Hoorcollege 2: Betrouwbaarheid

Betrouwbaarheid is te beschrijven aan de hand van de vraag 'In hoeverre zijn verschillen in testscores een functie van werkelijke individuele verschillen?'. Eigenlijk wil men dus weten in hoeverre men dezelfde uitkomst krijgt, wanneer men verschillende malen meet en in hoeverre deze score vrij is van random meetfouten. Testen kunnen namelijk nooit de werkelijke score meten, zij geven alleen de testscore weer. En op deze manier ontstaan deze *'errors'.*
In de klassieke testtheorie wordt hiervoor de volgende formule weergeven.
Xo = Xt + Xe. Waarbij de geobserveerde score (Xo) de optelling is van de ware score (Xt) en de error (Xe). De ware score (Xt) is dus niet direct observeerbaar. Daarom wordt dit een *latente variabele* genoemd, welke geschat moet worden. Bv. Wanneer men een IQ test afneemt, dan hoopt men dat de test de werkelijke score weergeeft. Echter, is dit nooit perfect. De geobserveerde score is daarom gelijk aan: t + error. Error = Xe – Xo – Xt (dit kan negatief of positief zijn).

Aan de klassieke testtheorie zitten drie assumpties vast. Aangezien Xo de som is van twee onbekende factoren, wat onoplosbaar is, geldt:

1. $\mu_e = 0$. De gemiddelde errorscore in de populatie is nul. Er is geen sprake van systematische over- of onderschatting van de ware scores.
2. $r_{et} = 0$. De correlatie tussen error en ware score is nul. Iedereen heeft een gelijke kans, de errors zijn dus ongecorreleerd met de ware scores. Bv. Het maakt niet uit of de werkelijke IQ score 100 of 60 is, voor alle mogelijke ware scores geldt nog steeds dat de gemiddelde error nul is.
3. $r_{eiej} = 0$. De error van persoon 1 zegt niets over de error van persoon 2 en deze scores zijn dus niet gecorreleerd; ze zijn volledig random en onafhankelijk van elkaar.

De variantie van X_O is $S_O{}^2 = S_T{}^2 + S_E{}^2$. In de ideale test is $S_T{}^2$ gelijk aan $S_O{}^2$ en zijn er dus geen random meetfouten gemaakt ($S_E{}^2=0$). In alle andere gevallen is er wel een error, welke negatief of positief kan zijn. Hoe kleiner deze error is, des te beter de scores een afspiegeling zijn van de ware scores.

Betrouwbaarheidscoëfficiënt

R_{xx} is de betrouwbaarheidscoëfficiënt, wat gelijk staat aan de proportie verklaarde variantie van X_O door X_T. $R_{xx} = S_T{}^2/ S_O{}^2$ of $R_{xx} = 1 – (S_E{}^2/ S_O{}^2)$. R_{xx} zit tussen de nul en één, en is verder gelijk aan de gekwadrateerde correlatie $r_{OT}{}^2$ (oftewel: $1- r_{oe}{}^2$).
Aangezien ware scores, de '*errors'* en varianties onbekend zijn, moeten er voor het schatten van de betrouwbaarheid minstens twee observaties zijn gedaan per persoon. Dit kan gedaan worden met parallelle metingen.

Parallelle metingen

Parallelle metingen zijn metingen waarbij de errorscores ongecorreleerd zijn en de varianties van de errorscores gelijk zijn. De metingen moeten ook dezelfde ware scores meten. De reden hiervan is dat alles wat de metingen van X en Y gemeen hebben, afkomstig is van de ware score. De correlatie tussen twee parallelle tests geeft een schatting voor de betrouwbaarheid van beide tests, want paralleltests hebben altijd dezelfde betrouwbaarheid.

Er zijn drie manieren van parallelle metingen: Alternate forms, test-hertest en split-half. Bij

‘*alternate forms’* zijn er twee verschillende test voor hetzelfde construct. Hierbij kan wel het ‘*carry-over effect’* ontstaan; test 1 beïnvloedt resultaat van test 2, wat tot een overschatting van de betrouwbaarheid leidt. Een ander probleem is dat men nooit zeker weet of de tests werkelijk parallel zijn.
Bij ‘*test-hertest’* wordt dezelfde test twee keer afgenomen op een ander tijdstip, maar ook hier kunnen carry-over effecten optreden. Daarnaast is er het probleem dat mensen veranderen over tijd.
Bij ‘*split-half’* worden er in één test twee parallelle helften gemaakt. De betrouwbaarheid voor de hele test wordt vervolgens berekend met de ‘*Spearman-Brown formule’*.

$$R_{xx-revised}\frac{n * R_{xx-orgineel}}{1 + (n-1)R_{xx-orgineel}}$$

N is in deze formule de factor waarmee een test vergroot of verkleind wordt, en dus niet het aantal items!

Er kunnen natuurlijk uitputtend veel tweedelingen worden gemaakt. Daarom kan er ook voor worden gekozen om alle items van de test als aparte test te zien. Er kunnen nu berekeningen worden uitgevoerd met de *Cronbach’s alpha*. Deze test is over het algemeen meer betrouwbaar bij een split-half test, aangezien alle items nu gecombineerd worden. Voor formules, zie de powerpoint op Blackboard.

Standaard meetfout

De *standaard meetfout* kan men als volgt berekenen: $S_E^2 = S_O^2 * \sqrt{(1-R_{xx})}$. De standaard meetfout wordt vaak afgekort met S_{em}. Deze geeft de nauwkeurigheid van individuele metingen aan en is ook de standaarddeviatie van de error. Er wordt vaak ook een 95% betrouwbaarheidsinterval bij gegeven. Als de scores negatief zijn, zitten de ware scores onder het gemiddelde.

Kritische kijk op gebruikte aannames:

Er zijn enkele feiten aangenomen om R_{xx} te schatten, echter zijn deze niet altijd (helemaal) waar:

- De drie CTT-aannames ($\mu_e = 0$; $r_{et} = 0$; = 0) – deze zijn in de praktijk niet altijd gelijk aan nul!
- Tau-equivalentie – in de praktijk meten de items van twee parallelle testen niet altijd precies dezelfde feiten: de testen zijn namelijk eigenlijk niet precies parallel te maken!
- Identieke error varianties - in de praktijk zijn de varianties van twee parallelle testen niet precies hetzelfde: de