Statistical significance and effect size tests in SEM: Common method bias and strong theorizing

Ned Kock
Texas A&M International University

Kevin Dow
University of Texas at El Paso

Full reference:
Kock, N., & Dow, K. E. (2025). Statistical significance and effect size tests in SEM: Common method bias and strong theorizing. *Advances in Management Accounting*, 37(1), 95–105.

Abstract

We generally acknowledge the problematic nature of classic statistical significance tests based on P-values or confidence intervals. In fact, we demonstrate based on an illustrative model for which we created simulated data, that with low and high statistical power, path coefficients in structural equation modeling whose true values are zero, routinely end up being found to be significantly different from zero at the $P < .05$ level. However, we argue that we should not do away with classic statistical significance tests, and that these tests can be useful but should be complemented by other methodological tools, including effect size tests, and tests of common method bias. We also argue that high quality theorizing is very important if we are to profitably use a combination of classic statistical significance, effect size, and common method bias tests.

KEYWORDS: Management Accounting Research, Statistical Significance Tests, *P*-values, Confidence Intervals, Structural Equation Modeling, Sampling Error

Introduction

We are thankful for the opportunity to write this response to the comprehensive and insightful article by Professor R. Murray Lindsay, which is aimed primarily at management accounting researchers. We write from the perspective of structural equation modeling (SEM) because most statistical significance tests can be conceptually seen as special cases of SEM (Kock, 2019a). Broadly speaking, SEM is a general multivariate data analysis method normally used for analyses of cross-sectional data obtained via questionnaires, but which can be used with other types of data, and which enables researchers to test structural and measurement models simultaneously (Kock, 2019a; 2023b).

We generally agree with Professor Lindsay's views about the problematic nature of statistical significance tests based on *P*-values or confidence intervals. We demonstrate based on an illustrative SEM model for which we created simulated data, that with both low and high statistical power, paths whose true values are zero routinely end up being found to be significantly different from zero at the $P < .05$ level. *P*-values have historically been used since as the generally preferred statistical method to summarize the results of a study, but have often been misused, misinterpreted, and misunderstood. This problematic state of affairs drives the need for complementary methods.

In SEM, the structural model usually involves a set of variables that cannot be measured directly without error, known as latent variables (LVs); as well as causal relationships among these LVs, which are usually represented through arrows. SEM offers insights that general linear models cannot (Dow et al, 2021; Dow et al, 2012, Teklay et al, 2023). For example, as illustrated in a recent management accounting study, insights into the complex interrelationships among LVs were revealed that were obscured using general linear models (Teklay et al, 2023). More often than not the structural model is aimed at representing a theory to be tested based on empirical data, where each LV-LV link is associated with one hypothesis to be tested. The measurement model is made up of variables that measure the LVs with error, known as indicators, typically as responses to question-statements on Likert-type scales in questionnaires.

For the SEM analyses presented in this paper, we used the factor-based algorithms implemented through the SEM software WarpPLS (Kock, 2023a), which also implements composite-based algorithms (a.k.a., classic PLS algorithms, not used here), as well as the factor-based algorithms implemented by the R package lavaan (Rosseel, 2012). This was done to double-check the results of our analyses. Some of the coefficients reported in this paper could only be generated from correlation-preserving LV estimates, which are provided by WarpPLS but not by lavaan. Other than that, the factor-based algorithms implemented in WarpPLS yielded virtually the same results as the covariance-based full information maximum likelihood algorithms implemented in lavaan, mirroring results from past research (Kock, 2019a; 2019b).

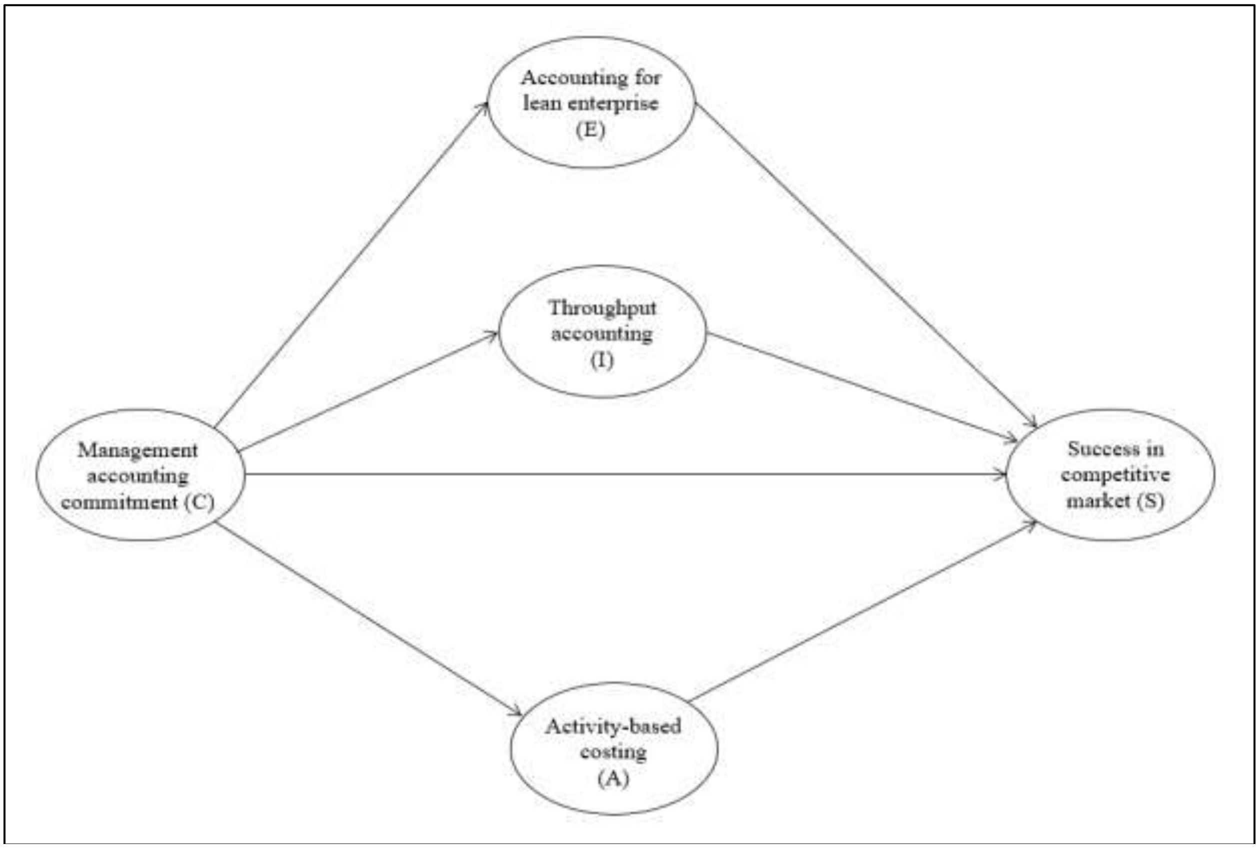
We argue that we should not do away with classic statistical significance tests, primarily because they are not entirely useless, and their usefulness increases when they are employed in combination with other tests. For instance, they allow us to correctly identify as likely to be zero at the population level, in SEM analyses, many path coefficients whose true values are zero. We also argue that classic statistical significance tests need to be complemented by other methodological tools, including effect size tests, and common method bias tests.

The latter, common method bias tests, are needed because of the strong distorting effect that common method variation may have on path coefficients in SEM, which is comparable in terms of its magnitude to the distorting effect of sampling error with very small sample sizes. This happens even when the amount of common method variation is relatively small, as we will see later. Finally, we argue and demonstrate that high-quality theorizing is very important if we are to profitably use a combination of classic statistical significance, effect size, and common method bias tests.

Illustrative SEM model

Our discussion is based on the illustrative SEM model shown in Figure 1, which is used to help accomplish the goal of making our discussion meaningful to those in the field of management accounting and related fields. The model contains five LVs, associated with the following constructs: management accounting commitment (C), accounting for lean enterprise (E), throughput accounting (I), activity-based costing (A), and success in a competitive market (S). These constructs are assumed to have been measured at the company level, and to have been measured with error through Likert-type question-statements in a questionnaire, each through five indicators.

Figure 1. Illustrative SEM model



The general idea underlying this model is that a company's management accounting commitment positively influences the company's success in competitive markets, both directly and indirectly through management accounting commitment's facilitation of the increased use of the techniques of accounting for lean enterprise, throughput accounting, and activity-based costing. This is a simplified model that is not meant to serve as the basis for future theoretical or empirical research. It is nevertheless a helpful model, as it aids us in our task of conducting a discussion that is not entirely conceptual or mathematical. From a methodological perspective, this model is also helpful because we can create simulated data assuming different population values of certain parameters, such as path coefficients.

The population values for loadings that we used were kept fixed across simulated datasets, and were as follows: C (.700, .700, .700, .700, and .700), E (.800, .800, .800, .800, and .800), I (.600, .600, .700, .700, and .800), A (.500, .600, .600, .700, and .700) and S (.550, .600, .650, .700, and .750). These loadings, while adding variety to our simulations (and thus generality to our findings), also ensured that all of the simulated datasets we created would pass common measurement model quality criteria used in SEM; e.g., for convergent and discriminant validity (Kock, 2014; Kock & Lynn, 2012).

Statistical significance tests in SEM

Table 1 illustrates the problems with classic statistical significance tests in SEM. It shows path coefficient estimates when true paths (path coefficients at the population level) are zero and N=100; a situation where we have low statistical power. It also shows path coefficient estimates when true paths are zero and N=100,000; a situation where we have high statistical power. Ironically, in both cases, with low and high statistical power, three paths whose true values are zero end up being found to be significantly different from zero at the $P < .05$ level. This problematic situation is not due to the use of *P*-values for significance testing, we would have similar results if we had used confidence interval tests (Kock, 2016).

Table 1. Estimated paths when true values are zero

Estimated paths (N=100)	C>E (.140), C>I (.102), C>A (.097), **C>S (.172)**, E>S (.064), **I>S (-.168)**, **A>S (-.204)**
Estimated paths (N=100,000)	C>E (.003), **C>I (.006)**, C>A (.000), C>S (-.004), E>S (.000), **I>S (.005)**, **A>S (-.006)**

Note: True paths = 0 that were found to be significant at $P < .05$ are bolded and underlined.

The reason for the problematic results in the low statistical power scenario (i.e., small sample size), is that with small samples we end up having more distortion of coefficients, because larger amounts of sampling error are present, leading to larger standard errors and thus more variability in the path coefficients across samples taken from the population (our results are based on the analysis of one such sample with N=100). For example, the path coefficient for A>S with N=100, whose true value was set as zero in our simulation, ended up being a rather strong and negative coefficient of -.204, entirely due to random sampling error.

The reason for the problematic results in the high statistical power scenario (i.e., large sample size), is that with large samples we end up having smaller amounts of sampling error, leading to smaller standard errors; and *P*-values tend to go down as standard errors decrease. *P*-values are very sensitive to sample size. The result is that, with large samples even tiny path coefficients, resulting from random sampling error, may become statistically significant.

Does the above mean that we should do away with classic statistical significance tests? It may seem odd for us to take this position, but we believe that the answer to this question is “no”, primarily because classic statistical significance tests (with $P < .05$) are not entirely useless, and arguably their usefulness increases when they are employed in combination with other types of tests. For instance, they allowed us to correctly identify 4 out of 7 path coefficients as likely to be zero at the population level. That is, classic statistical significance tests helped somewhat, but we argue that they need to be complemented by other methodological tools. One such tool is the use of effect size tests.

Effect size tests in SEM

Two main measures of effect size are commonly used in SEM. The most widely used is Cohen’s *f*-squared coefficient (Cohen, 1988; 1992), which is calculated as the error-adjusted incremental contribution of a predictor LV to the *R*-squared of the criterion LV to which it points. The other measure of effect size commonly used in SEM is the absolute contribution of the predictor LV to the *R*-squared of the criterion LV (Dow et al., 2008, Kock, 2014; Mandal et al., 2012). This second measure, calculated as the absolute value of the product between the path coefficient and the correlation among the two LVs (i.e., the predictor and criterion LVs), tends to yield lower results that take the influences of all predictor LVs into account simultaneously, thus being a more conservative and credible effect size estimate (Kock, 2014), and is therefore the one we use here.

By convention, effect sizes of .02, .15, and .35 are respectively termed small, medium, and large (Cohen, 1992; Kock, 2014). Because of this, one can use the threshold of .02 for effect sizes in SEM, rejecting path coefficients whose effect sizes are lower than .02; because they are *too small*, being below the threshold for small. This effect size test would assume that path coefficients associated with effect sizes lower than .02 are likely to be zero at the population level. Table 2 shows path coefficient estimates for our illustrative SEM model, as well as corresponding *P*-values and effect sizes, when true paths (path coefficients at the population level) are zero and N=300; a sample size level often used in empirical research employing SEM.

Table 2. Estimates when true paths are zero and N=300

True paths	C>E (.000), C>I (.000), C>A (.000), C>S (.000), E>S (.000), I>S (.000), A>S (.000)
Estimated paths	C>E (.048), C>I (.130), C>A (.209), C>S (.025), E>S (-.125), I>S (-.051), A>S (.024)
***P*-values**	C>E (.199), **C>I (.011)**, **C>A (.000)**, C>S (.332), **E>S (.014)**, I>S (.187), A>S (.338)
Effect sizes	C>E (.002), C>I (.017), **C>A (.044)**, C>S (.000), E>S (.017), I>S (.003), A>S (.000)

Note: True paths = 0 that were found to be significant at $P < .05$ or have effect size > .02 are bolded and underlined.

As we can see, three paths whose true values are zero end up being found to be significantly different from zero at the $P < .05$ level with N=300. As noted earlier, we would have similar results if we had used confidence interval tests. The paths and corresponding *P*-values are: C>I (.011), C>A (.000), and E>S (.014). The *P*-value shown as .000 is displayed as such because it is lower than .000, not because it is exactly zero. Only one path was found to have an effect size > .02. This path and its corresponding effect size are C>A (.044).

So, it seems that using effect sizes, and the effect size threshold of .02, allowed us to correctly reject, in several instances, the existence of effects that were actually zero (or nonexistent) at the population level. In fact, we would have been able to correctly reject 6 out of 7 paths, whereas the $P < .05$ level test allowed us to reject only 4 out of 7 paths. It should be clear to readers, based on the discussion in the preceding section, that effect size tests become more useful in terms of avoiding false positives as sample sizes increase.

However, the same effect size tests will tend to incorrectly reject paths associated with small but nonzero effects (i.e., commit false negatives) with large sample sizes. In large datasets sampling error will be minimized, thus ensuring that *all* weak effects are correctly estimated as weak, in turn leading to an even larger proportion of false negatives than if small samples are used. This bizarre situation would essentially mean that, with effect size tests applied to models with weak but nonzero effects, larger sample sizes would be associated with *lower* statistical power – i.e., a larger proportion of false negatives, as sample sizes increase.

Incorrect rejection of paths associated with small but nonzero effects could easily occur in studies of the impact of management accounting policies on the successful avoidance of rare adverse events, such as policies aimed at preventing factory explosions. In these cases, where path coefficients are expected to be weak but nonzero (due to the low amount of variation in the main dependent variable), it would be advisable to employ a logistic regression transformation of the main dependent variable, which could originally have stored the values 1 and 0 to respectively reflect the occurrence or not of an explosion, and then follow that dichotomous-to-probability variable conversion with a combination of classic statistical significance and effect size tests (Kock, 2023c). Based on this, it should be clear to the reader that our recommendation cannot simply be something like “use effect size tests and abandon classic statistical significance tests”.

As we can see from our results, at N=300 the effect size test incorrectly identified the path C>A as associated with a nonzero effect at the population level. This type of problem is likely to be more frequent at small sample sizes, which again calls for the combined use of effect size and classic statistical significance tests in these cases. Still, as we will show in the remainder of this paper, these must be combined with at least two other types of methodological tools, namely common method bias tests and strong theorizing. The latter, strong theorizing, is not always viewed as a methodological tool, but as we will see, it should be because it strongly influences the effectiveness of common method bias and other types of tests.

Common method bias in SEM

Common method bias is a phenomenon that is caused by the measurement method used in an SEM study, and not by the structural model. That is, common method bias is caused by sources that influence the LV-indicator associations, and not the LV-LV associations. For example, the instructions at the top of a questionnaire may bias the answers provided by different respondents in the same general direction, leading the indicators to share a certain amount of common variation. In the simplified case where common method variation is assumed to be uniform across all indicators, the amount of common variation is expressed by the corresponding method weight coefficient.

Two widely used tests for identification of common method bias in SEM analyses are Harman's single factor test (Kock, 2021a), which is based on a total variance explained (TVE) measure, and the more sensitive full collinearity variance inflation factors (VIFs) test (Kock, 2015; Kock & Lynn, 2012). Harman's single factor test is the most widely used of the two tests, even though it has been shown to have serious deficiencies due to low sensitivity in models of nontrivial complexity (Baumgartner et al., 2021; Kock, 2021a; Podsakoff et al., 2003).

Harman's single factor test usually entails assigning all indicators of all LVs in a SEM model to a single LV, for which the TVE is calculated and compared against the .5 threshold (Kock, 2021a). If the TVE is lower than .5, then the model is assumed to be free of common method bias. The full collinearity VIFs test entails calculating VIFs for all of the LVs in a model, and comparing each of them against the threshold of 10. This threshold is recommended in factor-based SEM analyses; in composite-based analyses, the lower threshold of 3.3 is normally used (Kock, 2015; Kock & Lynn, 2012). Should all of the full collinearity VIFs be equal to or lower than the threshold, the conclusion is that the model is free of common method bias.

In this section, we summarize the average results of simulations where data were generated for 100,000 model instances, and the sample size for each model was set at 100,000. These simulation settings allowed us to virtually eliminate the effect of sampling error. In other words, these settings allowed us to understand common method bias in the absence of sampling error. This is important because, as we have seen earlier in this paper, sampling error alone can significantly distort path coefficients in SEM analyses.

Table 3 shows the performance of Harman's single factor and full collinearity VIFs tests of common method bias under weak theory conditions, characterized by all true values of path coefficients being zero. We refer to this scenario as one associated with weak theory because all hypothesized effects are nonexistent, which characterizes poor quality theorizing – i.e., none of the hypothesized effects existed in reality. Various coefficients were estimated when we used a method weight that led at least one true path = 0 to be incorrectly estimated as > .197. In this situation, we have common method bias that is large enough to lead to problems in terms of classic hypothesis-testing based on path coefficients. The .197 value is used here because it has been proposed as a critical threshold, being the minimum absolute path coefficient that would be statistically significant with a sample size of 160 in an analysis with statistical significance set at .05 and statistical power of .8 (Kock & Hadaya, 2018; Kock et al., 2017).

Table 3. Common method bias tests with weak theory

Method weight	.344
True paths	C>E (.000), C>I (.000), C>A (.000), C>S (.000), E>S (.000), I>S (.000), A>S (.000)
Estimated paths	C>E (.177), **C>I (.211)**, **C>A (.242)**, C>S (.135), E>S (.122), I>S (.145), A>S (.151)
Harman's TVE	.191
Full collinearity VIFs	C (1.126), E (1.105), I (1.134), A (1.156), S (1.146)

Note: True paths = 0 that were estimated as > .197 are bolded and underlined.

As we can see, two paths were sharply distorted by common method bias alone (again, this has nothing to do with sampling error), and in ways that would not be truly identified by classic statistical significance or effect size tests, regardless of statistical power. And, this happened with a relatively small amount of common variation contamination. Here the method weight was only .344, the smallest needed for at least one true path = 0 to be incorrectly estimated as > .197. This method weight of .344 is much lower than the method weight of .6 frequently used in common method bias discussions (Kock, 2015; 2021a).

Both the Harman’s single factor and the full collinearity VIFs tests failed to recognize the existence of common method bias, in a model where the bias significantly distorted path coefficients. Harman’s TVE was calculated at .191, much lower than .5; and no full collinearity VIF was even close to 10. That is, under weak theory conditions, both common method bias tests performed quite poorly. Generally speaking, poor theory development is a source of a variety of problems; including, as we can see, methodological problems. This tends to happen in a stealthy way, and highlights the reality that there is no substitute for high quality theorizing in empirical research.

The importance of strong theorizing in SEM

In this section, we consider a scenario in which all true path coefficients (except for C > S) are set as .489, which would characterize situations where the theory was well developed (strong theory), in the presence of common method bias. The true path for the C > S link is set as .000 for methodological reasons. The path for the C > S link being .000 means that the mediation in the model is assumed to be full. In other words, the model provides a fairly complete view of the mediated overall effect of C on S, which is also an indication of strong (or high quality) theorizing.

The .489 value is a proxy for the midpoint between the path coefficients corresponding to the large (.35) and medium (.15) effect sizes proposed by Cohen (1988; 1992), calculated as: $\left(\sqrt{.35} + \sqrt{.15}\right)/2 = .489$. Table 4 shows various coefficients estimated when the method weight = .344. That is, in this scenario, we have common method contamination, as well as true paths that are strong enough for us to assume that the theory underlying the model was well developed.

Table 4. Common method bias tests with strong theory

Method weight	.344
True paths	C>E (.489), C>I (.489), C>A (.489), C>S (.000), E>S (.489), I>S (.489), A>S (.489)
Estimated paths	C>E (.583), C>I (.610), C>A (.599), C>S (.011), E>S (.410), I>S (.413), A>S (.399)
Harman's TVE	.361
Full collinearity VIFs	C (2.456), E (3.547), I (3.673), A (3.507), **<u>S (12.038)</u>**

Note: Instance of correct identification of existing common method bias is bolded and underlined.

Note that the estimated paths were higher than the true paths, which is due to the common method contamination. Harman's single factor test failed to recognize the existence of common method bias since the TVE was .361 and thus lower than the .5 threshold. Harman's single factor test would probably have succeeded if there was more common method variation present; i.e., if the method weight was much higher than .344. For example, Kock (2021a) showed that with strong theorizing Harman's single factor test works well with method weights of .6 and higher.

On the other hand, the test based on full collinearity VIFs, with a threshold of 10, succeeded in identifying the occurrence of common method bias. The highest full collinearity VIF was 12.038 and thus higher than the threshold of 10. Past research suggests that the success of the full collinearity VIFs test here was to be expected, as this test has been presented as more sensitive to common method bias than Harman's single factor test (Kock, 2015; 2023b). The correct identification of the presence of common method bias would call for some form of common variation removal (Kock, 2021b) to be employed before one can trust hypothesis assessment results based on classic statistical significance and effect size tests.

Conclusion

In this paper we provided a discussion of the problematic nature of statistical significance tests based on P-values or confidence intervals. Among other things, we demonstrated a misleading pattern in SEM analyses based on an illustrative model for which we created simulated data. The misleading pattern is that with both low and high statistical power, path coefficients whose true values are zero routinely end up being found to be significantly different from zero at the $P < .05$ level.

Nevertheless, we argued that we should not do away with classic statistical significance tests, chiefly because such tests can be useful despite their shortcomings. For instance, we showed that classic statistical significance tests allow us to correctly identify, in SEM analyses, many path coefficients whose true values are zero, as coefficients likely to be zero at the population level. We also argued that classic statistical significance tests need to be complemented by other methodological tools, including effect size tests, and common method bias tests.

We demonstrated that common method bias tests are needed because of the strong distorting overall effect that common method variation may have on path coefficients in SEM analyses, which is comparable to that of sampling error with very small sample sizes. We showed that this tends to happen even when the amount of common method variation is relatively small. Finally, we demonstrated that high quality theorizing is very important if we are to beneficially use a combination of classic statistical significance, effect size, and common method bias tests. We showed that common method bias tests are unlikely to be effective if theorizing is of low quality. Also, based on our discussion earlier in this paper, we showed evidence that low-quality theorizing is likely to have negative impacts on the effectiveness of classic statistical significance and effect size tests.

Good research requires high-quality theory and data, as well as sound analytical techniques. As outlined in this essay, there are a variety of statistical approaches that serve to complement classic significant tests by enhancing the interpretation of *P*-values through the augmentation of SEM. Regardless of the statistical method employed to analyze the data, it is important to recall the words of Sir Ronald Fisher from the early 20th Century: “To call in a statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of” (Fisher, 1938). Therefore, when selecting appropriate statistical approaches, we must always consider the aim and objective of the research, and the nature and type of the data used in the research. We must also be cognizant that a single study can only provide a limited amount of information. In the end, replication is critical if we are to truly understand and develop the field of management accounting.

References

Baumgartner, H., Weijters, B., & Pieters, R. (2021). The biasing effect of common method variance: Some clarifications. *Journal of the Academy of Marketing Science*, 49(1), 221-235.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Dow, K. E., Watson, M. W., Greenberg, R. H., & Greenberg, P. S. (2012). Understanding participation: Influence, situational participation, and intrinsic involvement. *Advances in Management Accounting*, 21(1), 25-47.

Dow, K. E., Askarany, D., Teklay, B., & Richter, U. H. (2021). Managers’ perceptions of justice in participative budgeting. *Advances in Management Accounting*, 33(1), 127-152.

Dow, K. E., Wong, J. A., Jackson, C., & Leitch, R. A. (2008). A comparison of structural equation modeling approaches: The case of user acceptance of information systems. *Journal of Computer Information Systems*, 48(4), 106-114.

Fisher RA. (1938). Presidential address to the First Indian Statistical Congress. *Sankhya*, 4(1), 14-17.

Kock, N. (2014). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration*, 10(3), 1-13.

Kock, N. (2015). Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration*, 11(4), 1-10.

Kock, N. (2016). Hypothesis testing with confidence intervals and P values in PLS-SEM. *International Journal of e-Collaboration*, 12(3), 1-6.

Kock, N. (2019a). From composites to factors: Bridging the gap between PLS and covariance-based structural equation modeling. *Information Systems Journal*, 29(3), 674-706.
Kock, N. (2019b). Factor-based structural equation modeling with WarpPLS. *Australasian Marketing Journal*, 27(1), 57-63.
Kock, N. (2021a). Harman's single factor test in PLS-SEM: Checking for common method bias. *Data Analysis Perspectives Journal*, 2(2), 1-6.
Kock, N. (2021b). Common structural variation reduction in PLS-SEM: Replacement analytic composites and the one fourth rule. *Data Analysis Perspectives Journal*, 2(5), 1-6.
Kock, N. (2023a). *WarpPLS User Manual: Version 8.0*. Laredo, TX: ScriptWarp Systems.
Kock, N. (2023b). Contributing to the success of PLS in SEM: An action research perspective. *Communications of the Association for Information Systems*, 52(1), 730-734.
Kock, N. (2023c). Using logistic regression in PLS-SEM: Dichotomous endogenous variables. *Data Analysis Perspectives Journal*, 4(4), 1-6.
Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, 28(1), 227–261.
Kock, N., & Lynn, G.S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.
Kock, N., Avison, D., & Malaurent, J. (2017). Positivist information systems action research: Methodological issues. *Journal of Management Information Systems*, 34(3), 754-767.
Mandal, P., Mukhopadhyay, S., Bagchi, K., & Gunasekaran, A. (2012). The impact of organisational strategy, culture, people and technology management on organisational practice and performance: an empirical analysis. *International Journal of Information Systems and Change Management*, 6(2), 160-176.
Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36.
Teklay, B., Dow, K. E., Shen, Y., Wong, J. A., & Askarany, D. (2023). Linkages between Service quality, customer satisfaction and financial performance in the US Airline Industry. *Advances in Management Accounting*, 34(1), 63-82.