



# Generalized reliability in industriële user studies

Foto: BoH, Kenterschalm

JAN ENGEL

*Love is the total absence of fear. Love asks no questions. Its natural state is one of extension and expansion, not comparison and measurement. – Gerald Jampolsky*

Echter, in alle andere gevallen dan liefde, zullen we toch wel meten. En dat niet alleen: het trekken van conclusies omtrent de werkelijkheid gaat het beste als we ook precies genoeg kunnen meten. Maar wat is precies genoeg? In dit artikel gaan we er van uit dat we in een industriële user studie een vergelijking willen maken tussen verschillende condities, bijvoorbeeld tussen verschillende producten. Dit wordt gedaan door participanten in een meetpanel die hun beoordelingen vastleggen in een of meer meetschalen.

Precies genoeg is dan in dit geval: kunnen de participanten voldoende goed onderscheid maken tussen de condities? Doen ze dat consistent?

Dit artikel heeft de volgende opbouw. Na een introductie van het begrip 'reliability' uit de psychometrie zullen we dit generaliseren naar een versie die bruikbaar is voor het bepalen van meetprecisie in user studies, een *generalized*

*reliability* die we G-reliability zullen noemen. Deze geeft antwoord op de vraag: hoe precies kan het panel, maar ook de individuele participant daarin, de verschillen tussen condities vaststellen? Vervolgens laten we aan de hand van een voorbeeld zien dat G-reliability een handige maat is voor de karakterisering van een meetpanel. G-reliability hangt nauw samen met het onderscheidingsvermogen van de toets op conditieverschillen en dit zullen we illustreren. Een discussie besluit dit artikel.

## Meetprecisie van items

Bij het kwantificeren in empirisch onderzoek wordt veelal de vraag gesteld naar meetprecisie. In de psychometrie, met name in de klassieke test theorie (CTT), hanteert men daarvoor het begrip 'reliability'. Hierbij wordt een eigenschap van de participanten in de studie gemeten op een een-dimensionale meetschaal met K items. Daarbij wordt het volgende meetmodel verondersteld (Spector, 1992):

$$Y = P + E$$

Hierin is  $P$  de echte waarde van een participant met variantie  $\sigma_P^2$ ,  $Y$  de gemeten versie ervan terwijl  $E$  het verschil aangeeft, de error met variantie  $\sigma_e^2$ .

Hoewel onderwerp van veel discussie, zie bijvoorbeeld Clarke and Watson (1995), is Cronbach  $\alpha$  (Cronbach, 1947) een standaardmaat voor reliability. Onder bepaalde veronderstellingen in het meetmodel kan worden aangetoond dat Cronbach  $\alpha$  een schatter is van de reliability

$$\frac{\sigma_P^2}{\sigma_P^2 + \sigma_e^2 / K}$$

De reliability is een relatieve maat voor meet-precisie. Deze beantwoordt de vraag: hoe goed kunnen we de verschillen in  $P$ -waarden van participanten, gemeten door  $\sigma_P^2$ , onderscheiden op grond van de gemiddelden van de meetwaarden  $Y$  aan de  $K$  items, met variantie  $\sigma_P^2 + \sigma_e^2 / K$ . Het bepalen van reliability gebeurt overigens veelal pas in de (observatie) studie zelf. In industriële user studies is het veelal mogelijk, alvorens een experiment uit te voeren, een beeld te krijgen van de precisie waarmee gegevens worden verkregen. We zullen nu reliability generaliseren naar de industriële setting.

## Generalisatie van reliability

Bij het bepalen van reliability in CTT vraagt men zich in feite af: hoe goed kan ik verschillen onderscheiden tussen participanten, gegeven de meetvariatie van items. Die participanten zijn hier de meettarget. In een industriële user studie zijn participanten geen meettarget, maar onder-

deel van de meetmethode. De meettarget wordt gevormd door de condities, de producten, die worden aangeboden aan de participanten. Het begrip reliability is dan nog steeds heel bruikbaar, maar we zullen het herformuleren. De vraag is dan: hoe precies kunnen we verschillen vaststellen tussen de meetcondities gegeven de variatie van items en van participanten? Jammer genoeg ziet men dat ook in dit geval nog steeds Cronbach  $\alpha$  wordt berekend, maar dan nu over de hele dataset van participanten en condities. Deze geeft helaas geen duidelijk, en een soms zelfs misleidend, beeld van de reliability waar het echt om gaat: het meten van verschillen tussen condities. De verschillen tussen participanten zijn helemaal niet interessant. Dat dit mis kan lopen zullen we laten zien in een voorbeeld in de volgende paragraaf.

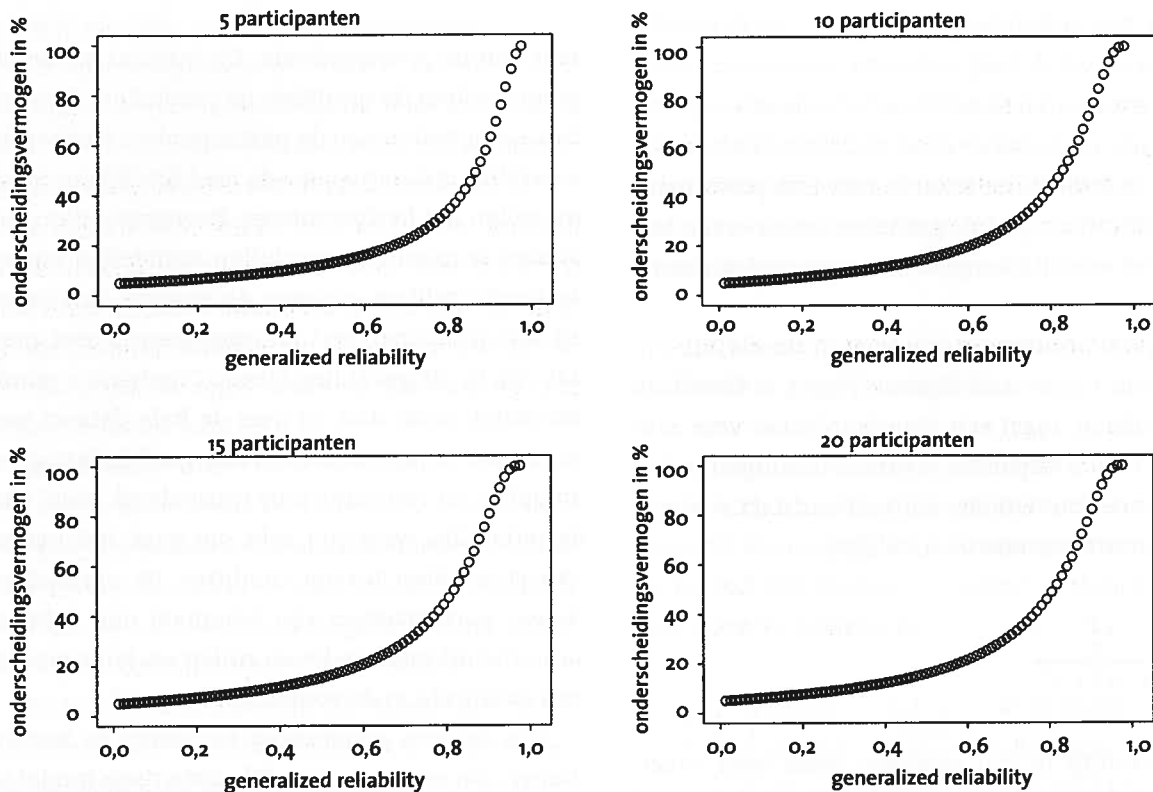
Om tot een generalized reliability te komen dienen we eerst een redelijk statistisch model te formuleren voor de gegevens van een user studie. We gaan nu uit van het geval dat we één factorcondities testen op elke participant in een Within Subject Design (WSD), een veel voorkomend type user studie. Dan zou een redelijk model als volgt kunnen zijn:

$$Y_{ijk} = \mu + P_i + C_j + PC_{ij} + \beta_k + e_{ijk}$$

In dit model vinden we de effecten terug van participanten  $P_i$ ,  $i = 1, \dots, I$ ; condities  $C_j$ ,  $j = 1, \dots, J$ ; de interactie van participanten en condities  $PC_{ij}$ ; de test items  $\beta_k$ ,  $k = 1, \dots, K$  en de residuele fout  $e_{ijk}$ . We veronderstellen voorts dat de  $P_i$ ,  $PC_{ij}$  en  $e_{ijk}$  stochastisch zijn en onderling onafhankelijk, met varianties  $\sigma_P^2$ ,  $\sigma_{PC}^2$  en  $\sigma_e^2$ . Voorts hanteren we

$$\sigma_C^2 = \Sigma(C_j - C)^2 / (J-1)$$

als maat voor de verschillen tussen condities. We gaan nu de reliability op twee manieren generaliseren.



Figuur 1. Onderscheidingsvermogen van de F-toets op condities tegen generalized reliability, voor vier panelgrootten. Resultaten zijn voor twee condities ( $J = 2$ )

Geval 1: de G-reliability per participant,

$$\frac{\sigma_C^2}{\sigma_C^2 + \sigma_e^2 / K}$$

Deze vorm kennen we al uit de vorige paragraaf. Het enige verschil is dat we participant vervangen hebben door condities. Hiermee karakteriseren we de meetprecisie van een participant in het panel. Hoe goed kan deze condities onderscheiden?

Geval 2: de G-reliability over het meetpanel van participanten,

$$\frac{\sigma_C^2}{\sigma_C^2 + \sigma_{PC}^2 / I + \sigma_e^2 / (I \cdot K)}$$

Deze expressie geeft het volgende weer: hoe goed onderscheiden we de verschillen tussen condities

zoals gemeten door  $\sigma_C^2$  door het gemiddelde per conditie van de data  $Y$  over items én participanten. Dit gemiddelde heeft variatie  $\sigma_C^2 + \sigma_{PC}^2 / I + \sigma_e^2 / (I \cdot K)$ . De aanpak is analoog aan het eerste geval, het resultaat is wat anders.

Aardig is nu dat een eenvoudige momentenschatter van G-reliability wordt gegeven door de grootte  $(F-1)/F$  waarbij  $F$  de waarde is van de F-toets van Anova voor het toetsen van condities. In het eerste geval toetsen we het effect van condities voor een bepaalde participant. In het tweede geval doen we dat voor alle participanten samen. De F-toets wordt gebruikt als schatter voor G-reliability. De resultaten volgen uit de verwachtingswaarden van de Mean Squares waaruit de F-toets is opgebouwd, met andere woorden uit de EMS tabel van de variantieanalyse.

Er is een verband met een eerder, en tamelijk onbekend resultaat. Hoyt (1941) liet al zien dat

Cronbach  $\alpha$  ook kan worden bepaald door  $(F-1)/F$  voor het geval dat in CTT wordt bestudeerd, zie de vorige paragraaf. Maar het geldt veel algemener! We kunnen dan eenvoudig G-reliability bepalen via de Anova F-toets, en standaard software is geschikt om dit uit te voeren.

## Onderscheidingsvermogen van de F-toets

G-reliability in geval 2 heeft een relatie met het onderscheidingsvermogen van de F-toets op condities. We kunnen dit als volgt inzien. We formuleren de nulhypothese  $H_0: C_1 = \dots = C_j$ , en het alternatief  $H_1$ : niet alle  $C_j$  zijn gelijk. Indien  $H_0$  waar is heeft de F-toets een centrale F-verdeling, als  $H_0$  onwaar is heeft de F-toets een niet-centrale F-verdeling, met niet-centraliteits (nc) parameter die we  $\lambda$  noemen. Maar die parameter  $\lambda$  kunnen we uitdrukken in de G-reliability, en is daarvan een monotoon stijgende functie:

$$\lambda = (J-1) (G\text{-reliability} / (1- G\text{-reliability}))$$

Hoe groter G-reliability, hoe groter het onderscheidingsvermogen; zie figuur 1. G-reliability heeft direct een betekenis voor de kwaliteit van toetsen.

Uit figuur 1 trekken we twee conclusies:

1. gegeven een waarde van de generalized reliability is het aantal participanten niet erg bepalend voor het onderscheidingsvermogen van de F-toets, en
2. voor een onderscheidingsvermogen van 80% is toch al snel een, relatief hoge, generalized reliability nodig van 0,90. Aan de Cronbach  $\alpha$  worden vaak lagere eisen gesteld.

## Voorbeeld: meetprecisie van een meetpanel

Het voorgaande geeft methoden om de reliability van een meetpanel na te gaan. In het volgende,

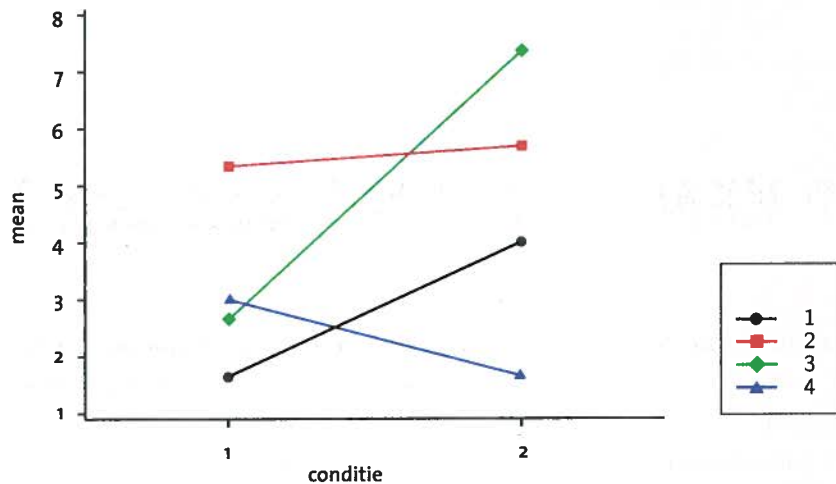
eenvoudige, voorbeeld heeft elk van vier participanten, onder twee verschillende condities, een oordeel gegeven op een 9-punts meetschaal met drie items. Zie tabel 1.

Na berekening blijkt Cronbach  $\alpha = 0,924$ . Er lijkt dus, op grond van Cronbach  $\alpha$ , niets aan de hand, en onze meetschaal lijkt het doet te doen. Maar we meten de verkeerde reliability: die van de totale variatie, dus van participanten, condities, de interactie, en niet van condities alleen! Voor de F-toets op condities vinden we  $F = 1,343$ , en daarmee de G-reliability  $(F-1)/F = 0,26$ , en dat is direct een stuk minder optimistisch. Kijken we naar de individuele G-reliability waarden, gebaseerd op de F-toets uitkomsten per participant, dan vinden we 0,98, -12,00, 0,96 en 0,75. Daaruit blijkt dat de tweede participant het wel heel slecht doet. Laten we deze weg, dan vinden we echter een G-reliability van 0,14.

Een en nader wordt wellicht duidelijk aan de hand van de interactieplot van figuur 2. We zien hierin dat participant 2 inderdaad slecht discrimineert: een non-discriminator. Maar er is meer: de interactie is groot en daarmee de meetprecisie van het panel gering. Nog sterker, door die grote interactie is Cronbach  $\alpha$  groot, maar voor het echte doel, het bepalen van verschil tussen condities, is die interactie juist funest! De waarde

Participant	Conditie	Item 1	Item 2	Item 3
1	1	2,00	1,00	2,00
1	2	4,00	3,00	5,00
2	1	5,00	6,00	5,00
2	2	6,00	4,00	7,00
3	1	3,00	2,00	3,00
3	2	6,00	7,00	9,00
4	1	3,00	4,00	2,00
4	2	1,00	2,00	2,00

Tabel 1. Meetgegevens op 9-punt schaal van vier participanten, twee condities en drie items



Figuur 2. Interactie plot van participanten en condities voor de gegevens van tabel 1

van de G-reliability geeft dit adequaat weer: die is heel laag. De grote variantie van de interactie vinden we terug in de noemer van G-reliability. Een verbetering in de situatie? Wanneer dit een reële situatie is zou men kritisch moeten kijken of de participanten wel geschikt zijn voor hun taak, en of selectie/opleiding niet duidelijk een noodzaak is. Zomaar meer participanten opnemen vergroot de waarde van I, en daarmee de G-reliability, maar is misschien niet de meest effectieve weg.

## Discussie

De G-reliability geeft nuttige informatie om de meetprecisie vast te stellen in een industriële user studie. We hebben dit laten zien aan de hand van een WSD met één factorcondities. Overigens is dit resultaat natuurlijk eenvoudig te generaliseren voor andere designs, zoals een WSD met meerdere factoren, of voor het Between Subject Design. De karakterisering van afzonderlijke participanten ziet men ook terug in de sensometrie (Brockhoff en Skovgaard, 1994). Hierbij wordt ook de F-waarde van de individuele participant gehanteerd als maat voor kwaliteit,

maar er is geen standaardisatie. Het aardige is dat nu de link wordt gelegd tussen deze waarde uit de sensometrie en de reliability uit de psychometrie. Ten slotte geldt dan ook een volgende spin-off: uit het betrouwbaarheids-interval voor de nc-parameter  $\lambda$  kunnen we een betrouwbaarheids-interval bepalen voor G-reliability. En nuttige resultaten hiervoor worden gegeven door bijvoorbeeld Steiger (2004).

## LITERATUUR

- Brockhoff, P.M. & Skovgaard, I.M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference*, 5, 215-224.
- Clarke, L. A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.
- Cronbach, L.J. (1947). Test "reliability": Its meaning and interpretation. *Psychometrika*, 12, 1-6.
- Hoyt, C. (1941). Test reliability by analysis of variance. *Psychometrika*, 6, 135-160.
- Spector, P.E. (1992). *Summated rating scale construction. An introduction*. Newbury Park: Sage publications.
- Steiger, J.H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164-182.

JAN ENGEL is senior consultant bij CQM in Eindhoven.  
E-mail: <Engel@cqm.nl>