commentary

Time is a jailer: what do alpha and its alternatives tell us about reliability?

Rik Crutzen

Maastricht University

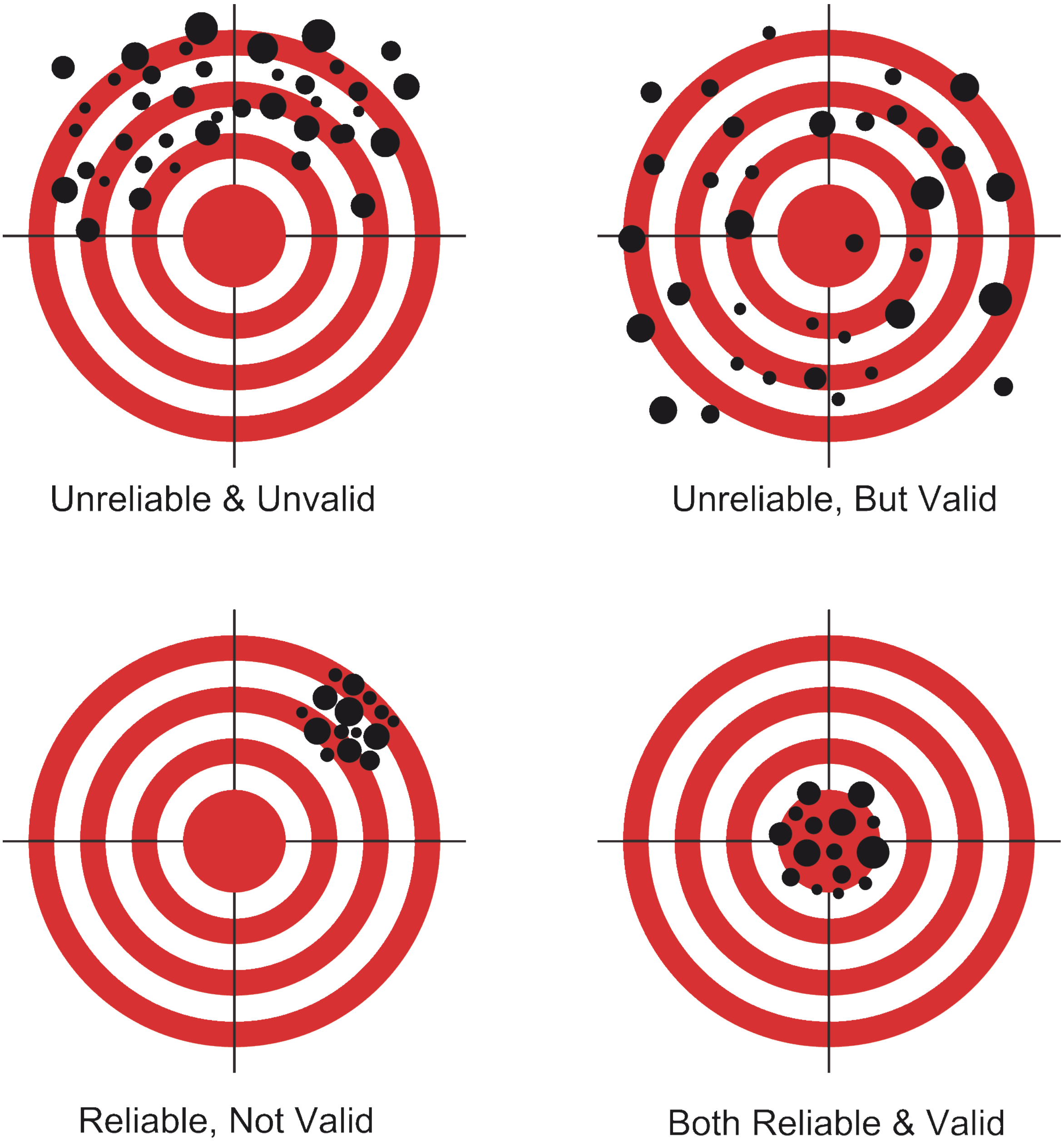
Psychologists do not have it easy, but the article by Peters (2014) paves the way for more comprehensive assessment of scale quality. I plead guilty to habitually reporting alpha[1] and – in some cases where I did not – reviewers were so kind to request this as well. Peters (2014) rightly states that alpha is a fatally flawed estimate of a scale's reliability. He presents readily available alternatives, such as the greatest lower bound (glb) or omega, as superior estimates of reliability. I agree with the suggestion by Peters (2014) to routinely generate a combination of diagnostics, but I think we are still missing out on an important aspect of reliability: test-retest reliability.

Figure 1 might bring flashbacks to your Statistics 101-course. My apologies if this side effect is an unpleasant experience. The figure is very useful, however, to explain test-rest reliability. Whereas Peters (2014) discusses items within a scale (e.g., attitude items), I will focus on the scale in its entirety (e.g., an attitude measure). So, each dot in Figure 1 represents, for example, a single administration of an attitude measure. The closer these dots are to the bull's eye, the more likely that the scale actually measures attitude. This concerns the validity of the scale. However, if we use the same measure repeatedly over time, we also want to be sure that we get the same score (if nothing has changed). So, the dots should be close to each other (or, ideally, overlap each other). This is an aspect of a scale's reliability.

A legitimate question to ask is why time is such an important factor contributing to reliability? The reason behind this is that over time both true scores and measurement error can fluctuate. The observed test score (e.g., a participant's score on an attitude measure) is the sum of the true score (e.g., a participant's actual attitude) and the measurement error. This measurement error does not only differ between participants or items within a scale, but also within participants over time (Guttman, 1945). At the same time, however, differences in the observed test can also be the result of actual changes in attitude.

Imagine an intervention targeted at the attitude towards use of protective clothing to prevent tick bites (see e.g., Crutzen & Beaujean, 2014 for brief background information). The efficacy of this intervention is tested in a two-arm randomized controlled trial (RCT) with a waiting-list control group. No change is expected in the control group, and differences in the intervention group should reflect differences in true scores regarding attitude. This does not mean that test–retest reliability is only desirable in measures of constructs that are not expected to change over time. It can be, for example, that a national health campaign about prevention of tick bites is launched during the trial period. This might lead to changes in attitude of the control group as well. Therefore, an important aspect of assessing scale quality is

[1] Cronbach considered it an embarrassment that the formula became conventionally known as Cronbach's alpha (Cronbach & Shavelson, 2004).



Unreliable & Unvalid

Unreliable, But Valid

Reliable, Not Valid

Both Reliable & Valid

Figure 1. Reliability and validity (© Nevit Dilmen).

to distinguish between changes in observed scores due to actual changes in, for example, attitude (even if they are unexpected) and changes in measurement error due to time (also known as transient error).

The magnitude of transient error in real data can range from non-existent to very large (Becker, 2000). Ignoring transient error can lead to inaccurate conclusions (Chmielewski & Watson, 2009). Even though I agree with the suggestions by Peters (2014), they are not sufficient to address transient error. It appears that high internal consistency does not indicate that a scale can measure change reliably, nor can it estimate stability of true scores. The opposite is also true; a low internal consistency does not attenuate stability (McCrae, Kurtz, Yamagata, & Terracciano, 2011).

Sijtsma (2009) questions the estimation of reliability on the basis of a single administration of a scale, even when using alternatives to alpha such as the glb. Nevertheless, there are indices that go beyond traditional correlate coefficients and that explicitly take transient error into account. Green (2003), for example, explains an index based on coefficient alpha, but susceptible to transient error, whereas Schmidt, Le, and Ilies (2003) present a procedure for estimating the coefficient of equivalence and stability (CES). Test-retest data is required for these indices. Huysamen (2006) argues that "the very reason for the original coefficient's popularity has been that it doesn't require a retest, and Green's coefficient has to forgo this luxury, as any other index that wishes to reflect transient error by definition has to do."

This leaves us at a crossroad. We more or less ignore transient error and simply go on[2] or we agree that test-retest analyses should be part of comprehensive assessment of scale quality. In case of the latter, we have to acknowledge that this brings additional workload. This additional workload does not only concern the need for test-retest data, but assessing test-retest reliability also brings along additional issues. For example, the choice of an appropriate retest interval[3] (Chmielewski & Watson, 2009; Green, 2003) and comparison of the indices between domains (Schmidt et al., 2003). I hope that the arguments presented in this article are convincing to take on this additional workload.

To make this workload as minimal as possible, I will now briefly explain how to easily compute these indices. First, install R and the package 'userfriendlyscience' (see Peters, 2014). Then, in your commonly used statistical environment (e.g., SPSS), create a data file that only contains the items of the two administrations of your scale. The order is important: the items of the first administration should come first, followed by, in the same order, the items of the second administration (e.g., "t0_item1", "t0_item2", and "t0_item3" followed by "t1_item1", "t1_item2", and "t1_item3"). Then, load this data file into R and compute the test-retest alpha coefficient and the CES with:

```
testRetestReliability();
```

R again opens a dialog to enable selection of the data file, after which output similar to the below will be shown.

Note that computing the single administration indices (e.g., original coefficient alpha, omega, and the glb, computed in Peters, 2014) yielded much higher values (.75-.85). This means that when using this scale and computing single-administration reliability indices, one might erroneously assume a negligible effect of transient error, which might have far-reaching consequences in non-experimental designs.

In the ideal situation, we choose to conduct test-rest analyses as part of comprehensive assessment of scale quality, but how do we achieve this? A (too) simple, but nonetheless recommendable, first step would be to conduct test-retest analyses whenever longitudinal data are available (e.g., after conducting an RCT). It would be far better to conduct a pre-test to assess test-retest reliability, using the indices mentioned above. In such a pre-test, the choice of retest interval should be grounded theoretically depending on the construct of interest (Chmielewski & Watson, 2009). This

[2] Following the suggestions by Peters (2014) is already a big step forward.
[3] E.g., a personality trait measure might be less likely to change over time in comparison with an attitude measure.

```
                         Items at time 1: t0_item1, t0_item2, t0_item3, t0_item4, t0_item5
                         Items at time 2: t1_item1, t1_item2, t1_item3, t1_item4, t1_item5
                             Observations: 250
           Test-retest Alpha Coefficient: 0.43
Coefficient of Equivalence and Stability: 0.45

To help assess whether the subscales (automatically generated using means) are parallel, here
are the means and variances:
Mean subscale a1, time 1: 82.19 (variance = 235.07)
Mean subscale a2, time 1: 51.95 (variance = 132.18)
Mean subscale a1, time 2: 83.57 (variance = 243.19)
Mean subscale a2, time 2: 52.09 (variance = 138.76)
```

might all sound as "yet another thing to do before I can run my study". However, if we agree on the importance of pre-testing our intervention materials to avoid counterproductive results (e.g., Whittingham et al., 2009), I think we should be as strict with regard to the measures of constructs we are interested in. After all, we draw our conclusions based on these measures and we should not try "to explain findings that result from transient error masquerading as true change" (Chmielewski & Watson, 2009).

Acknowledgements

I would like to thank Gjalt-Jorn Y. Peters for adding the test-retest alpha coefficient and the CES to the userfriendlyscience R package.

References

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5*(3), 370-379. doi:10.1037/1082-989X.5.3.370

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*(1), 186-202. doi:10.1037/a0015618

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

Crutzen, R., & Beaujean, D. (2014). Preventive behaviours regarding tick bites. *BMJ, 348*, g231. doi:http://dx.doi.org/10.1136/bmj.g231

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8*(1), 88–101. doi:10.1037/1082-989X.8.1.88

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282. doi:10.1007/BF02288892

Huysamen, G. K. (2006). Coefficient alpha: unnecessarily ambiguous; unduly ubiquitous. *SA Journal of Industrial Psychology, 32*(4), 34-40. doi:10.4102/sajip.v32i4.242

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15*(1), 28-50. doi: 10.1177/1088868310366253

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*(2), 54-67

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: an empirical examination of the effects of different sources of measurement error on realibility estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. doi:10.1037/1082-989X.8.2.206

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

Whittingham, J. R. D., Ruiter, R. A. C., Bolier, L., Lemmers, L., Van Hasselt, N., & Kok, G. (2009). Avoiding counterproductive results: an experimental pretest of a harm reduction intervention on attitude toward party drugs among users and nonusers. *Substance Use & Misuse, 44*(4), 532-547. doi:10.1080/10826080802347685

Rik Crutzen
is Assistant Professor at the Department of Health Promotion, Maastricht University, Maastricht, The Netherlands

rik.crutzen@maastrichtuniversity.nl