Over misverstanden rond Cronbachs alfa en de wenselijkheid van alternatieven

Cronbachs alfa wordt gebruikt als maat voor de interne consistentie van de items in een test, en is tevens de bekendste methode voor de schatting van de betrouwbaarheid van de testscore. In de psychometrie is echter bekend dat alfa juist geen goede maat is voor de interne consistentie, en eveneens dat alfa niet alleen de betrouwbaarheid onderschat, maar dat ook nog eens sterker doet dan de meeste andere methoden. In dit artikel wordt het falen van alfa uiteengezet en worden alternatieve methoden besproken en aangeraden.

Klaas Sijtsma

Foto: Herman Wouters

In de test- en vragenlijstconstructie is geen kwaliteitsmaat zo populair als Cronbachs alfa. Alfa wordt op twee manieren gebruikt: als maat van de interne consistentie van de items in de test en als schatting van de betrouwbaarheid van de testscores. Daar wringt meteen al de schoen, want als alfa twee kenmerken van een test kan beschrijven, dan zouden die kenmerken moeten samenvallen. Dus, interne consistentie zou hetzelfde moeten zijn als betrouwbaarheid. In dit artikel wordt uitgelegd dat het om verschillende zaken gaat. Interne consistentie is de mate waarin de items in de test samenhangen als uiting van een gemeenschappelijk psychologisch proces of gedragsprincipe. Interne consistentie is dus een aspect van de constructvaliditeit van de test. Betrouwbaarheid is de mate waarin testscores onder dezelfde condities herhaalbaar zijn.

De vraag lijkt dan te zijn wat Cronbachs (1951) alfa uitdrukt. Betoogd zal worden dat alfa geen maat is voor de interne consistentie, en één van de slechtst denkbare maten voor de betrouwbaarheid van een testscore. Betere methoden voor het schatten van de betrouwbaarheid bestaan al sinds de jaren veertig (bijvoorbeeld Guttman, 1945), en sommige zijn gewoon beschikbaar in SPSS, in de procedure waarin ook alfa is opgenomen. Andere zijn wel in softwarepakketten beschikbaar, maar die zijn doorgaans minder gemakkelijk te bedienen dan SPSS of vereisen specialistische kennis waarover niet elke onderzoeker beschikt.

Doel van dit artikel is aandacht te vragen voor de betekenis van Cronbachs alfa, en vervolgens een advies te geven over de schatting van de betrouwbaarheid van de testscore. Ook wordt ingegaan op de wijze waarop de interne consistentie van een test of vragenlijst kan worden onderzocht. Testconstructeurs en onderzoekers die met tests en vragenlijsten werken, wordt aangeraden kennis te nemen van de bestaande, goed toegankelijke en bovenal, betere methoden om de betrouwbaarheid te schatten. Om de boodschap optimaal voor het voetlicht te krijgen, beperken we technische uitwijdingen zoveel mogelijk, overigens zonder ze helemaal te kunnen vermijden.

Cronbachs alfa

We nemen aan dat een test of vragenlijst bestaat uit J items, die genummerd zijn als $j=1,\ldots,J$. De score op item j wordt genoteerd als X_j. Itemscores kunnen bijvoorbeeld 0 en 1 zijn voor respectievelijk foute en goede ant-

Tabel 1. Voorbeeld van een variantie-covariantiematrix voor vier items

	1	2	3	4
1	.25	.12	.16	.10
2	.12	.24	.08	.09
3	.16	.08	.25	.12
4	.10	.09	.12	.21

woorden op de items uit een intelligentietest, maar ze kunnen ook gelijk zijn aan bijvoorbeeld de scores 1, 2, 3, 4, 5, die aangeven hoe een respondent heeft geantwoord op een rating scale uit een persoonlijkheidsvragenlijst. Meestal wordt de som van de itemscores, ook wel testscore of totaalscore genoemd, gebruikt om testprestaties uit te drukken:

$$X_+ = \sum_{j=1}^{J} X_j.$$

De kwaliteit van de testscore (we laten nog even open wat met 'kwaliteit' wordt bedoeld) wordt vrijwel standaard in Cronbachs alfa uitgedrukt. Om alfa te schatten, zijn de covarianties of correlaties tussen de itemscores onderling nodig, en de variantie van de testscore X_+. De covariantie voor twee items, j en k, noteren we als S_{jk}, en hun product-momentcorrelatie als r_{jk}. De variantie van de itemscores op item j geven we aan als S_j^2, en de variantie van X_+ als $S_{x_+}^2$. Voor de boodschap van dit artikel is het onderscheid tussen populatie en steekproef niet van belang. We laten het dus achterwege.

Voor de uitleg van Cronbachs alfa is het handig om eerst de covarianties tussen alle mogelijk paren van items in een variantie-covariantiematrix bijeen te zetten. Tabel 1 geeft voor vier items ($J = 4$) een voorbeeld van zo'n matrix. De itemvarianties staan in de cellen op de hoofddiagonaal (van linksboven naar rechtsonder): dus, $S_1^2 = .25$, $S_2^2 = .24$, enzovoort. De covarianties staan in de andere cellen: bijvoorbeeld, $S_{12} = .12$, $S_{13} = .16$. Verder is de matrix symmetrisch in de hoofddiagonaal: $S_{21} = S_{12} = .12$, $S_{31} = S_{13} = .16$, enzovoort. Alle covarianties staan er dus twee maal in.

Cronbachs alfa is gedefinieerd als

$$alfa = \frac{J}{J-1} \times \frac{2 \times \sum_{j=1}^{j-1} \sum_{k=j+1}^{J} S_{jk}}{S_{x_+}^2}$$

Alfa hangt af van drie kenmerken van de test: het aantal items J, de variantie van de testscore $S_{x_+}^2$, en de covarianties tussen de itemscores S_{jk}. In het voorbeeld, waarin $J = 4$, is de eerste breuk dus gelijk aan 4/3. Het dubbele somteken in de teller van de tweede breuk zorgt ervoor dat alle covarianties die boven de hoofddiagonaal in de variantie-covariantiematrix staan bij elkaar worden opgeteld; dit levert 0.67. Daarna volgt vermenigvuldiging met 2, wat resulteert in 1.34. Dit is de som van alle covarianties in de variantie-covariantiematrix. Tot slot is de variantie van de testscore, $S_{x_+}^2$, gelijk aan de som van alle itemvarianties en alle itemcovarianties in de variantie-covariantiematrix (dus, de som van alle getallen in de matrix); dit levert 2.29. We vinden voor alfa

$$alfa = \frac{4}{3} \times \frac{1.34}{2.29} \approx 0.78.$$

Voor een concrete test heeft alfa een waarde tussen de theoretische extremen a en 1. Minimum a is negatief als de gemiddelde covariantie tussen de items negatief is. Zoiets kan gebeuren als verschillende items sterk verschillende eigenschappen meten, die negatief samenhangen, maar ook als een deel van de items verkeerd om gecodeerd is. Dit gebeurt wel eens bij negatief geformuleerde rating-scale-items. Het gaat dan uiteraard om een administratieve fout. De waarde 1 kan alleen worden bereikt onder theoretische condities die in de praktijk overigens nooit worden gerealiseerd. De uitleg hiervan valt buiten het bestek van dit artikel (maar zie bijvoorbeeld Nunnally, 1978).

Relatie alfa met interne consistentie

Het begrip interne consistentie wordt in de literatuur over testtheorie en testconstructie niet ondubbelzinnig gedefinieerd. Cronbach (1951, p. 320) stelde bijvoorbeeld dat een intern consistente test 'psychologically interpretable' is, hoewel dit niet betekent 'that all items be factorially similar'. Auteurs (bijvoorbeeld Cortina, 1993; Schmitt, 1996) die later een poging tot een definitie waagden, hadden vervolgens weinig invloed op het gebruik van de term. De grootste gemene deler van alle definities en opvattingen lijkt te zijn 'de mate waarin de

Al sinds de jaren veertig bestaan er betere methoden voor het schatten van de betrouwbaarheid

items in de test met elkaar samenhangen, als teken van wat zij gemeenschappelijk met elkaar meten' (Sijtsma, 2009). Cronbachs alfa wordt vaak opgevat als een maat voor de interne consistentie van een test. De vraag is nu of deze interpretatie wel correct is. Wat zijn de condities waaronder deze vraag beslecht kan worden?

Om te beginnen dient onafhankelijk van Cronbachs alfa te worden vastgesteld wanneer een test wel of niet intern consistent is. Is dat gebeurd, dan zou men eerst bij toenemende graden van interne consistentie alfa kunnen berekenen, en zou een toenemende waarde gevonden moeten worden. Daarna zou men voor verschil-

lende tests, waarvan sommige duidelijk intern consistent zijn en andere niet, alfa kunnen berekenen. Voor intern consistente tests zou dan een hogere alfawaarde gevonden moeten worden dan voor intern inconsistente tests. Als deze relaties niet worden gevonden, kan worden geconcludeerd dat alfa niet een geschikte maat voor de interne consistentie is.

Invloed van interne consistentie op alfa

We nemen voor het gemak aan dat een test intern consistent is als de samenhang tussen de items gelijk is en ook nog eens sterk. Dit wordt als volgt geoperationaliseerd. We gaan uit van vier items ($J = 4$), die alle een variantie hebben van .25 (dus, $S_j^2 = .25$, voor $j = 1,\ldots,4$), terwijl alle covarianties dezelfde positieve waarde c hebben (dus, $S_{jk} = c$, voor alle $j < k$, dus alle itemparen). Naarmate c groter is, is de interne consistentie van de 4-itemtest dus ook groter.

We bekijken vijf gevallen: c = .01, .02, .05, .08, .11. Toepassing van de formule voor alfa levert respectievelijk *alfa* = .14, .26, .50, .65, .76. Als het aantal items en de itemvarianties dus gelijk blijven, en de (gelijke) covarianties nemen toe, neemt ook alfa toe. Dat lijkt dus een goed begin.

Interne consistentie nader beschouwd: dimensionaliteit

Aangezien in de vakliteratuur een ondubbelzinnige definitie van interne consistentie ontbreekt, is het niet goed mogelijk om de relatie te onderzoeken van interne consistentie met eendimensionaliteit van de test of, in termen van factoranalyse, een 1-factoroplossing. Het voorbeeld van daarnet was echter zo geconstrueerd, dat alle variantie-covariantiematrices 1-factorieel waren. Dit is na te gaan door eerst de covarianties om te rekenen naar correlaties via $r_{jk} = S_{jk} / S_j S_k$ (bijv., $r_{12} = .02/\sqrt{.25}\sqrt{.25} = .08$), en vervolgens factoranalyse op de correlatiematrices te doen. Daarbij wordt slechts één factor berekend die zoveel mogelijk gemeenschappelijke variantie van de items verklaart. Dit wordt gerealiseerd met behulp van minimum-rank-factoranalysis (MRFA; Ten Berge & Kiers, 1991).

Alfa zegt niets over de interne consistentie van een test

Ondanks dat alle correlatiematrices in het voorbeeld verschillende graden van interne consistentie representeren, laat MRFA zien dat ze alle perfect 1-factorieel zijn. Het verschil is dat als c toeneemt, de invloed van meetfouten afneemt. In termen van signaal en ruis, neemt bij toenemende c de ruis (dat is de meetfout) af waardoor het signaal (dat is de factor) beter waarneembaar wordt,

maar er is in alle gevallen één signaal. Terzijde zij opgemerkt dat psychologen meestal hoofdcomponentenanalyse gebruiken. Deze methode veronderstelt dat de data geen meetfouten bevatten. Het gevolg is dat de eerste factor minder variantie verklaart dan bijvoorbeeld de eerste factor van een MRFA (Ten Berge & Sočan, 2004).

In het voorbeeld gaat een sterkere mate van interne consistentie dus samen met 1 factor, en de oplopende waarden van alfa horen dus steeds bij een 1-factortest. De toenemende alfawaarden weerspiegelen dat de items sterker samenhangen maar ook dat de invloed van meetfouten ('ruis') in de itemscores geringer wordt. Een 1-factortest kan dus sterk uiteenlopende waarden van alfa hebben, in principe tussen 0 en 1. Keren we de zaken om, en we zouden de variantie-covariantiematrix niet kennen, dan zouden we elke waarde van alfa kunnen vinden terwijl we steeds te maken hadden met een 1-factortest. Bij een waarde van *alfa* = .14 zou echter niemand beweren dat de test intern consistent is, maar toch is hij 1-factorieel en meten de items alle hetzelfde, zij het met veel ruis.

Men zou vervolgens kunnen redeneren dat het juist de 1-factortests zijn die niet intern consistent zijn, die een lage alfawaarde hebben. Deze redenering is te ondergraven door een langere test te nemen. Kiezen we bijvoorbeeld $J = 40$, itemvarianties van .25, en covarianties van .01, dan vinden we *alfa* = .625. Bij covarianties gelijk aan .02 vinden we *alfa* = .78. Ten opzichte van de test met vier items, is het enige verschil het grotere aantal items, wat dus voldoende is om alfa te laten toenemen van .14 naar .625, en van .26 naar .78. Als nu bijvoorbeeld alleen de alfawaarden .14 en .625 bekend zouden zijn, en we zouden geen andere voorkennis hebben, dan zouden we concluderen dat de tweede test een grotere interne consistentie heeft. Toch is dit niet zo: de covarianties zijn alle nog steeds .01, en de verschillende alfawaarden voor beide tests verwijzen naar dezelfde 1-factorstructuur. Het verschil zit hier dus alleen in het aantal items, maar dat heeft niets met interne consistentie te maken.

Interne consistentie en inconsistentie, en alfa

Tabel 2 laat twee variantie-covariantiematrices zien, waarvan de tweede 1-factorieel is ('intern consistent'; alle covarianties zijn gelijk aan $r = S_{jk} / S_j S_k = .06/\sqrt{.25}\sqrt{.25} = .24$), en alle correlaties zijn gelijk aan), maar de eerste een 2-factorstructuur heeft ('intern *in*consistent'): De eerste twee items en de laatste twee items hebben een covariantie van .16 (correlatie gelijk aan .64), terwijl de items uit de verschillende paren covarianties hebben van .01 (corrrelatie .04). In beide gevallen geldt $J = 4$ en zijn de itemvarianties, S_j^2 ($j = 1,\ldots, 4$), gelijk aan .25. Cruciaal is nu dat de som van de covarianties in beide gevallen gelijk is aan .72, en alfa daardoor aan .56.

Alfa hangt hier dus alleen af van de som van de covarianties, wat neerkomt op de *gemiddelde covariantie*, aangegeven als $\overline{S}$, terwijl de informatie over variatie tussen covarianties verloren gaat zodra men ze alle optelt. Beschikt men alleen over dit hogere aggregatieniveau, dan is de informatie over het lagere niveau, de afzonderlijke covarianties, niet meer terug te krijgen. Voor de volledigheid geven we ook de formule van alfa met de gemiddelde covariantie,

$$alfa = \frac{J^2 \overline{S}}{S^2_{x_+}}.$$

Dezelfde som van covarianties of, identiek, hetzelfde gemiddelde, kan dus evengoed verwijzen naar een 1- of 2-factorstructuur, en dat geldt ook weer voor elke andere som (of elk ander gemiddelde) en elke andere factorstructuur. Dus, *alfa* = .25 kan bij elke factorstructuur horen, maar dat geldt ook voor *alfa* = .90. Kent men alleen de waarde van alfa, dan zegt dit nog niets over de samenstelling van de test.

Zodra men begrijpt hoe alfa in elkaar steekt, wordt het gemakkelijk om voorbeelden te maken waarbij onder uiterst verschillende omstandigheden – bijvoorbeeld, 1, 2, 3, 4 factoren, verschillende aantallen items per factor – toch dezelfde alfa wordt gevonden. Daarbij moet de variantie-covariantiematrix wel steeds aan een technische voorwaarde voldoen, die ermee te maken heeft dat de covarianties onderling afhankelijk zijn, en de matrix ook bij echte data gevonden zou kunnen worden.

Conclusies over alfa en interne consistentie

We concluderen dat alfa sterk uiteenlopende waarden kan hebben voor

Tabel 2. Twee variantie-covariantiematrices met dezelfde waarde van Cronbachs alfa

	1	2	3	4	1	2	3	4
1	.25	.16	.01	.01	.25	.06	.06	.06
2	.16	.25	.01	.01	.06	.25	.06	.06
3	.01	.01	.25	.16	.06	.06	.25	.06
4	.01	.01	.16	.25	.06	.06	.06	.25

- verschillende 1-factortests die variëren in de hoeveelheid ruis (meetfouten); en
- tests met dezelfde interne consistentie, maar die bestaan uit verschillende aantallen items;

maar

- dat alfa dezelfde waarde kan hebben voor tests met verschillende factoriële samenstelling, waarvan sommige wel intern consistent zijn en andere niet.

Derhalve kan alfa niet worden opgevat als een maat voor interne consistentie.

Relatie alfa met betrouwbaarheid

De betrouwbaarheid is de correlatie tussen de testscores behaald op twee parallelle tests (Lord & Novick, 1968, p. 48). Testscores verkregen door middel van twee parallelle tests verschillen alleen van elkaar wat betreft de meetfouten, en zijn op te vatten als testscores behaald op twee onafhankelijke afnamen van dezelfde test. We geven de betrouwbaarheid aan met ρ. Voor een concrete test geldt dat $0 \leq \rho \leq 1$.

Alfa is gebaseerd op de gegevens van een enkele test, wat betekent dat we de gegevens van een tweede, parallelle test te kort komen. De helft van de benodigde gegevens ontbreekt dus, en dat gaat ten koste van de zuiverheid van de schatting van de betrouwbaarheid. Bewezen kan worden (bijvoorbeeld Drenth & Sijtsma, 2006, p. 215-220) dat alfa daardoor systematisch te laag is, dus dat $alfa \leq \rho$.

Gelijkheid wordt alleen gehaald onder theoretische condities waaraan in de praktijk van de psychologische test niet wordt voldaan. Van belang is de betekenis van de ongelijkheid: alfa geeft altijd een lagere waarde dan de echte betrouwbaarheid, en is daarmee een ondergrens voor de betrouwbaarheid. Dit betekent bijvoorbeeld dat als men pas tevreden is met een betrouwbaarheid van .85, een gerealiseerde waarde van $alfa = .85$ inhoudt dat de echte betrouwbaarheid hoger is.

Er bestaan diverse methoden die waarden geven die dichter bij de betrouwbaarheid liggen dan alfa. Een mogelijkheid is de methode die de grootste ondergrens voor de betrouwbaarheid levert. Deze grootste ondergrens wordt aangeduid als GLB (van greatest lower bound). Een andere mogelijkheid wordt geboden door kant en klare formules die een waarde geven die tussen alfa en de grootste ondergrens ligt.

De grootste ondergrens voor de betrouwbaarheid

Het probleem van het vinden van de GLB is in de jaren zeventig en tachtig opgelost. Het probleem is als volgt. We gaan uit van de vooronderstellingen van de klassieke testtheorie. Die komen erop neer dat testscores meetfouten bevatten, die nergens mee correleren. Dit wordt als uitgangspunt genomen voor een zoektocht naar de meetfoutenvarianties van de J items in de test, waarvan de som maximaal groot is bij de variantie-covariantiematrix die uit de data is berekend, onder de voorwaarde dat de berekeningen gebaseerd zijn op echte variantie-covariantiematrices voor true scores en meetfouten (Ten Berge & Sočan, 2004). Dit betekent dat gezocht wordt naar een 'worst-case scenario' voor de betrouwbaarheid: Gegeven de gevonden variantie-covariantiematrix van de items, wordt gezocht naar de laagst denkbare betrouwbaarheid. Dit is dan de grootste ondergrens, ofwel de GLB.

Is voor een bepaalde test de GLB gevonden, dan ligt de betrouwbaarheid van de testscore dus in het interval tussen de GLB en de theoretisch maximale betrouwbaarheid van 1: $GLB \leq \rho \leq 1$. Twee zaken vallen op. Ten eerste kan op basis van een enkele testafname alleen maar een *interval* voor ρ worden vastgesteld. De verklaring is dat er twee onafhankelijke testafnamen nodig zijn om ρ te kunnen schatten. Op basis van een enkele testafname legt het patroon van itemvarianties en -covarianties wel beperkingen op aan ρ, maar ρ in één waarde vastleggen gaat niet. Ten tweede kan men op basis van een enkele testafname niet uitsluiten dat de testscore perfect betrouwbaar is; immers, de waarde 1 behoort tot het interval. Deze onbepaaldheid volgt weer uit de beschikbaarheid van slechts één testafname, daar waar er twee nodig zijn.

Door alfa te rapporteren, doen testonderzoekers zichzelf tekort

Omdat alfa per definitie een kleinere ondergrens is dan de GLB, geldt dat $alfa \leq GLB \leq \rho \leq 1$. Alfa ligt dus buiten het interval van mogelijke waarden van ρ, en heeft derhalve altijd waarden die geen waarden van ρ kunnen zijn, voor zover dat op basis van een enkele testafname en onder aanname van de klassieke testtheorie valt te zeggen. Dus, als alfa voor een test gelijk zou zijn aan .8, dan weten we niet alleen dat de echte betrouwbaarheid ρ groter is, maar ook dat deze test geen betrouwbaarheid van .8 zou kunnen hebben. Op zich is dit weer geen ramp, als men ervan uit wil gaan dat een te lage schatting van de betrouwbaarheid door middel van alfa als aansporing kan worden opgevat om te streven naar een nog betere test.

De GLB zit niet in SPSS, maar is wel te schatten met behulp van software, zoals beschikbaar via http://www.ppsw.rug.nl/~kiers/ (Ten Berge & Kiers, 2003), http://www.citogroep.nl/oenw/onderzoek/psychometrie/eind_fr.htm (CITO, Arnhem), en http://www.mvsoft.com/eqs60.htm (het programma EQS; Bentler, 1995). Een nadeel is voorlopig nog dat de schattingen te hoog kunnen uitvallen in steekproeven kleiner dan duizend proefpersonen en tests groter dan tien items. Dat is een behoorlijke beperking, en het is te hopen dat de psychometrie met oplossingen komt voor dit probleem.

Andere alternatieven voor Cronbachs alfa

Reeds lang bestaan formules die net als alfa een ondergrens voor de betrouwbaarheid opleveren, maar gegarandeerd één die groter is dan alfa. We noemen Guttmans (1945) lambda2 en de mu-reeks van Ten Berge en Zegers (1978). De bijbehorende formules zijn net als alfa afhankelijk van de testlengte, J, de testscorevariantie, $S^2_{x_+}$, en de covarianties tussen de items, S_{jk}, maar ze zijn wel ingewikkelder wat betreft de structuur. Om die reden laten we ze weg, maar geven wel de relaties van deze ondergrenzen met alfa.

Ten eerste is bewezen dat *alfa* ≤ *lambda2*. Dus, als voor een concrete test *alfa* = .82, dan is lambda2 minstens zo groot. Alfa is in de mu-reeks van Ten Berge en Zegers (1978) het eerste element, mu0, en lambda2 het tweede, mu1. Deze auteurs toonden aan dat *mu0* (*alfa*) ≤ *mu1* (*lambda2*) ≤ *mu2* ≤ *mu3* ≤.... Zij raadden verder aan om mu2 of mu3 te berekenen voor tests die slechts uit enkele items bestaan. In het algemeen geldt echter dat alleen het verschil tussen mu0 (= alfa) en mu1 (=lambda2) de moeite waard is. Verder is elk van deze ondergrenzen kleiner dan de GLB, zodat we uiteindelijk hebben: *mu0* (*alfa*) ≤ *mu1* (*lambda2*) ≤ *mu2* ≤ *mu3* ≤ ... ≤ *GLB* ≤ ρ ≤ 1.

Het berekenen van mu2 en mu3 is dus alleen voordelig voor zeer korte tests, en de GLB is vooralsnog alleen te vertrouwen bij grote steekproeven en korte tests. Lambda2 kan zonder bezwaar altijd worden gerapporteerd in plaats van alfa. Ook al is lambda2 niet veel groter dan alfa, hij zit evenals alfa in SPSS, is altijd minstens zo groot, en ligt gegarandeerd dichter bij de echte betrouwbaarheid ρ.

Ten slotte kan vermeld worden dat er andere methoden bestaan om de betrouwbaarheid te schatten. Een mogelijkheid is om dit te doen via een confirmatief factormodel, zoals aanbevolen door Raykov (1997) en Bentler (2009). Revelle en Zinbarg (2009) stellen nog weer andere methoden voor. Hoewel deze methoden het alle beter doen dan alfa, hebben ze soortgelijke problemen in kleine steekproeven als de GLB (Ten Berge & Sočan, 2004).

Een rekenvoorbeeld

Alfa, lambda2 en de GLB werden berekend voor acht rating scale items, elk gescoord 0, 1, 2, 3, die waren afgenomen bij 828 respondenten. Met deze items werd 'coping behavior' gemeten van mensen die in de buurt woonden van industrie die stank verspreidde waar men regelmatig last van had (Cavalini, 1992). Er bestonden sterke aanwijzingen dat de acht items in twee viertallen konden worden verdeeld, die elk een verschillend aspect van coping behavior representeerden.

Zowel voor de totaalscore op alle acht items als de totaalscores op de twee viertallen geeft Tabel 3 de drie ondergrenzen. Noodzakelijkerwijs is lambda2 groter dan alfa, maar de verschillen zijn klein. De GLB werd berekend omdat de steekproef bijna duizend respondenten telde en het aantal items kleiner was dan tien. De verschillen van alfa en lambda2 met de GLB zijn spectaculair. Als we het mogelijke verschil tussen steekproef en populatie even vergeten en doen alsof het hier populatieresultaten betreft, dan zou de echte betrouwbaarheid dus weer groter zijn dan de GLB.

Tabel 3. Alfa, Lambda2 en de GLB voor drie totaalscores

	Aantal Items		
	8	4 (set 1)	4 (set 2)
Alfa	.778	.736	.640
Lambda2	.785	.746	.644
GLB	.852	.820	.696

Conclusies

We trekken twee conclusies uit dit onderzoek. Ten eerste is alfa geen maat voor de interne consistentie van een test, ondanks de vasthoudendheid waarmee dit in de literatuur wordt beweerd. Uitspraken zoals 'Voor deze test is alfa gelijk aan .88, en dus is de interne consistentie in orde' moeten ten zeerste worden afgeraden. De reden is dat een test met een alfa van .88, zowel intern consistent als intern *in*consistent kan zijn: alfa zegt hier gewoon niets over. Los van wat men er precies onder wil verstaan, kan men interne consistentie beter onderzoeken met behulp van factoranalyse en item-responstheorie.

Ten tweede is alfa bijna de kleinste ondergrens voor de betrouwbaarheid die bekend is. Door alfa te rapporteren, doen testonderzoekers zichzelf tekort. Voor steekproeven groter dan duizend waarnemingen en tests met hooguit tien items doet men er goed aan de GLB te schatten. Zolang voor kleinere steekproeven en langere tests het bias-probleem in de GLB-schatting niet is opgelost, kan men beter Guttmans lambda2 schatten. Het verschil met alfa is doorgaans klein, maar er is geen argument om deze kleine verbetering te negeren, zeker als men naast Guttmans lambda2 of de GLB ook alfa rapporteert. Dan is ook duidelijk dat alfa inderdaad een kleinere schatting oplevert dan de andere methoden.

Men zou kunnen beargumenteren dat men alfa alleen zou moeten berekenen voor testscores die zijn gebaseerd op items die hoog laden op dezelfde factor. Men gebruikt dan factoranalyse om de interne consistentie te onderzoeken en alfa om de betrouwbaarheid van de testscore per factor te schatten. Dat is een correcte werkwijze, maar staan blijft dat er betere methoden zijn dan alfa om de betrouwbaarheid te schatten. Die methoden zou men dan ook moeten gebruiken.

Testconstructeurs en testgebruikers lijken vaak niet te weten dat alfa geen maat is voor de interne consistentie van de test en ook niet dat alfa een onderschatting is van de betrouwbaarheid, maar in de psychometrie is deze kennis gemeengoed. Cortina (1993) en Schmitt (1996) bespreken op kritische wijze de vermeende relatie tussen alfa en interne consistentie, en Cronbach (1951) en Drenth en Sijtsma (2006, p. 215-222) geven het wiskundige bewijs dat alfa kleiner is dan de betrouwbaarheid. Ook Sijtsma (2009) bekritiseert het gebruik van alfa, terwijl de discussianten van dit artikel (Bentler, 2009; Green & Yang, 2009; Revelle & Zinbarg, 2009) het overwegend met hem eens zijn en hun bijdragen vervolgens wijden aan het schatten van de betrouwbaarheid met behulp van confirmatieve factoranalyse. Door bijvoorbeeld Cronbach (1954) en Borsboom (2006) is gewezen op de kloof tussen de theorie van de psychometrie en de praktijk van de testconstructie en het testgebruik. De bedoeling van dit artikel is het overbruggen van deze kloof voor wat betreft het begrip van Cronbachs alfa.

Prof.dr. K. Sijtsma is als hoogleraar verbonden aan het Departement Methoden en Technieken van Onderzoek, van de Faculteit Sociale Wetenschappen van de Universiteit van Tilburg, Postbus 90153, 5000 LE Tilburg. E-mailadres: <k.sijtsma@uvt.nl>.

Literatuur

Bentler, P.A. (1995). *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P.A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143.

Berge, J.M.F. ten & Kiers, H.A.L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. *Psychometrika, 56*, 309 -315.

Berge, J.M.F. ten & Kiers, H.A.L. (2003). *The minimum rank factor analysis program MRFA*. Internal report, Department of Psychology, University of Groningen, The Netherlands.

Berge, J.M.F. ten & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613-625.

Berge, J.M.F. ten & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*, 575-579.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Cavalini, P.M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries*. Ph.D. thesis, University of Groningen, The Netherlands.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L.J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika*, *19*, 263-270.

Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu van Loghum.

Green, S.A. & Yang, Y. (2009). Coefficient alpha: a cautionary tale. *Psychometrika*, *74*, 121-135.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184.

Revelle, W. & Zinbarg, R.E. (2009). Coefficients alpha, beta, omega and the glb. Comments on Sijtsma. *Psychometrika*, *74*, 145-154.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.

Misunderstandings concerning Cronbach's alpha

K. Sijtsma

summary

Cronbach's alpha is used for estimating both the internal consistency and the reliability of a test. It is demonstrated that alpha has no relationship to the test's internal consistency, considered as the degree to which items are associated as a sign of what they share in common. Instead of alpha, factor analysis and item response theory are recommended for investigating a test's internal consistency. Alpha is a lower bound to the reliability. Many alternative lower bounds are available, which exceed alpha and thus are closer to the test's reliability. A real-data example shows that the differences can be considerable. It is recommended to estimate the reliability using, for example, Guttman's lambda2 and the greatest lower bound instead of alpha.