original article

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

Gjalt-Jorn Y. Peters
Open University of the Netherlands

Health Psychologists using questionnaires rely heavily on Cronbach's alpha as indicator of scale reliability and internal consistency. Cronbach's alpha is often viewed as some kind of quality label: high values certify scale quality, low values prompt removal of one or several items. Unfortunately, this approach suffers two fundamental problems. First, Cronbach's alpha is both unrelated to a scale's internal consistency and a fatally flawed estimate of its reliability. Second, the approach itself assumes that scale items are repeated measurements, an assumption that is often violated and rarely desirable. The problems with Cronbach's alpha are easily solved by computing readily available alternatives, such as the Greatest Lower Bound or Omega. Solving the second problem, however, is less straightforward. This requires forgoing the appealing comfort of a quantitative, seemingly objective indicator of scale quality altogether, instead acknowledging the dynamics of reliability and validity and the distinction between scales and indices. In this contribution, I will explore these issues, and provide recommendations for scale inspection that takes these dynamics and this distinction into account.

Psychologists do not have it easy. Whereas researchers in chemistry, medicine, or physics can usually directly observe the objects of their study, researchers in psychology not only have to rely on indirect measurement of the variables of interest, but these measurements are also subject to a plethora of biases and processing quirks that are not yet fully understood. Whereas biological measures, using for example electroencephalograms or functional magnetic resonance imaging, provide direct access to what are generally considered proxies of psychological activity, most psychologists are limited to measuring behavior. Although behavior is sometimes the variable of interest itself, psychologists often use participants' behavior to measure psychological variables. For example, implicit association tasks present participants with various stimuli and measure how fast participants respond to different stimuli, with the aim of inferring how strongly hypothesized psychological variables are associated; and questionnaires present participants with various items and measure which answer options participants endorse, with the aim of inferring the value of hypothesized psychological variables.

The indirect nature of these measurements leaves much room for unknown sources of variance to contribute to participants' scores, which translates to a relatively low signal to noise ratio, or a proportionally large measurement error. This is detrimental to studies' power to draw conclusions as to associations between the variables under investigation. To ameliorate this situation, researchers often use multiple measurements that are then aggregated. This process decreases the error variance, because as the number of aggregated measurements increases, those parts of the error variance that are not systematic cancel each other out more and more (since, conveniently, researchers usually assume that error variance is random). Of course, this

approach requires *repeated* measurements; if a researcher devised three additional questionnaire items to strengthen the measurement of the construct tapped by a first original item, the three additional items must measure the same construct as the first item. If they measure something else instead, they will decrease the validity of the measurement by adding a source of systematic measurement error. Thus, because psychologists are condemned to indirect measurements of psychological variables, aggregating our measurements is a valuable instrument; but at the same time, caution is advised when aggregating separate measurements into a scale.

Most researchers understand this, and perhaps this is one reason why researchers routinely report Cronbach's Alpha, which is widely considered almost as a quality label for aggregate variables. Researchers and reviewers alike are satisfied by high values of Cronbach's Alpha (many researchers will cite a value of .8 or higher as acceptable), and in fact, inter-relations of items are rarely inspected more closely if Cronbach's Alpha is sufficiently high. This reliance on Cronbach's alpha is unfortunate, yet has proven quite hard to correct (Sijtsma, 2009). One of the reasons may be a combination of self-efficacy and a lack of clear guidelines. Articles addressing the problems with Cronbach's Alpha tend to be quite technical, and rarely provide a tutorial as to what to do instead of reporting Cronbach's Alpha (Dunn, Baguley, & Brunsden, 2013, being a notable exception). The current paper aims facilitate improved scale scrutiny by doing three things. First, a brief non-technical explanation is provided as to why Cronbach's Alpha should be abandoned. Second, alternatives are introduced that are easily accessible with user friendly, free tools, and a tutorial of how to compute these alternatives is provided. Third, a plea is made to step away from convenient quantitative measures as means of assessing scale quality.

Why abandon Cronbach's Alpha

Imagine that we want to measure 'connectedness with the European Healthy Psychology Society (EHPS)' with four items. Figure 1 shows these four items in the simplest possible situation: they are all exactly the same. Of course, this never happens; and Figure 2 shows a more realistic picture. The gray normal curves in the background depict the population distributions for each item. In addition, for each item, the scores of three individuals are shown. When an individual answers each item, each single measurement, depicted by a black dot, is determined by the individual's true score on that item, represented by vertical dotted lines, and measurement error, represented by normal curves that show the likelihood of obtaining given measurements. In Figure 2, "How do you feel about the EHPS?" has considerably more measurement error than "How many EHPS conferences have you attended?". This might be, for example, because factors such as mood and whether somebody happens to have just gotten a submission to Psychology & Health accepted or rejected are more likely to temporarily influence somebody's appreciation of the EHPS than their recollection of the number of attended EHPS conferences. Another difference between the items in Figure 2 are the means: for example, naturally the mean for "How often do you read the EHP?" is exceptionally high. Finally, the variance in some items (e.g. attended EHPS conferences) is higher than in others (e.g. EHP reading frequency).

The items in Figure 2 satisfy the assumptions of the so-called 'congeneric model' of reliability, and the items in Figure 1 satisfy the much more restrictive assumptions of the 'parallel model' of reliability. Just like the differing assumptions of

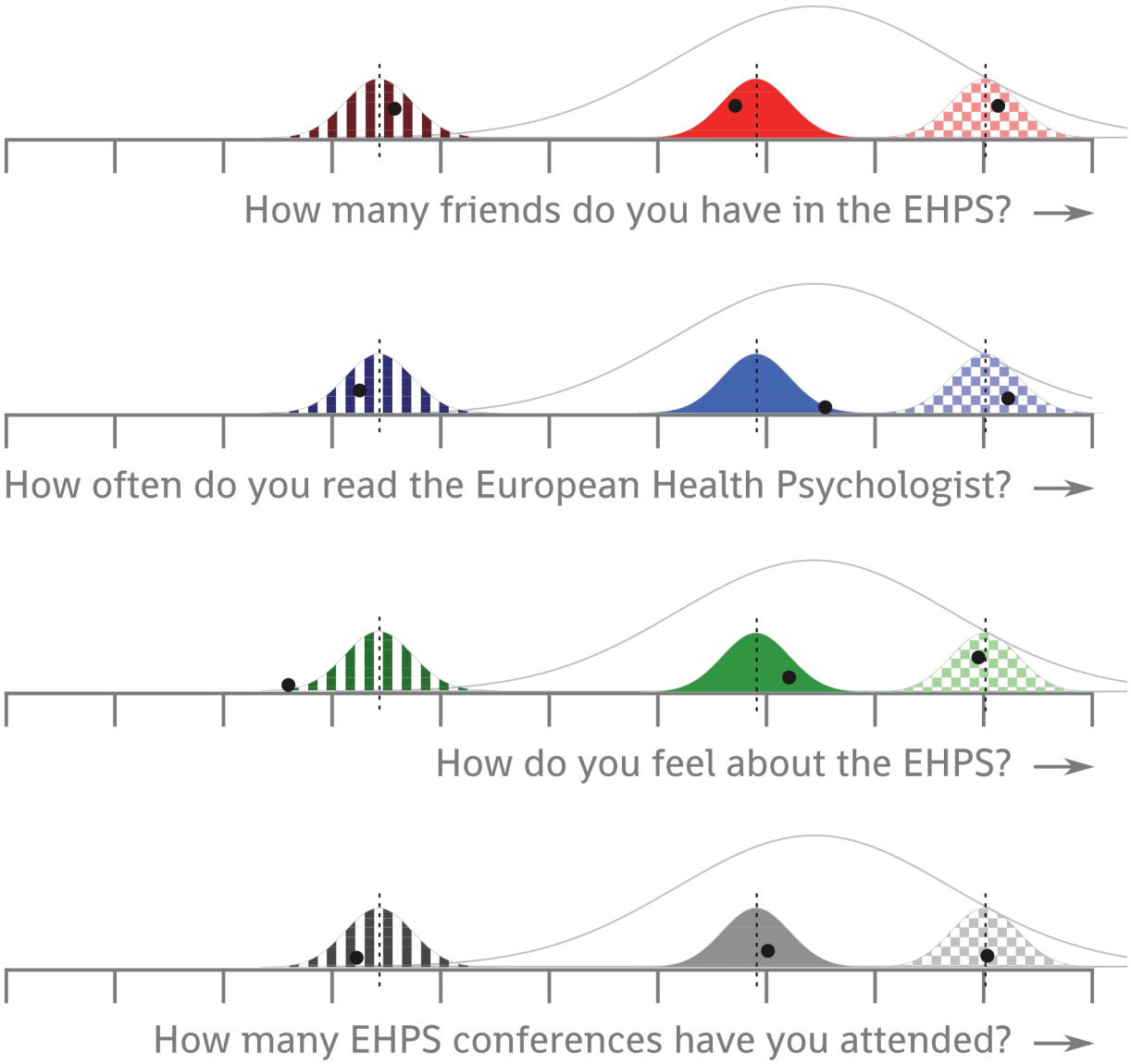

Figure 1: the scores of three individuals on four items that satisfy the assumptions of the 'parallel model of reliability'

the independent samples t-test and the paired samples t-test change the way the value of Student's t needs to be calculated, the assumptions of the different reliability models determine how a test's reliability can be estimated. A shared assumption of both of these models is that the items measure one underlying construct ('unidimensionality'), in this case connectedness with the EHPS. The congeneric model has no additional assumptions, but the parallel model also requires the items to have the same means, the same error variance, and the same variances in and covariances between items. In between this extremely restrictive parallel model and the much more liberal congeneric model lives the 'essentially tau-

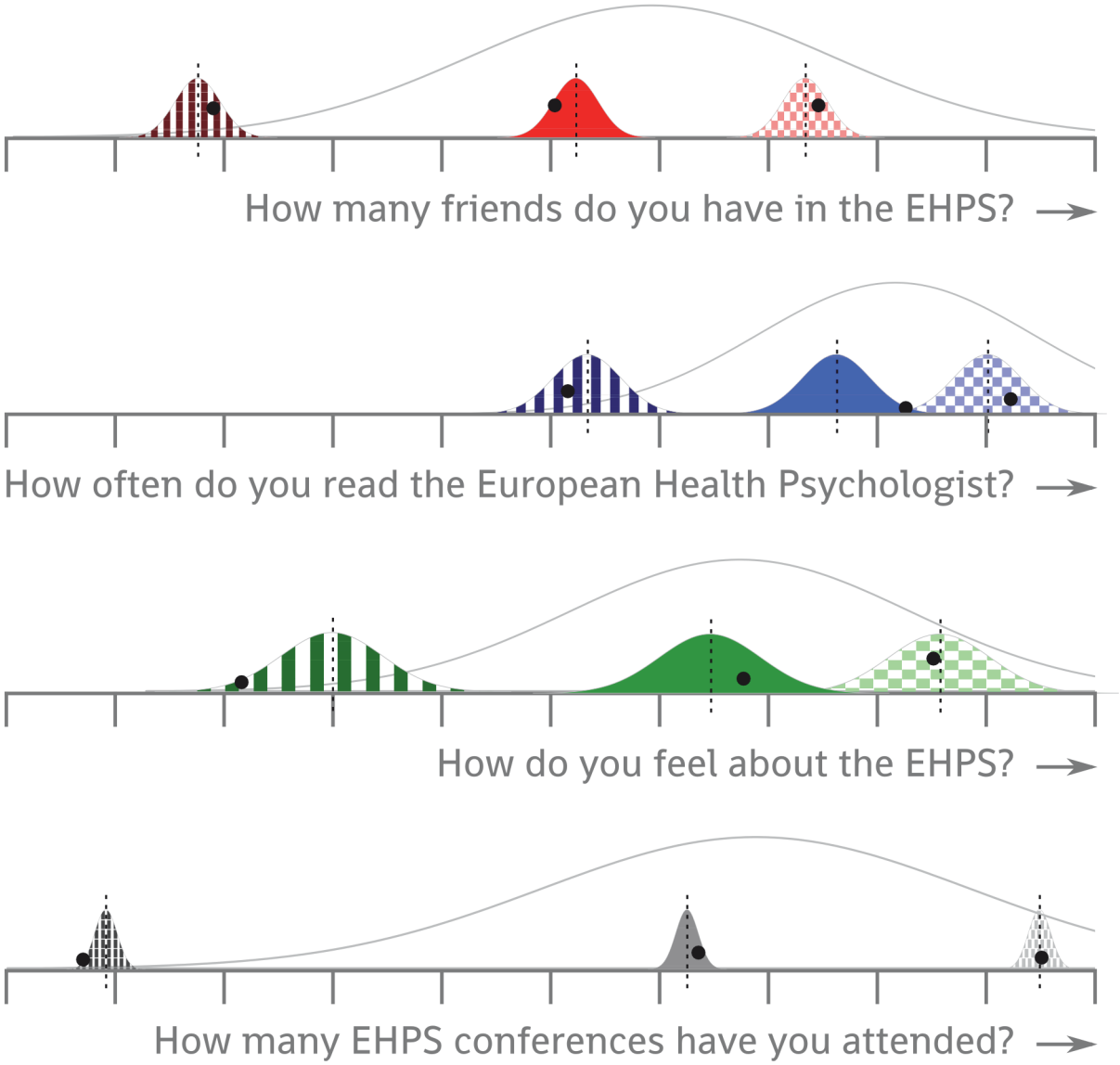

Figure 2: the scores of three individuals on four items that satisfy the assumptions of the 'congeneric model of reliability'

equivalent model', which assumes unidimensionality and equal variances of and covariances between items. This last model is the model relied on by Cronbach's Alpha (Cronbach, 1951). This essentially tau-equivalent model assumes that all items measure the same underlying variable, that they do so on the same scale, and that they are equally strongly associated to that underlying variable. In these situations, Cronbach's Alpha can be calculated as a measure of reliability of the scale; and conversely, violation of these assumptions means that Cronbach's Alpha is no longer a useful measure of reliability. In fact, it can be shown and has been shown that when essential tau-equivalence does not hold, it is

impossible that Cronbach's Alpha equals the reliability of the test (Sijtsma, 2009). Thus, when the assumptions of essential tau-equivalence are violated, the only thing you can be sure of when you know the value of Cronbach's Alpha, is that the test's reliability cannot possibly be that value. Unfortunately, these assumptions are almost always violated in 'real life' (Dunn et al., 2013; Graham, 2006; Revelle & Zinbarg, 2009; Sijtsma, 2009).

Cronbach's alpha is also seen as a measure of a scale's internal consistency, which is often loosely perceived as an indicator of the degree to which the items making up the scale measure the same underlying variable (interestingly, this is the assumption of 'unidimensionality' in the congeneric, parallel, and essentially tau-equivalent models of reliability). However, unfortunately, in addition to the fact that in most situations, Cronbach's Alpha is not a measure of reliability, Cronbach's Alpha has also been shown to be unrelated to a scale's internal consistency. 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). In other words (almost those of Sijtsma, 2009, p. 107, to be precise): Cronbach's Alpha has very limited usefulness. I therefore recommend that we abandon it.

Aside: note that I have kept these explanations deliberately conceptual. For example, I have conveniently neglected to even acknowledge the semantic swamp that one enters when trying to define reliability and internal consistency (instead, I worked from the assumption that many researchers use Cronbach's Alpha with a vague idea that it provides some information on reliability and/or internal consistency, whatever the precise definitions may be). However, for those readers interested in the technical background to these explanations, an extensive literature is available (Cortina, 1993; Dunn et al., 2013; Graham, 2006; Revelle & Zinbarg, 2009; Sijtsma, 2009). The goal of the current paper is not so much to provide yet another thorough argument of why Cronbach's Alpha should be abandoned; this has been done better than I can by people who understand the issues at hand much better. Instead, this paper is meant to make it easy to adopt a different approach than computing Cronbach's Alpha.

How to abandon Cronbach's Alpha

So, in most situations, we know that if we computed Cronbach's Alpha, the resulting value cannot possibly be the reliability of our scale. This of course begs the question of whether other measures exist that provide better estimates of a scale's reliability. The answer, of course, is yes[1]. Two have been recommended: the 'greatest lower bound' (glb; Sijtsma, 2009) and omega (Revelle & Zinbarg, 2009). Sijtsma (2009) argued that the glb is the lowest possible value that a scale's reliability can have. That means that when the glb is known, the reliability is by definition in the interval [glb, 1]. Revelle and Zinbarg (2009) argue that omega in fact provides a more accurate approximation of a scale's reliability, and that omega is almost always higher. For details on these measures, please see their respective papers; for now, we will focus simply on how to compute these superior estimates of reliability.

Both the glb and omega are available in the free and open source package R (R Development Core Team, 2014), and a step-by-step explanation of how to compute omega has even been published already (Dunn et al., 2013). However, this step-by-step explanation is not

1 Imagine, though, how awkward it would be if I would realise only at this point that none exist...

Open Access, limiting its accessibility to researchers and students. In addition, it involves using quite some R commands, and some researchers have become so accustomed to using SPSS that the idea of learning a new statistical package can seem somewhat daunting. Finally, as we health psychologists know, behavior change is facilitated by making the desired behavior easier to perform. This is where the current paper comes in: it introduces a so-called 'wrapper' function that enables researchers with no knowledge of R to compute a number of measures of reliability with one command. The minor catch is that before this function can be used, it needs to be downloaded and installed into R. However, like downloading and installing R itself, this needs to be done only once; and this, too, consists of only one command. The following paragraphs explain what R is, how to install it, how to install the required package, and how to request the glb and omega.

R is an open source statistical package. It has several advantages over SPSS, such as that it is free and that almost any existing statistical analysis is available. In addition, very accessible introductory texts exist (e.g. Field, Miles, & Field, 2012). It can be downloaded from http://r-project.org. Windows users who prefer to not install anything on their system (or are unable to) can download a portable version from http://sourceforge.net/projects/rportable/, which can even run from a USB stick. Once installed and started, R displays the console, an interface enabling users to input commands for R. The aptly named function 'install.packages' can be used to install packages. Specifically, to install the package we now require, run the following command:

```
install.packages('userfriendlyscience');
```

R will then ask the user to select a mirror. Simply select the geographically closest location, after which R will proceed to download the requested package 'userfriendlyscience' and all packages it depends on. Once the package 'userfriendlyscience' is installed, we need to tell R that we actually require it, using the function 'require', after which we can immediately compute the reliability estimates with 'scaleReliability':

```
require('userfriendlyscience');
scaleReliability();
```

R then presents a dialog where an SPSS datafile can be selected. The function 'scaleReliability' assumes that this datafile only

```
-- STARTING BOOTSTRAPPING TO COMPUTE CONFIDENCE INTERVALS! --
--    (this might take a while, computing 1000 samples)    --
-- FINISHED BOOTSTRAPPING TO COMPUTE CONFIDENCE INTERVALS! --
                 dat: dat.time1
                       Items: all
              Observations: 250
                       Omega: 0.8
Greatest Lower Bound (GLB): 0.85
          Cronbach's alpha: 0.75

Confidence intervals:
                     Omega: [0.74, 0.83]
         Cronbach's alpha: [0.71, 0.79]
```

contains items of one scale. Therefore, before heading into R, store an intermediate version of your datafile from SPSS by selecting the 'Save as …' option in the 'File' menu, in the resulting dialogue clicking the 'Variables…' button, and then using the 'drop' and 'keep' functionalities to select which variables to store. R then produces output similar to that showed at the bottom of the previous page.

Note that after having displayed the first two lines, R starts bootstrapping to generate the confidence intervals, which may take a while. The function scaleReliability has a number of other arguments that can be used, for example to specify which variables in the data should be used, whether to compute confidence interval in the first place, and how many samples to compute for the confidence interval bootstrapping. Interested readers can get more information by entering '?scaleReliability' in the R console. An example script that generates simulated data and computes these estimates (these exact estimates, in fact), as well as the output of the script, is provided at this paper's Open Science Framework page at http://osf.io/tnrxv.

As most researchers know, and as has been argued countless times before, the informational value of point estimates is negligible compared to the value of confidence intervals. However, SPSS does not normally provide confidence intervals for most of the statistics it reports, and this may have contributed to the phenomenon that researchers generally report only a point estimate for their reliability estimates. Hopefully, the fact that scaleReliability by default reports confidence intervals for Omega (and for the old-fashioned researchers among us, for Cronbach's Alpha) can contribute to a change in reporting standards for reliability estimates. Although it would be a huge improvement if researchers would from now on report confidence intervals for omega instead of, or in addition to, point estimates for Cronbach's Alpha, it might be even better to try and decrease our reliance on quantitative 'quality labels' for aggregate measures.

Multidimensional aggregated measures: indices

All measures of reliability discussed here share one important assumption: that of unidimensionality. Even this single assumption, however, is not always plausible. For example, many health psychology studies explore the relative importance of a variety of psychological determinants for predicting a given health behavior. Common determinants included in such studies are attitude, descriptive subjective norm, injunctive subjective norm, and perceived behavioral control. When the study is meant to inform the development of behavior change interventions, these determinants are usually defined as aggregate variables, measured with various items that each reflect a specific belief (Bartholomew, Parcel, Kok, Gottlieb, & Fernández, 2011; Fishbein & Ajzen, 2010). For example, beliefs underlying injunctive subjective norm reflect perceived approval or disapproval of social referents regarding the target behavior; beliefs underlying descriptive norm reflect perceived performance of the target behavior by social referents; and beliefs underlying perceived behavioral control reflect perceived environmental barriers and possessed skills. Imagine, for example, the following three items to measure descriptive norm: "My partner exercises [never-daily]", "My best friend exercises [never-daily]", "Of my colleagues, [none exercise-all exercise]", and the following three items to measure perceived behavioral control: "The sports facility is located [very far-very close] to my home", "For me, exercising

three times a week is [very hard-very easy]" and "A subscription to a sport club is [very expensive-very cheap]".

Most readers will probably feel it coming: these three descriptive norm items do not measure the same dimension, and neither do the perceived behavioral control items. Instead of being meant as repeated measurements of the same underlying unidimensional construct, these items are combined in one measure because aggregating the normative pressure experienced with regards to these different social referents provides a useful indicator of the total pressure experienced. If most of one's colleagues exercise, but one's partner and best friend rarely do, the descriptive norm is considerably lower than when one's partner and best friend also exercise. Similarly, there is no reason to assume that there is a correlation between the proximity of one's house to exercise facilities and one's assessment of the monetary costs of a membership at such facilities; but both measures likely contribute to a person's intention to exercise regularly and their subsequent behavior. Aggregating these measures despite the clear lack of unidimensionality is warranted on the basis of theory: for example, a theory might hold that a person's perceptions of social referents' behavior all influence that person's own intention and behavior in a similar fashion. If a researcher then wants to study the relative contribution of descriptive norms to the prediction of intention and behavior, aggregating these descriptive normative beliefs, which all exert their influence on intention and behavior in a similar manner, makes sense. This allows convenient comparison to the association strength of other determinants such as attitude and perceived behavioral control. To distinguish such deliberately multidimensional aggregate measures from intended unidimensional scales, I will refer to them as indices.

Although for indices, aggregation of the measures can be justified, computation of reliability or internal consistency measures cannot; after all, the assumption of unidimensionality has been violated. Nonetheless, it is not uncommon to see authors computing Cronbach's Alpha for variables such as subjective norm or perceived behavioural control that are measured with items assessing a variety of beliefs. Even worse, in the case of a low value, items might be removed to enhance Cronbach's Alpha, sometimes even causing authors to resort to single-item measures. This means the validity of the relevant measure is decreased on the basis of a flawed measure that should not have been computed in the first place. Of course, for indices, the assumption of the glb or omega would have been violated as well. And to make matters worse more challenging, to a degree this problem of multidimensionality holds for all psychological variables.

Reliability versus validity

The example given above used indices that are commonly adopted in health psychology, and showed how such measures are multidimensional, yet can still be useful aggregate measures. Other psychological variables, such as attitude, coping skills, or optimism, can more easily be argued to be unidimensional. However, even for these constructs, the different items used to measure them are usually not merely intended as exact replications of each other. Besides increasing reliability, a second reason for using multiple measurements to measure a construct is increased validity. Take for example these three items from the General Self-Efficacy (GSE) scale, all answered on a 4-point scale from "Not at all true" to "Exactly true": "I can always manage to

solve difficult problems if I try hard enough", "If someone opposes me, I can find the means and ways to get what I want" and "It is easy for me to stick to my aims and accomplish my goals" (see e.g. Luszczynska, Scholz, & Schwarzer, 2005). Each of these items taps quite different aspects of self-efficacy: the first item concerns self-efficacy regarding difficult problems, and imposes the condition of considerable investment of resources; the second item concerns general self-efficacy, but only under the circumstances where another person attempts to thwart goal-directed behavior; and the third item taps both self-efficacy and perceived self-regulatory skill. These three aspects are different, but all are part of the generic construct general self-efficacy. The GSE scale contains these items not to enhance reliability, but to enhance validity of the scale.

The fact that measures such as the GSE contain items that measure different aspects of a construct is not a weakness of the measure: rather, it is a strength. Very narrowly defined and measured psychological constructs have very limited applicability; in fact, most psychological constructs derive part of their usefulness from the generic level at which they are defined. For example, the Reasoned Action Approach recommends applying the principle of compatibility when measuring behavior and its determinants (Fishbein & Ajzen, 2010). This principle assumes that any behavior has four defining elements (action, target, context, and time), and dictates that behavior and its determinants must be measured with regards to the exact same action, target, context, and time. For example, when measuring EHPS conference attendance and its determinants, an intention item might be "Will you attend the EHPS conference in 2014? [absolutely not-absolutely]", a subjective norm item might be "How many of your colleagues will attend the EHPS conference in 2014? [none-all]", and a self-efficacy item might be "How easy or hard will it be for you to attend the EHPS conference in 2014? [very easy-very hard]". The measure of self-efficacy acquired this way will have extremely high applicability when predicting EHPS conference attendance in 2014, but it will be almost useless for anything else (such as predicting exercise behavior). By contrast, general self-efficacy is useful to predict a broad range of behaviors precisely because of its generic nature. Thus, many psychological constructs derive their usefulness from their relatively broad definition, and therefore, their relatively broad operationalization.

At the same time, the fact that different aspects of a psychological construct are measured means that the measure can never be perfectly unidimensional. Although an individual's response to each item should normally be determined mainly by the psychological construct of interest, other psychological constructs will have an influence as well; and accordingly, factor analysis may reveal that the first factor explains a disappointingly low proportion of variance. However, this does not have to be a problem: after all, if a set of items measures a very generic psychological construct, influence of related psychological constructs is to be expected. Scale diagnostics cannot be interpreted without taking into account how specific or generic the measured construct is defined. Therefore, scale inspection should entail more than computation and evaluation of a single quantitative measure.

A comprehensive assessment of scale quality

If we acknowledge that aggregate measures contain different items to enhance both

reliability and validity, and that more specific, more narrowly operationalized measures are not by definition better than more generic, more broadly operationalized measures, it becomes even harder to defend thresholds for estimates such as Cronbach's Alpha. Even when refraining from relying on tentative thresholds, Cronbach's Alpha, omega, and the glb provide only a very narrow view on the dynamics of a scale. In addition, it seems useful to examine the degree of unidimensionality of a scale by conducting a factor analysis (or principal component analysis, depending on the goal), and inspecting the Eigen values of each component, as well as the factor loadings. Furthermore, inspecting the distribution of each item, as well as the way the items are associated, can help identify anomalies in single measures. Therefore, I suggest that researchers routinely generate a combination of diagnostics:

1. Compute omega, the glb, and Cronbach's alpha, preferably with confidence intervals;
2. Conduct a factor analysis or principal component analysis and inspect all Eigen values and the factor loadings (at least for the first factor);
3. Inspect the means, medians, and variances for each item;
4. Generate a correlation matrix;
5. Inspect the scatterplots of the associations between all items;
6. Inspect histograms of each item's distribution.

These diagnostics should then be interpreted in conjunction with the separate measurements of the aggregate measure (e.g. the complete list of the items forming a scale in a questionnaire). Unfortunately, inspecting such a diverse combination of diagnostic information means that providing clear guidelines as to when a scale is acceptable becomes impossible. Of course, that was more or less the point of this contribution: because operationalization and measurement are so important to psychological science, assessment of successful operationalization deserves more attention than simple comparison to a quantitative threshold. Conveniently, the R package described above just so happens to contain another function called 'scaleDiagnosis', which provides most of these diagnostics. It can be used the same way 'scaleReliability' is used:

```
scaleDiagnosis();
```

The user can then select an SPSS datafile, after which the function produces output similar to that shown on the next page. The function also creates a plot similar to the one shown in Figure 3[2]. This so-called scattermatrix shows the (bivariate) scatterplots of the combinations of all items in the scale, as well as the univariate distribution of each item, and the point estimates for the correlation coefficients in the upper right half. This is useful for quick visual inspection of the nature of the associations between the items and their distributions. This output, the text as text file and the plot both as .png and .svg, is also available at this paper's Open Science Framework page at http://osf.io/tnrxv.

However, forgoing the comfort of a quantitative threshold means that decisions about scale construction become much more subjective. It seems wrong to on the one hand acknowledge the importance and complexity of these decisions, and on the other hand, forgo the convenient possibility of external scrutiny that quantitative measures such as Cronbach's Alpha seem to afford. And indeed, this would be wrong. The problems of the so-called 'researcher degrees of freedom' have been made painfully clear recently (Simmons, Nelson, & Simonsohn, 2011), and the solution is straightforward: as argued before in the European Health Psychologist, researchers should fully disclose

```
                dat: res$dat
                    Items: t0_item1, t0_item2, t0_item3, t0_item4, t0_item5
             Observations: 250
                    Omega: 0.8
Greatest Lower Bound (GLB): 0.85
          Cronbach's alpha: 0.75

Eigen values: 2.924, 0.64, 0.566, 0.463, 0.407
Loadings:
         PC1
t0_item1 0.76
t0_item2 0.78
t0_item3 0.75
t0_item4 0.78
t0_item5 0.75

                PC1
SS loadings    2.92
Proportion Var 0.58

         var   n  mean   sd median trimmed   mad   min   max range  skew kurtosis   se
t0_item1   1 250 17.94 2.92  18.00   17.97  2.92 10.93 24.68 13.75 -0.08    -0.54 0.18
t0_item2   2 250 31.70 9.52  31.05   31.66  9.06  2.62 58.49 55.87  0.04    -0.01 0.60
t0_item3   3 250 20.26 3.53  19.99   20.26  3.71 11.81 30.20 18.39  0.06    -0.51 0.22
t0_item4   4 250 34.48 5.99  34.14   34.32  5.80 20.14 56.36 36.22  0.32     0.21 0.38
t0_item5   5 250 29.78 9.73  29.63   29.69 10.09  4.60 56.98 52.38  0.08    -0.16 0.62
```

(Peters, Abraham, & Crutzen, 2012). In this case, such disclosure would mean making these diagnostics public, along with the complete questionnaires that were used. Preferably, these resources are published in an Open Access online repository such as the free Open Science Framework (see https://osf.io/), as this makes them available to the entire scientific community. This will be of considerable use to other researchers who are constructing similar measurement instruments. At the very least, researchers should publish these scale diagnostics as supplementary materials with their articles[3]. Publishing the scale diagnostics will enable reviewers to critically and thoroughly assess the integrity of the used measurement instruments, and can facilitate both interpretation of the findings and future meta-analysis.

The question then becomes, what do we know about the quality of measurement instruments of studies that only report Cronbach's Alpha? The answer is, very little. We know that the reliability is in any case not the value reported for Cronbach's alpha (but by definition something higher, although we have

2 To store a plot in R, the 'Save as' option in the 'File' menu can be used.

3. Although this is less desirable, as it will restrict access to this information if the main article is behind a paywall.

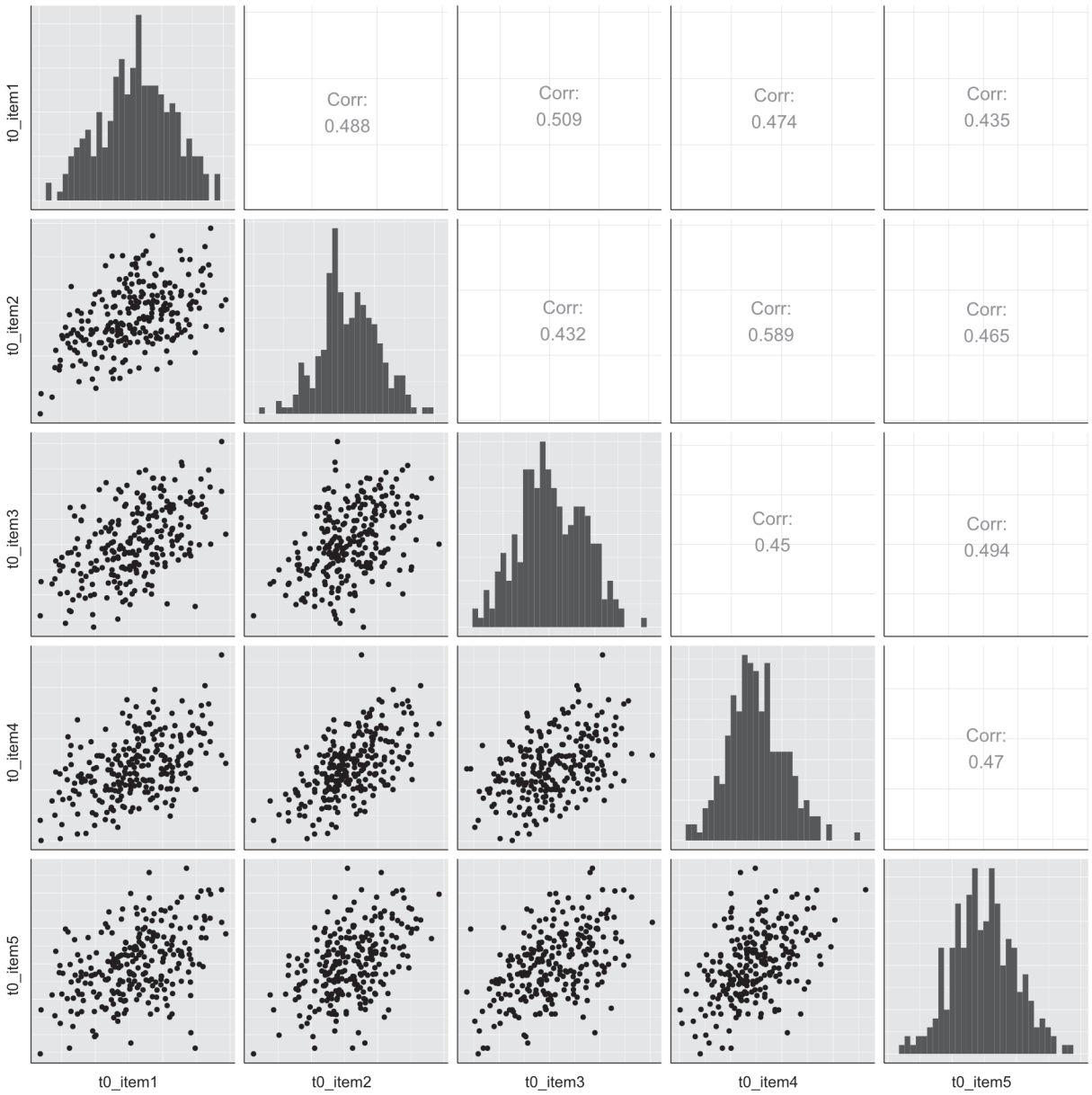

Figure 3: A scattermatrix as produced by the scaleDiagnosis() function in the userfriendlyscience package for R

no clue as to how much higher). We know nothing about the internal consistency of the scale. For those studies that published the questionnaires as appendices or supplemental materials, it is possible to inspect the items to establish the face validity (i.e. whether the items seem to tap cognitions/emotions that make up or contribute to the construct the scale intends to measure); and if correlation tables were published as well, a more thorough assessment of the measurement instruments becomes possible. However, without such information, we know almost nothing about the validity and reliability of the used measures. If we assume that the validity and reliability of the measurement instruments used in most

studies are acceptable, the only remaining problem is that we don't know which studies are the ones with unacceptable measures.

Conclusion

Researchers often compute and report Cronbach's Alpha to determine whether aggregate measures have acceptable reliability or internal consistency. Although most authors and reviewers seem content with this, Cronbach's Alpha is both unrelated to a scale's internal consistency and a fatally flawed estimate of its reliability. In addition, this reliance on one quantitative estimate fails to acknowledge the relationship between reliability and validity. Finally, some measures are deliberately multidimensional (indices), violating the assumption of unidimensionality underlying Cronbach's Alpha, omega and the Greatest Lower Bound. Scale diagnostics would be improved if researchers would assess, simultaneously, estimates and their confidence intervals for omega, the glb, and perhaps Cronbach's Alpha; Eigen values and factor loadings; individual item distributions; and a correlation- and scattermatrix of all items. These diagnostics should be assessed in conjunction with the raw measurement instrument (e.g. the items in a scale). This will enable researchers to base their decisions on a more complete picture of scale performance. In addition, publishing these diagnostics and the measurement instruments will enable reviewers and readers to closely scrutinize the reliability and validity of such measures. Finally, such a process will enable considerable acceleration of scale construction in general, as it will become possible to spot and study item formulations that consistently perform badly. It is important not to underestimate the importance of how we measure our psychological variables of interest, since psychologists do not have the luxury of the more objective measures that many other disciplines use (after all, even implicit and biopsychological measures are indirect and require many assumptions). Hopefully, this paper and the R functions described herein will have made it sufficiently easy for this more comprehensive assessment of scale quality to become commonplace.

References

Bartholomew, L. K., Parcel, G. S., Kok, G., Gottlieb, N. H., & Fernández, M. E. (2011). *Planning health promotion programs: an Intervention Mapping approach* (3rd ed.). San Fran-cisco, CA: Jossey-Bass.

Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and application. *Journal of Applied Psychology, 78(1), 98–104. doi:10.1037/0021-9010.78.1.98*

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555

Dunn, T. J., Baguley, T., & Brunsden, V. (in press). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*. doi:10.1111/bjop.12046

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics using R*. London: Sage Publications Ltd.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: the reasoned action approach*. New York: Psychology Press.

Graham, J. M. (2006). Congeneric and (Essentially) Tau-Equivalent Estimates of Score Relia-bility: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930–944. doi:10.1177/0013164406288165

Luszczynska, A., Scholz, U., & Schwarzer, R. (2005). The general self-efficacy scale:

multicul-tural validation studies. *The Journal of Psychology, 139*(5), 439–57. doi:10.3200/JRLP.139.5.439-457

Peters, G.-J. Y., Abraham, C. S., & Crutzen, R. (2012). Full disclosure: doing behavioural science necessitates sharing. *The European Health Psychologist, 14*(4), 77–84.

R Development Core Team. (2014). *R: A language and environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.r-project.org/

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: comments on Sijtsma. *Psychometrika, 74*(1), 145–154. doi:10.1007/s11336-008-9102-z

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–66. doi:10.1177/0956797611417632

Gjalt-Jorn Y. Peters
is Assistant professor of Psychology at the Open University of the Netherlands, Heerlen, the Netherlands

gjalt-jorn@behaviorchange.eu