Cognadev Technical Report Series

3

16th September, 2014

Theta, Alpha, and Omega estimates of reliability for the VO orientation scales

The three forms of internal-consistency reliability of the VO orientations, at the item level of analysis.

cognadev

Paul Barrett, PhD.
Chief Research Scientist

18B Balmoral Ave
Hurlingham Sandton 2196
Johannesburg
South Africa

Tel: +27 11 884-0878
Fax: +27 11 884-0910

19 Carlton Road
Pukekohe
Auckland 2120
New Zealand

Tel: +64 9 889-2630
Mob: +64 21-415625
Fax: +64 9 280-6121
Skype: pbar088

www.cognadev.com paul@cognadev.com

Executive Summary

1 Using the VO Technical Manual mixed-gender sample dataset of n=3,683 cases, four homogeneity (reliability) coefficients were computed for every values orientation 'scale':

- Armor's theta,
- Cronbach alpha,
- Heise and Borhnstedt's omega, and
- Hancock and Mueller's H-omega.

Each index varies between 0 (no reliability) to 1.0 (Maximum possible reliability).

2 The median, minimum and maximum reliability coefficients across both accepted and rejected Values Orientations are:

VO Colour Orientations	α	θ	ω	ω_H
Accepted				
Median	.87	.88	.88	.88
Min-Max	.84-.89	.85-.89	.84-.89	.86-.90
Rejected				
Median	.84	.86	.84	.87
Min-Max	.81-.87	.81-.88	.81-.88	.82-.88

α = coefficient alpha

θ = Armor's theta (PCA)

ω = Heise & Borhnstedt omega (MLFA)

ω_H = Hancock & Mueller H-omega (MLFA)

3 While these results are encouraging, it must be understood that when looking carefully at the VO scoring methodology *(detailed within the Technical Manual for the VO)*, it is apparent that no conventional psychometric approach to assessing single-occasion reliability is strictly applicable to the VO assessment data. The reasons are twofold:

- Although unit-weight core-items might be analysed for an orientation, the non-core items are weighted uniquely for each individual, relative to the scaled *coreprop* constant for that individual for a particular orientation. So, no consistent set of weights are applied to the non-core items. Orientation-score is the average of summed core + unique-per-case non-linearly scaled/weighted non-core items.
- Each item may contribute to the scoring across up to 4 orientations simultaneously, with the same item scored/weighted differentially for Accepted and Rejected orientations. This in part reflects the 3-dimensional spiral model for the orientations, where orientation boundaries are 'fuzzy' rather than discrete.

So, the results reported here are simply a basic confirmation that the items assessing each orientation are at least homogenous in content, with respect to how individuals interpret their meaning.

4 The most appropriate estimate of test reliability for the VO is retest reliability. Technical Report #2 (downloadable from: http://www.pbarrett.net/cognadev/#tech_series) provides the methodology and available evidence supporting the retest reliability of the VO over several between-assessment durations.

Contents

Tables

1. Reliability in the Context of the VO Scoring Methodology

This is a complex issue. Consider the fundamental definition of reliability found within any science as well as in everyday usage:

> Reliability is the extent to which a second, third and onwards observations of an event, measure, rating, or occurrence, deviates from the first or proceeding observations. If they are exactly the same, there is perfect reliability. In engineering terms, reliability is referred to as repeatability.

The repeatability problem for psychological assessment is noted by Guttman as far back as 1945 in his article entitled: "A basis for analyzing test-retest reliability"… p. 256:

> "The problem of reliability is of course not peculiar to psychology or sociology, but pervades all the sciences. In dealing with empirical data in any field, the question should be raised: if the experiment were to be repeated, how much variation would there be in the results?"

And on page 257:

> "(3) A major emphasis of this paper is that the reliability coefficient cannot in general be estimated from but a single trial-that items do not replace trials. If two trials are experimentally independent, then we show that the correlation between two trials is, with probability of unity, equal to the reliability coefficient.
>
> (4) As is well known, there may be great practical difficulties in making two independent trials; therefore our principal focus is on *what information can be obtained from a single trial.* We find that *lower bounds* to the reliability coefficient can be computed from a single trial. Six different lower bounds are derived, appropriate for different situations. Several of these bounds are as easy as or easier to compute than are conventional formulas, and all of the bounds assume less than do conventional formulas.
>
> (5) To prove that bounds can be computed from a single trial, we use essentially one basic assumption: that the errors of observation are independent between items and between persons over the *universe of trials.* In the conventional approach, independence is taken over *persons* rather than trials, and the problem of observability from a single trial is not explicitly analyzed."

Therein lies the major assumption required to be made in assessing reliability over a single occasion – we must invoke a hypothetical *universe of trials* from which a single occasion (trial) has provided observations. By making an assumption based upon statistical sampling theory, and invoking the concept of a universe of items which 'measure' a single attribute , from which a random sample has been drawn by the investigator *(the items in any particular test)*, it is possible to generate a variety of bounds for reliability, which is exactly what Guttman achieved in his article.

Cronbach (1951) extended Guttman's work and introduced the now famous Cronbach alpha coefficient. This made reliability assessment a routine feature of analysis, augmenting the simple definition of reliability as repeatability by using that key assumption of an investigator sampling items from a 'universe' of items, with the additional propositions stating that individuals can be administered a multitude of parallel tests drawn from that universe, and that each individual's observed score is an additive function of a hypothetical true score and error.

From this assumption-laden test-theory definition other kinds of reliability coefficients were constructed, such as omega, factor validity, and the complex indices that have been constructed recently in Structural Equation Modeling *(Cortina (1993), Schmitt (1996), Green and Yang (2009) provide good overviews)*. However, all these

indices depend for their validity upon assumptions made about test scores as quantities, hypothetical true scores, and hypothetical item universes.

It is worth summarising three critical assumptions which must true be 'as stated' for a coefficient alpha reliability coefficient to possess validity:

1. A person's observed score is a function of a hypothetical true score + measurement error.
2. A set of items within a test have been sampled randomly from a hypothetical universe of items.
3. The universe of items is unidimensional; that is, the items cannot simultaneously be measures of other dimensions.

Firstly, there is no empirical evidence that any individual actually possesses a 'true score' for any attribute. Indeed, as Borsboom and Mellenburg (2002) explain:

"Although providing a definition of psychological constructs and construct scores is generally difficult, the true score has a clear definition in classical test theory. This allows us to proceed from this definition to show that, upon any reasonable conceptualization of construct scores, these scores are not true scores. We first show that the identification of true scores with construct scores confounds issues of validity and issues of reliability. Second, we emphasize that such an identification is logically inconsistent with classical test theory itself. The true score is one of the central concepts of classical test theory; actually, the derivation of the theory begins with a definition of the true score (Lord & Novick, 1968, p. 30). The true score of a subject is defined as the expected value of the observed scores, where *the expectation is taken over an infinitely long run of independent repeated observations*. So, for a person taking a psychological test, that person's true score is defined as the expected value of the observed scores over an infinitely long run of repeated independent administrations of that test. Such a long run of observations is unrealistic, of course, because human beings typically learn, fatigue, and change in many other ways during repeated administrations. As a result, repeated observations will not be statistically independent.

Because the notion of independent replications is critical for the introduction of the probability model that Lord and Novick want to use, however, they introduce a thought experiment. In this thought experiment, the subject is brainwashed between each successive pair of measurements, so that the resulting observations may safely be considered independent. This allows Lord and Novick to define the true score as the hypothetical expectation over an infinite series of independent measurements. Of course, they readily admit that this definition is based on counterfactual premises and therefore has a limited interpretation. They use it primarily because it is mathematically convenient in defining some important concepts in classical test theory. For example, if observed scores are conceived of as composed of a true score plus random error (giving the classic equation Observed = t_{true} + E_{rror}), it follows from Lord and Novick's definition of the true score that the expectation of the error scores is zero. This yields mathematically simple expressions for concepts such as reliability. The above definition of the true score, as the expected value over replications, allows Lord and Novick to define reliability mathematically as the ratio of true score variance to observed score variance across the population: Reliability is the proportion of true score variance that can be linearly predicted from the observed scores in a population of subjects (Mellenbergh, 1996).

Thus, in classical test theory, the true score is defined as the expectation of a hypothetical series of observed scores. Consequently, within the framework of classical test theory, the true score does not necessarily reflect a construct score, and it is certainly not identical with it— either by definition, by assumption, or by hypothesis. Classical test theory does, as a matter of fact, not assume that there is a construct underlying the measurements at all. From the point of view of classical test theory, literally every test has a true score associated with it. " (p. 507)

So, we see that in reality, the entire 'true-score' assumption is unrealistic from a psychological perspective, but required from a statistical perspective in order to generate 'reliability' estimates from single-occasion observations. Instead of dealing with reliability directly in the context of human psychology, psychometricians

have instead invented a strange language around test scores and rather obtuse statistical procedures to assess what is really a very straightforward construct.

Secondly, the assumption that a test's items have been randomly sampled from a hypothetical universe of items is remarkable. No one in their right mind would claim that any psychological test (scale) consists of items randomly sampled from some (hypothetical) population of items. But, any scale constructor needs to say this in all seriousness before computing a coefficient alpha and claiming it is a valid measure of test reliability.

Thirdly, it is in principle valid to test a claim that items within a test form a unidimensional set. But, other procedures for assessing reliability are now required, based around factor analysis and the concept of a latent variable (see Revelle & Zinbarg, 2009); we are now replacing a hypothetical universe of items with an equally hypothetical latent variable (Maraun & Halpin, 2008). However, there is a simple underlying logic to the omega reliability coefficients that stands some scrutiny; that is, items which are claimed to be indicators of a construct should show a 'commonality' among themselves, as indexed by the magnitude of their factor loadings on a single component or latent common factor (Armor, 1974; Heise and Borhnstedt, 1970, Hancock and Mueller, 2001). This logic can also be seen within an item-response-theory perspective, where items are required to be located on a single latent variable continuum, ordered by 'facility/difficulty'.

However, when looking carefully at the VO scoring methodology *(detailed within the Technical Manual for the VO)*, it is apparent that no conventional psychometric approach to assessing single-occasion reliability is applicable to the VO assessment data. The reasons are twofold:

- Although unit-weight core-items might be analysed for an orientation, the non-core items are weighted uniquely for each individual, relative to the scaled *coreprop* constant for that individual for a particular orientation. So, no consistent set of weights are applied to the non-core items. Orientation-score is the average of summed core + unique-per-case non-linearly scaled/weighted non-core items.
- Each item may contribute to the scoring across up to 4 orientations simultaneously, with the same item scored/weighted differentially for Accepted and Rejected orientations. This in part reflects the 3-dimensional spiral model for the orientations, where orientation boundaries are 'fuzzy' rather than discrete.

As within any measurement framework, the most appropriate measure of Values Orientation reliability is retest reliability, especially as assessment results are not reported in numerical score-form at all, but as a series of selected 'class-category' orientations. Although retest score-agreement measures could be constructed from the mean orientation scores, the most appropriate measure of reliability will be a category pattern-match coefficient, which indexes agreement between reported category patterns/constituents. Technical Report #2 (downloadable from: http://www.pbarrett.net/cognadev/#tech_series) provides the methodology and available evidence supporting the retest reliability of the VO over several between-assessment durations.

But, for those who would still like to see an estimate of specific orientation reliability with which they may be most familiar, even though none are particularly meaningful for the VO items beyond approximately indexing item homogeneity, we have computed four kinds of scale reliability:

- Armor's theta,
- Cronbach alpha,
- Heise and Borhnstedt's omega, and
- Hancock and Mueller's H-omega.

Each index varies between 0 (no reliability) to 1.0 (Maximum possible reliability).

1.1 Sample Data

Using the n=3683-case mixed-gender dataset used within the technical manual for all basic statistical parameters and orientation analyses. The results of the reliability analyses are presented in Table 1.

1.2 Armor's Theta

Principal component analysis of raw item responses (from the slider responses to each item) were used to calculate the PCA eigenvalue for theta (Armor, 1974):

$$\theta = \left(\frac{k}{k-1}\right)\left(1-\frac{1}{eig_{pca}}\right)$$

where

$k =$ number of items in a scale

$eig_{pca} =$ the 1st PCA eigenvalue for the k items

1.3 Coefficient Alpha

This index was calculated using the same responses (Cronbach (1951)):

$$\alpha = \left(\frac{k}{k-1}\right)\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_y^2}\right)$$

where

$k =$ number of items in a scale

$\sum_{i=1}^{k}\sigma_i^2 =$ sum of the individual item variances

$\sigma_y^2 =$ the total sum-score (test-score) variance

1.4 Heise and Borhnstedt's omega

This was computed from the loadings on the first maximum likelihood factor extracted from a common factor analysis of the raw responses (Heise and Borhnstedt, 1970):

$$\omega = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2}{\left(\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i=1}^{k}\left(1-\lambda_i^2\right)\right)}$$

where

$k =$ number of items in a scale

$\lambda =$ maximum likelihood item loading on 1st factor

1.54 Hancock and Mueller's H-omega

This was also computed from the loadings on the first maximum likelihood factor extracted from a common factor analysis of the raw responses (Hancock and Mueller, 2001):

$$\omega_H = \frac{\sum_{i=1}^{k}\left(\frac{\lambda_i^2}{1-\lambda_i^2}\right)}{1+\sum_{i=1}^{k}\left(\frac{\lambda_i^2}{1-\lambda_i^2}\right)}$$

where

$k =$ number of items in a scale

$\lambda =$ maximum likelihood item loading on 1st factor

2. The Homogeneity Coefficients

Table 1: Reliability indices for the accepted and rejected values orientations

VO Colour Orientation	No. of Items	α	θ	ω	ω_H
Accepted					
Purple	31	.858	.868	.861	.875
Red	33	.871	.884	.878	.89
Blue	33	.88	.891	.886	.897
Orange	27	.871	.879	.876	.883
Green	33	.888	.893	.89	.897
Yellow	32	.865	.872	.869	.875
Turquoise	23	.84	.849	.841	.859
Rejected					
Purple	30	.863	.872	.869	.875
Red	22	.857	.863	.859	.87
Blue	25	.806	.813	.808	.817
Orange	25	.821	.828	.823	.835
Green	22	.825	.835	.825	.847
Yellow	28	.841	.855	.842	.868
Turquoise	29	.873	.878	.876	.883

α = coefficient alpha

θ = Armor's theta (PCA)

ω = Heise & Borhnstedt omega (MLFA)

ω_H = Hancock & Mueller H-omega (MLFA)

These are all high, with ω_H being the 'state of the art' recommended scale reliability computed directly from a Maximum Likelihood Factor Analysis (MLFA), single-factor solution. **All reliability estimates are > 0.80**. The estimates are a basic confirmation that the items assessing each orientation are at least homogenous in content, with respect to how individuals interpret their meaning.

References

Armor, D. J. (1974). Theta reliability and factor scaling. *Sociological Methodology*, 1973-1974, 5, 1, 17-50.

Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: a comment on Schmidt and Hunter. *Intelligence,* 30, 505-514.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 1, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Green, S.B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74, 1, 121-135.

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

Hancock, G.R., & Mueller, R.O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S.R. Du Toit, & D. Sörbom (Eds.). *Structural equation modeling: Present and future. A Festchrift in honor of Karl Jöreskog* (Chapter 10, pp 195-216). Illinois: Scientific Software International Inc.

Heise, D.R., & Borhnstedt, G.W. (1970). Validity, invalidity, and reliability. In E. F. Borgatta and G. W. Bohrnstedt (Eds.). *Sociological Methodology* (pp 104-129). San Francisco: Jossey Bass.

Maraun, M.D., & Halpin, P.F. (2008). Manifest and latent variables. *Measurement: Interdisciplinary Research & Perspectives*, 6, 1&2, 113-117.

Revelle, W., & Zinbarg, R.E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 1, 145-154.

Schmitt, N. (1996). Uses and Abuses of Coefficient Alpha. *Psychological Assessment*, 8, 4, 350-353.