Cronbach's α and the Person Separation Index (PSI)

Cronbach's α of traditional test theory and the Person Separation Index (PSI) from the Rasch model are indices formed as an estimate of the proportion of the true, or error free, variance of the distribution of person estimates relative to the sum of this variance and the error variance in these estimates. Within traditional test theory, this proportion is referred to as *reliability*. This is a central index in traditional test theory, but not in modern or item response theory, of which the Rasch model is a distinctive case. In modern test theory, the emphasis is on the precision of the estimate of each person. Nevertheless, the proportion of true variance relative to the true and error variance is useful as part of a comprehensive Rasch model analysis of any data set.

The traditional definition of reliability

Let the observed score of person *n* on item *i* be x_{ni}, $x_{ni} \in \{0,1,2,...m_i\}$ and let the total score of the person on the scale of I items be

$$y_n = \sum_{i=1}^{I} x_{ni} \ . \tag{1}$$

Further let

$$y_n = \tau_n + e_n \ , \tag{2}$$

where τ_n is the true score and e_n is the accumulation of the random errors of the responses to the items of person *n* on the scale. Eq. (2) is the defining equation of TRADITIONAL TEST THEORY. Across a population of persons, the error is taken to be distributed normally, homogeneously across items, and to be uncorrelated with the true scores:

$$e \sim N(0,\sigma_e^2)\,,\ \tau \sim N(\mu,\sigma_\tau^2)\,,\ COV[\tau,e] = 0. \tag{3}$$

From Eqs. (2) and (3), the total observed score variance is given by

$$\begin{aligned} &V[y] = V[\tau] + V[e] + 2COV[\tau,e] \\ &i.e\ \ \sigma_y^2 = \sigma_\tau^2 + \sigma_e^2 + 0 \\ &i.e\ \ \sigma_y^2 = \sigma_\tau^2 + \sigma_e^2. \end{aligned} \tag{4}$$

where $\quad \sigma_y^2 = V[y]; \quad \sigma_\tau^2 = V[\tau]; \ \sigma_e^2 = V[e].$ (5)

Then the reliability r_{yy} of a test in the above terms is given by

$$r_{yy} = \frac{\sigma_\tau^2}{\sigma_y^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} \tag{6}$$

Construction of α

It can be shown readily that with the above assumptions,

$$\alpha = \frac{I}{I-1}\left(\frac{V[y] - \sum_{i=1}^{I} V[x_i]}{V[y]}\right) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_i^2 / I} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2}, \tag{7}$$

where $\sigma_i^2 / I = \sigma_e^2$, (8)

and where σ_e^2 is defined in Eq.(5) and where σ_i^2 is the error variance associated with the responses of a single item. Eq. (8) shows how the error variance in the instrument, $\sigma_e^2 = \sigma_i^2 / I$, reduces as the number of items increases and how, for homogeneous error variances in the items, α increases correspondingly.

Writing $V[y] = V[\sum_{i=1}^{I} x_i]$ gives

$$\alpha = \frac{I}{I-1}\left(\frac{V[\sum_{i=1}^{I} x_i] - \sum_{i=1}^{I} V[x_i]}{V[\sum_{i=1}^{I} x_i]}\right) \tag{9}$$

showing that the structure of the difference in the numerator, $V[\sum_{i=1}^{I} x_i] - \sum_{i=1}^{I} V[x_i]$, is "the variance of the sum of the items – the sum of the variances of these items" divided by the variance of the sum of the items. The whole calculation is directly in terms of the observed scores.

Construction of PSI

With the same definitions of above, let

$$\Pr\{x; \beta, \delta_i, \underline{\tau}\} = \frac{1}{\gamma}\exp\left[-\left(\tau_{1i} + \tau_{2i} + \tau_{3i} + \tau_{xi}\right) + x\left(\beta_n - \delta_i\right)\right] \tag{10}$$

where

$\gamma = \sum_{k=0}^{m_i} \exp\left[-\left(\sum_{x'=0}^{k} \tau_{x'}\right) + k\left(\beta_n - \delta_i\right)\right]$ is the normalising factor.

Eq. (10) is the Rasch model for more than two ordered categories where β_n, δ_i are the proficiency and difficulty of person n and item i respectively, and τ_{xi}, $x = 1,2,...m_i$ are thresholds defining the categories. Eq. (10) reduces to the dichotomous Rasch model in the case of just two ordered categories. The parameter β above has the same role as the true score in traditional test theory. However, estimates of β are obtained by a non-linear transformation of the raw scores, and with each estimate $\hat{\beta}_n$, there is a standard error of this estimate $\hat{\sigma}_n$.

In parallel to Eq. (6), the PSI $r_{\beta\beta}$ is defined according to

$$r_{\beta\beta} = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_e^2} = \frac{\sigma_{\hat{\beta}}^2 - \sigma_{\hat{e}}^2}{\sigma_{\hat{\beta}}^2}$$

where $\hat{\sigma}_{\hat{\beta}}^2$ is the estimated variance of the locations of the persons, and $\hat{\sigma}_{\hat{e}}^2$ is the average error variance of measurement among the persons.

Some similarities and differences between α and the PSI

Similarities

(a) The two indices are formed as the proportion of variance of the estimated person estimates and the total variance including error.

(b) When the persons and items are well aligned, so that there are no extremes and when there is complete data, then the two indices are very close in value.

Differences

(a) α can be calculated with complete data only, while the PSI can be calculated with random missing data. With substantial missing data, the values for the two indices might be different.

(b) The PSI is based on the estimated locations of the persons which are non-linear transformations of the raw scores.

(c) α is calculated with all scores, including the maximum and minimum scores, while the estimate of the PSI requires extrapolated values for extreme scores. This is because there is no finite estimate for extreme scores.

(d) When the the item difficulties relative to the person proficiencies are misaligned, so that there is a skewed distribution with extreme raw scores, a difference emerges between α and the PSI: α remains more constant than the PSI. The reason for this difference is in their respective constructions: α is based on raw scores while the PSI involves a non-linear transformation of these raw scores. The error variance for persons increases as the scores become more extreme, so with scores close to the extreme, the error variance increases in the PSI while there is no such effect in the construction of α.

RUMM2020 and RUMM2030 and subtests

There may be differences which emerge between RUMM2020 and RUMM2030. These will generally occur when the distribution of the persons is not well aligned to the distribution of the items and floor or ceiling effects are evident. These indices are perhaps not as informative as when the persons are well aligned to the items. The reason is that in RUMM2020 some adjustments were made for the extreme scores and their estimates; in RUMM2030, this adjustment has been removed.

In the context of some form of violation of independence, it is considered meaningful to sum the scores of the dependent items and to form a new item with a maximum score equal to the sum of the maximum scores of the individual items. These items are called *subtests* in RUMM and, as a consequence, these subtests are considered higher order items. However, if there are random missing responses, then there will be even more random missing responses in the subtests. This is because if a person has a missing response to an item, that person's maximum possible score on the subset of items is different from the maximum possible score of persons who have responded to all items.

We now consider that it is best to be consistent with the paradigm of Rasch model analyses, and the features of RUMM, so that the user is as fully empowered as possible to make interpretations. With this purpose in mind, the following features (not available in RUMM2020) are now present in RUMM2030.

(a) RUMM2030 has an option to eliminate persons with random missing responses and to form a subset of responses with complete data. This permits the calculation of both α and the PSI for that subset of data.

(b) In forming subtests from this complete data subset, both indices can again be calculated.

(c) RUMM2030 now provides calculations of α and the PSI both when the extreme scores are included and when these are not included.

(d) In addition, the component parts used to calculate α [see Eq. (7)] are now displayed by RUMM2030 in the "Traditional-based Statistics" option of the "Complete Data Only" segment located on the main Display Specifications form (these complement the equivalent values for the PSI which were provided in RUMM2020).

This now provides the user with the information from which to make interpretations. In particular, if there are differences between the two indices in the complete data, it is very likely to be because of the floor or ceiling effects in the data, and in particular, the presence of extreme scores.

If complete data are not used, then there might be differences between RUMM2020 and RUMM2030 in the calculation of the PSI. The values in RUMM2030 should be the ones used and interpreted. However, with random missing data, especially when there are extreme scores as well, these indices should never be interpreted confidently.