Werkgroep 2: Betrouwbaarheid van testscores

Opdracht 1

1. *What is the difference between observed scores and true scores? Why are we unable to observe true scores directly?*
   Ware scores zijn onderdeel van de geobserveerde scores. Psychologische eigenschappen zijn niet direct zichtbaar. Men kan bijvoorbeeld niet direct intelligentie observeren. Dit noemen we de latente variabele, oftewel de 'true score'. Wanneer er bijvoorbeeld een IQ test wordt afgenomen, dan hoopt men dat dit ook de werkelijke score weergeeft. Ware scores zijn latente variabelen en moeten worden geschat. Echter, is deze schatting nooit perfect. Er is altijd een error.
2. *What is the theoretical meaning of reliability in psychometric theory? In which regard(s) is this different from the day-to-day meaning of reliability?*
   De theoretische betekenis van betrouwbaarheid is in hoeverre dezelfde uitkomst volgt wanneer men iets meerdere malen meet. Oftewel: in hoeverre is de test vrij van random error. Dit is een verschil wanneer men het heeft over betrouwbaarheid waarover men praat in het dagelijks leven. Er wordt meestal over een persoon gepraat die wel of niet betrouwbaar is. Dit komt echter wel in de buurt van de betrouwbaarheid waarover we spreken bij psychometrie. Iemand is namelijk betrouwbaar wanneer hij meerdere keren een geheim bijvoorbeeld niet doorvertelt. Iemand is minder betrouwbaar wanneer hij meerdere keren een geheim wel doorvertelt. Het gaat hierbij dan ook om meerdere "metingen" waar een overeenkomstige uitkomst volgt.
3. *Give two possible meanings for the reliability coefficient.*
   1) De proportie verklaarde variantie van geobserveerde scores door ware scores (Rxx). Er is 0 errorvariantie, en er blijft dan ook 1 over. Dit is dan ook perfect.

   $$R_{xx} = \frac{s_t^2}{s_o^2} = 1 - \frac{s_e^2}{s_o^2}$$

   2) De gekwadrateerde correlatie tussen geobserveerde scores en ware scores.
   $R_{xx} = r_{ot}^2 = 1 - r_{oe}^2$
4. *What is the standard error of measurement and for which purpose can we use it?*
   Standaard error of measurement is de standaarddeviatie van de error. Dit geeft duidelijkheid over individuele metingen.

   $$se_m = 15\sqrt{(1 - .90)} = 15*$$
5. *Furr & Bacharach mention three general methods to estimate the reliability of a test: alternate forms, test-retest, and internal consistency. Give a brief description of each method.*
   1) Alternate forms (parallel tests): Twee verschillende testen meten hetzelfde.
   2) Test-retest: Dezelfde test op verschillende momenten afnemen.
   3) Internal consistency: Het afleiden van de betrouwbaarheid van de correlatie tussen delen van de test, hiermee wordt bijvoorbeeld een split-half test bedoelt. Het nadeel hiervan is dat dit een willekeurige split is. Of wanneer men meer dan twee metingen doet en ieder item als afzonderlijke test beschouwd.
6. *What is the correction for attenuation?*
   De ware correlatie schatten als metingen perfect betrouwbaar zouden zijn. Oftewel: het "verzwakkingseffect" van meetfouten uit correlatiecoëfficiënt halen.

Opdracht 2

A test for depression among Dutch adults has a true score variance of st 2 = 75 and an error variance of se 2 = 25.

1. *Calculate the reliability of this test.*
   Rxx = 75/100 = 0,75
   De bijbehorende formule is: $Rxx = S^2t / S^2o = S^2t / (S^2t + S^2e)$
2. *Calculate the standard error of measurement in two different ways.*
   De standaardmeetfout heeft te maken met error. Je kunt de standaard meetfout op twee manieren berekenen:
   - So√1-Rxx = 10*√1-0.75 = 5
   - $\sqrt{Se^2} = \sqrt{25} = 5$
3. *Harry's observed score on the test is 100. Calculate the 95% confidence interval for Harry's true score.*
   Het 95% betrouwbaarheidsinterval voor Harry's score:

100 – 1.96*5 < Xb < 100 + 1.96*5
90,20 < Xb < 109,80
De bijbehorende formule is: Xo +/- 1,96*Sem, Sem = So√1-Rxx

4. *What is the reliability of this same test for children, for which the error variance is the same (se 2 = 25), but the true-score variance is lower (st 2 = 60)?*
   De betrouwbaarheid van de test met een lagere ware score variantie:
   S_o^2=85, Rxx = 60/85 = 0,71
5. *A test improvement team manages, by using entirely new questions, to reduce the error variance to se 2 = 15 (the true score variance remains the same). What is the reliability of the test for adults now? What about the test for children?*
   De betrouwbaarheid van de nieuwe test:
   S_e^2= 15, 60/75 = 0,80 (kinderen), 75/90 = 0,83 (volwassenen)

Opdracht 3

- *Via the split-half method, in which an r of .50 is found between the two test halves (which have exactly the same length and are perfectly parallel).*
  (2*0.5)/(1+0.5) = 0.667

$$R_{xx\text{- total}} = \frac{2R_{xx\text{- subtest}}}{1 + R_{xx\text{- subtest}}} = \frac{2r_{hh}}{1 + r_{hh}}$$

- *Via the alternate forms method, in which a correlation of .50 is also found, but this time between the total test X and a parallel version Y.*
  Alternate forms: Rxx = r_{xy} = 0.50

- *If you did this correctly, you will now have two different values for reliability. What is the most obvious explanation for this discrepancy?*
  Discrepantie: Mogelijk zijn test X en Y niet parallel.

Opdracht 4

- *Test 1 consists of five items with the following standard deviations: 2 / 1.5 / 2 / 1 / 1.5.*

$$R_{xx} = \left[\frac{k}{k - 1}\right]\left[1 - \frac{\sum s_i^2}{s_x^2}\right]$$

  (k/k-1)(1-((Σsi²)/(Sx²)) = 0.69
- *Test 2 consists of six dichotomous (correct/incorrect) items with the following proportions of correct answers per item: .3 / .5 / .4 / .6 / .7 / .5. The total variance of Test 2 is 5.*
  Er wordt gebruik gemaakt van KR20, want het gaat om '*binary scores'* (correct/incorrect) (Σpiqi²).
  (k/k-1)(1-((Σpiqi²)/(Sx²)) = 0.86
- *Test 3 consists of ten items. The dog ate the item data, but the average correlation between the items is .25.*

$$R_{xx} = \frac{k\bar{r}_{ii'}}{1 + (k - 1)\bar{r}_{ii'}}$$

  10 * 25 = 2.5
  1 + (9) * 0.25 = 3.25
  2.5 / 3.35 = **0.77**

Opdracht 5:

A test with four items has a standardised item alpha of .40.

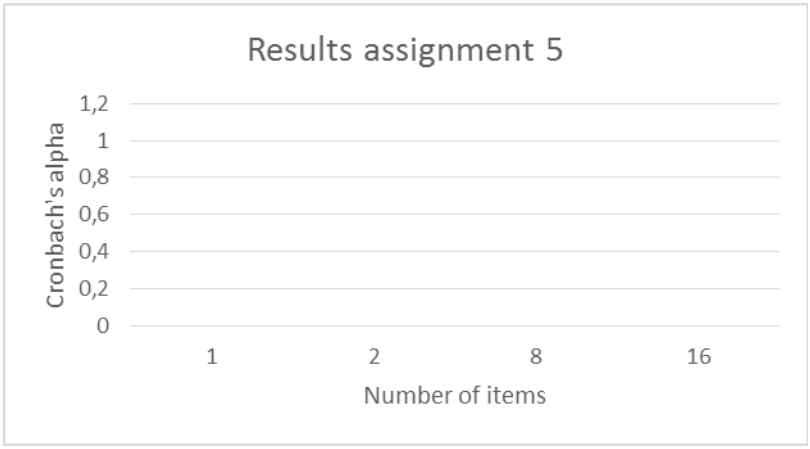
- *Calculate the reliability of this test if it is lengthened (or shortened) to respectively 1, 2, 8, and 16 items.*

$$R_{xx\text{- revised}} = \frac{nR_{xx\text{- original}}}{1 + (n - 1)R_{xx\text{- original}}} \quad n = \frac{k_{revised}}{k_{original}}$$

K = 1: betrouwbaarheid 0.14
K= 2: betrouwbaarheid: 0,25
K = 8: betrouwbaarheid: 0,57
K = 16: betrouwbaarheid: 0,73
De bijbehorende formules zijn:
$Rxx_{revised} = (n*Rxx_{original})/(1 + (n-1) Rxx_{original})$
$n = K_{revised} / K_{original}$

- *What is the average correlation between the items of this test?*
  rxoyo = rxtyt √rxx ryy = 1 * √ .70 * .80 = .75

- *Make a diagram of the results, with the number of items on the X-axis and Cronbach's alpha on the Y-axis.*
  Tabel :



Opdracht 6:

- De correlatie tussen onderliggende ware scores X en Y is 0,20/√0,7x0.8 = 0.27
- De maximale correlatie als Rx = 1 * √0,7x0.8 = 0,75

Opdracht 7:

- Betrouwbaarheid neemt toe bij een hogere gemiddelde inter-item correlatie (r).
  Dit is aanvaardbaar wanneer alle items inhoudelijk verschillen maar wel dezelfde construct meten. Dit is onaanvaardbaar als telkens dezelfde items gemeten worden maar in net iets andere woorden.
- Betrouwbaarheid neemt ook toe bij een groter aantal items.
  Een groter aantal items is aanvaardbaar als de test ondanks een lage r wel een 1 dimensionale structuur heeft (als de errorvariantie per item groot is vergeleken met de ware-score variantie, maar wel volledig random). Dit is onaanvaardbaar als de test meerdimensionale structuur heeft.
  Bijvoorbeeld een deel van de items (P) meet positieve gevoelens en een deel van de items (N) meet ontbreken van negatieve gevoelens. Ook als P-N correlaties nul zijn, is gemiddelde correlatie nog steeds positief (wegens positieve P-P en N-N correlaties) zo kan een groter aantal items 'compenseren voor de gebruikte dimensionaliteit).