Which is the best way to measure job performance: Self-perceptions or official supervisor evaluations?

Ned Kock

Full reference:

Kock, N. (2017). Which is the best way to measure job performance: Self-perceptions or official supervisor evaluations? *International Journal of e-Collaboration*, 13(2), 1-9.

Abstract

Among latent variables that can be used in e-collaboration research, job performance is a particularly important one. It measures what most e-collaboration tools in organizations aim to improve, namely the performance at work of individuals executing tasks collaboratively with others. We report on a comparative assessment of scores generated based on a self-reported job performance measurement instrument vis-à-vis official annual performance evaluation scores produced by supervisors. The results suggest that the self-reported measurement instrument not only presents good validity, good reliability and low collinearity; but that it may well be a better way of measuring job performance than supervisor scores.

Keywords: Job Performance; Measurement Instrument; Electronic Collaboration; Social Networking; Structural Equation Modeling; Partial Least Squares; Measurement Error.

Introduction

Structural equation modeling (SEM) methods and software tools make possible for researchers to simultaneously specify and test measurement and structural models involving latent variables. Mathematically, latent variables are aggregations of their indicators and measurement error. The indicators are quantitative responses, often provided along Likert-type scales, to question-statements in questionnaires.

Two main classes of SEM have been experiencing increasing use in e-collaboration research, as well as in empirical research in many other fields where multivariate statistical methods are typically used. One of these two main classes is SEM based on the partial least squares (PLS) method, a composite-based (as opposed to factor-based) method that does not explicitly account for measurement error. This class of SEM methods owes much of its existence to the work of Herman Wold, who devised a set of computationally efficient and nonparametric algorithms that serve as an alternative to the more restrictive covariance-based approach to SEM.

Covariance-based SEM accounts for measurement error and yields fairly precise parameter estimates; also, it has been widely used in the past, although it is arguably experiencing some decline in recent years. Among the reasons for this decline are that covariance-based SEM is somewhat difficult to use and restrictive in its assumptions (i.e., it assumes multivariate normality); and, perhaps more importantly, covariance-based SEM does not estimate factors as part of its iterative parameter convergence process, which can be seen as a major limitation (Kock, 2015a; 2015b).

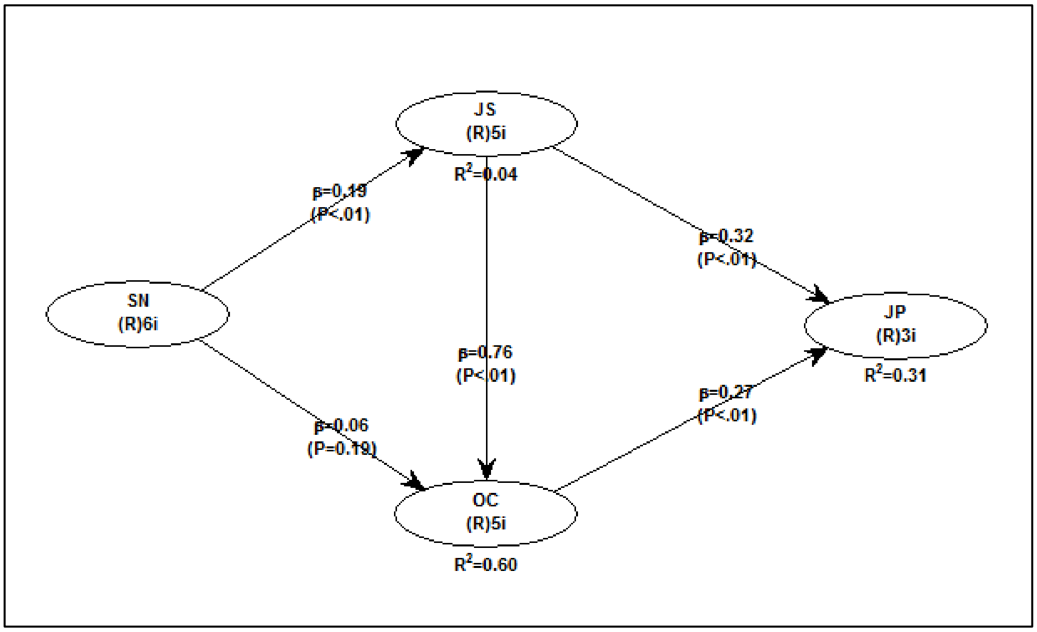
The other main class of SEM methods seeing increasing use is factor-based SEM (Kock, 2015b), which fully accounts for the measurement errors that are part of the factors in a model. This latter SEM class in some cases builds on coefficients generated by PLS algorithms, although it is very different from PLS-based SEM. The increasing use of this latter type of SEM is due in part to the ease-of-use and extensive features of software tools that implement it, such as WarpPLS (Kock, 2010; 2015a), which we use here in our analyses, building on an illustrative model. We use WarpPLS not only because it implements factor-based SEM, but also because it provides the most extensive set of features among comparable SEM software. Among these features is a comprehensive set of model fit and quality indices, as well as various coefficients that can be used in a variety tests – e.g., full collinearity variance inflation factors, used in multi-collinearity and common method bias tests.

Among latent variables that can be used in e-collaboration research, job performance is a particularly important one. After all, it measures what one usually wants to ultimately improve with the use of practically any e-collaboration tool in any organization – the job performance of individuals working in teams. In this study, we provide a comparative assessment of scores generated based on a self-reported job performance measurement instrument vis-à-vis official annual performance evaluation scores produced by immediate supervisors. The results discussed here suggest that the self-reported job performance measurement instrument is not only more than adequate, but may well be a better measure than official evaluation scores produced by supervisors.

Illustrative model and data

Our discussion is based on the illustrative model depicted in Figure 1. This illustrative model addresses the organizational effect of the use of social networking sites (SN), such as Facebook and LinkedIn, on job performance (JP). In the model, this effect (i.e., of SN on JP) is hypothesized to be indirect and mediated by intermediate effects on job satisfaction (JS) and organizational commitment (OC). These hypotheses are generally supported by the structural model coefficients; notable among these are the path coefficients and P values indicated next to each arrow in the model. This illustrative model is based on an actual study.

Figure 1: Illustrative model used



Notes: SN = social networking site use; JS = job satisfaction; OC = organizational commitment; JP = job performance; notation under latent variable acronym describes measurement approach and number of indicators, e.g., (R)5i = reflective measurement with 5 indicators.

Note that in the model the effect of SN on OC also appears to be primarily indirect and mediated by JS. Our illustrative model is consistent with theoretical developments and past empirical studies relating the use of social networking sites and job performance (Moqbel et al., 2013; Kock & Moqbel, 2016; Kock et al., 2016). The impact of social networking site use on job performance is a topic that can be seen as falling within the broader scope of the e-collaboration research area (Kock, 2005; 2008; Kock & Lynn, 2012).

The figure has been created with the SEM analysis software WarpPLS (Kock, 2010; 2015a). As such the figure displays the software's standard notation for summarized latent variable description. In this notation the alphanumeric combination under each latent variable's label (e.g., "JS") in the model describes the measurement approach used for that latent variable and the number of indicators. For instance, the alphanumeric combination "(R)5i" means reflective measurement with 5 indicators.

The latent variables have been measured through indicators based on the question-statements listed in Appendix A. The data was collected from 193 working professionals across the USA. These working professionals used social networking sites, with different degrees of use intensity. Use of social networking sites in general was considered in the data collection, not specifically use of social networking sites during business hours or while formally at work. The underlying theoretical basis of the model builds on the notion that social networking site use in general is likely to lead to positive emotions, with an overall positive effect on job performance. Therefore, the use of social networking sites is seen through this model as akin to mood-improving activities that could be performed within or outside work hours, such as recreational exercise and meditation, in terms of the activities' overall positive effect on one's performance at work.

Assessment of self-perceived measure: Validity, reliability and collinearity

In our analysis we employed a factor-based SEM algorithm that estimates parameters based on the two-stage process described by Kock (2015b), with the key difference that the factor-based SEM algorithm used here employs a consistent reliability estimate developed by Theo K. Dijkstra during his Ph.D. work conducted under Herman Wold. Wold is widely regarded as the originator of the classic PLS-based methods for SEM, which does not build on factors and thus does not take measurement error into account when estimating model parameters.

Our factor-based SEM algorithm is implemented in WarpPLS starting in version 6.0. Dijkstra's "consistent PLS" reliability estimate, which is utilized by this factor-based SEM algorithm, itself uses coefficients generated via the PLS Mode A algorithm employing the centroid scheme. This is the basic scheme for PLS Mode A, developed by Wold, and also implemented in WarpPLS. The factor-based SEM algorithm we used here yields results that are very similar to those previously implemented in WarpPLS by factor-based SEM methods employing Cronbach's alpha, a more widely used measure of reliability – which nevertheless tends to underestimate the true reliability when factor-indicator correlations (i.e., loadings) are highly heterogeneous. All of these factor-based SEM algorithms implemented in WarpPLS generally display the same precision as covariance-based SEM with full-information maximum likelihood.

In Table 1 we show several coefficients associated with JP. Some of these coefficients were calculated with respect to other variables, following standard approaches for measurement instrument quality assessment. They include the loadings associated with each of JP's three indicators (top-left of table), the correlations with other latent variables in the model (top-right), the composite reliability and Cronbach's alpha coefficients associated with JP (bottom-left), the average variance extracted and its square root (bottom-right), and the full collinearity variance inflation factor associated with JP in the context of our illustrative model (also bottom-right). These coefficients illustrate the quality of the self-perceived measure used for JP.

Table 1: Validation of self-perceived measure

Loadings		Correlations with LVs	
JP1	0.843	SN	-0.006
JP2	0.933	JS	0.487
JP3	0.927	OC	0.490
Composite reliability		AVE and square root of AVE	
0.929		0.814	0.902
Cronbach's alpha		Full collinearity VIF	
0.924		1.390	

Notes: LVs = latent variables other than JP; AVE = average variance extracted; VIF = variance inflation factor; the square root of the AVE is normally used in discriminant validity tests.

The loadings for the three indicators associated with JP were equal to or greater than the threshold value of 0.5, and the P values associated with these loadings were all significant at the 0.001 level. Moreover, the average variance extracted for JP was greater than the threshold value of 0.5. Taken as a whole, these results suggest that the measurement instrument used for JP presents good convergent validity (Kock, 2011; 2014; 2015a; Kock & Lynn, 2012). In fact, the lowest loading for JP was 0.843 and the average variance extracted was 0.814, both well above the generally accepted threshold of 0.5.

The square root of the average variance extracted for JP was 0.902. This is far greater than any of the three correlations involving JP and the other three latent variables in the model, which are shown in the table. Therefore we can safely conclude that the measurement instrument used for JP displays good discriminant validity in the context of our model (Kock, 2011; 2014; 2015a; Kock & Lynn, 2012). The composite reliability and Cronbach's alpha coefficients were 0.929 and 0.924 respectively, both greater than the conservative threshold value of 0.7, suggesting that the measurement instrument used for JP has good reliability (Kock, 2011; 2014; 2015a).

Finally, the full collinearity variance inflation factor associated with JP was 1.390. This coefficient is a model-wide measure of multi-collinearity, calculated in a way that incorporates the variations in the other variables in the model, and that allows us to test whether respondents viewed JP as conceptually different from all of the other latent variables. This 1.390 value is well below the recommended threshold of 3.3 (Kock & Lynn, 2012), suggesting that JP is not collinear with any of the other latent variables in the model, and thus is perceived as conceptually unique within our model.

Moreover, this low 1.390 value is also below the recommended threshold of 3.3 for common method bias; this is the same threshold used for multi-collinearity assessment (for details, see: Kock, 2015c; Kock & Lynn, 2012). This suggests that JP is not contaminated with pathological common method variance (Kock, 2014; 2015a; 2015c). As previously demonstrated (Kock, 2015c), common method variance tends to lead to high model-wide collinearity levels, leading to biases that are often not captured by discriminant validity or other classic common method bias tests; but that are captured by the test based on full collinearity variance inflation factors proposed by Kock (2015c) and Kock & Lynn (2012). As can be inferred from the above discussion, this latter test is also used for multi-collinearity assessment.

Assessment of self-perceived measure against official supervisor evaluations

The measurement instrument for JP was also validated against actual performance evaluation scores received by employees from immediate supervisors. Two steps were used in this validation. In the first step, we created two models containing two variables, JPself and JPsupv, and conducted linear and nonlinear analyses. JPself refers to self-perceived job performance, measured using the instrument for JP in Appendix A. JPsupv refers to official annual performance evaluation scores received from immediate supervisors. In the second step, we combined the indicators of these two variables into one latent variable, and again conducted validity and reliability assessments. These steps are discussed in more detail below.

First step: Segregated indicators. In the first step the following simple models, with only two variables each, were analyzed linearly and nonlinearly: JPself → JPsupv, and JPsupv → JPself. The reason for two models instead of one is that nonlinear results vary depending on the direction of a link; the linear results are the same. The linear and nonlinear path coefficients obtained were 0.502 (P<0.001) and 0.546 (P<0.001), respectively, for the model JPself → JPsupv. For the model with the link reversed, namely JPsupv → JPself, the linear and nonlinear path coefficients obtained were 0.502 (P<0.001) and 0.503 (P<0.001), respectively.

These results suggest that the association between the variables JPself and JPsupv is strong enough to justify using either of these two variables as a proxy for the other (Kock & Lynn, 2012). These results also suggest that, if only one of the two variables were to be used, the better choice would arguably be JPself, because it is the better predictor of the two (Kock & Gaskins, 2016); Pearl, 2009), as indicated by the nonlinear coefficients.

Second step: Combined indicators. In the second step we combined the indicators of the two variables into one new latent variable to further assess the compatibility of the two variables, and thus their proxy potential. The indicators of the variable JPself refer to the three corresponding question-statements in Appendix A. The variable JPsupv has only one indicator, namely the annual performance evaluation scores received from immediate supervisors.

The new latent variable with four indicators passed the criteria for validity and reliability assessment discussed earlier. Its composite reliability was 0.916, its Cronbach's alpha was 0.874, and its average variance extracted was 0.736. The loadings obtained for each of the indicators were the following: 0.879, 0.936, 0.929 and 0.658. The first three loadings refer to the three corresponding question-statements in Appendix A; the last loading refers to the annual performance evaluation scores received from immediate supervisors.

As can be seen, the composite reliability and Cronbach's alpha coefficients suggest acceptable reliability in connection with the new latent variable with four indicators. The average variance extracted and loadings, the latter all significant at the $P<0.001$ level, suggest acceptable convergent validity. These results provide further evidence in favor of the conclusion that JPself and JPsupv are fairly compatible, and thus could be used as proxies for each other.

However, the loading for 0.658 was below the more conservative threshold of 0.7 for convergent validity (Kock, 2014; 2015a), and also significantly lower than those for JP1, JP2 and JP3. Moreover, both the variability and range of variation for JPself were greater than for JPsupv, which suggests that the former variable provides a more fine-grained measure of job performance. Interestingly, the means for the indicators of JPself were lower than the mean for the single JPsupv indicator, suggesting that self-evaluations might have been more realistic. These results, combined with the results from the first step, suggest that it might be advisable to use only JPself (made up of JP1, JP2 and JP3) in SEM analyses, instead of the expanded latent variable with four indicators.

Conclusion

The results of the assessments discussed here not only suggest that the self-reported job performance measurement instrument used in our illustrative study is more than adequate, but also that it is probably a better measure than official annual performance evaluation scores received from immediate supervisors. In other words, anonymous self-evaluations of job performance may be better, from a measurement instrument quality perspective, than official annual performance evaluation scores produced by supervisors.

Therefore, we recommend that researchers employ the measurement instrument for job performance provided in Appendix A, or an extended measurement instrument that builds on the question-statements provided in Appendix A. An extended measurement instrument can be generated by adding question-statements whose meaning is likely to be seen as redundant with the three existing question-statements (see Appendix B). An example would be: "I am proud of my performance in my current job."

Having more redundant question-statements in the measurement instrument for job performance means that more redundant indicators would be available, which would likely lead to greater reliability values and thus lower measurement error magnitudes. It should be noted that redundant indicators present high levels of collinearity among themselves, which is a desirable property in a reflective latent variable measurement instrument such as the one we used here for job performance. This desirable measurement model property, at the individual latent variable level, is undesirable at the structural model (i.e., the inner model). The structural model includes the latent variable scores (or factor scores) and the links among latent variables. A high level of collinearity among latent variables, indicated by one or more full collinearity variance inflation factors above a threshold (e.g., 3.3 or, less conservatively, 5), is generally undesirable in part because it tends to lead to path coefficient distortion and instances of Simpson's paradox (Kock & Gaskins, 2016).

A key reason why anonymous self-evaluations of job performance may be better than official annual performance evaluation scores produced by supervisors may be that the former are anonymous while the latter are not. That is, the identities of the supervisors are known to the employees they evaluate. As such, supervisors may tend to generate evaluation scores that are a poorer reflection of their employees' performance than the employees own self-reported job performance evaluations. More specifically, supervisors may be tempted to overstate their employees' performance, perhaps to avoid conflict and ensure their employees' loyalty.

Acknowledgments

The author is the developer of the software WarpPLS, which has over 7,000 users in more than 33 different countries at the time of this writing, and moderator of the PLS-SEM e-mail distribution list. He is grateful to those users, and to the members of the PLS-SEM e-mail distribution list, for questions, comments, and discussions on topics related to the use of WarpPLS.

References

Kock, N. (2005). What is e-collaboration. *International Journal of e-Collaboration*, 1(1), 1-7.

Kock, N. (2008). E-collaboration and e-commerce in virtual worlds: The potential of Second Life and World of Warcraft. *International Journal of e-Collaboration*, 4(3), 1-13.

Kock, N. (2010). Using WarpPLS in e-collaboration studies: An overview of five main analysis steps. *International Journal of e-Collaboration*, 6(4), 1-11.

Kock, N. (2011). Using WarpPLS in e-collaboration studies: Descriptive statistics, settings, and key analysis results. *International Journal of e-Collaboration,* 7(2), 1-18.

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. (2015a). *WarpPLS 5.0 User Manual*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2015b). A note on how to conduct a factor-based PLS-SEM analysis. *International Journal of e-Collaboration*, 11(3), 1-9.

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

Kock, N., & Gaskins, L. (2016). Simpson's paradox, moderation, and the emergence of quadratic relationships in path models: An information systems illustration. *International Journal of Applied Nonlinear Science*, 2(3), 200-234.

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., & Moqbel, M. (2016). A six-stage framework for evolutionary IS research using path models: Conceptual development and a social networking illustration. *Journal of Systems and Information Technology*, 18(1), 64-88.

Kock, N., Moqbel, M., Barton, K., & Bartelt, V. (2016). Intended continued use of social networking sites: Effects on job satisfaction and performance. *International Journal of Virtual Communities and Social Networking*, 8(2), 28-46.

Moqbel, M., Nevo, S., & Kock, N. (2013). Organizational members' use of social networking sites and job performance: An exploratory study. *Information Technology & People*, 26(3), 240-264.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Appendix A: Latent variable measurement instrument

The question-statements below, answered on a Likert-type scale, were used for data collection related to the indicators of the latent variables in our illustrative model. The question-statements were answered on a Likert-type scale going from “1 – Strongly disagree” to “5 – Strongly agree”.

Social networking site use (SN)

SN1: My social networking sites’ account/s are/is a part of my everyday activity.
SN2: I am proud to tell people I’m on social networking sites such as Facebook.
SN3: Social networking sites have become part of my daily routine.
SN4: I feel out of touch when I haven’t logged onto social networking sites for a while.
SN5: I feel I am part of the social networking sites community.
SN6: I would be sorry if social networking sites shut down.

Job satisfaction (JS)

JS1: I am very satisfied with my current job.
JS2: My present job gives me internal satisfaction.
JS3: My job gives me a sense of fulfillment.
JS4: I am very pleased with my current job.
JS5: I will recommend this job to a friend if it is advertised /announced.

Organizational commitment (OC)

OC1: I would be very happy to spend the rest of my career with this organization.
OC2: I feel a strong sense of belonging to my organization.
OC3: I feel ‘emotionally attached’ to this organization.
OC4: Even if it were to my advantage, I do not feel it would be right to leave my organization.
OC5: I would feel guilty if I left my organization now.

Job performance (JP)

JP1: My performance in my current job is excellent.
JP2: I am very satisfied with my performance in my current job.
JP3: I am very happy with my performance in my current job.

Appendix B: Extended job performance (JP) measurement instrument

The first three question-statements below are from Appendix A. The other question-statements are suggestions for an extended set, which could be expected to increase latent variable reliability if used (without compromising loadings).

JP1: My performance in my current job is excellent.
JP2: I am very satisfied with my performance in my current job.
JP3: I am very happy with my performance in my current job.
JP4: I am proud of my performance in my current job.
JP5: I contribute a lot to my current organization in terms of job performance.
JP6: I am a high-performing member of my current organization.

These question-statements may be answered based on Likert-type scales going from “1 – Strongly disagree” to “5 – Strongly agree”, or from “1 – Strongly disagree” to “7 – Strongly agree”. Broader scales (e.g., 1 to 100) would probably not add much useful variation due to the human cognitive limitations associated with the “magical number 7” notion proposed by the cognitive psychologist George A. Miller.