

KM 16(1984)

pag 63 - 72

EEN DEFINITIE VAN BETROUWBAARHEID IN TERMEN VAN RUWE SKORES

J.M.F. ten Berge*

Samenvatting

In de klassieke testtheorie is de betrouwbaarheid van een meetinstrument een centraal begrip. De betrouwbaarheid is gedefinieerd als de verhouding van de betrouwbare op de geobserveerde variantie. Dat leidt ertoe dat meetinstrumenten in homogene populaties, waarin de personen weinig verschil in betrouwbare scores vertonen, doorgaans zeer onbetrouwbaar zijn. Toch kan het in dergelijke gevallen wenselijk zijn te weten in hoeverre de geobserveerde scores vrij zijn van meetfouten. Een grootte die dat laatste weergeeft is de verhouding van de verwachting van de gekwadrateerde betrouwbare score op de verwachting van de gekwadrateerde geobserveerde score. In onderstaand artikel wordt nagegaan wat de consequenties zijn van invoering van deze grootte, naast de 'klassieke' betrouwbaarheid, in de klassieke testtheorie.

De auteur is dank verschuldigd aan C. Lewis en F.E. Zegers voor nuttig commentaar op een eerdere versie.

* Vakgroep Persoonlijkheidspsychologie
Subfaculteit Psychologie, R.U. Groningen,
Grote Markt 31/32
9712 HV Groningen
tel.: 050-116204

Klassieke betrouwbaarheid en een alternatief

In de klassieke testtheorie is de betrouwbaarheid gedefinieerd als de verhouding van de varianties van betrouwbare en geobserveerde scores. Daardoor hangt de betrouwbaarheid in hoge mate af van de betrouwbare variantie in de onderzochte populatie. In een populatie waarin alle personen bijvoorbeeld een betrouwbare score 7 hebben, en de meetfouten tussen -1 en 1 liggen, is de betrouwbaarheid nul, terwijl de meetfouten in feite zeer klein zijn vergeleken met de betrouwbare score. Dit roept de gedachte op dat het de moeite waard kan zijn een grootheid in te voeren voor 'de mate waarin scores vrij zijn van meetfouten' naast de betrouwbaarheid als 'de mate waarin scoreverschil vrij is van meetfouten'. Voor variabelen op ratio-meetniveau lijkt zo'n grootheid zonder meer interessant. In het onderstaande wordt een schets gegeven van een testtheorie, waarbij de mate waarin scores vrij zijn van meetfouten een centrale rol vervult.

Beschouw de konstruktie van betrouwbare scores en meetfouten uit de klassieke testtheorie, cf. Lord & Novick (1974, Ch. II), uitmondend in de grondformule

$$X = T + E \quad (1)$$

waarin $E(X) = E(T)$; $E(E) = 0$, en E ongekorreleerd is met alle variabelen waar E géén komponent van is. Naast de 'klassieke' betrouwbaarheid

$$\rho_{xx} = \frac{\text{Var}(T)}{\text{Var}(X)} = \frac{E(T - E(T))^2}{E(X - E(X))^2}, \quad (2)$$

die in feite geheel afhangt van de afwijkings-scores op X en T , definiëren we de 'ruwe-scores betrouwbaarheid' van X als

$$\rho_{xx}^* = \frac{E(T^2)}{E(X^2)} = 1 - \frac{E(E^2)}{E(X^2)}. \quad (3)$$

Het is gemakkelijk na te gaan dat

$$\rho_{xx} = 1 - \frac{\text{Var}(E)}{\text{Var}(X)} \leq 1 - \frac{\text{Var}(E)}{E(X^2)} = \rho_{xx}^*. \quad (4)$$

Om ρ_{xx}^* meetbaar te maken, nemen we het begrip 'parallellie' uit de klassieke testtheorie ongewijzigd over, d.w.z. X en X' zijn parallel indien voor iedere persoon, over replicaties, geldt

$$\mathcal{E}(X) = \mathcal{E}(X') \quad \text{en} \quad \text{Var}(X) = \text{Var}(X') \quad (5)$$

Dit impliceert onder meer dat betrouwbare scores op X en X' dezelfde zijn en dat

$$\mathcal{E}(X^2) = \text{Var}(X) + (\mathcal{E}X)^2 = \text{Var}(X') + (\mathcal{E}X')^2 = \mathcal{E}(X'^2) . \quad (6)$$

Er zijn nu in principe twee manieren om de ruwe-skores-betrouwbaarheid ρ_{XX}^* te berekenen met behulp van parallelle metingen. Daartoe zullen we gebruik maken van een familie van koëfficiënten beschreven door Zegers en Ten Berge (1984). De koëfficiënten in deze familie zijn de produkt-moment korrelatie ρ_{XY} , de kongruentie $\phi(X,Y) = \mathcal{E}(XY)(\mathcal{E}(X^2)\mathcal{E}(Y^2))^{-\frac{1}{2}}$, cf. Burt (1948), de Additiviteit $A(X,Y) = 2 \text{Kov}(X,Y)(\text{Var}(X)+\text{Var}(Y))^{-1}$, en de Identiteit $I(X,Y) = 2\mathcal{E}(XY)(\mathcal{E}(X^2)+\mathcal{E}(Y^2))^{-1}$.

Het is gemakkelijk na te gaan dat de kongruentiekoefficiënt tussen parallelle metingen gelijk is aan

$$\phi(X,X') = \frac{\mathcal{E}(XX')}{\sqrt{\mathcal{E}(X^2)\mathcal{E}(X'^2)}} = \frac{\mathcal{E}(TT')}{\mathcal{E}(X^2)} = \frac{\mathcal{E}(T^2)}{\mathcal{E}(X^2)} = \rho_{XX}^* . \quad (7)$$

Evenzo kan de Identiteitskoefficiënt gebruikt worden om ρ_{XX}^* te bepalen:

$$I(X,X') = \frac{2\mathcal{E}(XX')}{\mathcal{E}(X^2)+\mathcal{E}(X'^2)} = \frac{2\mathcal{E}(T^2)}{2\mathcal{E}(X^2)} = \rho_{XX}^* . \quad (8)$$

Aangezien (7) óók geldt als X' op een scalaire vermenigvuldiging na parallel is met X verkiezen we (7) boven (8). Een gelijksoortig keuze-probleem treffen we ook in de klassieke testtheorie aan, wanneer we opmerken dat de Additiviteit tussen twee paralleltests ρ_{XX} oplevert. Immers

$$A(X,X') = \frac{2 \text{Kov}(X,X')}{\text{Var}(X)+\text{Var}(X')} = \frac{2 \text{Var}(T)}{2 \text{Var}(X)} = \rho_{XX} . \quad (9)$$

Ook in de klassieke testtheorie valt de produkt-moment-korrelatie tussen X en X' te prefereren boven $A(X,X')$ omdat de eerstgenoemde onder een

ruimere klasse van transformaties gelijk blijft aan ρ_{xx} dan $A(X, X')$. Een andere reden voor het prefereren van (7) boven (8) en van de produkt-moment-korrelatie boven (9) is dat ρ_{xx} resp. ρ_{xx}^* ook gedefinieerd kunnen worden als het kwadraat van resp. $\rho(X, T)$ en $\phi(X, T)$, hetgeen voor $A(X, T)$ en $I(X, T)$ niet opgaat.

Met behulp van (3) en (7) kan een aantal indices uit de klassieke testtheorie opnieuw worden afgeleid. We doen dat in een aantal afzonderlijke paragrafen.

De standaard-metfout

Invoering van de ruwe-skores-betrouwbaarheid heeft geen invloed op de standaard-metfout. Wel dient zich nu een tweede methode aan om de standaard-metfout te bepalen, namelijk als de wortel uit:

$$E(E^2) = E(X-T)^2 = E(X^2) - 2E(XT) + E(T^2) = E(X^2) - E(T^2) = E(X^2)(1 - \rho_{xx}^*). \quad (10)$$

Attenuatie-korrekties

Wanneer X en Y een willekeurig tweetal variabelen is, kan men zich afvragen op welke wijze de betrouwbare skores T_x en T_y samenhangen. Naast de gebruikelijke attenuatie-korrektie

$$\rho(T_x, T_y) = \rho_{xy} / \sqrt{\rho_{xx}} \sqrt{\rho_{yy}} \quad (11)$$

kan men beschikken over de Additiviteit

$$A(T_x, T_y) = \frac{2\text{Kov}(X, Y)}{\rho_{xx} \text{Var}(X) + \rho_{yy} \text{Var}(Y)}; \quad (12)$$

over de kongruentie

$$\phi(T_x, T_y) = \phi_{xy} / \sqrt{\rho_{xx}^*} \sqrt{\rho_{yy}^*}; \quad (13)$$

en over de Identiteit

$$I(T_x, T_y) = \frac{2E(XY)}{\rho_{xx}^* E(X^2) + \rho_{yy}^* E(Y^2)} \quad (14)$$

waarmee respectievelijk de lineariteit, additiviteit, proportionaliteit en identiteit van T_x en T_y wordt uitgedrukt, cf. Zegers & Ten

Berge (1984). In de terminologie van de klassieke testtheorie (Lord & Novick, 1974) is (12) een maat voor essentiële tau-equivalentie van X en Y , is (14) een maat voor tau-equivalentie van X en Y , en geeft (11) aan in welke mate X en Y congenerische tests zijn (Kristof, 1974).

Ondergrenzen aan de betrouwbaarheid

Zoals er tal van ondergrenzen zijn ontwikkeld voor ρ_{xx} , e.g., zie Guttman (1945), Jackson & Agunwamba (1977), Woodhouse & Jackson (1977), Ten Berge & Zegers (1978), zo kunnen er ook tal van ondergrenzen worden ontwikkeld voor ρ_{xx}^* . Bijvoorbeeld, analoog aan de ondergrens

$$\alpha = \frac{k}{k-1} \left(1 - \sum_{g=1}^k \frac{\text{Var}(X_g)}{\text{Var}(X)} \right) \leq \rho_{xx}, \quad (15)$$

waarin $X = \sum_{g=1}^k X_g$ de totaalscore op k items is, kan worden bewezen dat

$$\alpha^* = \frac{k}{k-1} \left(1 - \sum_{g=1}^k \frac{\mathcal{E}(X_g^2)}{\mathcal{E}(X^2)} \right) \leq \rho_{xx}^*. \quad (16)$$

Bewijs: Voor ieder itempaar X_g, X_h geldt

$$\mathcal{E}(T_g - T_h)^2 \geq 0 \quad (17)$$

en dus

$$2\mathcal{E}(X_g X_h) = 2\mathcal{E}(T_g T_h) \leq \mathcal{E}(T_g^2) + \mathcal{E}(T_h^2). \quad (18)$$

Sommeren en middelen over alle $k(k-1)$ paren levert

$$\sum_{g \neq h} \frac{\mathcal{E}(X_g X_h)}{k(k-1)} \leq \sum_{g=1}^k \frac{\mathcal{E}(T_g^2)}{k}. \quad (19)$$

Uitschrijven van ρ_{xx}^* als

$$\rho_{xx}^* = \frac{\mathcal{E}(T_g^2)}{\mathcal{E}(X^2)} = \frac{\sum_{g=1}^k \mathcal{E}(T_g^2) + \sum_{g \neq h} \mathcal{E}(T_g T_h)}{\mathcal{E}(X^2)} = \frac{\sum_{g=1}^k \mathcal{E}(T_g^2) + \sum_{g \neq h} \mathcal{E}(X_g X_h)}{\mathcal{E}(X^2)} \quad (20)$$

en invoegen van (19) in (20) levert:

$$\rho_{XX}^* \geq [\xi(X^2)]^{-1} \left[\sum_{g \neq h} \xi(X_g X_h) \right] \left(\frac{1}{k-1} + 1 \right) = \frac{k}{k-1} \left(\sum_{g \neq h} \frac{\xi(X_g X_h)}{\xi(X^2)} \right) \quad (21)$$

hetgeen ook kan worden geschreven als

$$\rho_{XX}^* \geq \frac{k}{k-1} \left(1 - \frac{k}{\sum_{g=1}^k \xi(X_g^2)} \right) = \alpha^*. \quad (22)$$

Daarmee is (16) bewezen. De ongelijkheid (16) gaat over in gelijkheid dan en slechts dan als $T_g = T_h$ voor alle paren (g, h) , d.w.z. dan en slechts dan als de k items tau-equivalent zijn, cf. Lord & Novick (1947, p. 49).

Aangezien $\rho_{XX} \leq \rho_{XX}^*$, cf. (4), is ook α een ondergrens voor ρ_{XX}^* . Er bestaat echter tussen α^* en α geen ordeningsrelatie. Er zijn gegevens denkbaar waarbij α een betere ondergrens voor ρ_{XX}^* is dan α^* , en omgekeerd.

De Spearman-Brown formule

Wanneer een variabele X , met ruwe-skores-betrouwbaarheid ρ_{XX}^* , met een factor k wordt 'verlengd' (door parallele delen toe te voegen c.q. weg te laten) ontstaat de variabele Y met

$$\rho_{YY}^* = \frac{\xi(T_Y^2)}{\xi(Y^2)} = \frac{k^2 \xi(T_X^2)}{k^2 \xi(T_X^2) + k \xi(E_X^2)} = \frac{k \rho_{XX}^*}{k \rho_{XX}^* + 1 - \rho_{XX}^*} = \frac{k \rho_{XX}^*}{1 + (k-1) \rho_{XX}^*} \quad (23)$$

waaruit op de bekende manier kan worden berekend met welke factor een variabele verlengd dient te worden om een bepaalde ruwe-skores-betrouwbaarheid te bereiken.

Betrouwbaarheid van heterogene somvariabelen

Laat $Y = X_1 + X_2 + \dots + X_k$ een somvariabele zijn van k heterogene variabelen, d.w.z. variabelen die in principe geen onderlinge samenhang hoeven te vertonen, bijv. prediktoren in een multiplere-

regressie probleem. Om iets over de betrouwbaarheid van Y te kunnen zeggen, is nodig dat iets van de betrouwbaarheid van X_j , $j=1, \dots, k$ bekend is. Laat $\rho_{jj}^* = \mathcal{E}(T_j^2)(\mathcal{E}(X_j^2))^{-1}$. Dan is de ruwe-scores betrouwbaarheid van Y te bepalen als

$$\rho_{yy}^* = 1 - \frac{\mathcal{E}(E_y^2)}{\mathcal{E}(Y^2)} = 1 - \frac{\sum_{j=1}^k \mathcal{E}(E_j^2)}{\mathcal{E}(Y^2)} = 1 - \frac{\sum_{j=1}^k \mathcal{E}(X_j^2)(1-\rho_{jj}^*)}{\mathcal{E}(Y^2)} \quad (24)$$

waarmee bij invulling van α_{jj}^* voor ρ_{jj}^* , een analogon voor de gestratificeerde alpha (Cronbach, Schönemann & McKie, 1963) is verkregen.

Restriction of range en restriction of size

Wanneer in een populatie selectie plaatsvindt door eliminatie van extreem scorende personen, zal de geobserveerde variantie ('range') dalen, alsook de betrouwbare variantie. Dientengevolge zal ook de betrouwbaarheid dalen. In de klassieke testtheorie wordt uiteengezet wat het effect van de 'restriction of range' op de betrouwbaarheid is, onder de veronderstelling dat de meetfout-variantie dezelfde blijft.

De ruwe-scores betrouwbaarheid hangt niet af van betrouwbare en geobserveerde variantie maar van de doorsnee grootte (size) van betrouwbare en geobserveerde scores. Derhalve is ρ_{xx}^* gevoelig voor 'restriction of size' i.p.v. 'restriction of range'. Selectie veroorzaakt vaak beide soorten 'restriction' tegelijk, maar de effecten op ρ_{xx} en ρ_{xx}^* kunnen geheel verschillend zijn, zoals aanstonds zal worden toegelicht.

Veronderstel dat de meetfoutenvariantie = meetfoutengrootte niet verandert door selectie. Dan is

$$\mathcal{E}(E^2) = \mathcal{E}(X^2) (1-\rho_{xx}^*) = \mathcal{E}(Y^2) (1-\rho_{yy}^*) \quad (25)$$

zodat

$$\rho_{yy}^* = 1 - \frac{\mathcal{E}(X^2)}{\mathcal{E}(Y^2)} (1-\rho_{xx}^*) , \quad (26)$$

een uitdrukking analoog aan formule (6.2.1) uit Lord & Novick (1974, p. 130).

Het is duidelijk dat de ruwe-skores-betrouwbaarheid perfect kan zijn zelfs als $\text{Var}(T) = 0$. In die zin is de ruwe-skores-betrouwbaarheid ongevoelig voor 'restriction of range'.

Voor de klassieke betrouwbaarheid maakt het in principe niet uit of bij selectie hoog skorende dan wel laag skorende personen uit de populatie worden verwijderd: In beide gevallen zal ρ_{XX} dalen. Bij verwijdering van laag skorende personen uit de populatie zal $E(T^2)$ en daarmee ρ_{XX}^* eveneens dalen. Bij verwijdering van laag skorende personen zal $E(T^2)$ echter toenemen, en zal ook ρ_{XX}^* toenemen. We lichten dit toe aan de hand van een fiktief voorbeeld.

In een populatie heeft de betrouwbare score de diskrete uniforme verdeling $p(T=8) = p(T=6) = p(T=4) = p(T=2) = .25$ en neemt de meetfout E de waarde 1 aan met kans .50 en de waarde -1 met kans .50. De betrouwbaarheid ρ_{XX} is $5/6 = .83$. Daarentegen is $\rho_{XX}^* = \frac{240}{248} = .97$. Selektieren we de personen met $T > 6$ dan is $\rho_{XX} = .50$ terwijl ρ_{XX}^* stijgt tot $200/204 = .98$. Selektieren we de personen met $T < 4$ dan is $\rho_{XX} = .50$ en $\rho_{XX}^* = 40/44 = .91$.

Diskussie

Het is uiteraard niet de bedoeling het begrip 'ruwe-skores-betrouwbaarheid' ter vervanging van de klassieke betrouwbaarheid te propagieren: de beide soorten betrouwbaarheid geven antwoord op verschillende vragen en kunnen dus heel goed naast elkaar worden gebruikt. De vraag is dan in hoeverre de ruwe-skores-betrouwbaarheid een aanwinst voor de klassieke testtheorie is. Hierover vallen de volgende opmerkingen te maken.

Allereerst biedt de ruwe-skores-betrouwbaarheid de mogelijkheid de sterkte van uiteenlopende vormen van samenhang tussen betrouwbare skores op twee variabelen uit te drukken in de termen van proportionaliteit of identiteit, zie de paragraaf Attenuatie-korrekties. Daarmee wordt de klassieke testtheorie in die zin geliberaliseerd dat de product-moment korrelatie niet langer de enige genormeerde samenhangsmaat is die er een rol in speelt.

Vervolgens biedt de ruwe-skores-betrouwbaarheid een tweede methode om de standaard-metfout te berekenen. Of deze methode bevredigender steekproefgedrag heeft dan de klassieke methode is een vraag

die nader onderzoek vereist.

Voorts is de ruwe-skores-betrouwbaarheid van belang in situaties waar extreme 'restriction of range' heeft plaatsgevonden. Wanneer de betrouwbare variantie (vrijwel) geheel verdwenen is door selectie, dan biedt de ruwe-skores-betrouwbaarheid nog de mogelijkheid situaties waarin de meetfouten groot zijn ten opzichte van de gebruikte schaal der geobserveerde skores te onderscheiden van situaties waarin de meetfouten relatief klein zijn.

Tenslotte komt de ruwe-skores-betrouwbaarheid ten dele tegemoet aan de kritiek op de klassieke testtheorie die o.a. door Samejima (1977, p. 196) is geuit: 'A fatal deficiency of classical test theory is that it cannot specify the standard error of estimation independently from the reliability coefficient, and thus independently from a specific group of examinees'. Deze kritiek treft de ruwe-skores-betrouwbaarheid minder hard, omdat deze niet afhangt van de betrouwbare variantie in de populatie. Bovendien is het mogelijk de ruwe-skores-betrouwbaarheid lokaal te definiëren, nl. als $T_i^2 / E(X_i^2)$, voor een willekeurige betrouwbare skore T_i . Daarmee is een grootheid verkregen die vergelijkbaar is met de -eveneens lokaal gedefinieerde- informatiefunctie uit de item-response theorie. In hoeverre er van meer dan oppervlakkige vergelijkbaarheid sprake is zal nader onderzocht moeten worden.

Literatuur

- Burt, C. (1948), The factorial study of temperament traits. British Journal of Psychology, Statistical Section, 1, 178-203.
- Cronbach, Schönemann & Mckie (1965), Alpha coefficients for stratified parallel tests. Educ. and Psychol. Measurement, 25, 291-312.
- Guttman, L. (1945), A basis for analyzing test-retest reliability. Psychometrika, 10, 225-282.
- Jackson, P.H. & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds. Psychometrika, 42, 567-578
- Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. Psychometrika, 39, 491-499.
- Samejima (1977), Weakly parallel tests in a latent trait theory with some criticisms of classical test theory. Psychometrika, 42, 193-198.
- Ten Berge, J.M.F. & Zegers, F.E. (1978), A series of lower bounds to the reliability of a test. Psychometrika, 43, 575-579.
- Woodhouse, B. & Jackson, P.H. (1977), Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II. A search procedure to locate the greatest lower bound. Psychometrika, 42, 579-591.
- Zegers, F.E. & Ten Berge, J.M.F. (1984), A family of association coefficients for metric scales. Psychometrika, 49 (in press).