

NIETS PRAKTISCHER DAN EEN GOEDE THEORIE: GENERALISEERBAARHEIDSTHEORIE ALS INSTRUMENT VOOR BETROUWBAARHEIDSTUDIES

C.P.M. Van der Vleuten
Projectleider
Project Evaluatie Studieresultaten

W.H.F.W. Wijnen
Hoogleraar
Onderwijsontwikkeling en -
research

Correspondentieadres:
Vakgroep Onderwijsontwikkeling
en -research
Rijksuniversiteit Limburg
Postbus 616
6200 MD Maastricht

Recente studies op het gebied van toetsing, vooral op het terrein van medische competentie, maken nogal eens gebruik van de generaliseerbaarheidstheorie: een methode voor de schatting van de betrouwbaarheid van scores verkregen met een meetinstrument. Houdt men zich in hoofdzaak bezig met toegepast onderzoek en is men niet specifiek opgeleid in psychometrische methodieken, dan zijn deze studies moeilijk te begrijpen. Soms worden ze zelfs afgedaan als hocus-pocus, aardig voor de toevallige getallenkraker, maar onbegrijpelijk voor de 'leek' geïnteresseerd in het onderwerp. De frequentie waarmee de theorie wordt toegepast, beweegt de meer geïnteresseerden ertoe eens te gaan napluizen wat de theorie inhoudt en hoe deze te gebruiken is in de eigen concrete situatie. Deze geïnteresseerden worden echter gemakkelijk gefrustreerd, omdat de bestaande teksten over de generaliseerbaarheidstheorie moeilijk toegankelijk zijn.

In dit artikel zal een poging worden gedaan de theorie nader te beschrijven. Gekozen is voor een bespreking, zoveel mogelijk in woorden - eventueel toegelicht met voorbeelden - zodat een teveel aan formules kan worden vermeden. Het ligt in de bedoeling de lezer ten minste een intuïtief begrip te laten krijgen van deze betrouwbaarheidstheorie. Wellicht worden daardoor sommige aspecten te zeer gesimplificeerd. Voor een meer statistische verantwoording wordt de lezer gewezen op de aan het einde van dit artikel opgenomen geannoteerde literatuur.

Ondanks de schijnbare complexiteit van de generaliseerbaarheidstheorie zijn de basisprincipes betrekkelijk eenvoudig. Wie eenmaal van de eerste schrik bekomen is, zal merken dat de generaliseerbaarheidstheorie uitermate geschikt is voor veel concrete onderzoekssituaties: het is een uitermate praktische theorie.

DRIE BETROUWBAARHEIDSTHEORIEËN

Betrouwbaarheid verwijst naar de precisie van gegevens verkregen met behulp van een meetinstrument. De precisie van instrumenten voor fysische metingen (bijvoorbeeld lengte) is doorgaans eenvoudig vast te stellen door een herhaalde meting uit te voeren en de uitkomsten te vergelijken. In de sociale weten-

schappen is een dergelijke benadering meestal niet mogelijk, omdat hier de meting zelf invloed heeft op het object van meting. Bij herhaling van bijvoorbeeld een studietoets zal een persoon zich vragen weten te herinneren en dientengevolge anders scoren. Bovendien geldt voor de sociale wetenschappen veel sterker dan voor de natuurwetenschappen, dat allerlei andere ruisfactoren van invloed zijn op de meting, zoals bijvoorbeeld de duur van de meting (vermoeidheid), de dag van de meting (iemand kan een slechte dag hebben), interpretatieruimte bij het nakijken van antwoorden, etcetera. De vaststelling of schatting van de betrouwbaarheid dient in de sociale wetenschappen op een aangepaste manier te worden verricht. Historisch gezien zijn er drie theorieën, die elk op een bepaalde manier betrouwbaarheid opvatten en uitwerken: de *klassieke test-theorie*, de *generaliseerbaarheidstheorie* en *item-respons theorieën*.

In de *klassieke testtheorie* wordt uitgegaan van het concept van parallelle metingen: een test wordt opgesplitst in delen en de mate waarin de gegevens betreffende deze delen met elkaar overeenkomen, zeg maar correleren, leidt tot een schatting van de betrouwbaarheid van de test als geheel. Een test kan in tweeën worden gedeeld, de zogenaamde *split-half methode*, maar het is ook mogelijk om elk item als een afzonderlijke test op te vatten. De meest

gangbare betrouwbaarheidsindex, Cronbach's alfa, is hierop gebaseerd.

De *generaliseerbaarheidstheorie* is in feite een uitbreiding van de klassieke testtheorie. De klassieke testtheorie let in feite maar op één foutenbron, namelijk de mate waarin de resultaten van items in een test met elkaar overeenkomen. De generaliseerbaarheidstheorie kan meerdere foutenbronnen identificeren, namelijk al die foutenbronnen waarover de onderzoeker informatie verzamelt. Bijvoorbeeld: stel dat we over een toets beschikken bestaande uit 10 open vragen die door meerdere beoordelaars wordt nagekeken. Volgens een klassieke testbenadering kan een alfa worden berekend. Deze alfa geeft de mate aan waarin de items als afzonderlijke subtoetsen met elkaar overeenstemmen. De beoordelaars vormen echter in deze situatie een andere potentiële foutenbron. Men kan natuurlijk berekenen in hoeverre de beoordelaars met elkaar overeenstemmen. Samen met de alfa levert dat twee betrouwbaarheidsindicaties op, die elk op hun waarde moeten worden gezien. Onduidelijk blijft echter hoe deze indicaties zich tot elkaar verhouden. Met behulp van de generaliseerbaarheidstheorie kunnen beide foutenbronnen worden geïntegreerd en kan één betrouwbaarheidsschatting worden verkregen.

Omdat meerdere foutenbronnen kunnen worden geïntegreerd in één model, wordt betrouwbaarheid niet opgevat als een enkel vaststaand gegeven (of als een enkele betrouwbaarheidsindex). Afhankelijk van de vraagstelling van de onderzoeker kunnen meerdere betrouwbaarheden relevant zijn. Vandaar dat minder de vraag centraal staat hoe betrouwbaar verzamelde gegevens zijn, maar eerder welke bedoelingen de onderzoeker heeft met de gegevens, of, in andere woorden, waar wil de onderzoeker naar generaliseren. De betrouwbaarheidsvraag verandert zo in een generaliseerbaarheidsvraag en daardoor kan de onderzoeker meer genuanceerde uitspraken doen over zijn

gegevens. De onderzoeker dient uitdrukkelijk aan te geven welke generalisatiewaarde toegekend kan worden aan de verkregen gegevens. We zullen deze kwestie later concreet uitwerken. Van belang is hier dat in de generaliseerbaarheidstheorie meerdere foutenbronnen toegelaten zijn, hetgeen de praktische bruikbaarheid aanmerkelijk vergroot, en de onderzoeker de mogelijkheid biedt genuanceerdere uitspraken te doen over de geldigheidswaarde van de verkregen gegevens.

Item-respons theorieën tenslotte vormen een geheel andere klasse van betrouwbaarheidstheorieën. Zowel de klassieke testtheorie als de generaliseerbaarheidstheorie nemen de totale testscore als vertrekpunt: deze wordt uitgesplitst in deelscores en het verband daartussen, de herhaalbaarheid, wordt nagegaan. Item-respons theorieën nemen het item als vertrekpunt: een theorie wordt geformuleerd over een item (bijvoorbeeld de kans om een item goed te beantwoorden wordt bepaald door een logistische relatie met de vaardigheid of competentie van een persoon) en nagegaan wordt of de in werkelijkheid gevonden gegevens overeenstemmen met de theorie. Is dat het geval, dan vormt een verzameling items die allen aan dezelfde eisen van de theorie voldoen een goede 'meetlat', waarmee betrouwbare uitspraken kunnen worden gedaan (de mate van precisie is eveneens in maat en getal uit te drukken). Er zijn meerdere item-respons theorieën, die elk een andere 'theorie' poneren over de wijze waarop antwoorden van kandidaten kunnen worden gemoduleerd. De eenvoudigste en bekendste item-respons theorie is het Rasch-model, maar het is eveneens het strengste model, waaraan veel items niet voldoen. Het voert op deze plaats te ver om dieper op item-respons theorieën in te gaan. Belangrijk is dat deze betrouwbaarheidstheorieën door hun sterke mathematische eigenschappen 'sterkere' modellen zijn dan de klassieke testtheorie en de generaliseerbaarheidstheorie. Daar staat tegenover dat zij 'strenger' zijn (of anders gezegd meer items afkeuren) en minder

praktisch bruikbaar zijn. Evenals de klassieke testtheorie, staan item-respons theorieën meerdere foutenbronnen niet toe, hetgeen eveneens hun praktische bruikbaarheid begrenst.

GENERALISEERBAARHEIDSTUDIES

Een betrouwbaarheidsstudie die gebruik maakt van de generaliseerbaarheidstheorie bestaat in feite uit een tweetal stappen: de generaliseerbaarheidsstudie (G-study) en de besluitvormingsstudie (decision-study; D-study). In de G-studie worden alle variantiebronnen waarin de onderzoeker geïnteresseerd is - zowel de foutenbronnen als de gewenste variantiebronnen - in kaart gebracht. In de D-studie worden de gegevens van de G-studie vervolgens gebruikt om concrete betrouwbaarheidsberekeningen te verrichten. Wanneer uit de G-studie bekend is hoe groot de verschillende variantiebronnen zijn, kan worden nagegaan met welke aantallen (items, beoordelaars, etcetera) gewerkt dient te worden voor het behalen van een aanvaardbare betrouwbaarheid. Bijvoorbeeld, wanneer de variantie tussen verschillende correctoren van opstellen groot is, zal een beoordeling door meerdere correctoren een grotere betrouwbaarheid opleveren. Op zowel de G-studie als de D-studie zullen we uitgebreider ingaan.

G-STUDIE

De basis voor een G-studie is een variantieanalyse (ANOVA). De variantie-analyse wordt gevolgd door het schatten van variantiecomponenten, waardoor de variantie van de verschillende factoren in verhouding tot elkaar

worden uitgedrukt. De precieze techniek daarvan laten we hier achterwege. Door middel van deze analyse wordt de variantie van alle in het onderzoek betrokken factoren in kaart gebracht. Welke factoren geschat kunnen worden hangt natuurlijk van de onderzoeker af en van de gegevens die verzameld zijn. De variantie behorend bij beoordelaars kan bijvoorbeeld alleen worden geschat als men beschikt over ten minste dubbele beoordelingen.

Laten we beginnen met een simpel voorbeeld. Gesteld we hebben een toets afgenomen bij 100 studenten bestaande uit een 10-tal open vragen. In dit geval hebben we voor elk van de 100 studenten 10 scores (op een 10-puntschaal) voor 10 items. Volgens de theorie is hier dan sprake van slechts één factor, de items, en in termen van de theorie hebben we hier te maken met een one-facet design. Hoewel er ook een factor studenten (Personen) is, wordt dit niet als een facet meegerekend. Het aantal variantiebronnen dat onderscheiden kan worden is dus ook beperkt: we kunnen variantie onderscheiden behorende bij de studenten (Personen), bij de items, en bij de combinatie van deze twee effecten, de interactie tussen personen en items. In tabel 1 zijn deze variantiebronnen weergegeven met de daarbij behorende geschatte variantiecomponenten. Omdat het om schattingen gaat, varieert de nauwkeurigheid van de schatting met de grootte van de steekproef: de variantieschatting berekend over 1000 items zal nauwkeuriger zijn, dan een schatting gebaseerd op 10 items. In kolom 2 is deze nauwkeurigheid van schatting weergegeven door vermelding van de 'standard error'. Aangezien we hier een schatting hebben verricht op slechts een beperkt aantal items, is de nauwkeurigheid van deze schatting niet erg groot. Dat blijkt dan ook uit de grote standard error van deze component. Hoewel het van belang is om de standard errors in publikaties te vermelden, zullen deze in het verdere artikel achterwege worden gelaten. In de rechterkolom van tabel 1 zijn de variantiecomponenten in procenten weergegeven.

Het persoonseffect geeft de variantie aan tussen de scores van de studenten en vormt met ongeveer 13 procent van de totale variantie het kleinste effect. De variantie behorende bij de itemcomponent is bijna drie keer groter en dit effect wijst op gemiddelde verschillen

Tabel 1.
G-studie van een one-facet design (Personen x Items) gebaseerd op een toets bestaande uit 10 open vragen (gemaakt door 100 studenten)

Variantiebron	Geschatte variantiecomponent	Standard error	Percentage van de totale variantie
Personen (P)	97.57	19.05	13.35
Items (I)	261.24	112.98	35.75
PI, error	371.97	17.60	50.90

tussen vragen, ofwel de variantie in de moeilijkheidsgraad van de items. Met ongeveer 36 procent van de totale variantie betekent dit dat de vragen nogal in moeilijkheid van elkaar verschillen. PI wijst op de mate waarin personen anders 'gerangordend' worden door de verschillende items. Dit effect is echter niet af te splitsen van alle andere foutenbronnen die hebben bijgedragen aan de scores op deze toets (vermoeidheid, tijdstip, etcetera) en wordt dan ook gerekend tot de algemene foutenterm. Ongeveer de helft van de totale variantie blijkt dus een algemene foutenterm te zijn. In vrijwel alle metingen die zo zijn opgezet (in jargon: 'met dit design') is de error term zo groot.

Een toets wordt afgenomen om onderscheid te kunnen maken tussen kandidaten; we zijn dan met name geïnteresseerd in de variantie tussen personen. Dit wordt de ware variantie genoemd of in meer specifieke termen van de theorie het object van meting. Het object van meting kan natuurlijk ook iets anders zijn, afhankelijk van de vraagstelling van de onderzoeker. Bijvoorbeeld, wanneer uitspraken gedaan moeten worden over de kwaliteit van een onderwijsprogramma met behulp van toetsgegevens, dan is men niet geïnteresseerd in wat individuele personen gescoord hebben, maar in gemiddelden van groepen personen (bijvoorbeeld, groepen die verschillende programma's hebben doorlopen). Deze groepen vormen in dat geval het object van meting.

Of en in hoeverre bepaalde bronnen tot de echte foutenbronnen moeten worden gerekend is afhankelijk van de uitspraken die men op basis van de gegevens wil verrichten. De algemene error term is altijd een daadwerkelijke foutenbron. Gezien de relatieve grootte ervan zal een redelijk aantal items noodzakelijk zijn om het effect van deze foutenbron te minimaliseren. Hoe dat in zijn werk gaat zullen we bij de berekening van de feitelijke betrouwbaarheid zien. De variantiebron behorend bij de items kan een foutenbron zijn. Gesteld dat niet elke student dezelfde vragen voorgelegd krijgt, maar verschillende vragen (bijvoorbeeld omdat de toets op verschillende tijdstippen wordt afgenomen en de inhoud bekend kan worden), in dat geval zullen verschillen in de moeilijkheidsgraad van de items, en daarmee van de toets in zijn geheel, sommige studenten bevor-

delen, anderen benadelen. De variantiebron behorend bij de items is dan een foutenbron.

De kern van de G-studie is het in kaart brengen van alle variantie gemoeid met een bepaalde meting. Het gebruikte voorbeeld vormt eigenlijk de eenvoudigste opzet van een meting, maar in een complexere opzet zijn de spelregels hetzelfde.

Gesteld dat eenzelfde toets wordt nagekeken door twee beoordelaars. Dat impliceert dat we eveneens informatie kunnen verzamelen over de variantie die veroorzaakt is door verschillen tussen beoordelaars. We beschikken dan over een twee-factoren design (two-facet design, items en beoordelaars; het object van meting, de studenten, wordt niet als aparte factor gerekend). Het aantal te schatten variantiebronnen neemt daarmee sterk toe. We hebben de hoofdeffecten Personen (studenten), Items, en Beoordelaars en alle interactie-effecten tussen die genoemde hoofdeffecten. In tabel 2 zijn deze weergegeven, met de variantiecomponenten.

De effecten P en I hebben dezelfde betekenis als in tabel 1. B is het hoofdeffect beoordelaars en staat voor de mate waarin de beoordelaars systematisch verschillend scoorden (streng of soepeler zijn). PI is de mate waarin de studenten anders worden 'gerangordend' door de items. In tegenstelling tot de vorige G-studie staat dit design toe het PI-effect te schatten onafhankelijk van de algemene error term. Zoals blijkt uit de grootte van de component, is dit een zeer fors effect. In vrijwel alle metingen van medische competentie blijkt dat het geval te zijn. Dit impliceert dat een score op het ene item weinig voorspellende waarde heeft voor (correleert met) de score op een ander item. Daarom zijn er veel items nodig om tot een betrouwbaar, of repro-

Tabel 2.
G-studie van een two-facet design (Personen x Items x Beoordelaars)

Variantiebron	Geschatte variantie component	Percentage van de totale variantie
Personen (P)	48.71	10.57
Items (I)	25.12	5.45
Beoordelaars (B)	15.00	3.26
PI	185.87	40.33
PB	33.18	7.20
IB	80.00	17.36
PIB, error	72.94	15.83

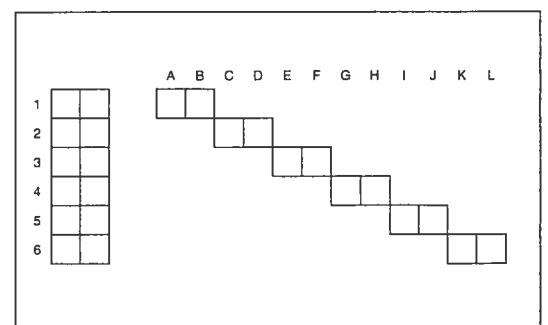
duceerbaar, resultaat te komen. In de literatuur over medische competentie wordt dit vaak aangeduid als het probleem van de inhoudsspecificiteit. PB duidt aan in hoeverre de beoordelaars de personen verschillend ordenen. Het hoofdeffect B geeft een algemeen verschil in de beoordeling weer (over alle personen), terwijl PB aangeeft in hoeverre beoordelaars sommige studenten anders beoordelen dan anderen. Beoordelaars kunnen gemiddeld genomen weinig van elkaar verschillen (gelijk aan B), maar tegelijkertijd kunnen zij de beoordeelde personen volledig anders 'rangordenen' (gelijk aan PB) en omgekeerd. Het IB-effect kan op dezelfde wijze worden opgevat, maar nu voor de mate van inconsistentie over items. Beide effecten zijn in deze studie niet onaanzienlijk. Tot slot geeft PIB de variantie weer van de interactie tussen personen, items, en beoordelaars, maar deze is opnieuw niet af te splitsen van de overige foutenbronnen en vormt daarom de algemene error term.

Variantiecomponenten vormen eigenlijk het hart van de generaliseerbaarheidstheorie. Hoewel betrouwbaarheid uitgedrukt wordt in velerlei betrouwbaarheidsindices, zijn het de variantiecomponenten waarop deze gebaseerd zijn. Zij vormen de basisgegevens voor alle verdere berekeningen en geven de onderzoeker (en de lezers) een tamelijk genuanceerd beeld over de betrouwbaarheid van een meting. Bijvoorbeeld, het is in het bovenstaande onderzoek mogelijk om een interbeoordelaarsbetrouwbaarheid uit te rekenen (bijvoorbeeld door middel van een correlatie, een percentage onderlinge overeenstemming, of een kappa), maar de variantiecomponenten geven een veel genuanceerder beeld. Zij geven aan of beoordelaars in het algemeen sterk van elkaar verschillen, dan wel of beoordelaars erg inconsistent zijn over verschillende items. In het eerste geval heeft men te maken met strenge en toegeeflijke beoordelaars en zou een remedie kunnen bestaan uit training of selectie van beoordelaars. In het tweede geval zou een (beter) beoordelingsvoorschrift bij sommige items uitkomst kunnen bieden. Bovendien kunnen al deze variantiebronnen in relatie tot elkaar worden beoordeeld. Uit tabel 2 blijkt duidelijk dat de variantie die wordt veroorzaakt door de beoordelaars, verhoudingsgewijs gering is ten opzichte van de overige varian-

tiebronnen. In publikaties waarbij van de generaliseerbaarheidstheorie gebruik wordt gemaakt is het dan ook van groot belang de variantiecomponenten te rapporteren.

Er is nog een andere belangrijke reden voor het publiceren van de variantiecomponenten. Het is niet altijd noodzakelijk om zelf G-studies uit te voeren. Men kan ook gebruik maken van de resultaten van eerder gepubliceerde gegevens. Men kan op grond van door andere onderzoekers gerapporteerde variantiecomponenten D-studies verrichten, specifiek gericht op de eigen situatie, of men kan gedeelten van de G-studies gebruiken voor een eigen G-studie. Wanneer bijvoorbeeld verschillende studies een consistent beeld geven over de variantie geassocieerd met een bepaald facet (zoals beoordelaars), dan is het mogelijk deze gegevens 'in te vullen' in een eigen G-studie waarin naast de 'geleende' componenten de overige componenten worden geschat op basis van de eigen gegevens. Bovendien is het mogelijk verschillende componenten van hetzelfde design te middelen over verschillende G-studies heen (meestal gewogen naar steekproefgrootte). Zo kunnen verschillende kleine studies worden gecombineerd tot een grotere, waardoor de precisie van de componentschattingen toeneemt.

De variantiebronnen die worden geschat in een G-studie zijn afhankelijk van de door de onderzoeker gewenste en verzamelde gegevens. Bovendien bepaalt de wijze waarop de gegevens zijn verzameld, welke bronnen op welke manier kunnen worden geschat. Bijvoorbeeld, wanneer in bovenstaande toets-



Figuur 1. Schematische weergave van een gekruist (links) en een genest design (rechts) van 6 vragen (1 t/m 6) nagekeken door verschillende beoordelaars (A t/m L).

situatie niet twee beoordelaars alle vragen zouden hebben nagekeken, maar 14 beoordelaars, waarvan elk paar één vraag voor alle studenten had nagekeken, dan had dit consequenties voor het design gehad. In het eerste geval is sprake van een 'gekruijsd' design (alle beoordelaars zien alle vragen van alle studenten), in het tweede geval zouden de beoordelaarsgegevens 'genest' zijn (figuur 1).

Het gevolg van nesten is dat een aantal bronnen niet meer te isoleren is. Bijvoorbeeld, de verschillen die worden gevonden tussen beoordelaars over items (PI) kunnen ook verklaard worden door algemene verschillen tussen paren beoordelaars (B). Door deze verstrengeling moeten beide componenten worden opgeteld en ontstaat een nieuwe term (B:I; beoordelaars genest binnen items). Het nesten van factoren kan een gunstige uitwerking hebben in praktische situaties. We zullen hier later op terugkomen. Op verschillen tussen designs en hun consequenties zal nu niet verder worden ingegaan. We volstaan met de opmerking dat men voor het verrichten van G-studies in het algemeen gebaat is bij gekruiste designs, die overigens in de praktijk niet altijd haalbaar zijn. Bij gekruiste designs kunnen echter meer variantiebronnen van elkaar worden onderscheiden, waardoor de meest genuanceerde informatie verkregen wordt.

D-STUDIES

De gegevens uit G-studies worden gebruikt voor het berekenen van betrouwbaarheidsindices. Daarvoor is het nodig expliciet te maken wat betrouwbaarheid nu eigenlijk is. De basale redenering hierover is dat iemands score op een test te zien is als een ruwe score, die feitelijk bestaat uit twee delen: de werkelijke of ware score en een deel dat moet worden gezien als 'ruis' of meetfout. Deze laatste component bestaat uit alle factoren die onbedoeld invloed hebben gehad op de meting. Op basis van deze redenering wordt de betrouwbaarheid uitgedrukt door een breuk waarin gewenste, ware variantie en ongewenste, foutenvariantie zijn opgenomen. In termen van de theorie: de breuk geeft de verhouding tussen ware variantie en geobserveerde variantie. Tot nu toe wijkt de generaliseerbaarheidstheorie in deze redenering

niet af van de klassieke testtheorie. De afwijking van de klassieke testtheorie is dat in de geobserveerde variantie meerdere variantiebronnen kunnen worden opgenomen. De ware variantie wordt bepaald door het object van meting; in de bovenstaande voorbeelden de persoonsvariantie. De geobserveerde variantie is afhankelijk van de uitspraken die de onderzoeker wenst te doen. Afhankelijk van het doel van de meting, worden alle variantiebronnen die voor dat doel verstorend zijn, in de geobserveerde variantie meegeteld. Populair gezegd is de betrouwbaarheidsberekening vergelijkbaar met de scoring bij bridge: wat wenselijk is komt boven de streep, wat onwenselijk is komt onder de streep. Beginnend met het simpele voorbeeld van hierboven (tabel 1) kan een betrouwbaarheidscoëfficiënt als volgt worden bepaald, waarbij n staat voor het aantal elementen dat men gebruikt binnen een bepaald facet (bijvoorbeeld aantal items):

$$\frac{P}{P + PI/n_i}$$

De variantie van de personen (P) komt in de teller te staan en vormt de ware variantie. Deze component wordt in de noemer herhaald zodat een breuk wordt verkregen resulterend in een getal tussen 0 en 1. Vervolgens worden die componenten in de noemer toegevoegd die als foutenvariantie worden beschouwd. Meestal zijn dat al die componenten die de rangorde van personen beïnvloeden: dat zijn alle bronnen waarin het element P voorkomt. In het simpele voorbeeld is dat PI . Deze component mag worden gedeeld door het aantal items waaruit de toets bestond. Dus:

$$\frac{97.57}{97.57 + 371.97/10} = 0.72$$

Deze betrouwbaarheidsindex wordt een generaliseerbaarheidscoefficiënt genoemd en kan als volgt worden geïnterpreteerd: het geeft de correlatie (rangorde) weer tussen de scores verkregen op deze toets en een denkbeeldige andere toets bestaande uit 10 willekeurige nieuwe items uit hetzelfde domein. Door te variëren met het aantal items, kunnen we eveneens nagaan welke betrouwbaarheid bereikt wordt met een grotere of kleinere

toets. Bij 5 vragen is deze 0.57, bij 15 0.80 en bij 30 0.90. Naarmate de toets langer wordt, neemt ook de betrouwbaarheid toe. Intuïtief is dat ook logisch. Immers, naarmate de steekproef een groter percentage van het aantal mogelijke items omvat, zal het minder uitmaken welke steekproef wordt voorgelegd: de steekproef zal steeds beter het domein dekken. Doorgaans wordt een betrouwbaarheidscoëfficiënt van 0.80 als een minimum beschouwd. Deze berekeningsprocedure kan worden gebruikt om voor een bestaande meting de betrouwbaarheid te berekenen, maar kan ook worden toegepast om besluiten te nemen over de samenstelling van nieuwe metingen.

In de berekening van de bovenstaande betrouwbaarheidscoëfficiënt zijn alleen de foutenbronnen opgenomen die de rangorde van de studenten beïnvloeden. Daarmee wordt een keuze gemaakt ten aanzien van de uitspraken die op de toets mogen worden gebaseerd: wanneer een andere toets met andere items wordt gebruikt, zijn verstoringen in de rangorde van personen niet gewenst, maar verstoringen door verschillen in de moeilijkheid van de items (toets) worden getolereerd. Betrouwbaarheid wordt in dit geval gezien vanuit een norm-georiënteerd perspectief, dat de scores van personen betekenis geeft in relatie tot elkaar (bijvoorbeeld, iemand hoort tot de 10% besten). Wanneer een absolute betekenis aan scores wordt toegekend (iemand moet ten minste 70% van een leerstofgebied beheersen), spreken we over een domein-georiënteerd perspectief (de term 'criterium-gerichte toetsing' wordt vaak ook gebruikt). Niet alleen verstoringen in de rangorde, maar ook de verstoringen in het algemeen niveau worden dan tot de foutenbronnen gerekend. In de bovenstaande formule dient in dat geval de variantiecomponent behorend bij de items in de noemer te worden opgenomen, eveneens gedeeld door het aantal items. De resulterende coëfficiënt wordt dependability-coëfficiënt genoemd. Deze is in het voorbeeld 0.44; een stuk lager dan de generaliseerbaarheidscoëfficiënt.

Dezelfde berekeningswijze zal moeten worden gevolgd wanneer de onderzoeker van plan is verschillende items aan (groepen) studenten voor te leggen. Ook dan zijn afwijkingen in moeilijkheidsgraad niet gewenst, omdat dat sommige personen bevoordeelt,

anderen benadeelt. De variantie van de items, in dit geval de verschillen in moeilijkheidsgraad van de vragen, wordt dan als een foutenbron opgevat. De waarde van de dependability coëfficiënt is de juiste waarde voor een norm-georiënteerd perspectief, onder de conditie dat studenten verschillende items krijgen voorgelegd.

Voor het meer complexe tweede voorbeeld (tabel 2) kan een volkomen analoge procedure worden gevolgd. De generaliseerbaarheidscoëfficiënt wordt bepaald door de verhouding tussen de persoonsvariantie en de geobserveerde variantie, waarbij deze laatste wordt bepaald door alle overige bronnen waardoor de rangordening wordt verstoord:

$$\frac{P}{P + PI/n_i + PB/n_b + PIB/n_i n_b} = \frac{48.71}{48.71 + 185.87/10 + 33.18/2 + 72.94/20} = 0.56$$

De betekenis: wanneer een denkbeeldige andere set van 10 vragen was afgenomen, nagekeken door twee denkbeeldige andere beoordelaars, mag men tussen de resultaten een verband verwachten van 0.56. Vanuit een domein-georiënteerd perspectief worden de overige variantiebronnen eveneens in de noemer opgenomen, hetgeen resulteert in een betrouwbaarheidscoëfficiënt van 0.48.

De betrouwbaarheidscoëfficiënt kan dus nogal variëren, afhankelijk van het doel van de metingen en van de generalisaties die een onderzoeker wenst te maken. Bijvoorbeeld, gesteld dat deze laatste toets een formatieve intreetoets is voor een klinische stage, bedoeld voor het zelf-inzicht van de studenten. Geheimhouding speelt dan geen rol. Men besluit dezelfde toets met dezelfde vragen bij alle instromende studenten af te nemen. Dit impliceert dat geen generalisatie hoeft plaats te vinden naar een denkbeeldige andere set van items en dat alle hiermee verbonden variantiebronnen komen te vervallen bij de vaststelling van de geobserveerde variantie: de betrouwbaarheid wordt dan 0.75. Denkbaar is ook een situatie, waarin verschillende vragen worden gebruikt in elke nieuwe toets, maar dat telkens dezelfde beoordelaars deze na-

kijken. In dat geval is de variantie afkomstig van de beoordelaars niet relevant. Ze verdwijnt in de berekening: de generaliseerbaarheidscoëfficiënt wordt dan 0.81.

De variantiecomponenten van de verschillende variantiebronnen vormen dus de basis voor de berekening van de betrouwbaarheid. De precieze berekening en de betekenis van de betrouwbaarheid is echter afhankelijk van de bedoelingen van de onderzoeker. Op grond van de variantiecomponenten kan de betrouwbaarheid van een bepaalde meting worden bepaald of kunnen besluiten voor de (toekomstige) inrichting van toetsen worden genomen. We kunnen bijvoorbeeld nagaan of en hoeveel de winst is van het toevoegen van items en/of beoordelaars. In de bovenstaande berekeningen worden simpelweg andere delingen uitgevoerd en kan het resultaat worden vergeleken. Ook is het mogelijk om de invloed na te gaan van andere afnameprocedures. Gesteld dat we de beschikking hebben over meer beoordelaars, maar dat we elke beoordelaar minder willen laten nakijken. In praktische situaties is dat een vaak voorkomend gegeven, omdat voor grote aantallen studenten het correctiewerk een groot beslag legt op de tijd van de beoordelaars. Opsplitsing van dit werk ligt dan voor de hand, maar de vraag is hoe dit het beste kan worden gedaan. Gesteld nu dat de opsplitsing plaatsvindt door elke beoordelaar één vraag te laten nakijken voor alle studenten. Deze procedure heeft consequenties voor de te onderscheiden variantiebronnen. Het is niet meer mogelijk om algemene verschillen tussen beoordelaars (B) te onderscheiden van inconsistente beoordelingen over items (PI), omdat deze ook tot stand kunnen zijn gekomen door verschillen tussen beoordelaars. Evenmin kan de inconsistentie van de beoordelaars over personen (PB) onderscheiden worden van de algemene error (PBI). Beoordelaars zijn door deze indeling 'genest' binnen items. Het gevolg is dat deze effecten samenvallen en bij elkaar opgeteld worden. De variantiecomponenten worden nu: $P = 48.71$; $I = 25.12$; $B:I = 95.00$ (B genest binnen items, is $B + IB$); $PI = 185.87$; $PB:I$ (is $PB + PBI$). Volgen we de regels voor de samenstelling van ware variantie en geobserveerde variantie, dan wordt de generaliseerbaarheidscoëfficiënt:

$$\frac{P}{P + PI/n_i + PB:I/n_i n_b} = \frac{48.71}{48.71 + 185.87/10 + 106.12/20} = 0.67$$

De coëfficiënt valt hoger uit dan de voorgaande berekeningen, omdat de PB component, die oorspronkelijk gedeeld werd door 2 beoordelaars, nu verbonden is met de PBI component. Deze PBI component mag in dit geval worden gedeeld door 20: het produkt van het aantal beoordelaars en het aantal items. Intuïtief is dat niet onlogisch. Immers, door beoordelaars te nesten binnen items komt het oordeel over de competentie van een student in feite tot stand door een veelvoud van verschillende beoordelaars. Fouten die geïntroduceerd worden door het toevallige paar dat een persoon krijgt toegewezen, worden door een veelheid aan beoordelaars gecompenseerd. Het is dan ook veel verstandiger een opsplitsing van beoordelaarswerk naar items door te voeren, dan een opsplitsing naar studenten.

Door variatie in de aantallen en door variatie in de toetsprocedure kunnen betere beslissingen worden genomen over de toekomstige inrichting van een meting. In tabel 3 wordt een overzicht gegeven van generaliseerbaarheidscoëfficiënten berekend volgens twee verschillende correctieprocedures en het aantal items en beoordelaars dat wordt gebruikt, gebaseerd op de variantiecomponenten uit tabel 2.

Tabel 3.
Generaliseerbaarheidscoëfficiënten als functie van het aantal items, beoordelaars en correctieprocedure

Aantal items	Dezelfde beoordelaar per item		Verschillende beoordelaars per item	
	Eén beoordelaar	Twee beoordelaars	Eén beoordelaar	Twee beoordelaars
10	0.45	0.56	0.63	0.67
20	0.51	0.64	0.77	0.80
30	0.54	0.67	0.83	0.86
40	0.55	0.69	0.87	0.89
50	0.56	0.70	0.89	0.91

Uit de tabel blijkt dat wanneer dezelfde beoordelaar alle items van alle personen corrigeert, het verlengen van de toets weinig invloed meer heeft. Het toevoegen van een tweede beoordelaar heeft blijkbaar meer nut. Een nog betere betrouwbaarheidswinst wordt verkregen wanneer verschillende beoordelaars worden ingezet per item. Daarentegen hebben dubbele beoordelingen per item, hetgeen nogal wat beoordelaars vergt, relatief weinig invloed. Toename van het aantal items levert hier meer winst op dan in de andere correctieprocedure. De beste samenstelling wordt in dit geval wellicht bereikt door ten minste een verdubbeling van het aantal items en door een zodanige opsplitsing van gegevens, dat elke beoordelaar één item corrigeert voor alle personen (ervan uitgaande dat men over 20 beoordelaars zou kunnen beschikken).

Het bovenstaande maakt duidelijk hoe belangrijk het is dat meerdere variantiebronnen in één betrouwbaarheidsberekening worden betrokken. De berekening van een klassieke betrouwbaarheid en daarnaast een berekening van de interbeoordelaarsbetrouwbaarheid, zoals in het voorgaande voorbeeld mogelijk zou zijn geweest, is op zich onvoldoende informatief. Het gaat vooral om de relatieve verhouding tussen de variantiebronnen onderling en de consequenties daarvan op de totale betrouwbaarheid. Het is best mogelijk dat een hoge beoordelaarsvariantie, resulterend in een lage interbeoordelaarsbetrouwbaarheid, toch een betrouwbare totale toets kan opleveren. Een en ander is afhankelijk van de overige bronnen, de wijze waarop de meting procedureel wordt ingericht, en de uitspraken die de onderzoeker wenst te doen.

ENKELE PRAKTISCHE TOEPASSINGEN

Om een en ander te concretiseren volgt hier een drietal voorbeelden van studies uit de Maastrichtse keuken, waarin gebruik is gemaakt van de generaliseerbaarheidstheorie. De eerste twee studies hebben betrekking op vergelijkingen van de betrouwbaarheid van verschillende toetsinstrumenten, en de derde op een specifiek betrouwbaarheidsprobleem gericht op het nemen van zak/slaag beslissingen.

Gesloten versus open vragen.

Gesloten vragen, zoals multiple choice en juist/onjuist vragen, worden vaak 'objectieve' vragen genoemd. Open vragen zijn subjectiever, omdat correctoren in het geding zijn, die nogal eens van mening blijken te verschillen. Vanwege dit verschil in objectiviteit wordt aan gesloten vragen vaak de voorkeur gegeven met een verwijzing naar de vaak lage interbeoordelaarsovereenstemming bij open vragen. Kijkend met het oog van de generaliseerbaarheidstheorie zal het duidelijk zijn dat beoordelaarsovereenstemmingslechts een deel van het verhaal vormt. Een faire vergelijking tussen de betrouwbaarheid van open en gesloten vragen is er een waarbij gekeken wordt naar de totale betrouwbaarheid. De beoordelaarsovereenstemming vormt daarin slechts één aspect en het is volstrekt afhankelijk van de overige facetten hoe deze totale betrouwbaarheid er uit zal zien.

Stalenhoef-Halling et al. vergeleken een toets met 7 open vragen met een toets met 114 juist/onjuist vragen - zoveel mogelijk gematched op inhoud - wat betreft betrouwbaarheid.¹ Omdat bleek dat de betrouwbaarheid van de juist/onjuist vragen uit deze toets wat lager uitviel dan gewoonlijk, werden alle toetsen die afgenomen waren in het studiejaar waarin het onderzoek plaatsvond, gecombineerd tot één betrouwbaarheidsschatting (door middelen van variantiecomponenten over toetsen heen). Voor een betere vergelijkbaarheid tussen vraagvormen werd de betrouwbaarheid niet uitgedrukt in het aantal items van de toets, maar in de benodigde toetstijd. In eenzelfde tijd kan men immers meer gesloten vragen stellen dan open vragen. Tabel 4 geeft de resultaten weer.

Tabel 4.
Generaliseerbaarheidscoëfficiënten van juist/onjuist vragen en open vragen^a

Toetstijd in uren	Juist/onjuist vragen ^b		Open vragen ^c	
	Alle toetsen van zelfde jaar	Experimentele toets	Eén beoordelaar	Twee beoordelaars
1	0.64	0.43	0.58	0.61
2	0.78	0.60	0.74	0.76
3	0.84	0.69	0.81	0.82
4	0.88	0.75	0.85	0.86
5	0.90	0.79	0.88	0.89
6	0.92	0.82	0.89	0.90

^aUit Stalenhoef-Halling et al.¹

^b90 vragen per uur

^c7 vragen per uur

De generaliseerbaarheidscoëfficiënten laten duidelijk zien dat de open vragen even betrouwbaar kunnen zijn als de juist/onjuist vragen. In de experimentele toets zijn de gesloten vragen zelfs minder betrouwbaar. Het toevoegen bij open vragen van een beoordelaar - de beoordelaars beschikten in dit geval over een correctievoorschrift - heeft weinig nut (de beoordelaars waren genest binnen items; er is hier dus steeds sprake van een veelvoud van beoordelaars).

Checklists versus globale beoordelingsschalen.

Eenzelfde verschil in betrouwbaarheid door objectiviteit/subjectiviteit als bij open versus gesloten vragen wordt verondersteld bij checklists en globale beoordelingsschalen bij metingen gebaseerd op directe observatie. Globale beoordelingen - zo meent men - zijn subjectiever en dus minder betrouwbaar dan de 'objectievere' checklists. Van Luijk en Van der Vleuten gebruikten in een observatietoets zowel globale beoordelingsschalen als checklists.² De globale beoordelingen bestonden uit twee delen: 1) de algemene indruk over de prestatie van de student en 2) specifieke globale oordelen over verschillende aspecten van de prestatie van de student. De observatietoets bestond uit een serie 'stations', waarin bij elk station een verschillende medische vaardigheid door de student werd gedemonstreerd en door een beoordelaar werd beoordeeld. De interbeoordelaarsbetrouwbaarheden van checklists en globale oordelen verschilden nogal: 0.81 voor checklists, 0.59 voor algemene indruk en 0.58 voor de specifieke globale oordelen. De totale betrouwbaarheid, weergegeven in tabel 5, liet echter een geheel ander beeld zien.

Opgemerkt moet worden dat ook hier elk station een andere beoordelaar kende - de beoordelaars zijn genest in stations - en dat het design van de studie het niet toeliet om het effect van meerdere beoordelaars op de totale betrouwbaarheid te schatten (niet in alle stations waren dubbele beoordelingen aanwezig). De totale betrouwbaarheid laat zien dat er nauwelijks verschillen zijn tussen beide methoden. Hoewel de subjectiviteit van globale beoordelingen tot uitdrukking komt in de interbeoordelaarsbetrouwbaarheid, is de invloed ervan op de totale score van een student over de gehele toets gering, of ten

minste vergelijkbaar met die van checklists. Omdat elk station door een andere beoordelaar wordt beoordeeld, en de totale score gebaseerd is op beoordelingen van meerdere beoordelaars (evenveel als er stations zijn), worden beoordelaarsfouten in individuele stations blijkbaar 'uitgemiddeld' over stations heen.

Beide studies laten zien hoe de toepassing van de generaliseerbaarheidstheorie de mogelijkheid biedt vergelijkingen te maken tussen betrouwbaarheden van verschillende instrumenten met soms verrassende uitkomsten. De studies dienen als illustraties voor de toepassing van de generaliseerbaarheidstheorie; er moet niet uit worden geconcludeerd dat het zinvol is massaal over te gaan tot subjectieve metingen. Daarvoor ligt een en ander blijkens de oorspronkelijke publikaties toch iets genuanceerder.^{1 2} Wel laten beide studies zien, dat de betrouwbaarheid van korte toetsen laag is. Ook in andere studies blijkt dat enkele uren toetstijd, met uitzondering van heel efficiënte of heel specifieke toetsen, over het algemeen noodzakelijk is.

Betrouwbaarheid van beslissingen.

De bespreking van de generaliseerbaarheidstheorie en de berekening van de betrouwbaarheid is tot nu toe beperkt gebleven tot de betrouwbaarheid van scores verkregen met een meetinstrument. In veel toetssituaties (en daarbuiten) is men vaak niet zozeer geïnteresseerd in de betrouwbaarheid van de scores, maar in de betrouwbaarheid van de beslissingen die men op basis van de toets heeft genomen. In dit beheersingsgerichte perspectief (mastery-oriented perspective) staat niet centraal 'hoeveel' iemand gescoord

Tabel 5.
Generaliseerbaarheidscoëfficiënten gebaseerd op checklists en globale oordelen in een observatietoets^a

Toetstijd in uren ^b	Checklists	Globale oordelen	
		Oordeel algemene indruk	Specifieke beoordelingen
1	0.44	0.45	0.47
2	0.61	0.62	0.64
3	0.71	0.71	0.73
4	0.76	0.76	0.78
5	0.80	0.80	0.82
6	0.83	0.83	0.84

^aUit Van Luijk en Van der Vleuten²

^bGemiddeld 4 stations per uur

heeft, maar of de kandidaat al of niet een bepaalde grenswaarde heeft gepasseerd, de cesuur. Hoeveel erboven of eronder is dan niet van belang. Hierin zit ook het verschil tussen het domein-georiënteerde perspectief en het beheersingsgerichte perspectief, die nogal eens verward worden. Vanuit het beheersingsgerichte perspectief wordt, evenals in het domeingerichte perspectief, een 'absoluut' standpunt ingenomen, maar de absolute betekenis geldt voor de score, niet voor de zak/slaag beslissing. In het domein-georiënteerde perspectief gaat het om 'hoeveel', in het beheersingsgerichte perspectief om 'of'. Het kiezen voor een beheersingsgericht perspectief kan veel betekenis hebben voor de eisen die gesteld moeten worden aan de betrouwbaarheid van een meting. Om dat duidelijk te maken hebben we het begrip van de standaardmeetfout nodig.

De betrouwbaarheid kan ook worden uitgedrukt door aan te geven hoe groot de fout is die gemaakt wordt als we de score van een persoon bepalen op een test. Dit wordt de standaardmeetfout genoemd. Wanneer een meting erg betrouwbaar is, is de standaardmeetfout klein en bij een lage betrouwbaarheid is deze groot. Met de standaardmeetfout kunnen we een betrouwbaarheidsinterval berekenen, waarmee we kunnen aangeven met welke zekerheid een score van een persoon binnen het berekende interval zal liggen. Bijvoorbeeld, een student scoort 70% van de items van een toets goed. We zullen de vereiste berekeningen achterwege laten, maar stel dat het betrouwbaarheidsinterval bij een matig betrouwbare toets 10% is. De score van de student (de ware score) ligt dan ergens tussen

de 60% en 80%. Wanneer men de betrouwbaarheid van de scores beschouwt (norm- en domeingeoriënteerd) is deze meetfout voor alle personen even 'erg'. Voor de betrouwbaarheid van de besluitvorming of de zak/slaag beslissing is dat niet het geval. Gesteld de zak/slaag grens ligt bij 65%. Voor de persoon die 70% heeft behaald weten we niet erg zeker of deze terecht geslaagd is, omdat de ware score wel eens onder de cesuur zou kunnen liggen. Maar voor iemand die 40% gescoord heeft, of 85%, is de beslissing zekerder. De betrouwbaarheid van de beslissing is dus in sterke mate afhankelijk van de grens die bij de beslissing wordt aangelegd. De generaliseerbaarheidstheorie biedt de mogelijkheid om de betrouwbaarheid van de beslissing te schatten. Ook hier weer zullen we de precieze techniek achterwege laten en het effect illustreren aan de hand van concrete gegevens.

In een betrouwbaarheidsonderzoek, opnieuw op het gebied van observatietoetsen, werden de gegevens gecumuleerd van een groot aantal toetsen.³ Betrouwbaarheidscoefficienten werden berekend voor verschillende toetslengtes en bij verschillende waarden voor de zak/slaaggrens. Tabel 6 vat deze gegevens samen.

De standaardmeetfout bij korte toetstijden is tamelijk groot, maar afhankelijk van de gekozen cesuur is toch een hoge betrouwbaarheid van de te nemen beslissing mogelijk. Naarmate de cesuur dichterbij het gemiddelde van de scoreverdeling komt, neemt de betrouwbaarheid af. Bij lange toetsen is de standaardmeetfout veel kleiner en zijn de verschillen in betrouwbaarheid tussen verschillende zak/slaaggrenzen kleiner. Het beheersingsgerichte perspectief en de berekening van de betrouwbaarheid van beslissingen, kan grote praktische consequenties hebben. In feite laat het zien dat bij een beslissing de betrouwbaarheid van een meting niet voor iedereen gelijk is, omdat die afhankelijk is van het vaardigheidsniveau van de personen. Men kan hiervan gebruik maken door verschillende personen kortere (minder betrouwbare) en langere (meer betrouwbare) toetsen voor te leggen, afhankelijk van hun prestatie. Dit wordt sequentiële toetsing genoemd: een korte (screenings-)toets wordt voorgelegd aan de gehele groep kandidaten,

Tabel 6.
Betrouwbaarheid van
beslissingen als functie
van het aantal stations^a

Toetstijd in uren ^c	Standaard meetfout	Cesuur ^b (Percentage goed)			
		60%	70%	80%	90%
1	8.90	0.80	0.48	0.29	0.71
2	6.30	0.89	0.69	0.54	0.84
3	5.14	0.93	0.78	0.66	0.89
4	4.45	0.94	0.83	0.74	0.91
5	3.98	0.96	0.86	0.78	0.93
10	2.82	0.98	0.93	0.88	0.96

^aUit Van der Vleuten & Van Luijk³

^bBehaalde gemiddelde score was 77%

^cGemiddeld 4 stations per uur

waarna de hele goede en slechte kandidaten worden uitgeselecteerd en de toetsing alleen wordt voortgezet voor de overblijvende groep (de groep die rond de cesuur scoort). Sequentiële toetsing kan een aanmerkelijke logistieke besparing opleveren.

Alle voorbeelden hebben studieprestaties als object van meting gehad. We zijn daarbij vooral geïnteresseerd in het onderscheid dat we kunnen maken tussen goed en slecht scorende personen, in de variantie tussen studenten ofwel de persoonscomponent. Het is ook mogelijk om de generaliseerbaarheidstheorie toe te passen op andere terreinen, bijvoorbeeld bij curriculum-evaluatie, zoals Wolfhagen et al. in hun studie hebben gedaan.⁴

GENOVA

Het berekenen van de variantiecomponenten is een ingewikkelde zaak. D-studies inrichten vereist het nodige denkwerk en het berekenen van betrouwbaarheidsindices is een tijdrovende aangelegenheid. Voor het berekenen van de variantiecomponenten kan gebruik gemaakt worden van het 8V-programma (General mixed model analysis of variance) uit het algemeen statistische pakket BMDP.⁵ Crick & Brennan hebben het programma GENOVA ontwikkeld, dat speciaal bedoeld is voor de generaliseerbaarheidsanalyse.⁶ Het is een zeer flexibel programma waarin zowel G-studies als D-studies kunnen worden uitgevoerd. Het programma neemt veel denken en rekenwerk van de onderzoeker over. GENOVA is ontwikkeld voor main frame computers en voor een gering bedrag verkrijgbaar. Voor de PC is een (minimaal) aangepaste versie beschikbaar (werkt niet op XT machines). Het main frame programma kan worden verkregen bij Dr. R.L. Brennan, The American College Testing Program, P.O. Box 168, Iowa City, Iowa 52243, USA. Voor de PC-versie kan men inlichtingen verkrijgen bij Dr. J.E. Crick, National Board of Medical Examiners, 3930 Chestnut Street, Philadelphia, PA 19104, USA.

GENERALISEERBAARHEIDSTHEORIE ALS PRAKTISCH INSTRUMENT

De generaliseerbaarheidstheorie is een praktische theorie doordat het een goed instrumentarium vormt voor de onderzoekspraktijk waarin de betrouwbaarheden van metingen moeten worden geschat. In de praktijk worden metingen veelal door meerdere factoren beïnvloed. De generaliseerbaarheidstheorie weet hiermee adequaat om te gaan. Vergelijkingen tussen metingen en instrumenten krijgen daardoor een grotere betekenis. Resultaten van G-studies kunnen worden gecumuleerd over een langere periode en uitgewisseld tussen onderzoekers. Met de generaliseerbaarheidstheorie stijgt het betrouwbaarheidsbegrip uit boven de statische interpretatie van betrouwbaarheid die men in de literatuur nogal eens tegenkomt: de betrouwbaarheid van een instrument bestaat niet, maar is afhankelijk van de gewenste reikwijdte en het doel van de meting. De onderzoeker dient hierin keuzes te maken en deze te expliciteren.

Met een goed begrip van de generaliseerbaarheidstheorie (en dat hoeft zeker niet tot in elk detail), zal men gaan ervaren dat de gedachten van de onderzoeker door de theorie worden gestructureerd: men gaat denken in termen van de theorie en dat kan in praktische zin erg vruchtbaar zijn. Kortom, er zijn weinig theorieën die een zo praktische betekenis hebben als de generaliseerbaarheidstheorie.

GEANNOTEEERDE LITERATUUR

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The Dependability of behavioral measurements: generalizability for scores and profiles. New York: John Wiley and Sons, 1972.

Dit boek is de bijbel van de generaliseerbaarheidstheorie. De theorie wordt hierin geponeerd en met alle mogelijke uitwerkingen uit de doeken gedaan. Helaas is het boek onleesbaar door de taai statistische behandelingswijze van een en ander. Wel is de inleiding tot de theorie de moeite waard.

Brennan RL. Elements of generalizability theory. Iowa: American College Testing Program, 1983.

Juist met het oog op de onleesbaarheid van het oorspronkelijke boek, heeft Brennan dit boek geschreven. Het is dan ook wat beter toegankelijk, maar vereist toch nogal wat kennis van statistiek en psychometrie. De uitleg van variantiebronnen in termen van Venn-diagrammen is erg inzichtelijk. Het bij het boek ontwikkelde computerprogramma GENOVA is uitermate praktisch (zie boven in tekst).

Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. American Psychologist 1989; 44: 922-32. Redelijk toegankelijke uitleg over de theorie, gebruikmakend van enkele voorbeelden. De verschillen tussen de klassieke testtheorie en de generaliseerbaarheidstheorie worden goed uiteengezet.

Thorndike RL. Applied psychometrics. Boston: Houghton Mifflin, 1982.

Het hoofdstuk over betrouwbaarheid bevat een korte, voor de leek zeer leesbare uitleg over generaliseerbaarheidstheorie. Aan de hand van een concreet voorbeeld wordt het rekenwerk toegelicht. Door de bondigheid is het verhaal natuurlijk ook wat onvolledig.

LITERATUUR

1. Stalenhoef-Halling BF, Van der Vleuten CPM, Jaspers TAM, Fiolet JFBM. The feasibility, acceptability and reliability of open-ended questions in a problem-based learning curriculum. In: Bender W, Hiemstra RJ, Scherpier AJJA, Zwierstra RP, eds. Teaching and assessing clinical competence. Groningen: BoekWerk, 1990; 552-7.

2. Van Luijk SJ, Van der Vleuten CPM (in press). A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, eds. More developments in assessing clinical competence. Montreal: Can-Heal.

3. Van der Vleuten CPM, Van Luijk SJ.

Betrouwbaarheid van observatietoetsen voor praktische vaardigheden in het medisch onderwijs. Tijdschrift voor Onderwijsresearch 1988; 13: 213-26.

4. Wolfhagen HAP, Van der Vleuten CPM, Gijssels WH, Essed CGM (in press). Reproducibility of scores for clerkship program evaluation ratings. In: Hart IR, Harden RM, eds. More developments in assessing clinical competence. Montreal: Can-Heal.

5. Dixon WJ, Engelman L, Hill MA, Jenrich RL. BMDP statistical software manual. Volume 2. Berkeley: University of California Press, 1988.

6. Crick JE, Brennan RL. Manual for GENOVA: A GENERALized analysis OF VARIance system. ACT Technical Bulletin 43, The American College Testing Program, Iowa City, 1983.

Na invoering van de twee-fasen structuur moest bij vele universitaire studies de opleidingsduur teruggebracht worden tot vier jaar. Slechts een klein gedeelte van de studenten van deze opleidingen mag een tweede fase opleiding volgen. Onder het motto "na vier jaar geneeskunde studeren ben je niets" hebben de faculteiten geneeskunde de overheid kunnen overtuigen dat de studie geneeskunde voor alle studenten zes jaar moet duren. Het komt niet erg consequent over, dat in gesprekken over de inrichting van de zesjarige opleiding vaak een nadrukkelijke scheiding tussen de eerste vier jaar en de laatste twee jaar wordt aangebracht. Ik bedoel met deze scheiding niet het doctoraal examen want dat blijft natuurlijk bestaan maar bijvoorbeeld uitspraken als "de laatste twee jaar van de opleiding zijn beroepsvoorbereidend" of "de eerste vier jaar zijn wetenschappelijk". Het lijkt mij consequent dat de gehele zesjarige opleiding beroepsvoorbereidend is. Het zelfde geldt voor de wetenschappelijkheid; de gehele zesjarige opleiding heeft een wetenschappelijk karakter. Bij de inrichting van het curriculum praten we dan als we consequent zijn, over de inrichting van zes jaar en dus niet over de inrichting van vier jaar voor het doctoraal en van twee jaar na het doctoraal. Er zijn toch voor de studie geneeskunde geen twee fasen; alle studenten volgen het zesjarig curriculum.

A.J.J.A. Scherpier

$$4 + 2 = 6$$