MIS Quarterly

METHODS ARTICLE

A MULTICOLLINEARITY AND MEASUREMENT ERROR STATISTICAL BLIND SPOT: CORRECTING FOR EXCESSIVE FALSE POSITIVES IN REGRESSION AND PLS

Dale L. Goodhue
Terry College of Business, MIS Department, University of Georgia,
Athens, GA 30606 U.S.A. {dgoodhue@terry.uga.edu}

William Lewis
{william.w.lewis@gmail.com}

Ron Thompson
School of Business, Wake Forest University,
Winston-Salem, NC 27109 U.S.A. {thompsrl@wfu.edu}

Multiple regression has a previously unrecognized "statistical blind spot" because when multicollnearity and measurement error are present, both path estimates and variance inflation factors are biased. This can result in overestimated t-statistics, and excessive false positives. PLS has the same weakness, but CB-SEM's estimation process accounts for measurement error, avoiding the problem. Bringing together partial insights from a range of disciplines to provide a more comprehensive treatment of the problem, we derive equations showing false positives will increase with greater multicollinearity, lower reliability, greater effect size in the dominant correlated construct, and, surprisingly, with higher sample size. Using Monte Carlo simulations, we show that false positives increase as predicted. We also provide a correction for the problem. A literature search found that of IS research papers using regression or PLS for path analysis, 33% were operating in this danger zone. Our findings are important not only for IS, but for all fields using regression or PLS in path analysis.

Keywords: Multicollinearity, measurement error, M+ME, multiple regression, partial least squares, PLS, CB-SEM, false positives, Type I error, statistical power, variance inflation factor, VIF, path estimate bias

Appendix A

Deriving Equations for M+ME Biases and for t-statistic Overestimations

Determining the Impact of VIF Bias on the t-statistic

Consider first the situation where we have no measurement error. Using standard equations for the estimated standard error of β_2 from any regression textbook, an unbiased and consistent estimate of the standard error is

$$s^2 = \Sigma\varepsilon_i^2 / (N - K) \tag{A1-1}$$

and the estimate of the variance of β_2 is

$$\text{Var}(\hat{\beta_2}) = [s^2 / \Sigma x_{2i}^2] * [1 / (1-\rho_{12\text{Underlying}}^2)] \tag{A1-2}$$

When the error terms are normally distributed, the following has a t distribution:

$$\text{t-stat} = (\hat{\beta_2} - \beta_2) / \{\text{Var}(\hat{\beta_2})\}^{.5} \quad \sim t_{N-K} \tag{A1-3}$$

Substituting in the equations for $\text{Var}(\hat{\beta_2})$ and s^2 from above (A1-1 and A1-2), we can see the impact of correlated predictor variables on the t-stat.

$$\text{t-stat} = (\hat{\beta_2} - \beta_2) \; / \; \{ [s^2 / \Sigma x_{2i}^2] * [1 /(1- \rho_{12\text{ Underlying}}^2)] \}^{.5}$$
$$\text{t-stat} = (\hat{\beta_2} - \beta_2) \; / \; \{ [\{\Sigma\varepsilon_i^2 / (N - K)\}] / \Sigma x_{2i}^2]) * [1 /(1- \rho_{12\text{ Underlying}}^2)] \}^{.5}$$

Rearranging terms:

$$\text{t-stat} = (\hat{\beta_2} - \beta_2) \; / \; \{ [\Sigma\varepsilon_i^2 /\{(N - K) * \Sigma x_{2i}^2 \}] * [1 /(1- \rho_{12\text{ Underlying}}^2)] \}^{.5}$$

Squaring both sides:

$$(\text{t-stat})^2 = (\hat{\beta_2} - \beta_2)^2 \; / \; \{[\Sigma\varepsilon_i^2 /\{ (N - K) * \Sigma x_{2i}^2 \}] * [1 /(1- \rho_{12\text{ Underlying}}^2)] \}$$
$$(\text{t-stat})^2 = (\hat{\beta_2} - \beta_2)^2 * \{ [(N - K) *\Sigma x_{2i}^2] / \Sigma\varepsilon_i^2] * [(1- \rho_{12\text{ Underlying}}^2) / 1] \}$$
$$(\text{t-stat})^2 = (\hat{\beta_2} - \beta_2)^2 * \{ [(N - K) * \Sigma x_{2i}^2 * (1 - \rho_{12\text{ Underlying}}^2)] \} \; / \Sigma\varepsilon_i^2$$

Rearranging terms:

$$(\text{t-stat})^2 = \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}^2 * (1) / \Sigma\varepsilon_i^2\}$$
$$- \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}^2 * (\rho_{12\text{ Underlying}}^2)] / \Sigma\varepsilon_i^2\} \tag{A1-4}$$

Note that the first half of equation (A1-4) is the squared t-statistic if there were no correlation between X_1 and X_2. The second half is the correction for when there is a correlation (at this point all assuming no measurement error).

As suggested by Goodhue et al. (2011), the problem with the above is that the estimate for the correlation ($\rho_{12\text{Underlying}}$) assumes perfect measurement (i.e., that the X_i values have no random measurement error). In regression, estimates of the two constructs are based on averaged or summed indicator scores, and any estimate of the $\rho_{12\text{Underlying}}$ correlation will be attenuated (reduced) by the random measurement error, as shown below:

$$\rho_{12\text{Underlying}} = \rho_{12\text{Overt}} \; / \; (\alpha_1 * \alpha_2)^{1/2} \qquad \text{(Equation 2 from the paper proper)} \tag{A1-5}$$

When we do have measurement error, $\rho_{12\text{-Overt}}$ is not equal to $\rho_{12\text{Underlying}}$. When the t-stat calculated by regression has not been corrected for attenuation, A1-4 becomes:

$$\begin{aligned}(\text{t-stat}_{\text{Overt}})^2{}_{(\text{incorrect})} = {} & \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2]/\Sigma\varepsilon_i{}^2\} \\ & - \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{-Overt2}})]/\Sigma\varepsilon_i{}^2\}\end{aligned} \quad \text{(A1-6)}$$

Below (in equation A1-7) we show how the t-statistic value is biased by the incorrect VIF, assuming we know the reliabilities for X_1 and X_2. The first two lines are equation A1-6, assuming there is no correction to the VIF for random measurement error. The third line adds back the amount that was incorrectly subtracted out to correct for the X_1 and X_2 correlation (incorrect because not taking into account random measurement error attenuation), and then the fourth line subtracts the proper amount to correct for X_1 and X_2 correlation (taking into account random measurement error attenuation).

$$\begin{aligned}(\text{t-stat}_{\text{Corrected}})^2 = {} & \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2]/\Sigma\varepsilon_i{}^2\} \\ & - \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{-Overt}}{}^2)]/\Sigma\varepsilon_i{}^2\} \\ & + \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{-Overt}}{}^2)]/\Sigma\varepsilon_i{}^2\} \\ & - \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{Underlying2}})]/\Sigma\varepsilon_i{}^2\}\end{aligned} \quad \text{(A1-7)}$$

The amount by which regression is overestimating the "t-statistic squared" due to the incorrect VIF is the following. This is the last two terms of equation A1-7 with the signs (and order) reversed:

$$\begin{aligned}(\text{t-stat})^2 \text{ overestimation} = {} & (\text{t-stat}_{\text{Overt}})^2 - (\text{t-stat}_{\text{Corrected}})^2 = \\ & + \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{Underlying2}})]/\Sigma\varepsilon_i{}^2\} \\ & - \{[(\hat{\beta_2} - \beta_2)^2 * (N - K) * \Sigma x_{2i}{}^2 * (\rho_{12\text{-Overt}}{}^2)]/\Sigma\varepsilon_i{}^2\}\end{aligned} \quad \text{(A1-8)}$$

By inserting the square of equation A1-5 for the $\rho_{12\text{Underlying}}{}^2$ term in A1-8, we get the following:

$$\begin{aligned}(\text{t-stat})^2 \text{ overestimation} = {} & [(\hat{\beta_2} - \beta_2)^2 (N - K) \Sigma x_{2i}{}^2 (\rho_{12\text{Overt}}{}^2) / (\alpha_1 * \alpha_2)] / \Sigma\varepsilon_i{}^2 \\ & - [(\hat{\beta_2} - \beta_2)^2 (N - K) \Sigma x_{2i}{}^2 (\rho_{12\text{Overt}}{}^2)] / \Sigma\varepsilon_i{}^2\end{aligned} \quad \text{(A1-9)}$$

Gathering common terms, the amount the t-statistic squared in regression is overestimated due to the VIF bias when there are both correlated predictors and random measurement error is[1]

$$(\text{t-stat})^2 \text{ overestimation} = [\,(1/(\alpha_1 * \alpha_2)) - 1] * [(\hat{\beta_2} - \beta_2)^2 \ (N - K) \ \Sigma x_{2i}{}^2 (\rho_{12\text{-0vert}}{}^2)\,] / \ \Sigma\varepsilon^2{}_i \quad \text{(A1-10)}$$

Equation A1-10 is helpful in seeing the impact of various factors on the t-statistic overestimation due to bias in the VIF. It was one of the insights that led to our hypothesis generation. Of course it does not tell the full story, which would also require taking into account the path bias embodied in equations 6 and 7 from the paper, and determining the corrected standard deviation for the corrected estimated underlying path. We explain the logic of that below.

Correcting the Path Estimates for the M+ME Bias

Johnston's (1972) equations. Consider the following regression equation:

$$Yi = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ . \ . \ . \ . \ + \beta_k X_{ki} + \varepsilon_i \quad \text{(A1-11)}$$

Though the essential results extend to the full equation above, for simplicity we will ignore all the $\beta_k X_{ki}$ terms above when k is greater than 2, giving

---

[1]We note that it would be incorrect to suggest that we can take the square root of both sides of equation (A1-10) and therefore determine that

$$\text{t-stat overestimation} = \{[(1/\alpha_1 * \alpha_2) - 1] * [(\hat{\beta_2} - \beta_2)^2 \ (N - K) \ \Sigma x_{2i} (\rho_{12\text{-Overt}}{}^2)\,] / \ \Sigma\varepsilon^2{}_i\,\}.5 \quad \text{(not correct)}$$

The $(\text{t-stat})^2$ overestimation term is not the same as $[\text{t-stat overestimation}]^2$.

$$Yi = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad \text{(A1-12)}$$

Following Johnston (1972, pp. 160-162), if we first look at the impact of multicollinearity on bias in the regression path estimates of β_1 and β_2 (that is, $\hat{\beta_1}$ and $\hat{\beta_2}$), we see the following:

$$\hat{\beta_1} - \beta_1 = \Sigma ux_1 - (\rho_{12} * \Sigma uv)/(1 - \rho_{12}^2) \quad \text{(A1-13)}$$

and

$$\hat{\beta_2} - \beta_2 = (\Sigma uv)/(1 - \rho_{12}^2) \quad \text{(A1-14)}$$

where u is the vector of error terms in Y; v is the vector of error terms in the equation for X_2 as a function of X_1; and ρ_{12} is the underlying correlation between X_1 and X_2.

It is not necessary for the reader to understand the two equations in depth, but only to understand that mathematically, when X_1 and X_2 are correlated and the error terms u and v are non-zero,[2] regression estimates of $\hat{\beta_1}$ and $\hat{\beta_2}$ will be biased away from the underlying value (the true value absent any systematic error). We note that as ρ_{12} increases, the size of the negative bias for $\hat{\beta_1}$ increases. Likewise, as ρ_{12} increases, the size of the positive bias for $\hat{\beta_2}$ increases. Johnston goes on to say that a large correlation between independent variables "is thus likely to produce large and opposite errors in $\hat{\beta_1}$ and $\hat{\beta_2}$; if $\hat{\beta_1}$ underestimates β_1, then $\hat{\beta_2}$ is likely to overestimate β_2 and vice versa. It is thus very important that the standard errors should alert one to the presence of multicollinearity" (p. 162).

Green and Kiernans's (1989, pp. 359-363) equations. Green and Keirnan carried the analysis of the impact of multicollinearity and random measurement error on regression path biases further than Johnston. In particular, Green and Kiernan make a clear distinction between what we are calling the "overt" ($\rho_{12overt}$) and the "underlying" ($\rho_{12underlying}$) correlation—the overt correlation is the correlation based on the "signal-plus-noise" or the correlation between the "error containing" measures of the two constructs, while the "underlying" correlation is the correlation based only on the "signal," without the noise (Green and Kiernan 1989, p. 360.)

Under fairly general conditions (equal reliability $\{\alpha_1 = \alpha_2\}$ for β_1 and β_2, $\beta_1 > 0$), Green and Keirnan gave equations for what they call "proportional inconsistencies" or (PI_i) defined as $(\beta_i - \text{plim } \hat{\beta_i}) / \beta_i$. If PI_i is positive, then plim $\hat{\beta_i}$ is less than β_i (i.e., an underestimation); If PI_i is negative, then plim $\hat{\beta_i}$ is greater than β_i; (i.e., an overestimation). We will reproduce two of their equations, folding in several other additional reasonable assumption that are appropriate to our analysis here.[3] With this assumption Green and Kiernan's equations for the proportional inconsistency show us more about the impact of M+ME on regression estimates than Johnston's treatment.

$$PI_{\beta 1} = (\beta_1 - \text{plim } \hat{\beta_1}) / \beta_1 = [(1 - \alpha_1)/(1 - \rho_{12}^2)] * [1 - (\rho_{12} * (\beta_2/\beta_1))] \quad \text{(Equation 6 repeated)} \quad \text{(A1-15)}$$

$$PI_{\beta 2} = (\beta_2 - \text{plim } \hat{\beta_2}) / \beta_2 = [(1 - \alpha_1)/(1 - \rho_{12}^2)] * [1 - (\rho_{12} * (\beta_1/\beta_2))] \quad \text{(Equation 7 repeated)} \quad \text{(A1-16)}$$

Here ρ_{12} is the "overt" correlation between X_1 and X_2, α_1 is the reliability of X_1 and of X_2 (α_1 is assumed be equal to α_2).[4] PI_i indicates the proportional amount the plim $\hat{\beta_i}$ estimate has been biased away from the true (or underlying) value of β_i. The "plim" indicates that plim $\hat{\beta_i}$ would be the estimate if there were an infinite number of data points. (As stated earlier in the paper, we found that with sample sizes of 100 to 200, the equations predicted the values of regression path estimates reasonably accurately, and for collections of 500 datasets at those sample sizes, quite accurately in the aggregate).

These equations can give us a feel for how M+ME will affect regression path estimates. Assume for the moment that β_1, β_2, and ρ_{12} are all positive and β_1 is greater than β_2 (a not uncommon situation). Under these conditions, given that α_1, ρ_{12}, and β_2/β_1 are all less than one, it can be seen that $PI_{\beta 1}$ will always be greater than zero. (Recall that: $PI_i > 0 \rightarrow \hat{\beta_i}$ is underestimated.) Therefore given our assumptions, when there is multicollinearity and random measurement error, the dominant β_1 will always be underestimated.

---

[2]The u and v error components of Johnston's equations might contain more than only random measurement error, but both would certainly increase as random measurement error in the X and Y values increased. Green and Kiernan's equations focus more precisely on measurement error and its implications.

[3]See footnote 15 in the paper proper.

[4]Note that when α_1 is not equal to α_2, we use the approximation of $\alpha_{combination} = (\alpha_1 * \alpha_2)^{1/2}$ so that we can continue to use Green and Kiernan's equations.

Depending on the values of ρ_{12} and β_1/β_2, the non-dominant β_2 could be under- or over-estimated. In Appendix C we present the logic that shows that except when β_1 and β_2 have close to the same value, M+ME will tend to push the $\hat{\beta_2}$ estimate higher than β_2 if β_2 is positive.

Equations for Correcting the Path Estimate Bias. Although the algebra is tedious, equations A1-15 and A1-16 can be turned around to allow us to calculate estimates of the underlying β values, given the $\hat{\beta_1}$ and $\hat{\beta_2}$ estimates, as follows:

Let $C = (1-\alpha_1) / (1-\rho_{12}^2)$ (Recall that when α_1 is not equal to α_2 we use $\alpha_1 = \alpha_{combination} = (\alpha_1 * \alpha_2)^{1/2}$. See footnote #16.)

$(\beta_1 - \hat{\beta_1})/\beta_1 = C * [1 - \rho_{12} * \beta_2/ \beta_1]$ (Green and Kiernan's Equation for $PI_{\beta 1}$)
$(\beta_1 - \hat{\beta_1}) = \beta_1 \{C * [1 - \rho_{12} * \beta_2/ \beta_1] \}$
$\beta_1 = \hat{\beta_1} + \beta_1 C * [1 - \rho_{12} * \beta_2/ \beta_1]$
$\beta_1 = \hat{\beta_1} + \beta_1 C - C * \rho_{12} * \beta_2]$
$\beta_1 - \beta_1 C = \hat{\beta_1} - C * \rho_{12} * \beta_2]$
$\beta_1 = \hat{\beta_1}/(1-C) - [C * \rho_{12} * \beta_2] /(1-C)$

Similarly

$$\beta_2 = \hat{\beta_2}/(1-C) - [C * \rho_{12} * \{\beta_1\}] /(1-C)$$

Substituting in (β_1) and collecting terms,

$\beta_2 = \hat{\beta_2}/(1-C) - [C * \rho_{12} * \{\hat{\beta_1}/(1-C) - [X * \rho_{12} * \beta_2] /(1-C)\}] /(1-C)$
$\beta_2 = \hat{\beta_2}/(1-C) - C * \rho_{12} * \hat{\beta_1}/(1-C)^2 + [(C * \rho_{12})^2 * \beta_2] /(1-C)^2$
$\beta_2 - (C * \rho_{12})^2 * \beta_2 /(1-C)^2 = \hat{\beta_2}/(1-C) - C * \rho_{12} * \hat{\beta_1}/(1-C)^2$
$\beta_2 [1 - (C * \rho_{12})^2 /(1-C)^2] = \hat{\beta_2}/(1-C) - C * \rho_{12} * \hat{\beta_1}/(1-C)^2$
$\beta_2 = [\hat{\beta_2}/(1-C) - C * \rho_{12} * \hat{\beta_1}/(1-C)^2] / [1 - (C * \rho_{12})^2 /(1-C)^2]$
$\beta_2 = \hat{\beta_2}/(1-C) - C * \rho_{12} * \hat{\beta_1}/(1-C)^2 / [(1-C)^2 - (C * \rho_{12})^2] / (1-C)^2$

$$\beta_2 = \frac{\hat{\beta_2} * (1-C) - C * \rho_{12} * \hat{\beta_1}}{\left[(1-C)^2\right] - (C * \rho_{12})^2} \quad \text{(A1-17)}$$

Similarly

$$\beta_1 = \frac{\hat{\beta_1} * (1-C) - C * \rho_{12} * \hat{\beta_2}}{\left[(1-C)^2\right] - (C * \rho_{12})^2} \quad \text{(A1-18)}$$

Correcting the Standard Deviation of the Corrected Path Estimate. One final insight is needed in correcting for the M+ME path bias. To determine the proper estimate of the path standard deviation, we have to recognize that the path bias created by the M+ME comes with a change in the standard deviation. To calculate the corrected standard deviation for the true path we need to, in a sense, undo that change. Once the variances of the $\hat{\beta_1}$ and $\hat{\beta_2}$ paths have been corrected for the VIF bias (Equation A1-2), the variance of the estimates for the above underlying β_1 and β_2 paths can be calculated using a standard result from statistics:

if $\beta_2 = a * \hat{\beta_2} + b * \hat{\beta_1}$
then $Var(\beta_2) = a^2 * Var(\hat{\beta_2}) + b^2 * Var(\hat{_1})$ (A1-19)

where in our case "a" is $(1-C) /\{ (1-C)^2 - (C * \rho_{12})^2\}$, and "b" is $C * \rho_{12} / \{ (1-C)^2 - (C * \rho_{12})^2\}$ from equation A1-17.

Appendix B

Will PLS with Bootstrapping Correct for M+ME?

For PLS with bootstrapping to correct for the M+ME blindspot seen in regression, it would need to overcome the deficiency noted in our equation 3 versus our equation 4. That is, it would need to somehow incorporate random measurement error into its estimate for the standard deviations of the X1 and X2 paths leading to Y1. We see two possible ways that PLS with bootstrapping could do this. First, bootstrapping could determine the reliability of the two constructs (in our case the X1 and X2 constructs). It could then adjust its (bootstrapping determined) standard deviation of the path using those reliabilities, similarly to our equation 4. However, we see no point in the PLS bootstrapping process where the reliability of the two construct measures is taken into account. For each bootstrapping resample, after the indicator weights are determined and the proxy construct scores calculated, all information about the indicator values is discarded. What occurs is that OLS regression is used with the proxy construct scores to determine another set of path values. Nowhere in the process is the information (for explicitly determining the measurement reliability of the constructs) used by PLS or its bootstrapping process.

A second possibility could be that bootstrapping automatically takes M+ME into account. The central assumption of bootstrapping in general (Mooney and Duval 1993) is that the variation contained in a given sample is representative of the variation existing in the larger population. If this were true for the M+ME blind spot, then bootstrapping could correctly incorporate the extra variation due to M+ME and suggest appropriately larger standard deviations.

If in some bootstrapping resamples M+ME led to additional *overestimations* of the path values, and in others M+ME led to additional *underestimations*, then the total distribution of the bootstrapping resample path values would be appropriately wider, and estimations for the path standard deviations would have increased accordingly. However, recall that each PLS bootstrapping resample is drawn from the original sample and therefore contains roughly the same underlying β_1, β_2, ρ_{12}, and other characteristics as the original sample. We showed in Appendix A that Green and Kiernan's equations (A1-15 and A1-16) clearly indicate that when there is M+ME and both path estimates are positive and not too close together, regression will tend to *systematically* underestimate $\hat{\beta_1}$ and systematically overestimate $\hat{\beta_2}$ based on the reliability and the values of β_1, β_2, and ρ_{12}.[5] The bias seen in Green and Kiernan's equations (6 and 7) for regression will then be apparent in each of the bootstrapping regression results.

Therefore, it is reasonable to assume that when PLS uses OLS to estimate paths for each of its bootstrapping samples, it will *tend to* systematically underestimate *each* of those $\hat{\beta_1}$ path estimates by roughly the same amount, and to systematically overestimate each of those $\hat{\beta_2}$ estimates by roughly the same amount. If that is true, instead of having the collection of bootstrapping $\hat{\beta_2}$ path estimates more widely dispersed, they will be biased but about as closely spaced as if there were no M+ME bias.

PLS's bootstrapping distributions should therefore incorporate the same variance inflation factor as seen in equation 3, mirroring the results seen in regression. We see no argument for a way in which the PLS bootstrapping resamples will incorporate the correction shown in our equation 4.

Appendix C

Further Exploration of Green and Kiernan's Equations: Impact of Negative Paths or Negative Correlations

Mela and Kopalle (2002) have argued that positive versus negative correlations would have very different (asymmetric) impacts on path biases and path estimate variances. For us (with multicollinearity and random measurement error) that question is answered by looking again at Green and Kiernan's (1989, p. 360) equations for the bias (or *proportional inconsistency*) in the estimates for $\hat{\beta_1}$ and $\hat{\beta_2}$:

$$PI_{\beta 1} = (\beta_1 - \text{plim}\ \hat{\beta_1}) / \beta_1 = [(1 - \alpha_1)/(1-\rho_{12}^2)] * [1-(\rho_{12} * (\beta_2/\beta_1))] \quad \text{(Eq. 6 repeated)}$$

$$PI_{\beta 2} = (\beta_2 - \text{plim}\ \hat{\beta_2}) / \beta_2 = [(1 - \alpha_1)/(1-\rho_{12}^2)] * [1-(\rho_{12} * (\beta_1/\beta_2))] \quad \text{(Eq. 7 repeated)}$$

[5]See Appendix C for the impact of relaxing these assumptions about the signs of $\hat{\beta_1}$, $\hat{\beta_2}$, and ρ_{12}.

Equations 6 and 7, repeated above, can tell us quite a bit about the behavior of $\hat{\beta_1}$ and $\hat{\beta_2}$. First, remember that if $PI_{\beta i}$ is > 0, then $|\hat{\beta_i}|$ is an underestimation of $|\beta_i|$. If $PI_{\beta i}$ is < 0, then $|\hat{\beta_i}|$ is an overestimation of $|\beta_i|$. If $|\hat{\beta_i}|$ is an underestimation of $|\beta_i|$, this means that $|\hat{\beta_i}|$ is closer to zero than $|\beta_i|$. Whether β_1 or β_2 are over- or under-estimated depends upon the sign of $PI_{\beta 1}$ or $PI_{\beta 2}$. Note that in both equations, the $[(1 - \alpha_1)/(1-\rho_{12}^2)]$ term before the asterisk is always positive. Therefore whether β_1 or β_2 are over or under-estimated depends upon the sign of $[1-(\rho_{12} * (\beta_2/\beta_1))]$ or $[1-(\rho_{12} * (\beta_1/\beta_2))]$. Table C1 shows the impact of different combinations of negative and positive signs on $[1-(\rho_{12} * (\beta_2/\beta_1))]$ or $[1-(\rho_{12} * (\beta_1/\beta_2))]$, and therefore on the value and sign of the proportional inconsistencies.

One interesting outcome is apparent in the column labeled "Decision Condition" of Table C1. For β_1 there is no decision: if $|\hat{\beta_1}| > |\hat{\beta_2}|$, we would say that $\hat{\beta_1}$ is dominant. In this case $|\hat{\beta_1}|$ will always be an underestimation of $|\beta_1|$. More interesting is the case for $|\hat{\beta_2}|$. If $|\hat{\beta_2}|$ is relatively large (i.e. , greater than $|\rho_{12} * \hat{\beta_1}|$), then $|\hat{\beta_2}|$ will also be an underestimate of $|\beta_2|$. Otherwise, $|\hat{\beta_2}|$ will be an overestimate of $|\beta_2|$. This last can lead to false positives.

Relative to Mela and Kopalle's arguments, in fact it can be seen that having zero or two negative signs for ρ_{12}, β_1,and β_2 creates a quite different configuration of the results than having one or three negatives. Having exactly two of the signs negative keeps the same configuration of results, though it does cause one or more path estimate biases to switch (symmetrically) from positive to negative. This all suggests a slightly different reading of the Mela and Kopalle paper. Although they focused on the impact of omitted but correlated variables, behavior similar to their findings can be created without omitted variables, by adding measurement error and collinearity. The impacts Mela and Kopalle seek to show have the same relationship as those pointed out above from Green and Kiernan's equations.

Table C1. Behavior of Proportional Inconsistencies (Green and Kiernans's PI) (Assuming $|\beta_1| > |\beta_2|$)

Focus on $PI\beta_1$ or $PI\beta_2$	Decision Condition	Relationship of $\lvert\rho_{12} * \beta_2/\beta_1\rvert$ to "1"	# Minus Signs among ρ_{12}, β_2, β_1	What Happens to $1 - (\rho_{12} * \beta_2/\beta_1)$	What Happens to $PI\beta_1$ or $PI\beta_2$	
$PI\beta_1 = (1-\rho_1)/(1-\rho_{12}^2) * [1-(\rho_{12} * \beta_2/\beta_1)]$	none	$\lvert\rho_{12} * \beta_2/\beta_1\rvert$ always < 1	0 or 2 Minus signs	$(\rho_{12}*\beta_2/\beta_1)$ subtracts from one	$PI\beta_1$ is always positive	$\lvert\hat{\beta_1}\rvert$ is an under-estimate of $\lvert\beta_1\rvert$
			1 or 3 Minus signs	$(\rho_{12}*\beta_2/\beta_1)$ adds to one	$PI\beta_1$ is more positive	$\lvert\hat{\beta_1}\rvert$ is a bigger underestimate of $\lvert\beta_1\rvert$
$PI\beta_2 = (1-\rho_1)/(1-\rho_{122}) * [1-(\rho_{12}*\beta_1/\beta_2)]$						
	$\lvert\rho_{12}*\beta_1\rvert < \lvert\beta_2\rvert$ ($\lvert\beta_2\rvert$ is relatively large)	$\lvert\rho_{12} * \beta_1/\beta_2\rvert$ always < 1				
			0 or 2 Minus signs	$(\rho_{12} * \beta_1/\beta_2)$ subtracts from one	$PI\beta_2$ is always positive	$\lvert\hat{\beta_2}\rvert$ is an underestimate of $\lvert\beta_2\rvert$
			1 or 3 Minus signs	$(\rho_{12} * \beta_1/\beta_2)$ adds to one	$PI\beta_2$ is more positive	$\lvert\hat{\beta_2}\rvert$ is a bigger underestimate of $\lvert\beta_2\rvert$
	$\lvert\rho_{12} * \beta_1\rvert > \lvert\beta_2\rvert$ (β_2 is relatively small)	$\lvert\rho_{12}*\beta_1/\beta_2\rvert$ always > 1				
			0 or 2 minus signs	$(\rho_{12} * \beta_1/\beta_2)$ subtracts from one	$PI\beta_2$ is always negative	$\lvert\hat{\beta_2}\rvert$ is an overestimate of $\lvert\beta_2\rvert$
			1 or 3 Minus signs	$(\rho_{12} * \beta_1/\beta_2)$ adds to one	$PI\beta_2$ is positive	$\lvert\hat{\beta_2}\rvert$ is an under-estimate of $\lvert\beta_2\rvert$

Appendix D

Correcting for M+ME Biases — Step by Step

When might an analysis be in the M+ME danger zone? For the sake of illustration, consider the model and correlations shown in Figure D1. Here we have nine constructs connected by hypothesized paths, showing selected (overt) path estimates and selected (overt) inter-construct correlations. How would a researcher recognize that any particular pair of these constructs is near enough to the M+ME danger zone to possibly require the correction? Note that Figures 3 through 7 and Figure 9 in the paper show underlying correlations. If a researcher compares their own results to Figures 3 through 7, they need to use equation 2 to convert overt correlations to underlying correlations. Note also that the M+ME Correction Application requires "overt" correlations, not underlying correlations.

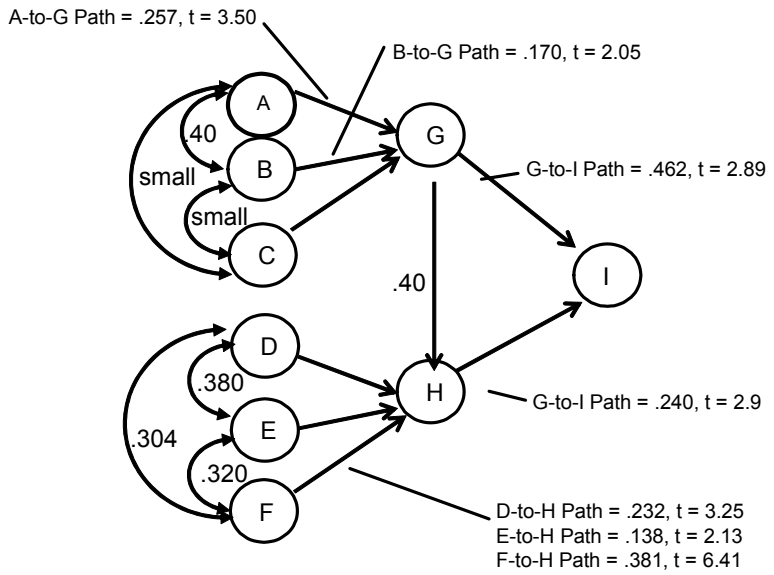


Figure D1. Overt Correlations in Hypothetical Model with N = 200, Reliability = .80

Step One: Identify Correlations of Interest. First, recognize that in Figure D1, only the seven overt correlations or paths with a value (or the word "small") added to the link are of concern to us. Note that the path between G and H is of concern even though it is modeled as a path rather than a correlation, because G and H are correlated, and both participate in the regression equation predicting construct I. Note also that even if the overt correlation between construct C and construct D were very high, say.72, that would not be of concern, because constructs C and D do not participate in the same regression equation.

Categorize all high correlation situations into two groups. Notice that the upper part of the figure shows two of what we will call an "isolated" high correlation—only two highly correlated constructs participating in the same regression. The bottom part of Figure D1 shows a different situation—three highly correlated constructs all participating in the same regression, specifically D, E, and F are correlated and all predict H. This latter situation we will call a "combination" of high correlations. These "combination" situations present more of a challenge than the isolated high correlation and will be dealt with in Appendix E.

Step Two: Assess and perhaps correct the "isolated" correlations of interest. There are two isolated correlations in Figure D1 that should be examined: A and B predicting G, and G and H predicting I (both having an overt correlation of .40).

To rule out obviously non-problematic correlated pairs, one can do a quick (very approximate) back of the envelope calculation to determine if there is even any cause for concern, as follows. For example, if the reliability of constructs A and B in Figure D1 is .80, using equation 2, we conclude that the underlying correlations are about .50 for A with B (and for G with H). To get a general feeling for whether A and B

predicting G might involve a substantive M+ME bias, we can refer to Figure 7 (rather than Figures 4A or 4B, because the N is 200). If the sample size were closer to n = 100, then Figures 4A or 4B might be appropriate.

In Figure 7 we see that an analysis with a β_1 about .292, sample size of N = 200, and an underlying correlation of .50 gives us about a 7% likelihood of excessive false positives, near the top of the 95% confidence interval around 5% false positives. This suggests that we should use the M+ME Correction Application to check the possibility of an M+ME bias.

The M+ME correction application will correct for the bias to the VIF, correct the path biases, and adjust the standard deviations for the corrected path values. This downloadable application is available on the MISQ web site (Online Supplements: M+ME Correction Application). Figure 8 of the article proper displays the front end of the application, showing the input required (on the left side) and the results of the correction calculations on the right side. First, be sure that you are using the standardized regression (or PLS) results. Though here we will assume that β_1 is the dominant path (the construct whose path has the highest absolute value), that is not necessary. With this understanding, enter the regression results for the two (overt) path estimates from regression or PLS, the two overt t-statistics, the two reliabilities (for X1 and X2) and the overt (or apparent) correlation between X1 and X2. The corrected path values, t-statistics and p values will be displayed by the M+ME correction application. Because it is relatively easy to use, the M+ME correction application can be used without referring to the figures in the paper as a "back-of-the-envelope" approximation.

Table D1 shows the results of applying the corrections. The first set of rows in Table D1 shows the A and B predicting G situation when it is entered into the M+ME correction application. The input values are to the left, and the corrected path values and t-statistics are shown to the right. In this case the path correction has increased the B → G path slightly[6] (to .184) and the recalculated standard deviation has decreased the t-statistic a good bit (to 1.561). The result is that the corrected B to G path is no longer statistically significant.

The second "isolated correlation" from Figure D1 is G and H predicting I, with G and H correlated at .40. This situation is depicted in the second set of rows of Table D1. There the H to I path is shown to be statistically significant even with the M+ME correction, though note that the t-statistic has dropped from 2.90 to 2.02. Finally, the third example in Table D1 has the same input as the second, but a correlation of .60 instead of .40. This increase in multicollinearity takes its toll, and under these circumstances the H to I path is no longer statistically significant after we apply the M+ME correction.

Note that in all three examples, the uncorrected results show that the questionable path is statistically significant (i.e., different from zero with 95% confidence). In two of those, the correction shows that the questionable path is not actually significant, and should not be considered different from zero.

We suggest that there is no place for optimists in questions of statistical significance. When M+ME for a particular path is not clearly ruled out by displays such as those in Figures 4A, 4B, or 7, the correction should be applied by inputting the relevant data into the M+ME correction application. Applying this correction to regression or PLS results when M+ME is not a problem will not create new problems. Instead, it will give the researcher a more accurate value for the t-statistic.

[6]As described in Appendix C and Table C1, this is an example where the true β_2 path is close enough to the true β_1 path that the M+ME bias will decrease both values, rather than decreasing the β_1 path and increasing the β_2 path. In this case the correction will result in an increase from the regression value to the true value for both paths. When the β_2 path is much smaller than the β_1 path, the correction will increase the β_1 path and decrease the β_2 path from what is seen in the regression results.

Table D1. Input and Output (Corrected Path Values and t-statistics) for the M+ME Correction Application

	Needed Input								Output				
Relation-ship	N	$\beta_1^{\wedge}$	$\beta_2^{\wedge}$	$\beta_1^{\wedge}t$	$\beta_2^{\wedge}t$	α_1	α_2	ρ_{12}	Path	Significance Without Corrections	Corrected Path Est	Corrected t-stat	Significance with path and VIF correction
A, B → G; $\rho_{a,b}$ = .40	200	.257	.170	3.50	2.05	.80	.80	.40	B→G	yes	0.184	1.56	**Not Sig!**
									A→G	yes	0.314	3.00	yes
G,H → I; $\rho_{g,h}$ = .40	200	.462	.240	2.89	2.90	.80	.80	.40	H→I	yes	0.243	2.02	yes
									G→I	yes	0.576	2.55	yes
G,H → I; $\rho_{g,h}$ = .60	200	.462	.240	2.89	2.90	.80	.80	.60	H→I	yes	0.179	1.01	**Not Sig!**
									G→I	yes	0.623	2.03	yes

Appendix E

The Challenge of Combinations of High Correlations

In the lower part of Figure D1 we see a "combination" of three constructs (D, E, and F) connected with one high and one only moderately high overt correlation, both of which affect the E to H path (that is .380 and .320 overt correlations, suggesting underlying correlations of .475 and .400). The third overt correlation (between D and F) is .304, suggesting an underlying value of .380. We acknowledge at the outset that we do not fully understand all the issues relating to multiple high correlations situations. The problem is that Green and Kiernan's (1989) equations do not extend to three intercorrelated constructs. In fact both pairs of correlated constructs have an impact on the estimated E to H path estimate and their impact could be additive or in some cases more than additive. This means that except for turning to CB-SEM, we do not have an effective way to correct for the path biases in regression or PLS, when faced with such a situation. This is an area where additional research would be quite valuable.

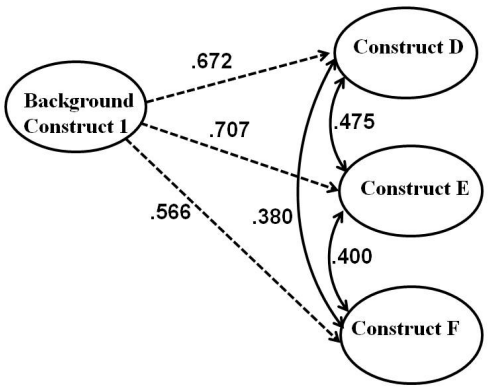
A key insight in understanding combinations of high correlations is that there are three general archetypes of the underlying "causes" of these three way configurations of correlations, as shown in Figure E1. Of course the cause may contain a mixture of several of these types, plus the possibility of direct causal links between the focal constructs (in this case Constructs D, E or F.

On the top left of Figure E1 (Panel A, single underlying cause) we see that the correlations between all three constructs are due to relationships with a single underlying construct. As it turns out, there is a reasonable correction approach when the combination of high correlations is produced by a single underlying construct. In those situations the M+ME biases can be considered additive, and it is possible to correct for the largest correlation as we have shown in the previous section for isolated high correlations. Once this is done, the researcher can determine (very roughly) from the size of the remaining correlation (for example using Figures 3, 4, or 7), whether that remaining correlation by itself would put the analysis into the M+ME danger zone. If not, then the single correlation correction method is sufficient. One definitely should not apply the M+ME correction application a second time.

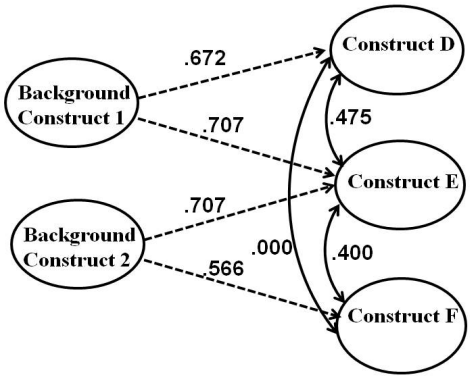
Unfortunately we know of no way to determine, from the data a researcher would have, whether a combination of correlations was caused by two, or even three, background constructs. Thus we have no way to safely assume a single underlying cause. This poses a difficult problem.

In Figure E2 we show the uncorrected results from data generated by each of the archetypal possible causes, along with the results after the highest correlation has been corrected. The figure does show that correcting for the dominant correlation always improves the situation. But unfortunately, even with the correction, some of the lines are well above the 95% confidence interval around .05.

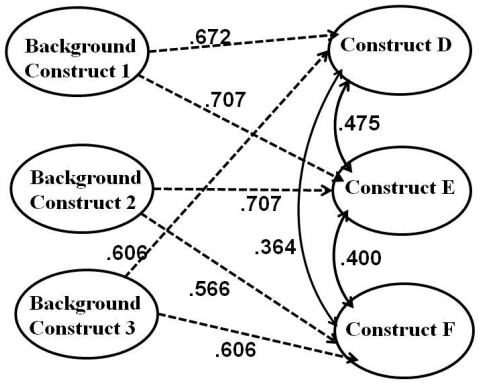
Since the researcher cannot know which situation they are working with, and since it is not appropriate to optimistically "assume the best possible situation" in hypothesis testing, we cannot recommend using the M+ME correction with combinations of high correlations.



Panel A. Single Underlying Cause

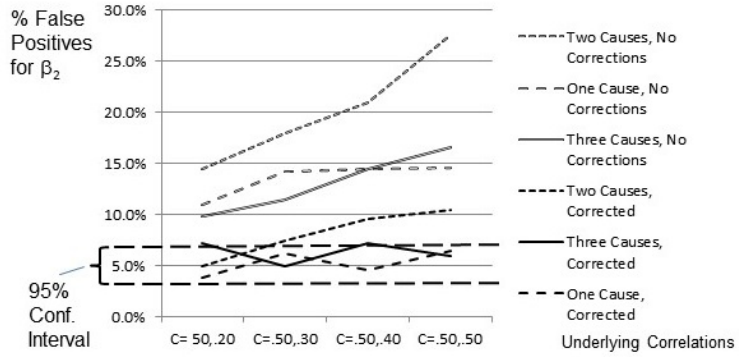


Panel B. Two Distinct Underlying Causes



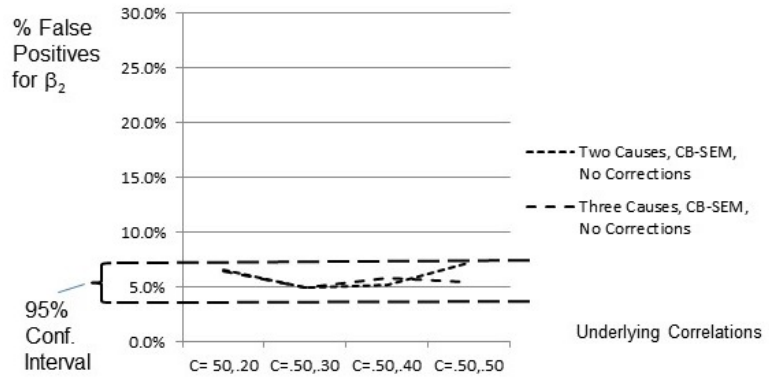
Panel C. Three Distinct Underlying Causes

Figure E1. Three Different Archetypal Causes of Combinations of High Correlations



No Correction Versus Dominant Correlation Corrected, α = .80, N = 200, X1 → Y1 = .420

Figure E2. Regression False Positives Results from Combinations of High Correlations



α = .80, N = 200, X1 → Y1 = .420

Figure E3. CB-SEM False Positives Results From Combinations of High Correlations

One safe method to use in combinations of high correlation situations would be to convert the analysis over to CB-SEM. Figure E3 shows the results from that method. We see that with CB-SEM, the large numbers of false positives never appear in the first place.[7] Though it may require extra work for those not well-versed in the use of CB-SEM, this clearly provides a solution to the combination of high correlation M+ME situation. Alternatively (or in conjunction), it may be appropriate to change the underlying model (e.g., using higher-order constructs).

As stated earlier, we do not fully understand the combinations of high correlation situation. Because combinations of high correlations do appear in IS research, additional research in this area could be helpful.

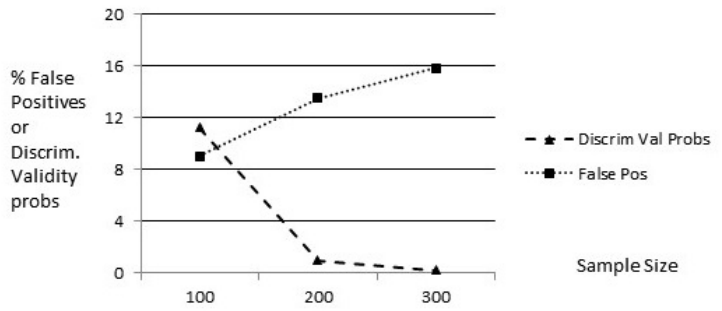
Appendix F

Testing Whether M+ME False Positives Could Be Due to Discriminant Validity Problems

One alternative explanation for the results we obtained might be framed within the larger context of measurement model misspecification, and specifically the presence of a lack of discriminant validity among constructs. To address this possibility, we conducted some ancillary analyses. We found the following. First, using chi square difference tests of discriminant validity and one of our Monte Carlo simulation datasets (500 samples each of N = 100, reliability = .80, ρ_{12} = .80) we did find evidence of discriminant validity problems. Specifically, 56 (or about 10%) of the samples had discriminant validity problems, as compared to 59 (also about 10%) samples with M+ME false positives. Although these numbers are quite close, only 4 samples had both discriminant validity problems and excessive false positives.

We note further that although both false positives and discriminant validity problems are exacerbated by increasing correlations and decreasing reliabilities, those two phenomena react quite differently to increasing sample size. As shown in Figure F1, larger sample sizes decrease discriminant validity problems to virtually zero, while they increase M+ME false positives substantially. Thus it is clear that the two phenomena share some causal factors but are actually quite distinct. Knowing that a dataset suffers from one of these phenomena does not provide much knowledge about the likelihood that it also suffers from the other.

[7]The reader may be concerned that in the .50/.50 combination correlation situation in Figure E3, the percent of false positives for CB-SEM (7.2%) is above the 6.9% limit for the confidence interval. We were concerned as well, but recognized that since we have used so many tests against that limit, it is highly likely that some values will be a little higher than the stated limit for an individual test. To give us more confidence in that explanation, we generated an additional 100,000 data points (500 new datasets of 200 cases each), and reran the analysis. In that run we found that CB-SEM resulted in 5.4% false positives. With 1,000 datasets (the combined total) we now have an average of 6.3% false positives, within the confidence interval of 3.65% to 6.35% for a sample size of 1,000.



(β_1 path = .292, N = 100, α = .80, ρ_{12} = .80)

Figure F1. Sample Size and Discriminant Validity Problems Versus M+ME False Positives

Appendix G

Pictorial Depiction: M+ME Impact on Regression and CB-SEM

Some readers may find it helpful to see a graphical representation of the impact of M+ME on regression and CB-SEM results for 500 samples as in our Monte Carlo simulations. Throughout we will be looking at an underlying X1 to Y1 path of .600, an X2 to Y1 path of zero, a sample size of N = 200, and reliability of .80 for both X1 and X2. For both regression and CB-SEM, the average path estimates across the 500 samples is shown below in Figures G1, G2, and G3, for three different scenarios. Along with the average path estimates, also shown is a representation of the distribution of those 500 estimates around that average estimate. The distribution curves shown comes from calculating the standard deviation of the 500 path estimates, and displaying them as a curve anchored at the average path estimate plus 2 times the standard deviation, and the average path estimate minus 2 times the standard deviation.

We first look at the results when the underlying reality is a zero correlation between X1 and X2. That is, when there is no M+ME. We will then shift our focus to the situation where M+ME is extreme, with a correlation between X1 and X2 of .90.

Figure G1 shows the results for β_2 (a zero path) when there is no correlation between X1 and X2. The average β_2 path estimates for regression and CB-SEM are both about zero, and the distribution curves for both span from about -1.3 to + 1.5. These results are what we might have expected.

In Figure G2 we show the results when the correlation between X1 and X2 is .90, a very high correlation. Though M+ME will rarely or never be this extreme in practice, this scenario allows us to see more clearly what the specific effects of M+ME are, and why this is a problem. Under these conditions, CB-SEM returns a path estimate for β_2 of near zero, and a standard deviation that is about 7 times as large as it was when ρ_{12} was .00 (.497 versus .068). The very large spread of the β_2 path estimations suggests that the high correlation between X1 and X2 makes the resulting path estimates very dependent upon random variance in the data. CB-SEM recognizes this and increases the standard deviation it uses appropriately.

For regression under these conditions, the average path estimate is quite skewed (from near zero at ρ_{12} = 0 to a value of .175 at ρ_{12} = .90. This is as predicted by the Green and Kiernan equations. The distribution around the .175 based on the standard deviation of the regression estimates is about as wide as in Figure G1, but now shifted up, so that most of the estimates are now significantly different from zero. This is of course misleading, since the true path is zero. This makes the M+ME bias very apparent.

In Figure G3, we have used the Green and Kiernan equations to correct for the M+ME path bias in regression, and at the same time corrected the standard deviation values in regression, also based on the VIF and Green and Kiernan equations. These corrections have removed the path bias, and now the bias and the dispersion of the path values are roughly equal for regression and for CB-SEM, and both reflect the uncertainly created by the large correlation between X1 and X2.

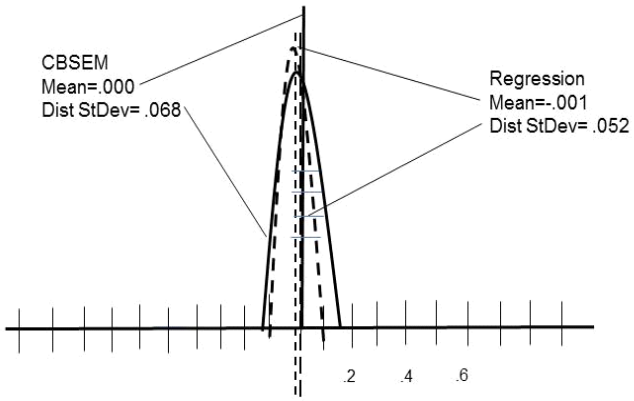


Figure G1. Regression and CB-SEM When Correlation Between X1 and X2 is 0.00

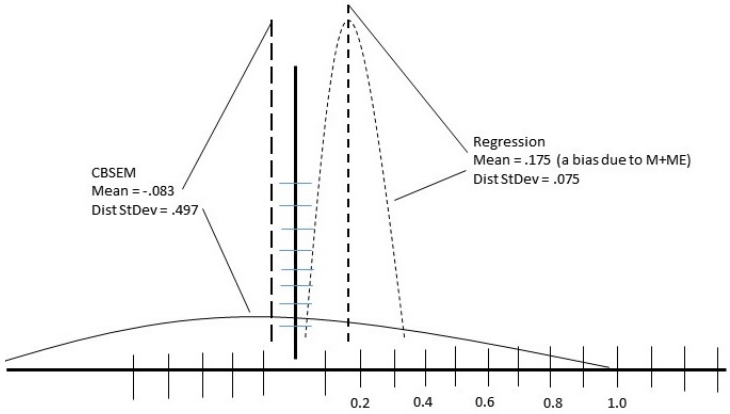


Figure G2. Regression and CB-SEM When Correlation Between X1 and X2 is 0.90

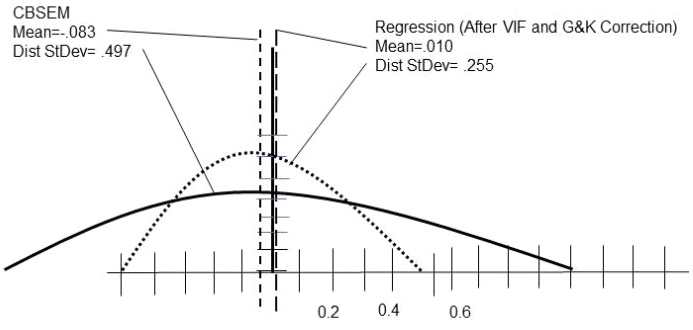


Figure G3. Regression and CB-SEM When Correlation Between X1 and X2 is 0.90, After VIF and Green and Kiernan Path Bias Corrections

References

Goodhue, D., Lewis, W., and Thompson, R. 2011. "A Dangerous Blind Spot in IS Research: False Positives Due to Multicollinearity Combined with Measurement Error," in *Proceedings of the 17th Americas Conference on Information Systems*, Detroit, MI, August 4-7.

Green, C. J., and Kiernan, E. 1989. "Multicollinearity and Measurement Error in Econometric Financial Modelling," *The Manchester School* (57:4), pp. 357-369.

Johnston, J. 1972. *Econometric Methods* (2nd ed.), New York: McGraw-Hill.

Mela, C., and Kopalle, P. 2002. "The Impact of Collinearity on Regression Analysis: the Asymmetric Effect of Negative and Positive Correlations," *Applied Economics* (34:6), pp. 667-677.

Mooney, C. Z., and Duval, R. D. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference,* Beverly Hills, CA: Sage Publications.