of specified prerequisites ... leads to negative consequences" (pg. 12). They argue that "users of any multivariate analysis method should be aware of its prerequisites and, if possible, comply with them ..." (pg. 12). Naturally we agree, particularly since any attentive reader should realize that in doing so, H21 fatally injure any claim that PLS is a valid method, because the advice boils down to recommending *against* the use of PLS when analyzing most real data sets; we elaborate below.

First, it is well established that correlated errors and cross-loadings are virtually inevitable in real-world applications of multi-item scale data (see Asparouhov, Muthén, & Morin, 2015; Marsh, Guo, Dicke, Parker, & Craven, 2020; Muthén & Asparouhov, 2012). It therefore seems nonsensical to recommend the use of a method where the presence of cross-loadings or correlated errors is a violation of 'specified pre-requisites'. Further, it is not at all clear where these 'pre-requisites' are specified, in a way that the typical user of PLS would realize they are (apparently) hardline assumptions that PLS is not robust to violations of. It is worth emphasizing this point, as Hair here is claiming that *PLS is not robust to departures from the assumption that items within and across scales do not have correlated errors or cross-loadings*. Therefore, unlike for example GSCA(m), the PLS framework is simply completely incapable of accommodating these ubiquitous features of real applications. If this is such a well-understood pre-requisite, we would certainly hope to see this assumption very clearly specified as one which PLS is not robust to departures of in the latest editions of core PLS primer texts. However, it does not seem to be obviously stated, at least to the best of our knowledge. For example, in the latest edition of *A Primer on PLS-SEM* (Hair et al., 2021a), the only mention of either of these we can find is a single unelaborated bullet in Exhibit 1.10 stating that one might consider using CB-SEM over PLS-SEM when “error terms require additional specification, such as the (*sic*) covariation”. Indeed, the term ‘assumptions’ does not even appear in the index.

In the same section, H21 also similarly critique our use of a two-composite model with a small or zero relationship between the composites as "clearly a boundary condition with no practical relevance" (p. 13). Again, this is a nonsense statement when considering the real conditions that typical PLS users face. It is obvious that researchers cannot *a priori* be certain of strong relationships between the composites in their models. Otherwise, why waste time testing them? Moreover, looking at findings of past PLS studies to argue that many or most identified relationships are statistically significant fails as evidence because of the bias that we note for small or zero relationships; any small or zero real correlations would be biased upwards by PLS, so could not serve as evidence that typical models examined using PLS do not contain small or zero correlations.

Moreover, the high prevalence of statistically significant results in the PLS literature is likely also due to the bootstrap sign-corrections implemented in some PLS software. Briefly, these procedures selectively flip the signs of the outputs (e.g., indicator weights and regression coefficients) within the bootstrap resamples to maintain consistency with the signs obtained from analyzing the entire, original data set. However, as demonstrated by Rönkkö et al. (2015), this ‘trick’ (as it is best described) will lead to drastically-inflated false positive rates. In fact, with the individual sign-change correction that reverses the signs of each and every bootstrap quantity, one achieves a 100% false positive rate! We are aware of only one publication stating that this approach "should be considered as deprecated" (Henseler, Hubona, & Ray, 2016, pg. 15, Note 3), but have no indication of how many applied PLS articles have actually heeded the warning. Therefore, there could be many contaminated results, hence many incorrect conclusions and recommendations, in the literature.

Therefore, H21's rebuttal here is an obvious example of fallacious reasoning. Specifically, it is *a priori* established as a premise that if a population correlation between composites is small, the PLS correlation is biased upwards. An observation that most correlations in published PLS studies are *not* small cannot then logically lead to the conclusion that published PLS studies are not biased. This seems to us at least to be an *inverse error*, or a denying of the antecedent. Hence, it is difficult to see how H21 can logically arrive at the conclusion that our example model is irrelevant.

Point 4: AVE And CR Are Not Useful To Assess PLS Models

The final major point we made in MoM is that the use of AVE and CR to assess the quality of PLS models is flatly wrong, and gives misleading results. Despite this, AVE and CR routinely appear in many PLS models in the marketing and other literatures, and remain part of the recommended assessment tools in even the latest instructional literature on PLS (e.g. Hair et al., 2021a). Given the very clear evidence of problems with using AVE and CR, as made clear in the methodological literature for many years now (e.g., Rönkkö & Cho, 2022), it is unsurprising that H21 struggle to find any way to critique our statement. In fact, they even seem to *agree* with us, while still appearing to attack us. For example, it certainly does not invalidate our point to state that it is "not new" (p.17), and indeed we wish that many more researchers would actually realise that AVE, CR, and the Fornell-Larcker criterion are useless for the purpose of model fit evaluation, and are unusable with PLS. That said, while these indices remain included in Hair et al.'s extremely popular primer text on PLS, and are still being advocated by them (Schuberth, 2021), it seems disingenuous for H21 to blithely discount our critical comments.

Further, we *agree* that HTMT is a better method for the most part than AVE for assessing discriminant validity. We also agree that coefficient (Cronbach's) alpha is a useful reliability measure. Importantly, both of these are general statistics that do not rely on PLS results, and make assumptions that are testable with factor analysis (Cho, 2016; Rönkkö & Cho, 2022). Thus, the logical conclusion that H21 do not want to state is that PLS results should not be used for model assessment and factor analysis should be used instead. However, we noted (fairly) that it HTMT is not without its own problems (see also Rönkkö & Cho, 2022), and has also been subject to some misleading comparisons in the marketing literature, based on incorrect applications of CFA. Either way though, we are *pleased* to see that HTMT performs better than AVE / CR, and we say so in MoM. That said, we note that HTMT is independent of any estimation procedure such as PLS. Perhaps H21 misunderstood our discussion of HTMT as an attack on PLS, which would seem to be the only reason for their extensive attack on a point which we actually seem to be in *agreement* over.

Summary of H21's Critique

There would have been a very simple way for H21 to effectively respond to our main concerns; simply use a combination of simulation studies and empirical data to show that PLS actually makes a substantive and beneficial difference to analysis results (i.e., improvements in unbiasedness, consistency, and efficiency), and provide a strong explanation of why PLS makes the difference. That H21 fail to do so indicates that situations where PLS makes an explainable beneficial difference in real-world analysis situations are either extremely rare or non-existent. Indeed, H21 could have used the very tool we provide – the

CEI – to show situations where there were substantive differences between composites created using PLS and other methods, if they could have done so.

As such, we conclude that the material in H21 does not offer a convincing rebuttal of the points we make in MoM. In fact, it actually *supports* the general claims of MoM, by providing yet more evidence that PLS is not a usable method for real-world sample data. In Table 2, we provide an overview of the points made in H21, aimed at countering the points we made in MoM. We show that in each case, H21 actually *strengthens* the claims against PLS that we make in MoM, and in places adds clear additional evidence speaking against PLS.

TABLE 2 HERE

RESPONSE TO YUAN'S COMMENTARY

We were surprised and intrigued to see that Yuan had written a commentary on MoM, and naturally wondered why he chose now to enter a discussion about PLS. Yuan's comment (Y21) contains a number of interesting points, most of which are drawn from his recent work in the area, some of which was unpublished at the time we wrote MoM. However, it is not correct for Y21 to claim that we were unaware of his results, given that we did not include them in MoM. We did not include Yuan's work in our original paper because it was not relevant for the points that we wanted to make. Yuan's comment does not offer any evidence to discount the points made in MoM. Moreover, some of his arguments are simply incorrect. That being said, Yuan's paper does bring up a number of broader issues, which can be discussed; but they are tangential to our more specific points. Still, below we attempt to group Yuan's main points into a set of overarching themes of relevance to our main points, and respond to Yuan's comments in relation to them.

Theme 1: Optimality

Y21 focuses much of his comments on various ways to expound on the 'optimality' properties of PLS. This issue is only peripherally related to our core points (which are already established above and in Table 1). According to Y21's derivations, when the full population is analyzed and a factor model holds for the data, the indicator weights under PLS Mode B are equivalent to those in the formulation of the Bartlett factor score. These findings are not entirely new to the PLS literature, but are a welcome formalization of previous results (Schuberth et al., 2022).

Y21 also makes two entirely new claims: That 1) PLS Mode B is equivalent to "the normal-distribution-based maximum likelihood (ML) estimator/predictor of the latent trait (Yuan & Deng, 2021)" (this issue) and that 2) "the composite following PLS-SEM mode B enjoys the optimal statistical properties of an ML estimator (see e.g., Casella & Berger, 2001)." (this issue) However, to the best of our knowledge at least, the first claim is not in fact made in Yuan & Deng (2021), at least directly. We also went through Casella & Berger, (2001), and did not find clear support for the second part of the claim there. Specifically, Casella & Berger (2001) presents consistency, efficiency and asymptotic normality as desirable properties of estimators (Chapter 10) and makes it clear that efficiency and normality do not follow from consistency (p. 473) but need to be proven separately. Indeed, there many estimators that are consistent but inefficient. However, at least how we understood Y21's claim, it may be correct in the sense that once we know the factor score weights, those weights can be used to calculate optimal predictions for individual observations. However, the claim does not mean that PLS Mode B would be an optimal way to calculate the weights themselves beyond being consistent. For example, efficiency would still need to be proven.

Leaving aside the fact that very few applications of PLS in marketing and related literature use either Bartlett scores or PLS Mode B, Y21's points – even if correct – do not demonstrate any substantive advantages of PLS. Indeed, even Y21 acknowledge that PLS is not an ideal technique for estimating Bartlett scores from sample data. Specifically, although both ML-based and PLS Mode B Bartlett scores may be biased by measurement error, there are many other reasons to prefer ML-based Bartlett scores, if one is to use them. Importantly, PLS Mode B Bartlett scores are susceptible to bias caused by capitalization on chance, as established over a decade ago by Rönkkö (2010). Thus there is seemingly no clear reason why a researcher in a typical situation of using multi-item scale data would choose this approach.

Further, Y21 states that his results hold only under models with no cross-loadings or correlated errors across the constructs (blocks of indicators). Here we repeat our point from above, namely that this situation is a) unlikely, and b) impossible to know *a priori* for real data. Thus, if a researcher insists on using a composite method, for computational simplicity perhaps, the obvious choice is not PLS, but GSCA(m), which has matrices to convey cross-block information at the indicator level, among other advantages. Indeed, there is no conceivable situation where PLS would be preferred, and as we noted earlier, even core PLS proponents such as Sarstedt and Ringle are moving towards advocating GSCA instead of PLS (e.g. Cho et al, 2020; 2021).

Theme 2: PLS Weights and Reliability

From the above, Y21 also derives the implication that – parallel to the Bartlett factor scores – composites under PLS Mode B are the most reliable among all weighted averages of the observed indicators (Yuan and Deng, 2021). This claim is true only if factor scores are calculated one indicator block at a time, which Y21 does not mention. In more general conditions, regression factor scores outperform Bartlett scores in terms of reliability, but this comes with the cost of producing scores that are biased by other factors.

Further, PLS Mode B composites and Bartlett scores are only asymptotically equivalent for correctly specified models. In finite samples (i.e., real samples, not theoretical infinite ones), PLS Mode B may be better than other composite scores, but all composite scores should perform *worse* than an ML-based Bartlett score. Indeed, the point of Bartlett scores is to produce scores that are not biased by other factors (univocality; see Harman, 1976, p. 387). Yet, PLS, regardless of whether Mode A or Mode B is used, weights the indicators based on their correlations with indicators of other factors virtually guaranteeing that the scores are biased in small samples (Rönkkö & Ylitalo, 2010; Rönkkö & Evermann, 2013; Rönkkö 2014). In this light, the point made here by Y21 is very weak with respect to applied uses of PLS.

Moreover, PLS Mode B models will almost invariably be misspecified because of the assumption that all cross-block indicator correlations (i.e., cross-loadings and error correlations) are channelled through the composite correlation, and PLS cannot handle a violation of that assumption. In common-factor models, such inherent data features can be accommodated (Asparouhov, Muthén, & Morin, 2015; Muthén & Asparouhov, 2012, 2020), which produces factor scores which are not biased by these misspecifications (although they will also contain measurement error). Moreover, the GSCAm procedure can take cross-loadings, correlated measurement errors, *and* measurement error itself into account in order to produce error-free composites (Choi & Hwang, 2020; Hwang, Takane, & Jung, 2017). Taken together, these points result in the conclusion that one should *not use PLS* in most real-world situations, because it cannot handle the inherent features of real-world data (i.e., cross-loadings) and is highly susceptible to capitalization on chance even under correct causal specification.

Y21 speculates that "The criticism against the weights in the PLS-fallacy article might be because analytical results regarding the reliabilities of the composites under PLS-SEM were available only recently." (p. 6) This incorrect for two reasons. First, Dijkstra (1981) proved already more than 40 years ago that asymptotically PLS Mode B produces the "most likely values" of latent variables, which is the same as maximizing reliability. Second, and more importantly as explained in MoM, there exists decades of research that demonstrate that practical advantages of differential indicator weighting are trivial even if ideal weights are known. Starting from Rönkkö & Ylitalo (2010), this has been demonstrated with PLS as well. The example by Y21 of using indicators with reliabilities of 0.16, 0.16, and 0.81 does not invalidate this. In practice the poor indicators would be just thrown away as explained in our response to H21, leaving us comparing a single indicator with reliability of 0.81 and composite with reliability of 0.823. The difference is trivial and in practice when the PLS weights are not calculated from population values like Y21 does, but estimated from sample data, PLS composites rarely outperform simple summed scales (or using a single indicator in this case) as explained in MoM.

Y21 also claims that although PLS Mode A composites can be less reliable than equally-weighted composites (Yuan, Wen and Tang, 2020), those Mode A weights can be transformed to Mode B weights using a non-iterative method. Thus, Y21 claims that these transformed weights enjoy the same statistical properties as Mode B weights. However, given the lack of situations where one would choose a PLS Mode B composite over a common factor-based method for obtaining scores, as detailed in the previous paragraph, there seems no real reason for readers to particularly care about this feature, whether or not it is correct. Finally, Y21's arguments here that weights sometimes make a difference, and sometimes do not, further support the use of the CEI to compare different composite methods.

Theme 3 Bias, Explained Variance, and 'Signal-to-Noise'

Y21 makes some quite interesting points as regards the idea of comparing different estimators as to their 'bias'. Unfortunately, the discussion in Y21 conflates two different issues, and in doing so makes some misleading points. First, we accept the point that without knowledge of 'true' values, we cannot technically speak about 'bias'. However, this does not appear to us to justify Y21's blanket rejection of the entire notion of quantifying the bias of estimators against population parameters in SEM methods, simply because choosing scales for latent variables is necessary for the models to be identified. In fact Y21's argument readily extends to composites as well because regression coefficients depend on the composite weights which too are chosen by a researcher. Taken to an extreme, the argument would also apply to physical measurements. For example, if we regress a persons weight on persons height, we get very different results depending on whether kilograms and centimeters or inches and pounds were used. We take the point that scaling choices and metrics are often specific to a particular simulation, but it is not clear why this should invalidate the notion of within-study comparisons for example.

In practice, latent variable scales are not arbitrary; by constraining the first indicator's loading to 1, the latent variable inherits the scale of the first indicator. (Little, Slegers, & Card, 2006). Nevertheless, we emphasize that researchers should carefully chose scaling methods, and be aware of the impact of this decision for analysis and interpretation. Numerous sources are easily available in the literature already to help researchers understand the implications of scaling constraints (e.g. Klopp & Klößner, 2021; Klößner & Klopp, 2019; Gonzalez & Griffin, 2001; Schweizer et al., 2019; Steiger, 2002). Further, it is worth noting that models, in social science especially, are only ever approximations to reality. If we are to discount the notion of bias due to the scales of latent variables being approximate representation or proxies for the true units of measurement of the corresponding attributes, the implication is that we must also discount statistical modelling entirely given that it is also only an approximation of the underlying reality.

Second, after rejecting the idea of bias as a meaningful concept for evaluating the performance of SEM estimators, Y21 promote using what he terms the 'signal to noise ratio', which he equates with 'effect size'. This argument has two major problems. First, we need to understand that the signal-to-noise ratio discussed by Yuan and Fang (2021) is nothing more than a *t* statistic, defined as the ratio of an estimate and its standard error. The *t* statistic is not an effect size measure according to any common definition (Kelley & Preacher, 2021). In contrast, the *t* statistic and the related *p* value are measures of statistical significance. Indeed, the recent ASA guidelines on using and interpreting statistical significance clearly state "A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result." (Wasserstein & Lazar, 2016, p. 132)

Y21 further claims that, in a situation with two latent variables, PLS Mode B always yields a greater signal to noise ratio than ML-SEM for estimating the regression coefficient between the two latent variables (Yuan and Fang, 2021). However, it seems to us that Y21's signal to noise analysis is incorrect, in that the standard errors are taken directly from the OLS regression analysis that is applied to the PLS composites to obtain the path coefficient, instead of being bootstrapped as per typical PLS practice (Streukens & Leroi-Werelds, 2016). More problematically, as Schuberth et al., (2022) demonstrate, in more realistic scenario with more than one predictor variable, the inconsistency of the PLS estimator can lead to incorrect results, failing to detect relationships that exist and detecting non-existent relationships, in contrast to ML-SEM that identifies the relationships correctly. As such, the claim by Y21 that an "inconsistent estimator can be more preferred if the purpose is to confirm a relationship between two constructs" (p. 3) is simply not true. Thus the claim that "effect sizes" obtained under PLS Mode B effect exceed those under ML is based on shaky results.

Still, the claims made in Y21 about signal to noise are interesting, and worth some more discussion, lest they lead to the creation of yet another PLS myth. Specifically, to avoid issues of scaling, which we point out above, Y21's indicator of 'effect size' is dimensionless. However, even leaving aside the problem noted above, that their effect size measure is in reality is not an effect size measure but a measure of statistical significance, it is still not clear what we can learn from Y21's discussion. Specifically, whether or not the effect size indicator promoted in Y21 and Yuan and Fang (2021) is the most appropriate benchmark, and therefore that all prior simulations are meaningless (which we believe is a conclusion without solid grounding), Y21 tells us what we already know; in their own words: "Our empirical results indicate that *PLS-SEM tends to have an inflated effect size* even with normally distributed data … Also, while PLS-SEM is statistically more powerful in detecting the relationship among theoretical constructs than CB-SEM, it still needs a large enough sample size and good quality of data for reliable model/parameter inference (Marcoulides & Saunders, 2006). In particular, samples with heavy-tails or data contamination can strongly affect the goodness of the estimates by the LS method" (Y21 this issue, emphasis added[2]). It is thus unclear why anyone would use PLS over another method, a point further emphasised even in the *abstract* of Y21: "PLS-SEM may have inflated type I errors and R-square values even with normally distributed data.". It is hard to reconcile any claim that PLS has some kind of 'optimal signal-to-noise ratio', with the claim that it *also* has 'inflated effect sizes' and/or 'inflated type I errors and R-square values'. It appears that Y21 anticipate this objection, as they state that "maximization of R^2 and capitalization on chance cannot be separated". This statement may of course be true, but that is the case for any optimization problem, so the statement is disingenuous at best. In fact, some methods are more robust and

[2] It is worth noting that the latest edition of Hair et al's *A Primer on PLS-SEM* contradicts this, and explicitly states that PLS-SEM is particularly suitable where "a small population restricts the sample size" and "distributional issues are a concern, such as lack of normality" (see Exhibit 1.10). It is hard to see how a typical researcher is to make sense of such contradictory advice.

less susceptible to this issue than others, and many are more robust than PLS (e.g., unit-weighted composites and sum scores, where the weights are fixed and thus immune to sampling fluctuations). Again, we return to the only logical conclusion; there is very little use for PLS in applied research. In Table 3 we provide a list of the main points made in Y21 (of relevance to MoM), and how they provide yet more evidence against the use of PLS.

TABLE 3 HERE

DISCUSSION AND CONCLUSIONS

In our paper MoM, we brought together well over a decade's worth of critique of PLS, using very simple examples, easily reproducible by anyone using software to run PLS or any other statistical analysis, to help illustrate these points to people who are not methodologically inclined. We did not aim to introduce any new points to the discourse; strong evidence for the problems of PLS abounds, but it is available mainly in methodological journals, and is therefore perhaps inaccessible to many applied marketing and management researchers. The existing critiques of PLS already provide more than enough evidence to conclude that PLS as it is used in marketing and related disciplines offers no meaningful advantages over ML SEM, 2SLS or GSCA(m); however, it has enough serious disadvantages that it should be avoided as a general rule. We cannot envisage a single realistic marketing research situation where PLS would be the preferred analytic method on any criterion other than convenience--although in such a case, one may as well use the *most* convenient option: sum scores and OLS regression[3]. Worse, as we have pointed out in MoM and here, PLS is currently used in marketing and related fields in such a way that it is harmful to scientific progress.

[3] Interestingly, in recent personal conversations, one of the present authors heard stories about occasions where researchers have tried to use OLS and sumscores, and reviewers have pushed back with the criticism that such methods are 'too simple', and thus the authors should use PLS. Such a situation is almost akin to a 'simplicity tax' on research. Of course, it should not need saying that simpler methods should actually be *preferred* when more complex results can offer no meaningful advantage. We hope that the material in MoM and this rejoinder

Our intention with MoM was to reach as broad an audience as possible in a clear way, and to reiterate that PLS is not a viable analysis method for typical marketing and related field research problems. Obviously, we expected to see some pushback from one or more of the authors who have for a long time advocated the use of PLS. However, although much of the commentary we received was frustrating in its *ad hominem* nature, employing straw man arguments and fallacious reasoning, we were surprised to note that our message here triggered *additional* evidence against the use of PLS, provided by its own strongest proponent, as well as by an established statistician. Specifically, it is now very clear, in the words of its *own defenders*, that PLS cannot handle typical real-world multi-item scale data features such as cross-loadings or correlated errors. Indeed, H21 go so far as to say that the absence of these features are hard-line pre-requisites of the method itself. Y21 states that relationships estimated in PLS tend to have inflated effect sizes, R^2s, and Type 1 errors (all obviously interconnected), even with normally-distributed data. It is obvious then, that applied users should just adopt better analysis methods, of which plenty are easily available. For example, GSCA(m) with fixed indicator weights would solve all three of the fatal problems of PLS that we point out: (a) capitalization on chance (by using sumscores rather than weighted composites), (b) measurement error (through the inclusion of the uniqueness terms), and (c) non-zero cross-loadings and error correlations (via the additional parameter matrices). Even using disattenuated regression with sum scores can only solve (a) and (b), while PLSc only solves (b). Again, PLS loses hands-down to readily available alternatives (e.g. see Choi and Hwang, 2020; Hwang et al., 2021; Hwang, Takane, and Jung, 2017).

We close our rejoinder with a brief digression, somewhat against our own better judgement, on the *ad hominem* attack by H21 that we have "cause[d] incalculable harm to management and marketing research, as well as the social sciences in general". This

provides enough material for researchers (and editors) to rebut such ill-informed criticism, and perhaps to convince reviewers to stop making it.

statement is both deeply ironic and logically inconsistent with their other arguments, but is worth some discussion because it exemplifies a number of key problems in the current discourse around both PLS and other controversial methods in applied fields like marketing.

Regarding the inconsistency; H21 begin their comment with a statement that PLS "enjoys widespread use in marketing research and is considered a standard procedure with high impact". It is obvious that just because a practice is used often does not make it correct. Even so, it is equally obvious that the PLS advocacy of Hair and others is extremely well cited. Even discounting the around 150,000 citations of Hair's *Multivariate Data Analysis* textbook, there are probably 100,000 citations of the core PLS advocacy papers, many of which have been published in highly-read applied research journals such as *Journal of Business Research*. In fact, the *International Journal of Research in Marketing* awarded its 2021 J-B Steenkamp Award for Long-Term Impact to Rienartz, Haenlein, and Henseler's (2009) empirical comparison of PLS and SEM (over 2500 citations so far). In comparison, even the most popular of the papers dedicated to critiquing PLS (Rönkkö and Evermann, 2013) has fewer than 500 (yes, five hundred) citations, and is published in *Organizational Research Methods* (Antonakis et al.'s *On Making Causal Claims* in *The Leadership Quarterly* has over 1,700, but PLS is only a minor aspect of this paper). It is not clear how any of us could have caused 'incalculable harm', given the evident paucity of our relative academic impact, as judged by people actually citing our critiques of PLS. Further, as stated repeatedly by Hair in his works, PLS is still very commonly used in many applied marketing (and related) research journals, although it must be said that it appears comparatively rarely in the very top rank of journals in marketing, or most other related fields.

We will leave the irony of H21's statement to readers to judge for themselves, but we would also like to make the important point that on too many occasions, authors appear to choose an analysis method using a small set of criteria: a) how many papers in the journals they read are advocating the method and using it, b) how important or well-known are the advocates versus the critics, and c) how powerful is the rhetorical argumentation for and against the method? None of these criteria are meaningful when it comes to methodological choices. Rather than rely on precedence for using a method in applied management journals, or second-hand advocacy and applied papers, researchers should consult *methodological* journals as well, to understand more thoroughly any method they wish to use. If a method has clear evidence pointing against its use, the onus is on the researcher (and the reviewer) to understand the limits of that evidence, and not to take counter-claims by obviously partial advocates at face value.

Moreover, none of the authors of this paper has anything materially to gain by the use or non-use of PLS. Our interest is simply in helping researchers do better work. Every one of the claims we make in MoM is well established in the methodological literature, and supported by strong simulation evidence. Any interested researcher should be able to at least gain some appreciation that support for PLS is far from strong in the methodological literature, and this should make researchers highly cautious about applying the method, no matter how many thousands of citations there are for PLS work in applied non-methodological journals. Finally, it is increasingly obvious that even those who in the past have strongly advocated for PLS, are now either explicitly advocating against it, or instead advocating other methods as being more capable. For example, we cited recent work of Sarstedt and colleagues as recommending the use of GSCA for modelling composites (Cho, Sarstedt, and Hwang, in press). Further, recently Henseler has stated explicitly that PLS is not suitable for models based on reflectively-measured variables (Henseler and Schuberth, this issue), and even more strongly in his recent book: "It is particularly worrying that some quite influential researchers such as Hair continue to spread outdated views on PLS even against their better judgment (see, for instance, the relatively recent publications Hair, Sarstedt, &

Ringle, 2019; Hair, Howard, & Nitzl, 2020)" (Henseler, 2021, pp. 96). Such comments are particularly noteworthy considering H21's dismissal of MoM because it "rest[s] on dated concepts and understandings of composite-based SEM in general…rather than considering the most recent research on PLS-SEM" (H21, this issue). To be clear, we have considered the recent research on 'PLS-SEM' and, like Henseler (who we must not forget was among the most vocal advocates of PLS until recently), we have found it wanting.

In conclusion, surely it is an absolute minimum standard of scientific integrity that we actually understand the tools we use to draw conclusions about the world we are studying. It is obvious that when considering the use of PLS, the best-case scenario is that researchers are either a) unaware of, or b) do not understand, the clear methodological evidence pointing against its use. The objective of MoM was to remedy this situation by showing simple examples that anyone can replicate with their own data, so that no longer can marketing researchers reasonably claim either lack of awareness or understanding. The worst-case scenario is that PLS continues to be used and promoted by advocates in spite of the methodological flaws clearly demonstrated here and in prior works, which truly would be a source of 'incalculable harm to marketing and management research'. Sadly, we can offer no remedy for that.

REFERENCES

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The leadership quarterly*, *21*(6), 1086-1120.

Asparouhov, T., & Muthén, B. (2020). Advances in Bayesian model fit evaluation for structural equation models. Structural Equation Modeling. doi: 10.1080/10705511.2020.1764360

Asparouhov, T., Muthén, B.O. & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. Journal of Management, 41(6), 1561-1577. doi: 10.1177/0149206315591075

Asparouhov, T., Muthén, B.O. & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. Journal of Management, 41(6), 1561-1577. doi: 10.1177/0149206315591075

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Cho, G., Hwang, H., Sarstedt, M., & Ringle, C. M. (2020). Cutoff criteria for overall model fit indexes in generalized structured component analysis. *Journal of Marketing Analytics*, *8*(4), 189-202.

Cho, G., Sarstedt, M., & Hwang, H. (2021). A comparative evaluation of factor- and component-based structural equation modeling approaches under (in)correct construct representations. British Journal of Mathematical and Statistical Psychology. doi: 10.1111/bmsp.12255

Choi, J.Y., & Hwang, H. (2020). Bayesian generalized structured component analysis. British Journal of Mathematical and Statistical Psychology, 73(2), 347-373. https://doi.org/10.1111/bmsp.12166

Choi, J.Y., & Hwang, H. (2020). Bayesian generalized structured component analysis. British Journal of Mathematical and Statistical Psychology, 73(2), 347-373. https://doi.org/10.1111/bmsp.12166

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cook, R.D. and Forzani, L. (2020), “Fundamentals of Path Analysis in the Social Sciences”, *arXiv*, No. 2011.06436.

Dijkstra, T. K. (1981). Latent variables in linear stochastic models: Reflections on “maximum likelihood” and “partial least squares” methods [Doctoral dissertation]. Rijksuniversiteit te Groningen.

Dijkstra, T. (2017). A perfect match between a model and a mode. In H. Latan, & R. Noonan (Eds.), *Partial Least Squares Path Modeling: Basic concepts, methodological issues and applications* (pp. 55-80). Springer Verlag.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. Psychological Methods, 6, 258-269. http://www-personal.umich.edu/~gonzo/papers/sem.pdf

Hair, J. F., & Sarstedt, M. (2019). Factors versus Composites: Guidelines for Choosing the Right Structural Equation Modeling Method. *Project Management Journal*, 50(6), 619–624. https://doi.org/10.1177/8756972819882132

Hair Jr, Joseph F., Hult, G.T.M., Ringle, C.M.,& Sarstedt, M. (2021a) *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications. London: UK.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021b). Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook. Springer International Publishing. https://doi.org/10.1007/978-3-030-80519-7

Hair, Joseph F., et al. (2009), Multivariate Data Analysis: A Global Perspective. 7th ed. Upper Saddle River: Prentice Hall.

Hair, Jr, J. F. et al. (this issue). The fallacies of simple but flawed demonstrations: A comment on "Marketing or methodology? Exposing the fallacies of PLS with simple demonstrations" *European Journal of Marketing*

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017). Mirror, mirror on the wall: A comparative evaluation of composite-based structural equation modeling methods. Journal of the Academy of Marketing Science, 1–17. https://doi.org/10.1007/s11747-017-0517-x

Hanafi, M. (2007). PLS path modelling: computation of latent variables with the estimation mode B. *Computational Statistics*, *22*(2), 275-292.

Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago Press.

Henseler, J. (2021). *Composite Based Structural Equation Modeling: Analyzing Latent and Emergent Variables*. New York: NY. The Guilford Press.

Henseler, J., Hubona, G. and Ray, P.A. (2016), "Using PLS path modeling in new technology research: updated guidelines", *Industrial Management & Data Systems*, Vol. 116 No. 1, pp. 2-20. https://doi.org/10.1108/IMDS-09-2015-0382

Henseler, J., & Schuberth, F. (this issue). Partial least squares as a tool for scientific inquiry: Comments on Cadogan and Lee. *European Journal of Marketing*

Hwang, H., Cho, G., Jung, K., Falk, C., Flake, J., Jin, M. J., & Lee, S. H. (2021). An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis. Psychological Methods, 26(3), 273–294. https://doi.org/10.1037/met0000336

Hwang, H., Takane, Y., & Jung, K. (2017). Generalized structured component analysis with uniqueness terms for accommodating measurement error. Frontiers in Psychology, 8, Article 2137. https://doi.org/10.3389/fpsyg.2017.02137

Hwang, H., Takane, Y., & Jung, K. (2017). Generalized Structured Component Analysis with Uniqueness Terms for Accommodating Measurement Error. Frontiers in Psychology, 8, 2137. https://doi.org/10.3389/fpsyg.2017.02137

IJRM (2021). Announcement: Winner of 2021 Jan-Benedict Steenkamp Award for Long-Term Impact. *International Journal of Research in Marketing* 38 A2–A3 J

Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100: Historical Developments and Future Directions* (1st ed., pp. 47–77). Routledge Academic.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137-152.

Klopp, E., & Klößner, S. (2021). The Impact of Scaling Methods on the Properties and Interpretation of Parameter Estimates in Structural Equation Models with Latent Variables. Structural Equation Modeling, 28(2), 182-206. doi: 10.1080/10705511.2020.1796673

Klößner, S., & Klopp, E. (2019). Explaining constraint interaction: How to interpret estimated model parameters under alternative scaling methods. Structural Equation Modeling, 26, 143-155. https://doi.org/10.1080/10705511.2018.1517356

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A Non-arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS Models. *Structural Equation Modeling*, 13(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

Marcoulides, G. A., & Saunders, C. (2006). Editor's comments: PLS: a silver bullet?. *MIS quarterly*, iii-ix.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Marsh, H.W., Guo, J., Dicke, T., Parker, P.D., & Craven, R.G. (2020). Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM), and Set-ESEM: Optimal Balance Between Goodness of Fit and Parsimony. Multivariate Behavioral Research, 55(1), 102-119. doi: 10.1080/00273171.2019.1602503

Muthén, B. O, & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. Psychological Methods, 17(3), 313-335. doi: 10.1037/a0026802

Muthén, B. O, & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. Psychological Methods, 17(3), 313-335. doi: 10.1037/a0026802

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of research in Marketing*, *26*(4), 332-344.

Rigdon, E.E. (2012), “Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods”, *Long Range Planning*, Vol. 45 No. 5-6, pp. 341-358.

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6-14. https://doi.org/10.1177/1094428120968614

Rönkkö, M. (2014). The Effects of Chance Correlations on Partial Least Squares Path Modeling. *Organizational Research Methods*, *17*(2), 164–181. https://doi.org/10.1177/1094428114525667

Rönkkö, M., & Evermann, J. (2013). A Critical Examination of Common Beliefs About Partial Least Squares Path Modeling. *Organizational Research Methods, 16*, 425-448 https://doi.org/10.1177/1094428112474693

Rönkkö, M., McIntosh, C. N., & Aguirre-Urreta, M. I. (2016). Improvements to PLSc: Remaining problems and simple solutions. Unpublished Working Paper. http://urn.fi/URN:NBN:fi:aalto-201603051463

Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. Journal of Operations Management, 47–48, 9–27. https://doi.org/10.1016/j.jom.2016.05.002

Sarstedt, M., Hair, J.F., Ringle, C.M., Thiele, K.O., & Gudergan, S.P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, 69(10), 3998-4010. https://doi.org/10.1016/j.jbusres.2016.06.007.

Schuberth, F. (2021) Confirmatory composite analysis using partial least squares: setting the record straight. *Review of Managerial Science* **15,** 1311–1345 (2021). https://doi.org/10.1007/s11846-020-00405-0

Schuberth, F., Rosseel, Y., Rönkkö, M., Trinchera, L., & Henseler, J. (2022). Relationships between latent variables are neither arbitrary nor equivalent to relationships between proxies: A Comment on Yuan and Deng (2021). Working Paper.

Schweizer, K., & Troche, S. (2019). The EV scaling method for variances of latent variables. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 15(4), 175–184. https://doi.org/10.1027/1614-2241/a000179

Sharma, P., Liengaard, B., Sarstedt, M., Hair, J., & Ringle, C. (2022). Extraordinary Claims Require Extraordinary Evidence: A Comment on “Recent Developments in PLS.” Communications of the Association for Information Systems.

Steiger, J.H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. Psychological Methods, 7, 210-227.

Streukens, S., & Leroi-Werelds, S. (2016). Bootstrapping and PLS-SEM: A step-by-step guide to get more out of your bootstrap results. European Management Journal, 34(6), 618-632. https://doi.org/10.1016/j.emj.2016.06.003.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Yuan, K.-H. (this issue). Comments on the article ``Marketing or methodology? Exposing the fallacies of PLS with simple demonstrations" and PLS-SEM in general. *European Journal of Marketing*

Yuan, K.-H., & Deng, L. (2021). Equivalence of partial-least-squares SEM and the methods of factor-score regression. *Structural Equation Modeling*, 28(4), 557–571. doi: 10.1080/10705511.2021.1894940.

Yuan, K.-H., & Fang, Y. (2021). Which method delivers greater signal-to-noise ratio: Structural equation modeling or regression analysis with composite scores? Under review

Yuan, K.-H., Wen, Y., & Tang, J. (2020). Regression analysis with latent variables by partial least squares and four other composite scores: Consistency, bias and correction. Structural Equation Modeling, 27(3), 333–350. doi: 10.1080/10705511.2019.1647107

TABLE 1: TECHNICAL POINTS IN MoM, AND ADDITIONAL POINTS MADE IN THIS REJOINDER

Point[1]	Additional Evidence in Existing Literatre[2]	Counter-Evidence in Existing Literature[2]	Conclusion
PLS does not maximize R^2 or explained variance.	Simulation evidence in Rönkkö (2020a)	*None.* Existing PLS literature does not provide evidence to support this claim. Indeed, it is simple to show many methods that can outperform PLS on any specific criteria of maximization.	PLS does not maximise R^2 or explained variance. The claim itself makes little sense and no supporting proofs exist.
Improving reliability by differential indicator weighting is not a reason to use PLS.	Simulation evidence in Rönkkö and Yitalo (2010), Rönkkö and Evermann (2013), and Rönkkö et al. (2016). The simulations in Henseler et al (2014) also show that in most situations PLS leads to a loss of reliability Decades of evidence show that differential indicator weights generally provide only trivial advantages at best.	*None under likely real-world analysis conditions*. PLS may offer small reliability improvements in simulation studies that are designed with conditions ideally favourable to PLS, such as extremely low inter-item correlations e.g.: Simulations in Henseler et al (2014) show a <1% improvement in reliability for situations expressly designed to favour PLS.	It is unclear why a researcher should favour a method which shows a trivial reliability improvement only in situations of very low inter-item correlation, and at the expense of proven serious drawbacks. Standard scale-development procedures recommend against items with low inter-correlations, and where they *should* be included (e.g. formative indices); internal consistency is irrelevant.
PLS weights bias composite correlations: a) if scales are weakly correlated; b) where there are cross-loadings or correlated errors between items in different scales; and c) particularly when sample size is small	Simulations by (Goodhue et al., 2015; Rönkkö, 2014; Rönkkö & Evermann, 2013)	*None*. Rigdon claims weakly-correlating composites are a known violation of PLS assumptions However, this is clearly not a well-known violation of specified PLS pre-requisites, since we are not aware of any published guidelines in primer or introductory PLS literature that state this should be tested.	It is impossible for researchers to know composites are weakly correlated *a priori*. That PLS is not robust to departures from this assumption should be pointed out in PLS introductory texts.
AVE and CR should never be used with PLS / PLS should not be used to validate measures	Simulations by Evermann and Tate (2010), Rönkkö and Evermann (2013), Rönkkö and Cho (2022). These results are corroborated even by PLS advocates' research, e.g. (Henseler et al., 2014; McIntosh et al., 2014)	*None*. HTMT has been proposed as an improvement, but it is not a PLS-specific method, and CFA works better more generally. Evidence which suggests HTMT generally outperforms CFA (e.g. Vorhees et al., 2016) is based on incorrect use of CFA (Rönkkö and Cho, 2022).	HTMT is a better method than using AVE with PLS. However, HTMT is not a PLS method. PLS introductory texts should remove mention of AVE and CR as measure validation and model assessment tools. Factor analysis should be used to test the assumptions of HTMT.
Additional Point not in MoM: The bootstrap "sign-change" options in PLS programs can produce 100% false positive rate.	Simulation evidence in Rönkkö et al. (2015).	*None*. In fact, Henseler et al. (2016) recommend abandoning the sign-change corrections.	Unfortunately, the damage to statistical decision-making has likely already been done, and is perhaps still continuing. A

			warning should be issued as an editorial. The method should be explicitly banned. Of course, if researchers abandon the PLS method entirely, this issue will be resolved.

[1]All points made in MoM are supported by numerical illustrations, in order to prioritise understandability for non-methodological readers.

[2]Sources of evidence and counter evidence are considered in terms of a hierarchy of strength. While we recognize that for different purposes, different forms of evidence are more or less appropriate, the strength of evidence for or against the sort of claims we make in MoM can be ascertained according to the following hierarchy: The strongest evidence is a mathematical proof, followed by appropriate simulations, followed by numerical illustrations (e.g. using real data). Rhetoric alone is not considered to be evidence for or against these claims, and hence we do not include sources who only rely on rhetoric here.

Note: we also make the conceptual point that PLS is not a latent variable method at all. This point is not countered in existing literature on PLS, and in fact, when pushed, PLS advocates often appeal to the argument that PLS is not really intended to estimate common factor-based population models and is in fact most suitable for examining "composite-based population models" (see Dijkstra, 2017; Hair & Sarstedt, 2019; Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016).), even though we show in MoM that it is common practice to explicitly and implicitly use PLS to examine factor-model based conceptualizations. Indeed, it is clear that recent PLS work still claims that PLS can estimate reflective models (Schuberth, 2021), and even the most recent edition of Hair et al's (2021) PLS primer text clearly indicates that PLS can handle reflective models, which from a measurement theory perspective are essentially equivalent to factor models, and certainly are not composite models (e.g. Markus and Borsboom, 2013).

TABLE 2: EVIDENCE *AGAINST* PLS PROVIDED BY H21

Point made in MoM	Counter Claim in H21	Conclusion
PLS weights bias composite correlations: a) if scales are weakly correlated b) where there are cross-loadings or correlated errors between items in different scales c) particularly when sample size is small	Conditions a), b), and c) constitute 'well known' violations of PLS pre-requisites	PLS should not be used where scales may be weakly correlated, or where there may be correlated errors or cross loadings between items in different scales. *Therefore*: PLS is not suitable for most real-world applications
PLS offers no meaningful advantage in reliability for composites under realistic scale-development scenarios. Any demonstrated advantage is only exhibited where there are very low inter-item correlations, and only to a trivial extent.	PLS shows a small advantage in internal consistency when there are very low inter-item correlations. Summed scores are more sensitive to low-quality indicators.	Using appropriate scale development procedures, as recommended by Hair (e.g. Hair et al., 2009), should lead to removal of low-quality indicators *before* creating composites. Situations where low inter-item correlations are possibly good practice (e.g. formative indices) are irrelevant to reliability / internal consistency assessment. *Therefore*: PLS is not suitable for real-world scale-development applications, and offers no advantage in other real-world applications.
CR and AVE should not be used either for measurement or model assessment with PLS.	H21 agree with us, and also agree that this is 'not new'. (Unfortunately, it is *also* evident that advice on using CR and AVE appears even in the latest PLS primer texts [e.g. Hair et al., 2021]).	CR and AVE should not be used with PLS results, either to assess models or measures. HTMT and coefficient (Cronbach's) alpha are general techniques that can be applied after their assumptions have been verified with factor analysis.
PLS does not maximise explained variance or R^2. The claim is ambiguous when there are more than one R^2 in the model and not supported by any evidence.	H21 claim that "PLS estimates are *coherent* in the sense that all the residual variances are minimized *jointly*" (emphasis added, pg. 5). However, this statement does not rebut the original claim in MoM, since the terms 'coherent' and 'joint' are not clear, and do not provide an optimization criterion. H21 also claim that GSCA does provide an optimization criterion.	PLS does not maximize explained variance or R^2 other than in the trivial sense that the OLS stage does in the second stage. However the latter is not specific to PLS. GSCA does provide an optimization criterion. *Therefore:* Because GSCA can also deal with cross-loading and measurement error, GSCA should be preferred to PLS when modelling composites. However, CEI should be always used to compare weighted composites against unweighted ones.

TABLE 3: EVIDENCE *AGAINST* PLS PROVIDED BY Y21

Claim in Y21	Conclusion
Indicator weights from PLS Mode B are equivalent to Bartlett factor scores, but only when assuming no cross-loadings or correlated errors across scales, which ML-SEM and Bayesian SEM can accommodate.	In real-world analysis situations, we cannot assume cross-loadings or correlated errors are non-existent. *Therefore:* PLS should not be used in situation where cross-loadings or correlated errors may exist, as it has no way to account for these features.
PLS Mode B composites are most reliable among all weighted averages of observed indicators, for a) correctly specified models, and b) where all cross-loadings and error correlations are completely channeled through the composite correlation(s).	Even if a model was correct, in real-world finite samples, advantages of differential item weighting are trivial as long as very bad items are first dropped from the data. *Therefore*: Differentially weighted composites should be always compared against unit weighted ones using the CEI. Unless meaningful differences are found and can be explained, unit-weighted composites should be chosen for their simplicity. .
In a situation with two latent variables, PLS Mode B always yields a greater signal to noise ratio than ML-SEM for estimating the regression coefficient between the two latent variables. *However:* “PLS-SEM tends to have an inflated effect size even with normally distributed data …” “PLS-SEM may have inflated type I errors and R-square values even with normally distributed data.”.	Y21 confuses effect size with statistical significance. While PLS may lead to higher statistical significance, and thus a greater likelihood of finding and effect, this comes at the expense of a higher chance of false positives. The claim by Y21 does not hold in more realistic models with more than one predictor variable, where the inconsistency of PLS can lead to incorrect conclusions about the existence of an effect. *Therefore*: Results from a PLS analysis are more likely to be false positives than those from other methods such as ML-SEM, *ceteris paribus*.
“[PLS] needs a large enough sample size and good quality of data for reliable model/parameter inference (Marcoulides & Saunders, 2006). In particular, samples with heavy-tails or data contamination can strongly affect the goodness of the estimates by the LS method”.	PLS should especially not be used with small samples or low quality data.