Parallel, tau-equivalent, and congeneric measures; their relationship to Cronbach's alpha; and a reprise of split-half reliability

I have been telling us to think about the value of a measure as equal to:

$$\text{observed} \;=\; \text{true score} \;+\; \text{measurement error}$$

If the measure is unbiased, then the expected value of this measurement error is zero. Let's translate the above into some notation. If we have a summated rating scale, we are going to be summing up a bunch of items, so if Y_p is our scale score for person p and y_i is the observed measure for item i, then:

$$Y_p = \sum_{i=i}^{k} y_{ip}$$

where k is the number of items and the value of y_i is determined by the sum of true score τ and error e:

$$y_{ip} = \tau_{ip} + e_{ip}$$

Assuming that the errors are independent across items means that we assume that $\mathrm{Cov}(e_i,e_j) = 0$ if $i \neq j$. If τ_{ip} and τ_{jp} are measures of the same thing, then the only differences between the two between items should be a matter of scaling or item difficulty. You can imagine that we could talk in terms of τ_p that is the same for items i and j but where you can think of a constant being added to τ_p for some items or, more importantly for our purposes, multiplying τ_p by different coefficients for different items. If we make such transformations to τ_p that are different for items i and j, then obviously $\tau_{ip} \neq \tau_{jp}$, even though they are both measures of the same thing and are only imperfect measures to the extent that they contain measurement error.

Parallel measures

The measures (items) comprising a scale are parallel if the following two conditions hold:

1. $\tau_{ip} = \tau_{jp}$ for all i and j
2. $\mathrm{Var}\,(e_i) = \mathrm{Var}\,(e_j)$ for all i and j

This implies that the amount of variation in the item score that is determined by the true score is the same for all items. It also implies that the expected value of each of the items will be the same. The easiest practical example one can imagine of something like this would be a situation where you employ the exact same measure on something on multiple occasions where we have no reason to expect any kind of testing effect or change in the true score over the period in which the multiple measures were administered.

We can simulate this easy enough in Stata. The first part of the simulation will create a dataset of 10000 observations, a random and standard normally distributed `truescore` for each of these observations, and a random and standard normally distributed `error` for each of the four items that will comprise our scale:

```
. clear
. set seed 8675309
. set obs 10000
obs was 0, now 10000

. gen truescore = invnorm(uniform())

. gen error1 = invnorm(uniform())
. gen error2 = invnorm(uniform())
. gen error3 = invnorm(uniform())
. gen error4 = invnorm(uniform())
```

The four items in our scale are all the sum of the true score (possibly subject to some linear transformation) and the error term (possibly multiplied by some factor, but with an expected value of 0). Let's pretend that the item scores are all 1.3 times the true score that we specified above; this still implies that $\tau_{1p} = \tau_{2p} = \tau_{3p} = \tau_{4p}$. Let's also mutliply all the errors by 1.6, which since they were all drawn from normal distributions with the same variance to begin with, does not change the fact that $\mathrm{Var}(e_1) = \mathrm{Var}(e_2) = \mathrm{Var}(e_3) = \mathrm{Var}(e_4)$. First we generate the values for the item scores and a variance `scalescore` that is equal to their sum:

```
. gen item1 = (1.3*truescore) + (1.6*error1)
. gen item2 = (1.3*truescore) + (1.6*error2)
. gen item3 = (1.3*truescore) + (1.6*error3)
. gen item4 = (1.3*truescore) + (1.6*error4)

. gen scalescore = item1 + item2 + item3 + item4
```

Now let's compute summary statistics on the individual item scores:

```
. sum item* scalescore

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
       item1 |   10000    .0022208   2.059244  -8.497154   8.105398
       item2 |   10000   -.0200165   2.063002  -7.658219   7.289746
       item3 |   10000   -.0069085   2.066514  -8.362303   8.142123
       item4 |   10000   -.0194049    2.03565  -7.145496   8.357056
  scalescore |   10000   -.0441092   6.108824  -22.18944   23.20498
```

Notice that the standard deviations of the items are the same (except for sampling variability), meaning that the variances of the items are the same. If we examine the correlations among the items, we can see that they are all the same (except for sampling variability):

```
. cor item*
(obs=10000)

             |    item1    item2    item3    item4
-------------+------------------------------------
       item1 |   1.0000
       item2 |   0.3947   1.0000
       item3 |   0.4082   0.4151   1.0000
       item4 |   0.3914   0.4040   0.4000   1.0000
```

What about the correlation between the score of the scale we created by summing these items and the true score? What about Cronbach's alpha for the scale?

```
. cor scalescore truescore
(obs=10000)

             | scales~e truesc~e
-------------+------------------
  scalescore |   1.0000
   truescore |   0.8541   1.0000

. alpha item*, casewise

Test scale = mean(unstandardized items)

Average interitem covariance:     1.700576
Number of items in the scale:           4
Scale reliability coefficient:     0.7291
```

If we square the correlation between `scalescore` and `truescore` ($r = .854$,so $r^2 = .729$), we get Cronbach's alpha ($\alpha = .729$), our measure of reliability. This is consistent with our definition of reliability as the proportion of the variance in the observed score that is "explained by" variation in the true score.

Tau-equivalent measures

When measures are tau-equivalent, $\tau_{ip} = \tau_{jp}$ for all i and j, as in the case of parallel measures, but we relax the assumption that $\text{Var}(e_i) = \text{Var}(e_j)$ for all i and j. You can imagine this conceptually as generating items as we did before; however, while we still multiply truescore by 1.3 in calculating the contribution of the true score for each item, we multiply the error terms by different amounts (instead of multiplying them all by a constant):

```
. gen item1 = (1.3*truescore) + (1.1*error1)
. gen item2 = (1.3*truescore) + (1.4*error2)
. gen item3 = (1.3*truescore) + (1.8*error3)
. gen item4 = (1.3*truescore) + (2.1*error4)

. gen scalescore = item1 + item2 + item3 + item4
```

What does this do? We have made the error variances differ while still maintaining that $\tau_{1p} = \tau_{2p} = \tau_{3p} = \tau_{4p}$. If we run the summary statistics, we will notice that the standard deviations for the items are no longer the same:

```
. cor item*
(obs=10000)

             |    item1    item2    item3    item4
-------------+------------------------------------
       item1 |   1.0000
       item2 |   0.5191   1.0000
       item3 |   0.4591   0.4164   1.0000
       item4 |   0.3933   0.3617   0.3075   1.0000
```

Moreover, when we look at the correlations among the items we will note that they are no longer the same, but instead the greater the amount of error variance (as indicated by the magnitude of the scaling factor of the errors that we used when generating these items), the lower the observed correlation:

```
. cor item*
(obs=10000)

             |    item1    item2    item3    item4
-------------+------------------------------------
       item1 |   1.0000
       item2 |   0.5191   1.0000
       item3 |   0.4591   0.4164   1.0000
       item4 |   0.3933   0.3617   0.3075   1.0000
```

We can compute the correlation between the `scalescore` and the `truescore` and Cronbach's alpha:

```
. cor scalescore truescore
(obs=10000)

             | scales~e truesc~e
-------------+------------------
  scalescore |   1.0000
   truescore |   0.8475   1.0000


. alpha item*, casewise

Test scale = mean(unstandardized items)

Average interitem covariance:     1.698276
Number of items in the scale:            4
Scale reliability coefficient:      0.7184
```

Notice that once again that the squared correlation between `scalescore` and `truescore` ($r = .8475$, so $r^2 = .718$), we get Cronbach's alpha ($\alpha = .718$).

Congeneric measures

Congeneric measures relax both the assumption that $\tau_{ip} = \tau_{jp}$ for all i and j and that $\mathrm{Var}\,(e_i) = \mathrm{Var}\,(e_j)$ for all i and j. We can imagine this as being like multiplying the underlying true score by different amounts (not 1.3 for every item) as well as multiplying the variance of the errors by different items.

```
. gen item1 = (2.1*truescore) + (1.1*error1)
. gen item2 = (.7*truescore) + (1.4*error2)
. gen item3 = (.9*truescore) + (1.8*error3)
. gen item4 = (1.7*truescore) + (2.1*error4)

. gen scalescore = item1 + item2 + item3 + item4
```

As when items are tau-equivalent, the items comprising our scale will now have different variances and different correlations among pairs of items.

```
. su item* scalescore

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------
       item1 |   10000   -.0076973   2.382428  -9.471482   8.845874
       item2 |   10000   -.0141689   1.561849  -5.558605   5.920243
       item3 |   10000   -.0034707    2.00838  -8.639221   8.325666
       item4 |   10000   -.0254212   2.667827  -9.366914   10.95329
  scalescore |   10000   -.0507581   6.322144  -22.75672    23.7716

. cor item*
(obs=10000)

             |    item1    item2    item3    item4
-------------+------------------------------------
       item1 |   1.0000
       item2 |   0.3956   1.0000
       item3 |   0.4092   0.2175   1.0000
       item4 |   0.5528   0.2884   0.2845   1.0000
```

We can compute the correlation between the `scalescore` and the `truescore` and Cronbach's alpha:

```
. cor scalescore truescore
(obs=10000)

             | scales~e truesc~e
-------------+------------------
  scalescore |   1.0000
   truescore |   0.8563   1.0000

. alpha item*, casewise

Test scale = mean(unstandardized items)

Average interitem covariance:     1.725273
Number of items in the scale:            4
Scale reliability coefficient:      0.6906
```

But, wait: the squared correlation between `scalescore` and `truescore` ($r = .8563$,so $r^2 = .718$) is now greater than Cronbach's alpha ($\alpha = .691$). This is because Cronbach's alpha is only a lower-bound estimate of the true reliability of our scale when measures are congeneric.

Split-half reliability (reprised)

Our intuition for split-half reliability was that you have divided your items into two halves and computing the correlation between subscales created by summing responses to the two halves. Let's go back to our example of parallel items. If you look above, our computed Cronbach's alpha was .7291, and this could be interpreted as the proportion of variance in the scale score that is "explained by" variance in the true score. If we compute the correlation between two halves of a scale, the result will be less than the reliability as calculated by Cronbach's alpha. However, the formula is simple for converting a split-half reliability into a reliability that has the Cronbach's alpha interpretation (estimate of reliability for parallel and tau-equivalent measures; lower-bound estimate of reliability for congeneric measures).

$$\text{scale reliability} = \frac{2r_{\text{split-half subscale scores}}}{1 + r_{\text{split-half subscale scores}}}$$

With four items, there are three possible ways that we could divide our scale into two halves (1 & 2 vs. 3 & 4; 1 & 3 vs.2 & 4; 1 & 4 vs. 2 & 3). The set of these three possible divisions is very much like the randomization set that we talked about when we talked about experiments. Let's create all these subscales and compute all three subscale correlations.

```
. gen scale12 = item1 + item2
. gen scale13 = item1 + item3
. gen scale14 = item1 + item4
. gen scale23 = item2 + item3
. gen scale24 = item2 + item4
. gen scale34 = item3 + item4

. cor scale12 scale34
(obs=10000)

             |  scale12  scale34
-------------+------------------
     scale12 |   1.0000
     scale34 |   0.5793   1.0000

. cor scale13 scale24
(obs=10000)

             |  scale13  scale24
-------------+------------------
     scale13 |   1.0000
     scale24 |   0.5694   1.0000

. cor scale14 scale23
(obs=10000)

             |  scale14  scale23
-------------+------------------
     scale14 |   1.0000
     scale23 |   0.5725   1.0000
```

Now let's plug these three correlations into our formula:

$$\begin{aligned}
\frac{2(.5793)}{1 + .5793} &= .7336 \\
\frac{2(.5694)}{1 + .5694} &= .7256 \\
\frac{2(.5725)}{1 + .5725} &= .7281
\end{aligned}$$

What is the average of these three reliability estimates?

$$\frac{.7336 + .7256 + .7281}{3} = .7291$$

Which was our estimated Cronbach's alpha. So you can think about using randomization to divide items into halves to compute a split-half reliability as being like making taking a draw from the randomized set, where the mean estimate of reliability is Cronbach's alpha.