Routledge
Taylor & Francis Group

Health Psychology Review

ISSN: 1743-7199 (Print) 1743-7202 (Online) Journal homepage: https://www.tandfonline.com/loi/rhpr20

Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all

Rik Crutzen & Gjalt-Jorn Ygram Peters

To cite this article: Rik Crutzen & Gjalt-Jorn Ygram Peters (2017) Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all, Health Psychology Review, 11:3, 242-247, DOI: 10.1080/17437199.2015.1124240

To link to this article: https://doi.org/10.1080/17437199.2015.1124240

HEALTH PSYCHOLOGY REVIEW, 2017
VOL. 11, NO. 3, 242–247
https://doi.org/10.1080/17437199.2015.1124240

Routledge
Taylor & Francis Group

Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all

Rik Crutzen[a] and Gjalt-Jorn Ygram Peters[b,c]

[a]Department of Health Promotion, Maastricht University/CAPHRI, Maastricht, The Netherlands; [b]Faculty of Psychology and Education Science, Open University of the Netherlands, Heerlen, The Netherlands; [c]Department of Work & Social Psychology, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

ABSTRACT

Cronbach's alpha is a commonly reported estimate to assess scale quality in health psychology and related disciplines. In this paper, we argue that alpha is an inadequate estimate for both validity and reliability – two key elements of scale quality. Omega is a readily available alternative that can be used for both interval and ordinal data. More importantly, we argue that factor-analytic evidence should be presented before assessing the internal structure of a scale. Finally, pointers for readers and reviewers of manuscripts on making judgements about scale quality are provided and illustrated by examples from the field of health psychology.

ARTICLE HISTORY
Received 1 May 2015
Revised 11 November 2015
Accepted 21 November 2015

KEYWORDS
Scale quality; alpha; internal consistency; validity; reliability

Cronbach's alpha[1] is a commonly reported estimate to assess scale quality in health psychology and related disciplines. To illustrate this, we have screened all articles published in Psychology & Health last year (see: http://osf.io/v7jxe). A total of 288 scales were reported in 88 articles. For 233 of these scales (80.9%), an estimate of scale quality was reported, which was alpha for 210 scales (90.1%). These figures demonstrate that reporting alpha is a widespread habit in health psychology. In this paper, we argue that alpha is an inadequate estimate for both validity and reliability – two key elements of scale quality – and that one of the readily available alternatives should be used. More importantly, we argue that also for these alternatives, factor-analytic evidence should be presented first when assessing scale quality.

Validity is a key element of scale quality and refers to the degree to which evidence and theory support the *interpretation* of scale scores. In other words, validity cannot be seen as a characteristic of a scale as such, but concerns the interpretation of scale scores in a specific study (AERA, APA, & NCEM, 1999). Therefore, evidence from previous studies can (and should) be used to substantiate the choice for a certain scale, but is not informative regarding the validity of the scale when it is used in a new study. This does not mean that assessing validity is necessarily an explorative endeavour, but only that assessing validity needs attention in each study – even when commonly used scales are employed.

Although traditionally many types of validity were distinguished (e.g., construct validity, criterion validity), nowadays validity is viewed as a unitary concept (AERA et al., 1999). Several sources of

CONTACT Rik Crutzen rik.crutzen@maastrichtuniversity.nl

evidence for validity are still acknowledged (e.g., evidence based on test content, internal structure, or relations to other variables), but these are no longer considered to present distinct types of validity. A comprehensive assessment of scale quality integrates these sources of evidence into a coherent account (AERA et al., 1999; Peters, 2014).

From alpha to omega

The focus of the current paper is on the analyses of the internal structure, which can indicate the degree to which the relationships among measurement items conform to the construct on which the proposed interpretation of scale scores is based. For example, the degree to which self-efficacy items used in a certain study reflect an underlying construct – in this case self-efficacy. Alpha, despite being frequently reported as such, is unrelated to a scale's internal structure. Sijtsma clearly shows that 'both very low and very high alpha values can go either with unidimensionality or multidimensionality of the data' (Sijtsma, 2009, p. 119). Therefore, in line with many others, we have previously argued to abandon alpha (Peters, 2014).[2] Instead, we recommend reporting alternative estimates such as omega, which provides a more accurate approximation of a scale's internal structure (Revelle & Zinbarg, 2009). Peters (2014) has published a step-by-step explanation on how to compute omega using the free and open source R (R Development Core Team, 2014) package 'userfriendlyscience' (Peters, 2015). Geldhof, Preacher, and Zyphur (2014) provide Mplus program code and extend the discussion to assessing and reporting reliability at multiple levels of analysis (e.g., individuals within groups).

The conventional approach to computing alpha and omega, using a Pearson correlation matrix, assumes interval data, whereas many scales in health psychology aggregate items using Likert-type responses (e.g., five answer options ranging from 'totally disagree' to 'totally agree'). This might yield either interval or ordinal data. In the case of ordinal data, there are alternative versions of alpha and omega that are calculated using a polychoric correlation matrix (Gadermann, Guhn, & Zumbo, 2012), which have recently been added to the function 'scaleStructure' in the free and open source package referred to before (Peters, 2015). Unless researchers have examined their items' measurement levels (e.g., using multidimensional scaling), it is recommended to report both omega based on a Pearson correlation matrix and omega based on a polychoric correlation matrix.

Factor-analytic evidence

Before reporting omega, however, researchers should confirm that for their sample (and by implication, their population), their measurement instrument retained its intended structure. In other words, we need to know whether a single latent variable is being measured in case of a unidimensional construct (Revelle & Zinbarg, 2009), or in the case of a multidimensional construct, whether the constructs dimensions are consistent with the exhibited factor structure. In case of a multidimensional construct, we assume that the scale is split into subscales. For example, the Strengths and Difficulties Questionnaire (SDQ) consists of five subscales (He, Burstein, Schmitz, & Merikangas, 2013). Subsequently, omega is reported per subscale. Hence, dimensionality should first be verified in order to know whether the measurement instrument retained its intended structure, because if not, the measurement instrument's validity is compromised, relegating reliability assessment to a secondary concern. In order to do so, the set of analysis techniques known as exploratory factor analysis (EFA) is available.[3] For example, the Very Simple Structure (Revelle & Rocklin, 1979), Velicer's Minimum Average Partial criterion (Velicer, 1976), and Parallel Analysis (Horn, 1965) approaches provide indicators that can help with this decision. Each of these methods can be easily applied using the 'vss' and 'fa.parallel' functions in the R package 'psych' (Revelle, 2015). Note, though, that if items are ordinal, Item Response Theory, latent class, or generalised latent variable models

may be more appropriate, as EFA may overestimate the number of factors (for an introduction, see Van der Eijk & Rose, 2015).

Despite the availability of methods to verify dimensionality, such analyses rarely seem to accompany reports of alpha. Of the 288 scales we surveyed, authors assessed dimensionality for 10 scales (2.4%). Although in some cases authors may have explored dimensionality but not reported it, this is unlikely to be true for the entire remaining 97.6%. Therefore, in the vast majority of cases, readers (and likely, reviewers) have no information as to the performance of the used scales. This means that the validity of the operationalisations of measurements cannot be verified. Of course, unexpectedly discovering a multidimensional scale structure can have implications for the interpretation of the data. This is exactly why conducting and reporting these analyses are so important. If a supposedly unidimensional scale turns out to have a two-dimensional structure in a given study, then this affects the interpretation of the scale's internal structure. Note that factor-analytic evidence can be considered to be evidence of internal structure (Cizek, Rosenberg, & Koons, 2008). Therefore, we recommend that factor-analytic evidence should be presented first when assessing the internal structure of a scale.

Validity and reliability

We introduced the focus of the current paper on the analyses of the internal structure in the context of validity. This is because evidence based on internal structure is one of the sources of evidence for validity (AERA et al., 1999). Yet, validity and reliability cannot be seen as two independent elements of scale quality. Reliability refers to the consistency of scale scores when a construct is assessed multiple times or in multiple ways. Please note, however, that reliability does not assume a unidimensional structure. A multidimensional scale can be perfectly reliable. The level of reliability has implications for the validity of the interpretation of scale scores (AERA et al., 1999). Kaplan and Saccuzzo (2013, pp. 154–155) state in a short and powerful way that 'we can have reliability without validity. However, it is logically impossible to demonstrate that an unreliable test is valid'. For example, a set of measurement items that is supposed to assess social desirability can lead to consistent scale scores, but this does not mean that these items necessarily reflect social desirability in line with contemporary social standards (Crutzen & Göritz, 2010). However, if the scale scores would vary dramatically over time and the construct being assessed (in this case social desirability) is expected to be stable over time, then it is impossible that the items conform to the construct on which the proposed interpretation of scale scores is based. Therefore, it is important to take reliability into account when assessing scale quality.

Traditionally, two types of reliability estimates have been applied in health psychology. The first type is estimates based on the relationships among scores derived from individual items. The same estimates as discussed above are used (e.g., omega in the case of a general factor). These estimates only require a single administration of a scale, but they ignore changes in measurement error due to time (also known as transient error). Ignoring transient error can lead to inaccurate conclusions (Chmielewski & Watson, 2009). The coefficient of equivalence and stability (CES) is a test–retest estimate that takes transient error into account (Schmidt, Le, & Ilies, 2003). We have previously explained how to compute CES using the previously mentioned R package (Crutzen, 2014). This second type of estimates (i.e., test–retest estimates) requires multiple administrations of a scale (Huysamen, 2006).

It should not be inferred, however, that test–retest estimates are always superior, because multiple administrations of a scale also have its limitations (besides participant burden). For example, hourly estimates of a stockbroker's anxiety throughout a business day are not likely to be stable and might even be related to the Dow Jones index (Cohen & Swerdlik, 1999, pp. 161–164). And even just administering a scale, for example the measurement of intention at baseline, can lead to changes over time (Mankarious & Kothe, 2015). The point is, however, that this might reflect actual changes over time, which does not necessarily indicate a lack of reliability. Such changes might even constitute the

phenomenon of interest. These considerations should be kept in mind when making judgements about scale quality.

Making judgements about scale quality: pointers for readers and reviewers

Both readers and reviewers of manuscripts often want to make a judgement about scale quality in a specific study. For commonly reported estimates, such as alpha, there are cut-off values available (e.g., Nunally & Bernstein, 1994) that we often consider to be fixed values. For example, 'values greater or equal to 0.9, 0.8, 0.7, 0.6, and 0.5 were considered as excellent, good, acceptable, questionable, and poor, respectively' (Crutzen & Kuntsche, 2013, p. 223). Such cut-off values are then used to interpret the estimates in terms of reliability of validity. Although this is a convenient way to quantify the interpretation of reported estimates, the minimum acceptable level for these estimates remains a matter of professional judgement (AERA et al., 1999). We understand the convenience of an answer to the question 'how high should these estimates be?' In all fairness, the only answer to this question is 'it depends'.

The interpretation of estimates used as an indicator of validity, for example, depends on the breadth of a construct. The General Self-Efficacy scale, for example, taps quite different aspects of self-efficacy (Luszczynska, Scholz, & Schwarzer, 2005). This is not a weakness, but a strength of the scale. Many psychological constructs derive their usefulness from their relatively broad definition, and therefore, their relatively broad operationalisation. A very narrow operationalisation of a construct might have high applicability in a specific situation (Fishbein & Ajzen, 2010; Peters, 2014), but hardly in any other situation. So, the broadness of the operationalisation of a construct affects scale quality estimates, meaning that ironically, in some situations, very high estimates are indicative of low validity.

The interpretation of estimates when used as an indicator of reliability, for example, depends on the expectations regarding stability of the construct (e.g., recall the stockbroker's anxiety throughout a business day). Some constructs (e.g., extraversion) are expected to be more stable over time than others (e.g., mood states). In case of the latter, estimates of stability are less appropriate when making a judgement about scale quality regarding such a construct in a specific study (Walsh & Betz, 2001).

Finally, the importance of high values for all of these estimates also depends on the consequences of interpretation of scale scores. For example, assessing psychosocial problems among adolescents to make a decision on future consultations (e.g., Crutzen, Bosma, Havas, & Feron, 2014) has more severe consequences than assessing ease of use of an intervention targeted at the same age group (e.g., Crutzen, Peters, Dias Portugal, Fisser, & Grolleman, 2011). In case of the former, high values for scale quality estimates are more important.

These examples concerning breadth and stability of constructs and consequences of interpretation of scale scores are not exhaustive. They do show, however, the importance of professional judgements regarding scale quality. To allow readers or reviewers of a manuscript to make such a judgement, it is crucial that authors disclose scale quality estimates *based on their own data* (Crutzen, Peters, & Abraham, 2012; Peters, Abraham, & Crutzen, 2012). Estimates from previous studies can be used to substantiate the choice for a certain scale (and therefore can be included in the 'Methods' section), but are not informative regarding the values of such estimates in a new study. We therefore recommend that authors include, and editors and reviewers demand, a paragraph in the 'Results' section where authors justify their implicit claim as to the validity of their operationalisations. This would include, for example, EFAs and estimates such as omega for each variable measured using items in a questionnaire. In the supplemental materials, authors ideally include (and editors and reviewers request) the full text of each item, accompanied by correlation matrices and scatter matrices of the items in each scale. Not only will this facilitate replication and provide the opportunity to optimise our scales over time, it will also enable editors, reviewers, and readers to make a judgment about the validity and reliability of the used scales. After all, scale quality – contrary to what the term might seem to suggest – is not a characteristic of a scale as such, but depends on the

interpretation of scale scores in a specific study. By facilitating this process of interpretation and judgement, we can move health psychology and related disciplines forward. The tools to do so are available and described above, it is up to all of us to take this step towards more insight into scale quality.

Notes

1. We are aware that Cronbach considered it an embarrassment that the formula became conventionally known as Cronbach's alpha (Cronbach & Shavelson, 2004) and, therefore, we will refer to it as alpha.
2. An extensive literature is available explaining the technical background for this argument (Cortina, 1993; Dunn, Baguley, & Brunsden, 2014; Graham, 2006; Revelle & Zinbarg, 2009; Sijtsma, 2009). In short, alpha relies on the essentially tau-equivalent model, which assumes unidimensionality and equal variances of and covariances between items. Unfortunately, these assumptions are almost always violated in 'real life' (Peters, 2014). McDonald (1978) decomposes variance into a general factor, a set of group factors (factors common to some but not all of the items), and specific factors unique to each item. $Omega_{total}$ is based upon the sum of squared loadings on all the factors, while $omega_{hierarchical}$ is based upon the sum of the squared loadings on the general factor (McDonald, 1999). It is possible that a scale is unidimensional or contains a general factor but that the factor common to all of the items is weakly saturated in the items. Therefore, the proportion of variance due to a general factor (i.e., $omega_{hierarchical}$) provides important information about the extent to which a scale score estimates a latent variable common to all items (Revelle & Zinbarg, 2009).
3. Alternatively, confirmatory factor analyses (CFAs) provide useful fit indices to compare pre-defined factor structures in situations where competing structures might fit the data (e.g., in the case of conflicting theories).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

AERA, APA, & NCEM. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*, 186–202.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, *68*, 397–412.

Cohen, R. J., & Swerdlik, M. E. (1999). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mountain View, CA: Mayfield Publishing Company.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.

Crutzen, R. (2014). Time is a jailer: What do alpha and its alternatives tell us about reliability? *The European Health Psychologist*, *16*, 70–74.

Crutzen, R., Bosma, H., Havas, J., & Feron, F. (2014). What can we learn from a failed trial: Insight into non-participation in a chat-based intervention trial for adolescents with psychosocial problems. *BMC Research Notes*, *7*, 824.

Crutzen, R., & Göritz, A. S. (2010). Social desirability and self-reported health risk behaviors in web-based research: Three longitudinal studies. *BMC Public Health*, *10*, 720.

Crutzen, R., & Kuntsche, E. (2013). Validation of the four-dimensional structure of drinking motives among adults. *European Addiction Research*, *19*, 222–226.

Crutzen, R., Peters, G. J. Y., & Abraham, C. (2012). What about trialists sharing other study materials? *BMJ*, *345*, e8352.

Crutzen, R., Peters, G. J. Y., Dias Portugal, S., Fisser, E. M., & Grolleman, J. J. (2011). An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study. *Journal of Adolescent Health*, *48*, 514–519.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York: Taylor & Francis Group.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*, 1–13.
Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.
Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930–944.
He, J. P., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The strengths and difficulties questionnaire (SDQ): The factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology*, *41*, 583–595.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.
Huysamen, G. K. (2006). Coefficient alpha: Unnecessarily ambiguous; unduly ubiquitous. *SA Journal of Industrial Psychology*, *32*, 34–40.
Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological assessment and theory: Creating and using psychological tests* (8th ed.). Boston, MA: Cengage Learning.
Luszczynska, A., Scholz, U., & Schwarzer, R. (2005). The general self-efficacy scale: Multicultural validation studies. *The Journal of Psychology*, *139*, 439–457.
Mankarious, E., & Kothe, E. (2015). A meta-analysis of the effects of measuring theory of planned behaviour constructs on behaviour within prospective studies. *Health Psychology Review*, *9*, 190–204.
McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement*, *38*, 75–79.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
Nunally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist*, *16*, 56–69.
Peters, G. J. Y. (2015). Userfriendlyscience: Quantitative analysis made accessible. R package version 0.3-0.
Peters, G. J. Y., Abraham, C., & Crutzen, R. (2012). Full disclosure: Doing behavioural science necessitates sharing. *The European Health Psychologist*, *14*, 77–84.
R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Revelle, W. (2015). Psych: Procedures for psychological, psychometric, and personality research. R package version 1.5.4.
Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403–414.
Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.
Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.
Van der Eijk, C., & Rose, J. (2015). Risky business: Factor analysis of survey data – assessing the probability of incorrect dimensionalisation. *PLOS ONE*, *10*, e0118900.
Velicer, W. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.
Walsh, W. B., & Betz, N. E. (2001). *Tests and assessment* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.