UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Savalei, V., & Reise, S. P. (2019). Don't Forget the Model in Your Model-based Reliability Coefficients: A Reply to McNeish (2018). *Collabra: Psychology,* 5(1): 36. DOI: https://doi.org/10.1525/collabra.247

PERSPECTIVE/OPINION

Don't Forget the Model in Your Model-based Reliability Coefficients: A Reply to McNeish (2018)

Victoria Savalei* and Steven P. Reise†

McNeish (2018) advocates that researchers abandon coefficient alpha in favor of alternative reliability measures, such as the 1-factor reliability (coefficient omega), a total reliability coefficient based on an exploratory bifactor solution ("Revelle's omega total"), and the glb ("greatest lower bound"). McNeish supports this argument by demonstrating that these coefficients produce higher sample values in several examples. We express three main disagreements with this article. First, we show that McNeish exaggerates the extent to which alpha is different from omega when unidimensionality holds. Second, we argue that, when unidimensionality is violated, most alternative reliability coefficients are model-based, and it is critical to carefully select the underlying latent variable model rather than relying on software defaults. Third, we point out that higher sample reliability values do not necessarily capture population reliability better: many alternative reliability coefficients are upwardly biased except in very large samples. We conclude with a set of alternative recommendations for researchers.

Keywords: Coefficient alpha; Cronbach's alpha; Coefficient omega; Reliability; Psychological measurement

In psychological measurement, internal consistency reliability is theoretically defined as the proportion of variance in the observed score (typically, the sum or average of all items on a scale) that is due to the variation in the underlying construct of interest. For many years, methodologists have been lamenting the near universal and uncritical use of coefficient alpha (Cronbach, 1951; Kuder & Richardson, 1937) as a sole measure of internal consistency reliability in psychological research (e.g., Sijtsma, 2009; Schmitt, 1996). The main criticism of alpha is that it assumes a *unidimensional* (i.e., 1-factor) and a *tau-equivalent* (i.e., equal unstandardized factor loadings) model for the items on the scale.[1] This model is deemed unrealistic for most psychological scales. Despite the efforts to encourage the use of alternative reliability coefficients, few, if any, competing coefficients have been successful at capturing the attention of substantive researchers. In another attempt to "exorcise" psychology of alpha, McNeish (2018) advocated the use of several alternatives. These include coefficient omega (McDonald, 1999; referred to as "omega total" by McNeish), which assumes unidimensionality but, unlike alpha, does not assume tau-equivalence; a total reliability coefficient ("Revelle's omega total"), which captures the reliability due to multiple extracted factors (one general and several specific) obtained from an exploratory bifactor analysis; and the "greatest lower bound" (glb; Jackson & Agunwamba, 1977, Woodhouse & Jackson, 1977), which estimates the minimum dimensionality of the true score space under the assumption that the observed covariance matrix is the true covariance matrix. McNeish argues that these alternative coefficients should be preferred because they are based on fewer assumptions than alpha. He supports this argument by showing that these coefficients "outperform" alpha, in the sense that they produce larger sample values, in several empirical examples (p. 423).

In this response, we offer three main criticisms of the original article and the advice therein. First, while we agree with McNeish in principle that omega, the 1-factor reliability coefficient, is more theoretically appropriate than alpha, we argue that the underestimation of reliability by coefficient alpha when the assumption of tau-equivalence is violated, *but the assumption of unidimensionality holds*, is typically very small. Researchers working with well-established unidimensional scales who continue reporting alpha are rarely more than trivially misrepresenting their measures' reliability. Second, we argue that when unidimensionality is violated, alternative reliability coefficients should be based on a carefully selected model for the data. We show that simply relying on software defaults to pick *some* alternative multidimensional model can lead to nonsensical reliability estimates, such as those reported by McNeish in his main examples. Third, we disagree with McNeish that *higher* reliability estimates necessarily better capture

* Department of Psychology, University of British Columbia, Vancouver, British Columbia, CA

† Department of Psychology, University of California, Los Angeles, California, CA

Corresponding author: Victoria Savalei (v.savalei@ubc.ca)

true reliability. Sample reliability coefficients that require the estimation of complex multidimensional models can be positively biased due to capitalization on chance (i.e., overfitting) if the model contains non-existent latent dimensions. Even if the proposed multidimensional model matches the reality, a large sample size may be required to estimate it accurately, leading to inflated estimates in smaller samples. We now elaborate on these three points, and conclude with our own recommendations.

If a scale is unidimensional, alpha is adequate

Both coefficient alpha and coefficient omega assume unidimensionality (a single latent factor underlies the data).[2] Coefficient alpha makes an additional assumption of tau-equivalence. Tau-equivalence is the equality of the unstandardized factor loadings, or, stated differently, of item covariances but not necessarily of item variances. Importantly, this equality only needs to hold in the population, and sample loadings may be quite variable. How robust is alpha to the violations of tau-equivalence? Citing Green and Yang (2009a), McNeish suggests that under the violations of tau-equivalence, alpha could be in the mid 0.50s when true reliability is around .7, which indeed sounds serious (p. 415). However, these numerical values have no support in Green and Yang, and, as far as we can tell, are not possible under any realistic scenario. The discrepancy between alpha and omega is a function of both the average loading and the scale length. This discrepancy is maximized when most loadings are low and one is high, and for shorter scales (Raykov, 1997) – in other words, for scales that have *very low* reliability to begin with. Among the conditions examined by Green and Yang, the maximum discrepancy was observed for a 6-item scale with five loadings of .2 and one loading of .8, which led to omega of .386 and alpha of .343.[3] Such poor relationships between the items and the factor (i.e., most loadings are .2) are rare in practice, but even in this case, the two coefficients are similar.

In more realistic scenarios, expected discrepancies between alpha and omega would be much smaller. In fact, this point is illustrated well in McNeish: alpha and omega estimates never differ by more than .01 in all six of his examples (see his Tables 2 and 4). For another example, consider Table 2 in Hussey and Hughes (2018), where alpha and omega never differ by more than .01 across 16 different (sub)scales presumed to be unidimensional.[4] Finally, one of us has contributed to the development of a Shiny app[5] that plots the discrepancy between population values of alpha and omega for unidimensional scales, showing how hard alpha is to "break" under the conditions of unidimensionality. The misplaced emphasis by methodologists on the importance of tau-equivalence may be one of the reasons substantive researchers are slow to abandon alpha. This fixation never yields a convincing example where omega, the coefficient that does not assume tau-equivalence, is substantially higher than alpha, the coefficient that does.

However, if unidimensionality is severely violated, neither alpha nor omega are appropriate, and further modeling is required. Thus, rather than trying to convince researchers to report omega instead of alpha, methodologists should focus on advocating careful modeling of the scale's factor structure, particularly if the scale is new or not well studied. The correct reliability coefficient will depend on the dimensionality and the interpretation of the final chosen model.

The Alternative Model is Important!

The assumption of both alpha and omega is that a single latent factor is responsible for the covariation among the items. Violations of unidimensionality can have serious consequences for both coefficients (Cortina, 1993; Green & Yang, 2009; Raykov, 1998, 2001; Zinbarg, Revelle, Yovel, & Li, 2005). Researchers can use common factor analysis[6] to assess whether unidimensionality holds by inspecting indices of exact and approximate model fit (e.g., Kline, 2015). If the data are found to violate unidimensionality to a considerable extent (Reise, Bonifay, & Hoviland, 2013), the researcher must then specify a multidimensional model for the data before alternative reliability coefficients can be computed.[7] If the correct model is not the 1-factor model, then what is it? This question is necessary to answer because reliability is the proportion of variance in the scale score explained by the construct of interest, which in the model is represented by one or more latent dimensions. Thus, determining how many latent variables underlie the scale, and what they represent, is essential before reliability can be estimated in a meaningful way.

Lack of careful modeling when selecting a reliability coefficient can lead to problems. To illustrate this, we examine more closely how McNeish computes and interprets reliability coefficients based on an exploratory bifactor model, obtained using the *psych* package in R (Revelle, 2018). The bifactor model has become very popular in psychology in recent years, particularly for describing the multidimensional structure of psychological scales (Reise, 2012). This model contains the general factor, intended to capture the construct of interest, as well as several secondary or "group" factors, which are posited to be orthogonal to the general factor. If this model is found to represent the data well, the researcher can then compute two different reliability estimates: "total" and "hierarchical." These coefficients differ in whether the variance due to *all of the factors* or due to *only the general factor* counts towards reliable variance. Group factors are typically thought to represent irrelevant sources of variability (Gustafsson & Aberg-Bengtsson, 2010; Reise, Bonifay, & Haviland, 2013; Zinbarg, Yovel, Revelle, & McDonald, 2006; Zinbarg, Revelle, Yovel & Li, 2005; Zinbarg et al., 2006; Revelle & Zinbarg, 2009), so it is unclear that their variance should be counted towards an improvement in reliability. Thus, most methodologists recommend the more conservative strategy of interpreting only the hierarchical coefficient; for example, Kelley and Pornprasertmanit (2016) write, "specific factors should not be used to compute the reliability because they are not intended to measure the construct of interest" (p. 72). Yet McNeish reports only the total reliability coefficient ("Revelle's omega total")

for the examples in his Table 4, even though the *psych* package prints both. The total reliability coefficient will be higher than the hierarchical coefficient, and it will usually be higher than alpha, creating the illusion of improved reliability, but this is not a good reason to prefer it.[8]

There is an even more fundamental problem with McNeish's computations of reliability estimates based on the bifactor model. **Figure 1** illustrates the exploratory bifactor solution obtained for the first dataset in his Table 4, which contains data on 5 agreeableness items from the BFI (the same problem occurs for all the other datasets illustrated in this table). This plot was produced by applying the "omega" function from the *psych* package with its default settings to this dataset, as was done by McNeish (see his Appendix). By default, the "omega" function extracts three group factors, which for this small dataset means a model that actually has negative degrees of freedom![9] Despite this, the function does produce a "Revelle's omega total" value of .77, which is higher than alpha (here, .71), but we hope we do not have to convince the reader that this coefficient is completely uninterpretable. For comparison, the hierarchical coefficient is .68 in this example, even though it is equally uninterpretable.

We believe it is crucial for methodologists to stress the importance of careful modeling in any recommendations to applied researchers about computing alternative reliability coefficients. Doing so will avoid a literature flooded with nonsensical and inflated estimates of reliability. Methodologists should also make sure that researchers know it is up to *them* to select the most interpretable reliability coefficient associated with a particular model—many reliability estimates are possible based on the same complex latent variable model, and they differ in terms of which latent dimensions are considered to be part of the construct and which are not.[10] In the case of bifactor modeling in particular, the hierarchical coefficient is often the most interpretable one. It also protects the researcher from drawing unreasonably optimistic conclusions about the scale's reliability if the additional latent dimensions improve model fit but do not have a clear interpretation (as often happens with bifactor modeling).

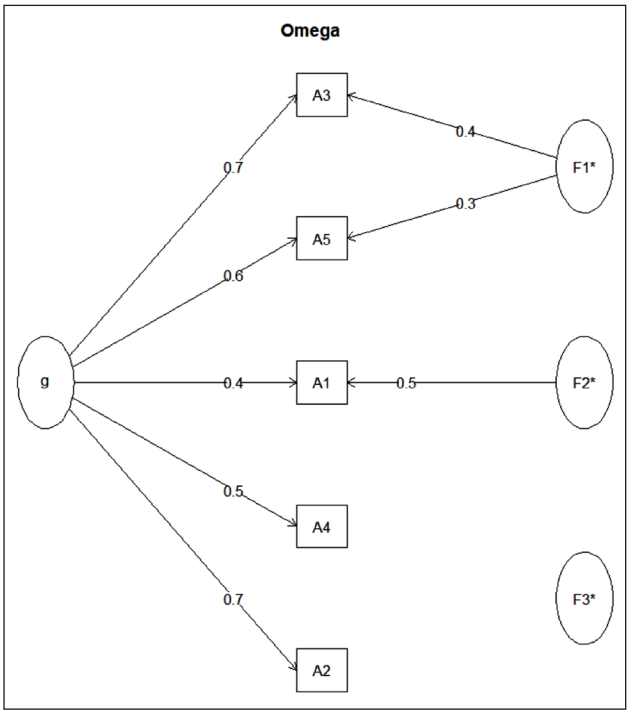

Figure 1: The default bifactor EFA solution for the 5-item agreeableness subscale of the "bfi" dataset (available from the *psych* package), estimated using the "omega" function.

Note: This diagram was produced automatically by version 1.8.12 of the *psych* package. (The estimated solution in earlier versions is different, but it is no more interpretable than this one.) Loadings smaller than .2 are not shown. This model has negative degrees of freedom.

Higher Sample Estimates Are Not Always Better

McNeish argues that because alternative reliability coefficients "have been shown to routinely exceed Cronbach's alpha," "there is no situation where Cronbach's alpha is the optimal method for assessing reliability" (p. 422). Our last point of disagreement is that "optimal" in the context of reliability assessment should not be equated with "highest." First, reliability coefficients based on more complex latent variable models may require large samples before they become accurate.[11] Some of the reliability coefficients recommended by McNeish, such as the glb, are indeed known to be very biased in all but very large samples (e.g., Sijtsma, 2009). For others, such as "Revelle's omega total", only limited research is available (e.g., Zinbarg, Revelle, & Yovel, 2007). In our opinion, more simulation work is necessary to determine the sample size requirements for many of the alternative reliability coefficients, even when the more complex model they are based on is correct.

Second, the existence of sampling fluctuations means that it is quite easy, in practice, to extract more latent dimensions than are present in the true model (i.e., to overfit the data), which also leads to inflated total reliability estimates (Yang & Green, 2010). A good analogy here is between reliability and the R-squared in regression. Adding predictors to the regression model will always improve the R-squared, but these additional relationships may not replicate in a new sample, because they are not real. In fact, in a latent variable model, a total reliability coefficient is exactly the proportion of variance in the observed composite explained by all the latent factors, so it *is* the R-squared for the latent regression of the scale score on the factors.

In Appendix A, we present a small simulation that illustrates the impact of overfitting on "Revelle's omega total." In this simulation, the true model is a 1-factor model, and the default computation of "Revelle's omega total" (which posits three group factors) by the "omega" function in the *psych* package severely overestimates true reliability at N = 100. Interestingly, the hierarchical coefficient severely *under*estimates reliability at this sample size, so it, too, is affected by extracting factors that are not present in the population. This performance may be due to the default rotation method used by the

"omega" function; using a bifactor EFA procedure with a different rotation method has been found to lead to better performance (Murray, Booth, Eisner, Obsuth, & Ribeaud, 2018). In any case, this small simulation reveals that it may be premature to recommend the default exploratory bifactor procedure in the "omega" function within the *psych* package to applied researchers. Promising alternatives exist (Abad, Garcia-Garzon, Garrido, & Barrada, 2017; Waller, 2018) that may perform better. There is also some evidence that a confirmatory factor analysis (CFA) approach may do better at providing accurate estimates of reliability under the bifactor model than the EFA approach (Yang & Green, 2010; Murray, Booth, Eisner, Obsuth, & Ribeaud, 2018).[12]

Conclusion and Recommendations

Our recommendations are as follows. For existing scales that have gathered sufficient empirical support for unidimensionality, coefficient alpha is simple to compute and reasonably accurate even if tau-equivalence is violated. While coefficient omega can always be reported instead, it will rarely make a meaningful difference, and we should not reprimand researchers about their choice of unidimensional coefficient. For scales that have an established structure with several substantive dimensions that are not strongly correlated, we support the common practice of reporting reliability coefficients (alpha or omega) for the subscales. For scales with strongly correlated latent dimensions and where the total score has substantive meaning, the best model representation for the scale structure will determine the appropriate reliability coefficient. If the structure of the scale is best conceptualized as a bifactor model, omega hierarchical is usually preferred, because group factors are rarely interpretable in such a case (Reise, Bonifay, and Haviland, 2013; Rodriguez, Reise, & Haviland, 2016; Revelle & Zinbarg, 2009). If the structure of the scale is best conceptualized as a correlated factors model or a second-order factor model, omega total will capture the variance due to all the extracted factors. In the case of the second-order factor model, omega hierarchical will capture the variance explained by the general factor only, which may also be meaningful. Everything hinges on how the factors are interpreted substantively, and what the researcher wants to know when computing reliability.

If previous research into the scale's structure is lacking, and the researcher has access to a sufficiently large sample, explicit modeling of the latent structure of the scale is recommended. Decisions about scale structure—such as how many latent factors underlie the data, whether any correlated errors are present, and whether the appropriate multidimensional model should be a correlated factors model, a second-order factor model, or a bifactor model–must be made on substantive, not just statistical grounds. The estimated solution for the final chosen model must be carefully inspected for interpretability, which includes the examination of size and sign of all factor loadings and the interpretation of the latent factors (Yang & Green, 2011). In order to select a reliability coefficient, the researcher also needs to decide which latent dimensions contribute to reliable variance and which do not. Substantive interpretation of all extracted latent dimensions is particularly important for the interpretation of total reliability coefficients.

A reviewer on an earlier version of this article commented that recommending that applied researchers engage in this type of careful modeling is not useful, as most applied researchers are not sufficiently familiar with latent variable modeling. Unfortunately, *there are no shortcuts for obtaining model-based estimates of reliability*. As we have demonstrated, obtaining coefficients from software packages that automatically select a model for the researcher via default settings can lead to nonsensical estimates. Of the two "evils"—thoughtlessly reporting alpha versus thoughtlessly reporting "default" multidimensional coefficients—we are not sure which is the lesser one. Particularly as it can produce inflated estimates, the second approach can give researchers a false sense of security, making scales "passable" that have questionable factor structure. Thus, to avoid "garbage in garbage out," researchers working with new psychological scales must obtain sufficient methodological training or seek assistance (see also Flake, Pek, & Hehman, 2017 for a similar point).

Latent variable modeling techniques can require large samples. The sample size recommendations vary and depend on the number of items on the scale, the strength of the correlations in the data, the number of extracted factors, the number of indicators per factor, and so on. Because reliability estimates are themselves statistics, a good way to capture the impact of sample size is to report confidence intervals alongside point estimates of reliability, where those are available. Such confidence intervals are not readily available for all coefficients, but for example, the confidence interval for coefficient omega can be obtained from the MBESS package in R (Kelley, 2007; Kelley & Pornprasertmanit, 2016). Another advantage of the CI approach is that it brings the distinction between sample and population reliability into focus: reliability estimates are themselves statistics, and can be inaccurate in small samples.

To keep this critique focused, we have omitted detailed discussion of other points of disagreement with McNeish, well covered in other sources. We briefly mention two. First, "coefficient H," also recommended by McNeish, is fundamentally different from the other coefficients he recommends because it gives the reliability of a weighted rather than unweighted scale score. A recent article by Aguirre-Urreta, Ronkko, and McIntosh (2018) provides a detailed discussion and study of this coefficient, expressing many concerns that we share. Second, the use of the polychoric correlation matrix when items are ordinal to estimate reliability is not without controversy (Chalmers, 2018; Green & Yang, 2009b). It is important not to interpret the resulting estimate as the reliability of the observed scale score; rather, it is an estimate of the proportion of reliable variance in the composite that *would have been obtained had the data been continuous*. Alternative reliability computations exist that may capture

the concept of reliability for categorical data better (e.g., Green & Yang, 2009b), but they require further study. Finally, for a more technical critique of McNeish, see Raykov & Marcoulides (2019).

Data Accessibility Statement

Code to reproduce the simulation study summarized in the Appendix is available at https://osf.io/dhejr/.

Appendix A

This appendix presents a small simulation to illustrate that reliability coefficients can capitalize on chance. McNeish used the "bfi" dataset in the *psych* package for illustration; this dataset has N = 2,800 participants. For such a large sample size, capitalization on chance would be expected to be minimal. In this simulation, we include two sample size conditions: N = 100 and N = 2,800. Code is available at https://osf.io/dhejr/, which can be used to extend the simulation to any other sample size.

We generated the data from a 1-factor (congeneric) model with nine items with unit variances. Factor loadings were set to {.4,.5,.6}, repeated three times. The population reliability under this model is .752. We drew 1,000 samples of each of the two sample sizes from this population. For simplicity, we estimated all of the reliability coefficients using the scaleStructure function in the *userfriendlyscience* package (Peters, 2018), following McNeish. We do not report "coefficient H," because it estimates the reliability of a weighted composite, and thus its values cannot be compared to the target population reliability value of .752. In addition to alpha, omega, and the glb, this package obtains "Revelle's omega total" and "omega hierarchical" (i.e., the total and hierarchical reliability coefficients under a bifactor EFA model) via the "omega" function of the *psych* package. The defaults of this function are to conduct an EFA extracting 3 factors, using minres as the extraction method, and with the correlation matrix as input; the solution is then rotated to a bifactor solution using the Schmid-Leiman rotation.

All coefficients but alpha should approach the correct reliability value asymptotically. We expect alpha to perform very similarly to omega, despite the heterogeneity in factor loadings. Because "Revelle's omega total" will estimate reliability from a model that will contain non-existent group factors, it may overestimate true reliability due to overfitting. Similarly, glb may extract non-existent latent dimensions in small samples.

The results are presented in the **Table 1** below.

As expected, at N = 100, both "Revelle's omega total" and glb severely overestimate reliability. Even the 25th percentile values for these coefficients are higher than then true reliability of .752. In the large sample condition, the glb estimate is now quite accurate, but "Revelle's omega total" is still upwardly biased. On the other hand, "omega hierarchical," which counts only the variance due to the general factor as reliable, is severely biased downward at N = 100, and remains biased even at N = 2,800. The bias in these EFA-based coefficients may be a function of the default choice of rotation for the bifactor procedure.

Notes

[1] Readers only familiar with classical test theory (CTT) may recognize the following equivalent formulation of the tau-equivalence assumption: Each item measures the same true score to the same extent.

[2] Technically, these coefficients also require local independence (error terms associated with each item are uncorrelated). Taking into account correlated error terms in the computation of reliability will tend to lower the resulting estimates (Raykov, 2001).

[3] Green and Young (2009a) considered scales of two lengths (6 vs. 12 items) and 14 different conditions of heterogeneous loadings (various combinations of .2, .5, and .8). For mathematical expressions relating the difference between alpha and true reliability (omega) under the 1-factor model, see Raykov (1997) and Zinbarg, Revelle, Yovel, and Li (2005).

[4] That is, we are referring to all rows in this table that contain "1" in the "Factors" column, and we are comparing the coefficients reported in the columns labeled "α" and "ω_t" for those rows. This paper is available at: https://psyarxiv.com/7rbfp/.

[5] https://semlab.shinyapps.io/breakalpha/.

[6] One can fit the 1-factor model using either exploratory or confirmatory factor analysis (EFA or CFA) software, with largely equivalent results. For example, the *psych* package in R (Revelle, 2018) will perform EFA, and the *lavaan* package (Rosseel, 2012) will perform CFA.

[7] A notable exception is the greatest lower bound (glb) coefficient (Jackson & Agunwamba, 1977; Woodhouse

Table 1: Mean, standard deviation (SD), and 25th and 75th percentile of the empirical distribution (1000 replications) of five reliability coefficients in the simulation study.

	N = 100				N = 2,800			
	Mean	**SD**	**25%**	**75%**	**Mean**	**SD**	**25%**	**75%**
alpha	0.74	0.04	0.72	0.77	0.75	0.01	0.74	0.75
omega	0.75	0.04	0.72	0.78	0.75	0.01	0.75	0.76
glb	0.82	0.03	0.79	0.84	0.75	0.01	0.74	0.76
"Revelle's omega total"	0.80	0.03	0.78	0.84	0.78	0.02	0.77	0.79
"omega hierarchical"	0.56	0.09	0.51	0.63	0.72	0.06	0.70	0.74

& Jackson, 1977; see also Bentler, 1972), also discussed by McNeish. This coefficient is "atheoretical" in the sense that the number of latent variables and what they represent substantively is not a priori specified. This method obtains a decomposition $S = S_T + E$, where the trace of E is maximized. However, the rank of S_T (which is the number of latent dimensions) can still be quite high. This coefficient assumes the sample covariance matrix S is the population covariance matrix, and thus overestimates the dimensionality of the latent space (and consequently, overestimates the reliability) by extracting "noise" dimensions from the data, unless the sample size is very large (Sijtsma, 2009).

[8] This is not to say that total reliability coefficients are never meaningful, but the researcher should have some idea what the additional factors represent conceptually, and why their variance should count towards the scale's reliability. This may be easier to do when the hierarchical or a CFA model is found to be the best representation for the data, rather than the bifactor model, because the multiple factors are not orthogonal to each other. The solution (e.g., estimated loadings) should also be examined for interpretability.

[9] The number of extracted factors is equal to the number of group factors; the solution based on k extracted factors is then rotated (using the Schmid-Leiman rotation) to a bifactor solution containing one general and k group factors. The number of correlations for p = 5 variables is p(p–1)/2 = 10, while the number of freely estimated loadings in an EFA model with k = 3 factors is k(p)–k(k–1)/2 = 5(3)–3(2)/2 = 12, so that df = 10–12 = –2. This number matches the df reported by the "omega" function. In defense of the "omega" function, clear warning messages are printed when one runs this analysis.

[10] Some coefficients also differ in terms of what is contained in the denominator, but a detailed discussion is beyond the scope of this commentary (see, e.g., Reise, Bonifay, & Haviland, 2013).

[11] This is an instance of the well known bias-variance trade-off, whereby in smaller samples, a simpler computation based on a slightly incorrect model may be more accurate than a computation based on a more complicated model, even if the latter is correct in the population.

[12] We recommend the *lavaan* package in R (Rosseel, 2012) for model fitting, and the "reliability" function in the *semTools* package (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018) to obtain total and hierarchical reliability estimates.

Competing Interests

No competing interests exist. Victoria Savalei is a Senior Editor at Collabra: Psychology. She was not involved in the review of this article.

Author Contributions

Victoria Savalei wrote the first draft and conducted the simulation study. Steven Reise contributed to multiple draft revisions.

References

Abad, F. J., Garcia-Garzon, E., Garrido, L. E., & **Barrada, J. R.** (2017). Iteration of partially specified target matrices: Application to the bi-factor case. *Multivariate Behavioral Research*. Advance online publication. DOI: https://doi.org/10.1080/00273171.2017.1301244

Aguirre-Urreta, M. I., Rönkkö, M., & **McIntosh, C. N.** (2018). A cautionary note on the finite sample behavior of maximal reliability. *Psychological Methods*. Advance online publication. DOI: https://doi.org/10.1037/met0000176

Bentler, P. M. (1972). A lower-bound method for the dimension-free measurement of internal consistency. *Social Science Research*, *1*, 343–357. DOI: https://doi.org/10.1016/0049-089X(72)90082-8

Chalmers, P. R. (2018). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement*, *78*, 1056–1071. DOI: https://doi.org/10.1177/0013164417727036

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*, 98. DOI: https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. DOI: https://doi.org/10.1007/BF02310555

Flake, J. K., Pek, J., & **Hehman, E.** (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370–378. DOI: https://doi.org/10.1177/1948550617693063

Green, S. B., & **Yang, Y.** (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121–135. DOI: https://doi.org/10.1007/s11336-008-9098-4

Green, S. B., & **Yang, Y.** (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167. DOI: https://doi.org/10.1007/s11336-008-9099-3

Gustafsson, J. E., & **Aberg-Bengtsson, L.** (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association. DOI: https://doi.org/10.1037/12074-005

Hussey, I., & **Hughes, S.** (2018, November 19). *Hidden invalidity among fifteen commonly used measures in social and personality psychology*. DOI: https://doi.org/10.31234/osf.io/7rbfp

Jackson, E. H., & **Agunwamba, C. C.** (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika*, *42*, 567–578. DOI: https://doi.org/10.1007/BF02295979

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & **Rosseel, Y.** (2018). *semTools: Useful tools for structural equation modeling. R package version 0.5–1*. Retrieved from https://CRAN.R-project.org/package=semTools

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*, 979–984. DOI: https://doi.org/10.3758/BF03192993

Kelley, K., & **Pornprasertmanit, S.** (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological methods*, *21*, 69. DOI: https://doi.org/10.1037/a0040086

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4th Ed). New York, NY: Guilford Press.

Kuder, G. F., & **Richardson, M. W.** (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160. DOI: https://doi.org/10.1007/BF02288391

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it for here. *Psychological Methods*, *23*, 412–433. DOI: https://doi.org/10.1037/met0000144

Murray, A. L., Booth, T., Eisner, M., Obsuth, I. & **Ribeaud, D.** (2018). Quantifying the strength of general factors in psychopathology: A comparison of CFA with maximum likelihood estimation, BSEM, and ESEM/EFA bifactor approaches. *Journal of Personality Assessment*. Advance online publication. DOI: https://doi.org/10.1080/00223891.2018.1468338

Peters, G. (2018). *userfriendlyscience: Quantitative analysis made accessible. R package version 0.7.2*, https://userfriendlyscience.com. DOI: https://doi.org/10.17605/osf.io/txequ

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*, 329–353. DOI: https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*, 375–385. DOI: https://doi.org/10.1177/014662169802200407

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76. DOI: https://doi.org/10.1177/01466216010251005

Raykov, T., & **Marcoulides, G. A.** (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*, 200–210. DOI: https://doi.org/10.1177/0013164417725127

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. DOI: https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & **Haviland, M. G.** (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*, 129–140. DOI: https://doi.org/10.1080/00223891.2012.725437

Revelle, W. (2018). *psych: Procedures for Personality and Psychological Research*. Evanston, Illinois, USA: Northwestern University. https://CRAN.R-project.org/package=psych Version = 1.8.12

Revelle, W., & **Zinbarg, R. E.** (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. DOI: https://doi.org/10.1007/s11336-008-9102-z

Rodriguez, A., Reise, S. P., & **Haviland, M. G.** (2016). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137–150. DOI: https://doi.org/10.1037/met0000045

Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. DOI: https://doi.org/10.18637/jss.v048.i02

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, *8*, 350. DOI: https://doi.org/10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. DOI: https://doi.org/10.1007/s11336-008-9101-0

Waller, N. (2018). Direct Schmid-Leiman transformations and rank-deficient loadings matrices. *Psychometrika*, *83*, 858–870. DOI: https://doi.org/10.1007/s11336-017-9599-0

Woodhouse, B., & **Jackson, P. H.** (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*, 579–591. DOI: https://doi.org/10.1007/BF02295980

Yang, Y., & **Green, S. B.** (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling*, *17*, 66–81. DOI: https://doi.org/10.1080/10705510903438963

Yang, Y., & **Green, S. B.** (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29, 377–392. DOI: https://doi.org/10.1177/0734282911406668

Zinbarg, R. E., Revelle, W., & **Yovel, I.** (2007). Estimating ω_h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement*, *31*, 135–157. DOI: https://doi.org/10.1177/0146621606291558

Zinbarg, R. E., Revelle, W., Yovel, I., & **Li, W.** (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133. DOI: https://doi.org/10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & **McDonald, R. P.** (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*, 121–144. DOI: https://doi.org/10.1177/0146621605278814

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.247.pr

How to cite this article: Savalei, V., & Reise, S. P. (2019). Don't Forget the Model in Your Model-based Reliability Coefficients: A Reply to McNeish (2018). *Collabra: Psychology,* 5(1): 36. DOI: https://doi.org/10.1525/collabra.247

Senior Editor: Simine Vazire

Editor: Eiko Fried

Submitted: 22 March 2019 **Accepted:** 17 June 2019 **Published:** 02 August 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS