PSYCHOMETRIKA—VOL. 86, NO. 4, 869–876
DECEMBER 2021
https://doi.org/10.1007/s11336-021-09800-2

A TEST CAN HAVE MULTIPLE RELIABILITIES

JULES L. ELLIS
RADBOUD UNIVERSITY NIJMEGEN

It is argued that the generalizability theory interpretation of coefficient alpha is important. In this interpretation, alpha is a slightly biased but consistent estimate for the coefficient of generalizability in a subjects x items design where both subjects and items are randomly sampled. This interpretation is based on the "domain sampling" true scores. It is argued that these true scores have a more solid empirical basis than the true scores of Lord and Novick (1968), which are based on "stochastic subjects" (Holland, 1990), while only a single observation is available for each within-subject distribution. Therefore, the generalizability interpretation of coefficient alpha is to be preferred, unless the true scores can be defined by a latent variable model that has undisputed empirical validity for the test and that is sufficiently restrictive to entail a consistent estimate of the reliability—as, for example, McDonald's omega. If this model implies that the items are essentially tau-equivalent, both the generalizability and the reliability interpretation of alpha can be defensible.

Key words: true score, stochastic subject, domain sampling, latent variable, generalizability, reliability, indeterminacy.

It is an honour that I have this opportunity to comment on the article of Sijtsma and Pfadt (in press). It is interesting to see how the opinion of Sijtsma developed over time. For the superficial reader, the message of the 2009 article seems to be "bad alpha", while the message of the current article seems to be "hail alpha". Sijtsma and Pfadt nicely point out that the message of Sijtsma was rather that alpha is a poor index of unidimensionality, but an acceptable lower bound of reliability. Nevertheless, my impression is that Sijtsma and Pfadt current opinion over alpha is more favourable than Sijtsma's (2009) opinion, especially in comparison with the greatest lower bound.

I agree with the conclusion of Sijtsma (2009) and Sijtsma and Pfadt (in press) that alpha should not be used as an index of unidimensionality. Hoekstra et al. (2019, table 3) show that, in a sample of 534 corresponding authors of nine top-tier journals from four disciplines, 80% made the wrong inference about alpha, suggesting that they interpreted alpha as an index of unidimensionality. I believe the frequent use of the term "internal consistency" is related to this. Using the term "internal consistency" for alpha is misleading, and use of the term in this meaning should be banned from future academic publications. I also agree with the claim of Sijtsma and Pfadt that alpha still has many advantages as an index of reliability, and that the lower bound theorem of Guttman (1945; attribution by Lord and Novick 1968 , p. 87) is valid and useful, which is also the conclusion of Raykov and Marcoulides (2019). However, I do not agree with the conclusion of Sijtsma and Pfadt, that "this is really all there to say about coefficient alpha". An important perspective on alpha is missing here, namely that it can be interpreted as an estimate of the coefficient of generalizability in a subjects x items design (Cronbach et al. 1972, pp. 80–82; Webb et al. 2006). I will discuss this in the next section. In Sect. 2, I will argue that this interpretation has advantages over the interpretation of alpha as a lower bound of reliability.

Correspondence should be made to Jules L. Ellis, Behavioural Science Institute, Radboud University Nijmegen, P.O.B. 9104, 6500 HE, Nijmegen, The Netherlands. Email: jules.ellis@ru.nl

© 2021 The Author(s)

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

1. Alpha as an Estimate for the Coefficient of Generalizability

The typical design where alpha is applied is a subjects x items design, where the items are questions or parts of a psychological test. Two different interpretations of alpha are possible here: one based on classical test theory (CTT) (Lord and Novick 1968) and one based on generalizability theory (GT) (Cronbach et al. 1972). Let me briefly explain the difference for readers who are unaware of the difference; see Vispoel et al. (2018) for a more elaborate discussion. In CTT, as described by Lord & Novick (1968, pp. 82-88) in their development of alpha, it is assumed that the test items are, in ANOVA terms, a fixed factor. That is, if the measurement is repeated, it would necessarily be based on exactly the same test items. In GT, in contrast, it is assumed that the items are, in ANOVA terms, a random factor. That is, it is assumed that the items are sampled from a large pool of item, and that if the measurement is repeated, one might use another sample of items, at least in theory. For an example of items that form a random factor, consider an examiner who possesses a pool of 1000 exam items, and who construes an exam each year by drawing 40 items randomly from the pool. For standard psychological tests, it is usually more difficult to replace the items, but still one can imagine that the test constructors could have ended with slightly different items.

In a design of subjects x items, where the items are drawn from a large pool, the item pool would be called the *domain* or *universe* (Nunnally 1978, p. 194), and the *universe score* of a subject would be defined as the expected value of its score across all items in the item pool (Webb et al. 2006). This universe score is also called a true score (Nunnally 1978, p. 194;), and it will be called the *domain sampling true score* here. The *coefficient of generalizability* is defined as the variance of the universe scores divided by the variance of the test scores (Cronbach et al. 1972, p. 17; Webb et al. 2006). This coefficient of generalizability can be estimated by alpha (Cronbach et al. 1972, pp. 80-82, 98). This then is another advantage of alpha that I would like to add to the discussion of Sijtsma and Pfadt: *coefficient alpha tells you how representative the test scores are for scores that would be obtained with the full domain of admissible items.*

The next question is: how good is alpha as an estimator for the coefficient of generalizability? While Sijtsma discuss the error and bias of alpha when subjects are sampled, in GT alpha will also have an error and bias due to the sampling of items. In the example of the 40 items sampled from a pool of 1000, if alpha is computed each year in the 40-items sample, then the average value of alpha over years would be approximately correct according to Cronbach et al. (1972, p. 98; Cronbach and Shavelson 2004 , p. 402). Thus, alpha is claimed to be an approximately unbiased estimate of coefficient of generalizability even if the items are not essentially τ-equivalent. McDonald (1978), however, showed that alpha is a lower bound to the coefficient of generalizability if the domain has a finite number of common factors, with equality only if the item covariances are equal.

The fact that coefficient alpha can be used as an estimate for the coefficient of generalizability even in the absence of unidimensionality, albeit biased, is certainly an advantage. In some cases, the selection of test items should be based on a broad domain definition rather than a theoretical analysis of dimensionality, and focussing on unidimensionality may lead to tests that are too narrow in content. As Cronbach and Shavelson (2004, p. 413) put it in their comment on the quest for unidimensionality: "A contrary position emphasizes that one needs to represent all aspects of the variable that is the focus of measurement, not narrowing it to a single focal topic". This is another advantage of alpha that I would like to add to the discussion of Sijtsma and Pfadt: *if the item domain is broad and heterogeneous, a high value of coefficient alpha tells you that you have enough items to cover it.*

Ironically, Sijtsma and Pfadt (in press) defend alpha as an index relevant to reliability by virtue of the lower bound theorem based on CTT true scores, whereas Cronbach (1951, p. 306) adopted the domain sampling interpretation of alpha ("α is therefore an estimate of the correlation

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

expected between two tests drawn at random from a pool of items like the items in this test") and later disagreed with the lower bound interpretation:

> My 1951 article embodied the randomly parallel-test concept of the meaning of true score and the associated meaning of reliability, but only in indefinite language. Once Lord's (1955) statement was available, one could argue that alpha was almost an unbiased estimate of the desired reliability for this family of instruments. The *almost* in the preceding sentence refers to a small mathematical detail that causes the alpha coefficient to run a trifle lower than the desired value.
> This detail is of no consequence and does not support the statement made frequently in textbooks or in articles that alpha is a lower value to the reliability coefficient. (Cronbach and Shavelson 2004, p. 402)

Alpha is a consistent estimator for the coefficient of generalizability in the subjects x items design (Webb et al., p. 16), but it is not directly clear how large the sampling errors can be under the sampling of items, and how this depends on further assumptions such as dimensionality. To illustrate this, I simulated items with a 2-dimensional 2PL model $P(X_i = 1|\Theta_1, \Theta_2) = (1 + \exp(-\alpha_{1i}\Theta_1 - \alpha_{2i}\Theta_2 - \beta_i))^{-1}$ with Θ_1, Θ_2 bivariate standard normal with correlation 0, and $\alpha_{1i}, \alpha_{2i} \sim Unif(0.5, 2.5)and\beta_i \sim Unif(-2, 2)$, where each item loaded on exactly one dimension: $\min\{\alpha_{1i}, \alpha_{2i}\} = 0,\ \max\{\alpha_{1i}, \alpha_{2i}\} > 0$. In 1000 samples, each with a test of 5 randomly selected items and 10,000 subjects, alpha ranged from 0.16 to 0.75. The 5^{th} percentile of alpha was 0.23, and the 95^{th} percentile was 0.57. The conclusion is that alpha is not always close to the coefficient of generalizability: for a small number of multidimensional items, the estimation error in alpha may be sizable even if the subject sample is large. Confidence intervals for alpha in the two-way random model are available for normally distributed variables (McGraw and Wong 1996; Demetrashvili et al. 2016), and bootstrap methods are developed for the variance components (Brennan 2007; Tong and Brennan 2007) and their ratio (Gilder et al. 2007; Ye et al. 2020). Still, it would be helpful for test makers in explorative research to have a simple guideline on the a priori minimum test length required for accurate estimation of the coefficient of generalizability with alpha.

2. The Indeterminacy of True Scores

Sijtsma and Pfadt (in press) assume in their article the existence of true scores as defined by Lord and Novick (1968). This is common practice in treatments of reliability, and I can see the merits of it, but in the present section, I will argue that these true scores are not well defined in most applications of coefficient alpha, and that this obscures the meaning of the concept "reliability". The domain sampling true score has a stronger connection to observations, and for this reason, the interpretation of coefficient alpha as an estimate of a coefficient of generalizability should be preferred.

The definition of true scores by Novick (1966) and Lord and Novick (1968, p. 30) can be paraphrased as follows if it is applied to item scores:

Definition 1. The true score of a subject on an item is the expected value of the observed score, where the observed score is drawn randomly from a probability distribution that depends on the subject and the item.

This definition assumes that there is some probability distribution of scores within a subject, and I will refer to this probability distribution as the *within-subject distribution*, and to the resulting true scores as *stochastic subject true scores.* The term stochastic subject was coined by Holland (1990).

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

Note that this definition does not specify the nature of the randomness of the observed scores. The true score is not necessarily defined by instantaneous replications under the same circumstances, where the subject is "brainwashed" between replications, although Lord and Novick (p. 29) cite this fictitious example of Lazarsfeld. An entirely different example of definition 1, that does not involve stochastic processes inside the subject, is this: if we measure the height of a child on one randomly sampled day from a five-year period, the corresponding true score would not be the momentary height, but rather the average height of the child over the entire period of five year. Indeed, Lord and Novick acknowledge the existence of multiple true scores:

> Finally, with respect to the syntactic definition of true score we have adopted here, it should be evident that a person's true score will depend on the various kinds of conditions under which the measurements are taken. For example, of all the conditions which affect measurement, we might choose to control lighting conditions. Suppose we set up two lighting conditions, one called "good" lighting and the other "bad". Then, over repeated experimentally independent observations for each condition, a true score for each person will be definable, and presumably these true scores will differ for each person over the two conditions. Also, if a third condition which involves a random sampling of the first two conditions is considered, a third true score can be defined. In Chapter 7 the first two of these true scores will be called a specific true score and the third, a generic true score. (p. 43)

The conclusion is that for the same test item, different true scores exist, depending on the within-subject probability distribution of the circumstances. Lord and Novick (p. 29) require that the within-subject distribution is "well-defined", and in that case the true scores are well defined. However, the within-subject distribution is often *not* well defined. Lord and Novick (p. 30) write that this distribution "is a hypothetical one because as we noted in Chapter 1, it is not usually possible in psychology to obtain more than a few independent observations". Even if such independent observations were possible, alpha is routinely applied in situations where only one observation per item per subject is available, and the possibility to do this is often presented as the main advantage of alpha. Thus, in most applications of alpha, the within-subject distribution is ill-defined, and then, the true scores are ill-defined.

That the within-subject distributions are ill-defined is not a logical necessity. In some cases, one can draw a random sample of observations from the same subject, as in the cited example given by Lord and Novick, and in that case the true scores can be properly defined. But in the typical application where coefficient alpha is being used outside GT, such explicit sampling schemes are absent. For example, suppose that a test is administered on one day in one location, and the answers are scored by one rater, and alpha is computed from this. Are the true scores now defined by sampling days within this fixed location and this fixed rater, or by sampling locations within this fixed day and fixed rater, or by sampling raters, or by a combination of days, locations and raters? Each of these possibilities may yield different true scores and different reliabilities, even though they all satisfy Lord and Novick's definition. The sampling scheme in reliability assessment in CTT is usually not explicit in these facets, leaving the reliability ill-defined. In contrast, GT requires explicit sampling schemes and solves the problem by adopting the domain sampling true score.

This limitation of stochastic subject true scores does not mean that they are useless under all circumstances. Ellis (1993) showed that a factor model for stochastic subject true scores predicts measurement invariance, while this prediction does not follow if the factor model holds merely for latent variable true scores. However, in that case there are additional data and additional restrictions about subpopulations. The point here is that stochastic subject true scores are ill-defined in a subject x items design without additional data or restrictions.

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

Let me now take this to the extreme by adding a second definition of true scores, in which they are considered as latent variables, possibly in factor analysis or item response theory models:

Definition 2. For a set of random variables $X_1, X_2, ..., X_J$, true scores are any set of (possibly latent) random variables $T_1, T_2, ..., T_J$ such that, with $E_i := X_i - T_i$, it holds that $\mathbb{E}(E_i|T_j) = Cov(E_i, T_j) = 0$ for all $i, j = 1, ..., J$.

I will call this *latent variable true scores*. It should be emphasized here that there is no claim here that these true scores are uniquely defined; on the contrary, many different true score variables may fit this definition with the same observed score variables, similar as in factor score indeterminacy (McDonald 1977; Steiger 1979) and indeterminacy of latent variables in the Rasch model (Ellis and Junker 1997, p. 508). The stochastic subject true scores of Lord and Novick are latent true scores with the additional restriction that $T = \mathbb{E}(X|S)$, where S is a variable that identifies the subject. Other latent variable true scores can defined by a random sampling formulation (Holland 1990), meaning that there is no within-subject variability; the additional restriction then is that $X = \mathbb{E}(X|S)$.

We need one more definition:

Definition 3. For a set of random variables $X_1, X_2, ..., X_J$ with latent variable true scores $T_1, T_2, ..., T_J$, we say that *the errors are uncorrelated* if $Cov(E_i, E_j) = 0$ for all $i, j = 1, ...J$ with $i \neq j$.

Latent variable true scores with uncorrelated errors will always exist, however. Suppes and Zanotti (1981)showed that if $X_1, X_2, ...X_J$ are binary random variables with a joint distribution, there exists a random variable Θ such that $X_1, X_2, ..., X_J$ are conditionally independent given Θ. They claim that this can easily be extended to continuous variables. For such Θ, we can define $T_i := \mathbb{E}(X_i|\Theta)$ and $E_i := X_i - T_i$, to obtain $\mathbb{E}(X_i - T_i|T_j) = \mathbb{E}(\mathbb{E}(X_i - T_i|T_j, \Theta)|T_j) = \mathbb{E}(0|T_j) = 0$, and uncorrelated errors follow in a similar fashion from conditional independence given Θ. In other words, the assumption that latent variable true scores exist with uncorrelated errors is always true.

For given latent variable true scores with uncorrelated errors, the lower bound theorem, attributed to Guttman (1945) by Lord and Novick (1968, p. 87), says that alpha is less than or equal to the reliability. The proof has been given in many texts and will not be repeated here. The following proposition illuminates that the latent variable true scores are so ambiguous that we can always assume that they are such that the reliability is greater than or equal to alpha; the restriction of uncorrelated errors is trivially satisfied if there are no further restrictions on the true scores. Let $X_+ := \Sigma_i X_i$ and $T_+ := \Sigma_i T_i$ and $Rel(X_+) := Var(T_+)/Var(X_+)$.

Proposition 1. *If a set of random variables $X_1, X_2, ..., X_J$ has covariance matrix C and $Var(X_+) > 0$, then there exist latent variable true scores $T_1, T_2, ..., T_J$ such that the errors are uncorrelated and such that $\alpha \leq Rel(X_+)$ holds for these true scores.*

Proof. Define $T_i := X_i$, then $E_i = 0$ and therefore $Cov(E_i, E_j) = 0$ for all $i, j = 1, ..., J$. Because $E_i = 0$, we have $Rel(X_+) = 1$, and it follows that $\alpha \leq Rel(X_+)$. The existence C of is merely required in order to assure that all relevant moments exist. □

Note that it can even be assumed that the reliability is equal to 1 without contradicting the CTT restrictions; there is no way to disprove this assumption. The additional restriction that the true scores be stochastic subject true scores, i.e. $T = \mathbb{E}(X|S)$, does not change this if the within-subject distribution is left unspecified or not estimable. Thus, in practical applications outside GT, after only a single test administration, without further model restrictions, one can always assume true scores such that alpha is a lower bound to the thus defined reliability.

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

3. Discussion

Coefficient alpha is typically used in a single test administration with a subjects x items design, and it was argued that in these designs the definition of true scores is ambiguous if the "syntactic" definition of Lord and Novick (1968, p. 30) is being used. Various CTT true score concepts have been discussed, and they all have limitations in this design (these points are certainly not new, and they are paraphrases of arguments brought up by psychometricians in personal communication, including the late Roderick P. McDonald):

- The stochastic subject true scores, which are used by Lord and Novick (1968) are ill-defined. They are defined by a within-subject distribution of scores on one item, while we typically have only one observation from this distribution. Consequently, they are indistinguishable from latent variable true scores in this design.
- The latent variable true scores suffer from indeterminacy. For items with finite range, one can always assume the existence of true scores as latent variables T_j with the restrictions $\mathbb{E}(E_i|T_j) = 0$ and $Cov(E_i, T_j) = 0$ and $Cov(E_i, E_j) = 0$. These true scores become meaningful only if further restrictions are added, as in factor analysis, or if further data are added, as in GT.

Consequently, a test can have multiple reliabilities, depending on the universe of generalization—which remains implicit in CTT—and the model restrictions—such as the number of common factors. In contrast, the domain sampling true score, called the universe score in GT, is based on a distribution from which multiple observations are available, namely the various item scores of the subject. The domain sampling true scores thus have a much stronger empirical anchor than the other true scores. This true score also has a disadvantage:

- The domain sampling true scores assume that items are randomly sampled from a large domain, which is usually "not true in any strict sense" (Cronbach and Shavelson 2004, p. 404)

Cronbach and Shavelson dismiss this criticism on the same grounds as used by Sijtsma and Pfadt (in press) to dismiss criticism on the use of alpha despite violations of essential -equivalence, namely that models "never fit the data perfectly".

Everything considered, my evaluation is that the best true scores are either 1) latent variable true scores in the context of a latent variable model with undisputed empirical validity for the given test and with sufficient restrictions to allow consistent estimation of the reliability or 2) domain sampling true scores. In the context of an undisputedly valid latent variable model, one should presumably prefer the measure entailed by the model, such as McDonald's omega. If the undisputed model implies essential -equivalence, this measure can be coefficient alpha, and in that case both the GT and the CTT interpretation of alpha are defensible. Otherwise, if the model implies uncorrelated errors, then the lower bound theorem would apply, but why would anyone use alpha if the undisputed model entails a consistent estimate that is different from alpha? Such a model can, in theory, have correlated errors, and then coefficient alpha can be greater than the reliability. Coefficient alpha would be useful in situations without such undisputed models, and in these situations, the domain sampling true scores would be best. *For this reason, I advocate the interpretation of coefficient alpha as an estimate for the coefficient of generalizability in a subjects x items design, and I recommend to take this as the default interpretation in teaching and empirical research where the coefficient is reported.*

The discussion of Sijtsma and Pfadt (in press) about the Bentler (2009) model can be analysed by noting that both the common factor and the sum of the common factor and the specific factor are latent variable true scores. Therefore, I agree with Sijtsma and Pfadt's conclusion that reliability

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

based on the common factor model is a CTT reliability. It is a nice illustration of the fact that a test can have two different reliabilities that both fit within the CTT definition of reliability.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1

Brennan, R. L. (2007). Unbiased estimates of variance components with bootstrap procedures. *Educational and Psychological Measurement, 67*(5), 784–803. https://doi.org/10.1177/0013164407301534

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. Wiley.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Demetrashvili, N., Wit, E. C., & Heuvel, van den, E. R. (2016). Confidence intervals for intraclass correlation coefficients in variance components models. *Statistical Methods in Medical Research, 25*(5), 2359–2376.

Ellis, J. L. (1993). Subpopulation invariance of patterns in covariance matrices. *British Journal of Mathematical and Statistical Psychology, 46*(2), 231–254. https://doi.org/10.1111/j.2044-8317.1993.tb01014.x

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62*(4), 495–523. https://doi.org/10.1007/bf02294640

Gilder, K., Ting, N., Tian, L., Cappelleri, J. C., & Choudary Hanumara, R. (2007). Confidence intervals on intraclass correlation coefficients in a balanced two-factor random design. *Journal of Statistical Planning and Inference, 137*(4), 1199–1212. https://doi.org/10.1016/j.jspi.2006.03.002

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282. https://doi.org/10.1007/BF02288892

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2018). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology, 22*(4), 351–364. https://doi.org/10.1080/13645579.2018.1547523

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*(4), 577–601. https://doi.org/10.1007/bf02294609

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

McDonald, R. P. (1977). The indeterminacy of components and the definition of common factors. *British Journal of Mathematical and Statistical Psychology, 30*(2), 165–176. https://doi.org/10.1111/j.2044-8317.1977.tb00736.x

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability". *Educational and Psychological Measurement, 38*(1), 75–79. https://doi.org/10.1177/001316447803800111

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Pfadt, J. (in press). Invited review part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*. https://doi.org/10.1007/s11336-021-09789-8

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika, 44*, 157–167. https://doi.org/10.1007/BF02293967

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? Synthese, 48(2), 191–199. http://www.jstor.org/stable/20115657

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67*(5), 804–817. https://doi.org/10.1177/0013164407301533

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods, 23*(1), 1–26. https://doi.org/10.1037/met0000107

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C.R. Rao, & S. Sinharay (Eds): *Handbook of statistics, volume 26* (pp. 81–124). Elsevier. https://doi.org/10.1016/s0169-7161(06)26004-8

Ye, R., Ge, W., & Luo, K. (2020). Bootstrap Inference on the variance component functions in the two-way random effects model with interaction. *Journal of Systems Science and Complexity, 34*(2), 774–791. https://doi.org/10.1007/s11424-020-9216-7

Manuscript Received: 7 JUN 2021
Final Version Received: 27 JUL 2021
Accepted: 30 JUL 2021
Published Online Date: 8 SEP 2021

https://doi.org/10.1007/s11336-021-09800-2 Published online by Cambridge University Press