Zwarte Dozen

Faculteit Toegepaste Onderwijskunde, Universiteit Twente

Rede uitgesproken bij de aanvaarding van het ambt
van Bijzonder Hoogleraar Psychometrie
op donderdag 9 juni 1994

N.D. Verhelst

Zwarte dozen

De titel van mijn rede, *Zwarte Dozen*, heeft niets te maken met de luchtvaart en met de rampspoed die deze af en toe treft, en waarbij de zwarte doos, na de feiten, een cruciale rol speelt. De zwarte doos in een vliegtuig is een registratie-apparaat, en bevat voor de ingenieurs van de luchtvaartmaatschappij geenszins raadsels of mysteries. Het grootste probleem is vaak het terugvinden van de zwarte doos. De zwarte doos waar men in de psychologie mee wordt geconfronteerd, heeft meestal net de tegengestelde eigenschappen. Het vinden is gemakkelijk genoeg, men zou losweg kunnen zeggen dat ze zich bij ieder mens tussen zijn twee oren bevindt. Het openmaken echter levert ons in vele gevallen weinig soelaas, zeker als we geïnteresseerd zijn in het verloop van cognitieve processen en emoties. Niemand, denk ik, hangt nog het extreme reductionistische standpunt aan dat dergelijke processen alleen hun ultieme verklaring vinden wanneer ze kunnen worden teruggevoerd op de anatomische structuren van onze hersenen en de fysiologische en biochemische processen die zich daar afspelen. Dit betekent niet dat de psychologie niet geïnteresseerd is in deze processen. Ons geheugen bijvoorbeeld heeft een materieel substraat. Wij geloven niet dat herinneringen puur immaterieel of geestelijk zijn, maar dat ze in een of andere materiële vorm in onze hersenen liggen opgeslagen. Maar het betekent evenmin, dat de enige manier om een wetenschappelijke studie te maken van het geheugen het biochemisch onderzoek zou zijn. Het geheugen kent zijn eigen wetmatigheden, die bestudeerd kunnen worden zonder ook maar iemand zijn schedeldak te lichten, zoals Ebbinghaus ons reeds meer dan honderd jaar geleden leerde (Ebbinghaus, 1885).

Een ander voorbeeld dat geschiedenis heeft gemaakt, is de ontdekking en de studie van de geconditioneerde reflex, door Pavlov in het begin van deze eeuw (Pavlov, 1904). Als een hond een stuk brood in zijn muil heeft, gaan de speekselklieren reflexmatig en onvoorwaardelijk speeksel afscheiden. Pavlov had echter opgemerkt dat zijn proefdieren ook speeksel gingen afscheiden bij het zien van brood, of bij het horen van de voetstappen van de naderende laboratorium-assistent, hoewel het geluid van voetstappen niet in alle omstandigheden leidt tot speekselafscheiding. Die reactie wordt alleen ontlokt onder bepaalde voorwaarden: de voetstappen krijgen de functie van signaal voor voedsel. Ontneemt men ze die functie, door herhaaldelijk na binnenkomst de hond geen voedsel te geven, dan dooft de reflex uit.

Dit verhaal is genoegzaam bekend en het is niet nodig er verder over uit te weiden. Wat voor ons interessant is, is te zien hoe Pavlov zelf op zijn ontdekking reageerde. Ik citeer (in tweedehandse vertaling) uit zijn rede bij het in ontvangst nemen van de Nobelprijs in 1904: "*Wij waren geenszins van plan de fysiologie op te geven*

ten faveure van de psychologie (die in die tijd de bewustzijnsinhouden tot voorwerp van haar studie had); *we kozen in onze experimenten met de zogenaamde psychische fenomenen voor een zuiver objectief standpunt. We spanden ons vooral in om de nodige discipline aan onze gedachten en uitspraken op te leggen om ons niet bezig te houden met de zogenaamde mentale toestand van onze proefdieren."* Let wel, Pavlov ontkende niet het bestaan van psychische of mentale toestanden, maar hij weigerde deze als voorwerp van wetenschappelijke studie te zien: "*We beperkten onze taak tot de exacte observatie en beschrijving van speekselafscheiding als antwoord op een prikkel die deze reactie teweegbrengt van op afstand en niet door direct contact met de mondslijmvliezen.*"

De historische betekenis van Pavlov ligt niet in zijn inzichten of zijn werk op het gebied van de fysiologie van de spijsvertering - hoewel dat de reden was om hem de Nobelprijs toe te kennen - maar in de ontdekking van de conditionele reflex en zijn inzicht in het belang van deze ontdekking. In plaats van te zoeken naar een verklaring voor dit verschijnsel, is het verschijnsel, omdat het zo vaak in zoveel verschillende omstandigheden volgens bepaalde wetmatigheden blijkt op te treden, zelf tot een belangrijk verklaringsprincipe geworden van menselijk en dierlijk gedrag. De verklaring van het verschijnsel in termen van bewustzijnsinhouden of andere 'psychische' categorieën werd overbodig geacht. Het bewustzijn werd niet ontkend maar kreeg de status van een 'black box', een zwarte doos die we niet per se open willen maken. Een 'black box' is dus geen mysterie waar we diepzinnige beschouwingen aan wijden; het is een methodisch principe waarbij we de inhoud, de structuur van de zwarte doos laten voor wat ze is, maar ons in de eerste plaats richten op haar functioneringsprincipes. Kunnen we wetmatige verbanden vinden tussen input en output, die ons iets kunnen leren over de werking van de 'black box'?

Hoewel we de zwarte doos niet openmaken, handelen we niet alsof ze leeg is, en dus evenzogoed zou kunnen worden verwijderd. De zwarte doos is een formeel model, die gevuld is met de gedachtenconstructies van de wetenschappelijke onderzoeker(s): wetenschappelijke concepten of hypothetische constructen zoals ze soms ook wel worden genoemd (MacCorquodale & Meehl, 1948). Deze concepten zijn geen losliggende elementen maar onderhouden verbanden met elkaar in een min of meer samenhangend bouwwerk dat net zo functioneert als het materiële voorwerp van onze studie: de mens, zoals hij zich gedraagt in zijn omgeving. Althans, dat is wat we wel zouden willen, maar zo ver zijn we nog lang niet. De psychologie is niet een schitterend gebouw dat alleen nog om een beetje verfraaiing vraagt, dat nog niet helemaal af is; integendeel, ze lijkt soms meer op een verzameling schuurtjes - lees theorietjes - die kriskras door elkaar worden gebouwd, met weinig of geen onderlinge

verbindingen en waarvan sommige buitengewoon bouwvallig zijn. Voor deze troosteloze aanblik zijn vele redenen en oorzaken op te noemen. Ik wil mij echter beperken tot één aspect dat mijns inziens van groot belang is, omdat het direct te maken heeft met de fundering van de theoriehuisjes. Het gaat om het probleem van de status van de wetenschappelijke concepten, in concreto om hun meetbaarheid.

Psychometrie

Psychometrie is, breed gedefinieerd, de leer van het meten van psychologische variabelen. Deze definitie is niet zeer nauwkeurig, maar nauwkeurig genoeg als algemeen kader voor de problemen waarover ik het wil hebben. Om de gedachten te bepalen beschouwen we een voorbeeld waar we allen goed mee vertrouwd zijn, en dat een grote mate van vanzelfsprekendheid lijkt te bevatten, namelijk de meting van de intelligentie. Een maat voor onze intelligentie is de uitslag die we behalen op een intelligentietest, waarbij aangenomen wordt dat een hogere uitslag een weerspiegeling is van een grotere intelligentie. Maar hoe komt men erbij deze aanname te maken? Laat ik U twee problemen voorleggen die illustreren dat deze aanname misschien toch wel minder vanzelfsprekend is dan op het eerste gezicht lijkt.

(1) Een intelligentietest bestaat uit een redelijk groot aantal vragen waarop de onderzochte persoon naar beste vermogen antwoord dient te geven. De uitslag op de test is het aantal vragen dat juist beantwoord wordt. In de meeste intelligentietests echter zijn de vragen die gesteld worden nogal divers. Er moeten bijvoorbeeld rijtjes getallen worden aangevuld, en geometrische figuren moeten met blokjes worden nagelegd. Bovendien zijn niet alle vragen even moeilijk. Als we alleen het aantal juiste antwoorden tellen verliezen we het onderscheid tussen de verschillende soorten vragen en het onderscheid tussen gemakkelijke en moeilijke vragen. We nemen dus bij toepassing van de intelligentietests het (impliciete) standpunt in dat dit soort onderscheid er eigenlijk niet toe doet. Een belangrijke vraag voor de psychometrie is hoe we zo een standpunt kunnen rechtvaardigen.

(2) Er bestaan vele intelligentietests waarvan de auteurs allemaal pretenderen dat ze intelligentie meten. Als de intelligentie van Jan en Piet gemeten wordt met intelligentietest A, en Jan blijkt een hogere uitslag te hebben dan Piet, dan zijn we geneigd te zeggen dat Jan intelligenter is dan Piet. Dan zouden we kunnen besluiten dat Jan ook een hogere uitslag moet behalen dan Piet op

> intelligentietest B, en dat het dus overbodig is de intelligentietest B ook nog eens aan Jan en Piet op te dringen. Alleen, er zijn psychologen die dat doen, en die vinden dat de twee tests Jan en Piet verschillend rangordenen? Hoe verlegen zijn wij, of horen we te zijn met zo'n uitkomst?

Om duidelijk te illustreren hoe moeilijk deze problemen zijn en hoe de psychometrie ermee omgaat, zal ik in plaats van intelligentie een eigenschap gebruiken waarmee we allen zeer bekend zijn en waarvan de meting bijzonder eenvoudig is, namelijk de lichaamslengte. Als we de lichaamslengte van Jan en Piet bepalen met behulp van een duimstok, en Jan blijkt langer te zijn dan Piet, dan zal het bij niemand opkomen om dit resultaat bevestigd te willen zien door de lichaamslengte van beide personen nog een tweede keer te bepalen met behulp van een meetlint. Een belangrijk verschil tussen deze procedure en de procedure die we volgen bij het meten van intelligentie is hierin gelegen dat de bepaling van de lichaamslengte een enkelvoudige operatie is - het aflezen van een getal op de duimstok. Bovendien is een begrip als lichaamslengte zo direct verbonden met onze meest basale kennis van de wereld, namelijk uitgestrektheid in een bepaalde richting, dat we lichaamslengte niet zo gauw een hypothetisch construct zullen noemen. Laten we echter eens kijken welk soort problemen we krijgen als we de lichaamslengte zouden willen bepalen op een wijze die lijkt op het bepalen van de intelligentie met een test.

Lumsden (1976) heeft bij wijze van gedachtenexperiment een test bedacht, de zogenaamde flogging wall test, de 'ranselmuurtest', waarvan het eerste ontwerp weergegeven is in figuur 1. In een lange muur zijn op verschillende hoogten gaten aangebracht waaruit stokken steken, die een langzame doch krachtige op en neergaande beweging maken; de amplitude van de beweging hoeft niet voor alle stokken dezelfde te zijn. De personen waarvan we de lichaamslengte willen bepalen staan rechtop op een wagentje dat met behoorlijke snelheid langs de muur rijdt. Het resultaat van deze gecompliceerde meetprocedure wordt samengevat in een enkele getal, het aantal stokslagen dat de persoon heeft gekregen. De vraag die voor ons belangrijk is, is of dit aantal een goede indicatie is van de lichaamslengte, of nog eenvoudiger uitgedrukt: kunnen we in vertrouwen zeggen: "hoe meer stokslagen, hoe langer de persoon"?

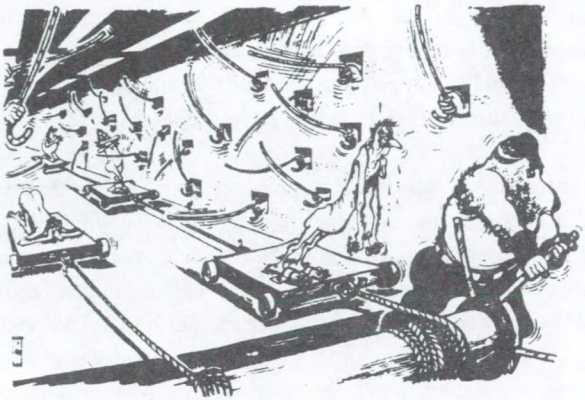


Fig.1. De ranselmuurtent (Lumsden, 1976)

De analogie met een intelligentietest of in het algemeen met een test of toets die in de psychologie of de onderwijskunde wordt gebruikt is de volgende. De 'ruwe' meetuitslag is niet het resultaat van een unitaire operatie, zoals het aflezen van de uitslag van een wijzer op een wijzerplaat, doch wordt samengesteld uit een aantal elementaire scores, die behaald worden op onderdelen van de test. Deze onderdelen worden items genoemd, en de meest elementaire vorm van een item is het zogenaamde binaire of dichotome item: wordt het item correct beantwoord (in de analogie: wordt men geraakt door de stok) dan krijgt men één punt, in het andere geval nul punten. De eenvoudigste samenstellingsregel zegt dat de testuitslag gewoon de som is van de itemscores, in dit geval dus het totaal aantal correct beantwoorde items.

Bij het gebruik van de ranselmuurtest zijn een aantal belangrijke en intrigerende vragen te stellen. Bijvoorbeeld, maakt het uit op welke hoogte de gaten zijn aangebracht waaruit de stokken steken, of in een terminologie die meer gangbaar is in de psychometrie, maakt het uit hoe moeilijk de items zijn? De vraag die mij hier echter het meest interesseert luidt: waarom laat Lumsden in zijn originele ontwerp de stokken zwiepen, in plaats van ze op een vaste hoogte te fixeren?

Stel dat we een verbeterde versie uitbrengen van Lumsdens test, door inderdaad de stokken op een vaste hoogte te fixeren. In dit geval zal een persoon die twee maal de test ondergaat, en in de veronderstelling dat hij in de tussentijd niet gegroeid is, twee maal dezelfde score halen. Bij Lumsdens test is dit niet het geval, omdat er, door de op en neergaande beweging van de stokken, een toevalselement wordt ingebouwd, dat er voor zorgt dat de testscore niet volkomen stabiel is. Men zegt in de psychometrie dat het meetinstrument niet volkomen betrouwbaar is, en in de analogie

van Lumsdens test zien we dat de testscore niet perfect betrouwbaar is omdat de items elk afzonderlijk niet volkomen betrouwbaar zijn.

In de analogie van de Lumsdenmachine zijn twee aspecten heel belangrijk. (i) Het attribuut of de eigenschap die we willen meten -de lichaamslengte- en de belangrijkste eigenschap van het item dat we voor dit doel gebruiken - de hoogte van het fixatiepunt van de stok in het voorbeeld - zijn een en dezelfde eigenschap, hoewel ze meestal complementair worden geformuleerd. In de psychometrie gebruiken we vaak het algemene begrip vaardigheid om de eigenschap van de persoon aan te duiden die we willen meten, en voor het item spreken we van moeilijkheid. Maar moeilijkheid is een hoeveelheid vaardigheid, namelijk de hoeveelheid vaardigheid die nodig is om het item correct op te beantwoorden. (ii) De meetinstrumenten die we in de psychologie construeren hebben in het algemeen niet de eigenschap van de verbeterde Lumsdenmachine, maar van de oorspronkelijke. Er is een zekere -en meestal grote- mate van instabiliteit en voor deze instabiliteit moeten we een verklaring zien te vinden. Die verklaring vinden we meestal door zelf - in min of meer abstracte termen - een mechanisme te verzinnen waarvan de effecten goed overeenkomen met wat we feitelijk observeren. En dat is de essentie van de black-box of de zwarte doos. In de Lumsdenmachine wordt dit voorgesteld door de handjes die de stokken vasthouden en op en neer bewegen. Maar dat is niet voldoende: hoewel we niets hoeven te veronderstellen over de eigenschappen van de personen aan wie die handjes toebehoren, moeten we wel veronderstellingen maken over de algemene kenmerken van de bewegingen van de stok. Omdat we ons op onzeker terrein bevinden proberen we zuinig te zijn met die veronderstellingen om ons op voorhand niet al te zeer vast te leggen. Twee belangrijke principes, die we in vrijwel elk model of theorie over het meten tegenkomen, zijn, vertaald in de analogie van de Lumsdenmachine, (a) dat de stokken onafhankelijk van elkaar bewegen en (b) dat de beweging van de stokken niet beïnvloed wordt door welke eigenschap dan ook van de langsrijdende slachtoffers.

Samenvattend kunnen we zeggen: of een persoon door een stok geraakt wordt is afhankelijk van zijn lengte, van de complementaire eigenschap van de stok, namelijk de hoogte van het fixatiepunt en verder van het toeval.

Maar, zult u misschien opperen, zo kan men wel alles verklaren. Noem gewoon alles wat je niet kunt verklaren toeval, en zeg vervolgens dat het toeval de verklaring is. Zo eenvoudig is het echter niet: hoe paradoxaal het misschien mag klinken, toeval kent ook zijn beperkingen en wetmatigheden. Het zijn echter geen wetmatigheden van alles of niets, waarvan zonder discussie duidelijk is of ze al dan niet geschonden zijn. In die discussie kan men bovendien standpunten innemen en maatregelen nemen die de allesverklarende kracht van het toeval bevoordelen of juist gaan tegenwerken. Met

andere woorden, we hebben een zwarte doos geconstrueerd, door niet te specificeren wat de precieze aard of bedoelingen zijn van de mannetjes achter de ranselmuur, maar bovendien hebben we de functionaliteit van de zwarte doos, de regels volgens welke de stokken worden bewogen, zeer onvolledig beschreven. Kunnen we hier wel iets mee, of we hebben we de flexibiliteit van het concept 'zwarte doos' dermate misbruikt dat we alles en dus niets kunnen verklaren?

De zwarte doos van het toeval

Om het voorafgaande algemeen geformuleerd probleem te illustreren, grijp ik even terug naar het eerste probleem dat ik signaleerde bij de intelligentietests: hoe kunnen we weten of de diversiteit aan vragen die in een intelligentietest worden opgenomen allemaal iets te maken hebben met hetzelfde concept. Misschien zitten er wel vragen tussen die een heel andere eigenschap aanspreken, die niets met intelligentie te maken heeft. Iets algemener geformuleerd luidt de vraag dan: hoe maakt men in de psychometrie aannemelijk dat de items van een toets dezelfde latente vaardigheid aanspreken? Als we over perfect betrouwbare meetinstrumenten zouden beschikken, zou het probleem makkelijk zijn op te lossen: wie in de Lumsdenmachine door een hoog bevestigde stok wordt geraakt, dient ook door alle lager bevestigde stokken te worden geraakt, en bovendien kunnen we, zonder dat we over een duimstok beschikken, gemakkelijk bepalen welke stok het hoogst is bevestigd. Door een redelijk grote groep personen de test te laten ondergaan, weten we dat de stok die het kleinst aantal personen geraakt heeft het hoogste is bevestigd. Stel nu dat de test een beetje van slag raakt, door dat de machinist af en toe wat zenuwachtig wordt en bij tijd en wijle een mep uitdeelt, zonder aanzien des persoons. Als we een rake mep meetellen in de toetsscore, dan zullen we er met het hierboven geschetst controlemechanisme gauw achterkomen dat het machinist-item niet homogeen is met de andere items. De situatie is heel wat minder comfortabel als we met items werken die niet helemaal betrouwbaar zijn: in theorie is namelijk alles mogelijk, in de Lumsdenmachine kan de langste persoon de hele test doorlopen zonder één keer geraakt te worden en de kortste kan door alle stokken geraakt worden. Om de homogeniteit van de items met betrekking tot de te meten eigenschap te onderzoeken kunnen we niet meer volstaan met de strenge procedure die hierboven werd beschreven; we moeten uitwijken naar statistische procedures.

Een psychometrisch model is een hypothese waarin de kansverdeling van een antwoordvariabele (het antwoord van de persoon op een item) beschreven wordt gegeven de waarde van de latente trek en de moeilijkheid van het item. Daarbij komen

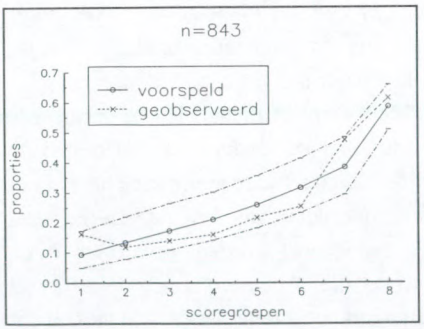
onzekerheden kijken op twee niveaus: we kennen de waarde van de latente trek van de persoon niet en we kennen de moeilijkheid van het item niet. Met bepaalde statistische procedures kunnen we echter een schatting maken van deze waarden, en we kennen ook bij benadering de nauwkeurigheid van deze schattingen. Maar zelfs als we de waarde van de latente trek van elke persoon en de moeilijkheid van elk item zouden kennen, dan blijft er nog een onzekerheid over met betrekking tot het antwoord: het model geeft alleen aan hoe groot de kans is op een juist antwoord, en zegt niet welk antwoord moet optreden. Toch is er een mogelijkheid om de aanvaardbaarheid van het model te beoordelen als we alleen kansen kennen. Stel dat een toets bestaat uit 100 items die allemaal even moeilijk zijn, en we vinden een persoon die een kans van precies 1/2 heeft om elk item correct te beantwoorden, dan verwachten we dat die persoon ongeveer de helft van de items juist zal beantwoorden. Constateren we echter dat die persoon bij het maken van de toets 85 van de 100 items juist beantwoordt, dan zullen we - hoop ik - de wenkbrauwen fronsen en denken: hier klopt iets niet. Maar omdat de psychometrie een empirische wetenschap is, en de observaties dus heilig, kan ons besluit niet anders zijn dan dat het model niet klopt: we verwerpen het model.

De inferentiële statistiek is de wiskundige theorie die ons vertelt hoe we op een rationele manier tot dat 'wenkbrauw fronsen' komen, en ze heeft voor de toepassingen die hier aan de orde zijn, twee boodschappen. Het kan gebeuren dat het model o.k. is, maar dat de statistische procedures ten onrechte het signaal afgeven dat er iets niet in de haak is. De kans op zo'n signaal hebben we zelf in de hand, en gewoonlijk wordt die kans vastgesteld op 5% of 1%. De andere boodschap is als het ware het omgekeerde: het kan gebeuren dat er wel iets aan de hand is, doch de statistische procedure geeft geen signaal af. De kans dat dit gebeurt hebben we maar voor een deel in de hand. De belangrijkste manier om deze kans te manipuleren is het aantal observaties dat we gebruiken om de statistische procedures toe te passen: hoe groter dat aantal des te groter de kans dat we een defect in het model zullen ontdekken. Maar dit brengt ons in een dilemma dat alles te maken heeft met de zwarte doos.

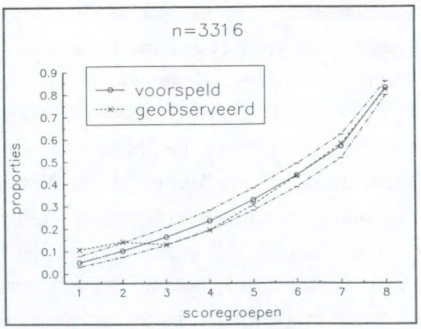
Een psychometrisch model, en eigenlijk elk formeel model, is een gestileerde beschrijving van de werkelijkheid. Men kan dus niet verwachten dat het model de werkelijkheid tot in de kleinste details beschrijft. Men geeft dus in principe toe dat het model fout is. Gaat men nu het model statistisch controleren aan de hand van zeer veel observaties, dan is het welhaast zeker dat het model moet worden verworpen. In de praktijk betekent dit meestal dat modelcontroles die gebaseerd zijn op heel veel observaties eigenlijk niet zo erg serieus moeten worden genomen- men wist toch al dat het verkeerd zou gaan. Neemt men met minder observaties genoegen, dan is de kans

dat de statistische procedure de detaildefecten ontdekt erg klein. Daarbij is dan impliciet de redenering aanwezig, dat er pas iets serieus aan de hand is als het met een klein aantal observaties mis gaat, en als het niet mis gaat, dat er niets aan de hand is. Het is natuurlijk de vraag of dit een wijze en vruchtbare benadering is van het probleem.

We lichten even toe met een voorbeeld. In zijn onderzoek naar leesvaardigheid en leesbaarheid heeft Staphorsius (1994) items gemaakt die naar zijn inzicht en overtuiging de leesvaardigheid van kinderen uit het basisonderwijs meten. Hij heeft voor dat doel meer dan 250 items geconstrueerd en deze items voorgelegd aan enkele duizenden kinderen, met dien verstande dat elk kind tussen de 30 en de 60 items heeft beantwoord, en dat elk item door 3 à 4000 kinderen werd beantwoord. Deze items werden op het Cito geanalyseerd met een model dat in grote lijnen overeenkomt met de principes die ik hierboven heb uiteengezet. In de figuren 2 en 3 is op grafische manier een onderdeel weergegeven van de statistische modelcontrole op item 246, het meest 'weerbarstige' item uit de hele verzameling. De totale groep van leerlingen wordt ingedeeld in 8 zogenaamde scoregroepen, dit zijn subgroepen met een ongeveer gelijk aantal items juist. Op grond van de parameterschattingen kunnen we de proportie juiste antwoorden in elke groep voorspellen (de cirkeltjes in de figuren). De twee buitenste stippellijnen in de figuren stellen de zogenaamde 95%-betrouwbaarheids-enveloppes voor en de x-en geven de geobserveerde proportie juiste antwoorden weer op dit item. De statistische procedure komt grosso modo hierop neer: zolang de geobserveerde proporties binnen de betrouwbaarheidsenveloppe vallen kunnen de afwijkingen tussen geobserveerde en voorspelde proporties verklaard worden als puur toevallig; in het andere geval worden we gewaarschuwd dat de afwijkingen toch wel wat te groot zijn om aan het toeval te kunnen worden toegeschreven.



Figuur 2
toetsing van een psychometrisch model bij een kleine steekproef



Figuur 3
Toetsing van een psychometrisch model bij een grote steekproef

Aan de hand van figuur 2, het resultaat van een analyse op een steekproef van 843 personen, zouden we ons dus op het formeel standpunt kunnen stellen dat 'er niets aan de hand' is: de overeenkomst tussen observaties en modelvoorspellingen is niet schitterend, maar kan aan toeval, aan de 'zwarte doos' worden toegeschreven. En let wel, figuur 2 biedt meer dan de routinematige toepassing van de statistiek doorgaans biedt, waar het plaatje in figuur 2 meestal is samengevat in één enkel getal dat (bij benadering) aangeeft of de observaties al dan niet binnen de 95%-betrouwbaarheidsenveloppe vallen. In figuur 2 worden we in staat gesteld een kleine residu-analyse te doen: we zien dat de afwijkingen een systematisch patroon vertonen: de extreme scoregroepen doen het beter dan voorspeld, de middelste slechter. We komen dus in een min of meer dubbelzinnige situatie waar we niet-passing van het model kunnen toeschrijven aan het toeval, dat we verder - per definitie - onverklaard laten, of we kunnen gealarmeerd zijn door de systematische niet-passing van het model en op zoek gaan naar een verklaring. In figuur 3 krijgen we een gelijkaardige afbeelding van de statistische procedure, maar nu na een analyse die gebaseerd is op meer dan 3000 observaties. Het afwijkingspatroon tussen geobserveerde en voorspelde proporties in beide figuren is gelijklopend, maar de statistische toetsingsprocedure is veel gevoeliger, en de alarmbel slaat aan: we kunnen redelijkerwijs geen beroep meer doen op toevalsprocessen als verklaring voor de geconstateerde afwijkingen.

De situatie zoals weergegeven in figuur 3 is echter nogal uitzonderlijk in psychometrisch onderzoek. Steekproeven van dergelijke omvang heeft men doorgaans niet, althans niet in Nederland, en als men ze al heeft dan worden ze vaak niet in hun volle omvang gebruikt, want verwerping van het model betekent meestal dat het onderzoek niet voor publikatie in aanmerking komt. Niemand hoeft echter toe te geven dat hier de schoen knelt, want het is waar dat kleine defecten in het model bij grote steekproeven tot statistische significantie leiden. In bepaalde opzichten kan het dus voordelig zijn om kleine steekproeven te gebruiken: de systematiek in de afwijkingen blijkt niet zonneklaar, de afwijkingen worden toeval genoemd en verdwijnen onder die naam in de zwarte doos.

Het is echter de vraag of dit vanuit wetenschappelijk oogpunt een vruchtbare aanpak is. Mijns inziens zijn er een aantal strategieën die ons in staat stellen om ook bij kleine steekproeven een interessantere structuur in de zwarte doos aan te brengen dan het ongestructureerde toevalsbegrip.

Specifieke modeltoetsen

De meest gebruikte techniek om statistische toetsen te bouwen is de zogenaamde likelihood ratio-toets, waarmee nagegaan kan worden of een speciaal geval van een meer algemeen model houdbaar is, d.w.z. of het hanteren van een spaarzamer model niet wordt tegengesproken door de observaties. Dit is een prima benadering indien het algemene model niet in twijfel moet of kan worden getrokken, zoals het hanteren van een verzadigd model in de log-lineaire analyse. Toepassing van de LR-toets om bijvoorbeeld de aanvaardbaarheid van restricties op de parameterruimte in het Raschmodel te toetsen is echter problematischer, omdat het algemene model hier zelf reeds een zeer gespecialiseerd model is, waarvan de geldigheid in de meeste gevallen niet zonder meer kan worden aangenomen. Toepassing van de LR-toets in een geval waarin de alternatieve hypothese - het algemene model - flagrant onjuist is, is een twijfelachtige onderneming en het gebruik van LR-toetsen dient mijns inziens vermeden te worden als niet eerst de aanvaardbaarheid van de alternatieve hypothese overtuigend is aangetoond.

De afgelopen jaren heb ik, in nauwe samenwerking met Kees Glas (Glas en Verhelst, 1989; Verhelst en Eggen 1989; Glas en Verhelst, 1993; Glas en Verhelst, in voorbereiding) een groot deel van mijn tijd besteed aan het ontwerpen van een klasse van statistische toetsen, een veralgemening van Pearsons chi-kwadraattoets, waarbij op lokaal niveau, d.w.z. op he niveau van de detailstructuur het model statistisch kan worden getoetst. De figuren 2 en 3 zijn daarvan een voorbeeld: de totale toets waarop de analyse is uitgevoerd bestond uit 250 items, doch de statistische toetsing is zo opgebouwd dat elk item afzonderlijk op zijn deugdelijkheid kan worden getest. De idee van itemgerichte toetsen in IRT is niet nieuw, wat wel nieuw is, is het rigoureuze mathematische bewijs van de asymptotische verdeling van de toetsingsgrootheden en de grote algemeenheid en flexibiliteit van de benadering. Het construeren van statistische toetsen die gevoelig zijn voor itemonzuiverheid, het verschillend functioneren van items in verschillende populaties van personen, bleek een uiterst eenvoudige toepassing te zijn van de klasse van veralgemeende Pearsons chi-kwadraattoetsen (Verhelst, 1992).

Meer in het algemeen echter is het belang van deze benadering hierin gelegen dat met een relatief geringe inspanning een deugdelijke statistische toets kan geconstrueerd worden die gevoelig is voor specifieke defecten van het psychometrisch model. Een mooie toepassing, die helemaal in de lijn ligt van het werk dat op de vakgroep OMD, mijn tweede thuis, door Kelderman is verricht, is het aftasten van de modelruimte die gedefinieerd is door Keldermans loglineaire Raschmodellen (Kelderman, 1984). Elk model in Keldermans systeem dat meer algemeen is dan het gewone Raschmodel kan als alternatieve hypothese in de veralgemeende Pearson's chi-

kwadraat-toets worden ingebouwd; het rekenwerk is beperkt tot het uitrekenen van de itemparameterschattingen onder het gewone Raschmodel, en de toetsingsgrootheid is een kwadratische vorm, waarvan het uitrekenen heel wat minder tijd vergt dan het schatten van de parameters onder het alternatieve model. Wellicht een interessant samenwerkingsproject tussen de VU en de UT.

Het lineair logistisch testmodel (LLTM)

Door Fischer (1973) werd een modelbenadering ontworpen waarbij hypothesen over de moeilijkheid van een item kunnen worden vertaald als lineaire combinaties van een relatief klein aantal basisparameters, die als het ware de moeilijkheid weerspiegelen van de meest basale operaties of principes die nodig zijn voor de correcte oplossing van een item. In een test voor mechanisch inzicht (Spada, Fischer & Heyner, 1973) werden items gepresenteerd waarin een of andere constructie met tandwielen en aandrijfriemen grafisch werd voorgesteld (zie figuur 4), en waarbij gevraagd werd of het donkergekleurde (tand-)wiel wel kan draaien, en zo ja, in welke richting het draait, rekening houdend met de bedoelde draairichting van een ander wiel, zoals door de pijl aangegeven in de figuur. Om dit item correct op te lossen dienen 4 mechanische principes te worden toegepast:

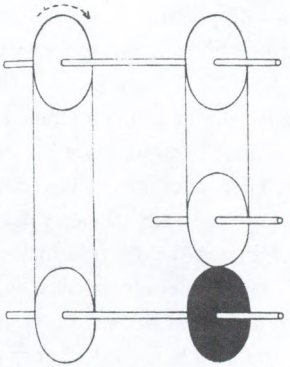


Fig. 4. Een voorbeeld-item uit een test voor mechanisch inzicht

(1) twee wielen waarvan de randen elkaar raken, draaien in tegengestelde richting;

(2) twee wielen die op dezelfde as zijn bevestigd draaien in dezelfde richting;

(3) twee wielen die verbonden zijn door een niet-kruisende aandrijfriem draaien in dezelfde richting;

(4) een wiel waarop twee even grote krachten inwerken die tegengesteld zijn in richting, kan niet draaien.

Het LLTM, toegepast op het item in figuur 4, komt nu neer op het volgende: elk van de vier basisoperaties heeft een bepaalde moeilijkheid, zeg α_1, α_2, α_3 en α_4 respectievelijk; de moeilijkheidsgraad van het item kan geschreven worden als

$$\text{moeilijkheid item} = c + 1\times\alpha_1 + 2\times\alpha_2 + 2\times\alpha_3 + 1\times\alpha_4 \qquad (1)$$

waarin c een constante voorstelt die voor onze uiteenzetting verder niet belangrijk is. Wat wel belangrijk is, zijn de getallen waarmee de onbekende parameters α worden vermenigvuldigd: ze zijn gelijk (of in het algemeen: proportioneel) met de frequentie waarmee de vier onderscheiden principes voorkomen en dus moeten worden toegepast in het item.

Dit voorbeeld is in meerdere opzichten interessant: (i) Het is een voorbeeld waarbij een poging wordt ondernomen de zwarte doos wat meer structuur te geven: de moeilijkheid van een item wordt teruggebracht op de moeilijkheid van meer basale componenten. Indien de theorie juist is, kunnen we uit de antwoorden van leerlingen op dit en soortgelijke items de parameters α schatten, waarbij de nauwkeurigheid van de schatting in eerste instantie afhankelijk is van het aantal personen dat de items beantwoordt. Maar de opbrengst is groter dan dat: we kunnen ook - alweer als de theorie juist is - de moeilijkheid voorspellen van elk nieuw te construeren item dat een beroep doet op dezelfde basisprincipes. (ii) De theorie is complex en bevat verschillende aspecten die in twijfel kunnen worden getrokken. Het hanteren van de vier hierboven genoemde basisprincipes lijkt wellicht voor de hand liggend voor wie een elementair leerboek over mechanica moet schrijven, maar de theorie is niet bedoeld als leidraad voor het ontwikkelen van cursusmateriaal doch als verklaring voor het gedrag van leerlingen die de test maken. M.a.w., de theorie is bedoeld als psychologische theorie en niet als didactische theorie. De gewogen optelling van de elementaire moeilijkheden in formule (1) veronachtzaamt bijvoorbeeld het elementaire principe van de Gestaltpsychologie dat het geheel meer is dan de som van de delen. Het zou goed kunnen zijn dat de combinatie van twee principes het item als geheel moeilijker maakt dan uit de moeilijkheid van de samenstellende principes valt af te leiden, of misschien juist gemakkelijker. De zwakste schakel in de hele theorie is mijns inziens echter de exacte specificatie van de coëfficiënten van de lineaire combinatie: door het aannemen van een toename in moeilijkheid die proportioneel is met de frequentie waarmee de basisprincipes in het item voorkomen doet men heel sterke uitspraken waardoor de theorie uiterst kwetsbaar wordt. Het zal dan ook geen

verwondering wekken dat de auteurs op statistische gronden hun theorie moesten verwerpen. (iii) De wijze van aanpakken is heel modern: het LLTM blijkt tal van interessante toepassingen te hebben (zie Fischer, in voorbereiding, voor een overzicht), hoewel het succes, in termen van goede overeenkomst tussen data en theorie eerder aan de magere kant is: het model moet bijna steeds worden verworpen.

Ik denk dat de tweespalt tussen enerzijds een theoretisch zeer aantrekkelijke en flexibele aanpak en anderzijds de bijna continue frustratie dat de modellen niet door de empirie worden gesteund, grotendeels kan worden opgeheven door de meest rigide component van het model, het fixeren van de coëfficiënten van de lineaire combinatie, iets te versoepelen. Dit kan op twee verschillende manieren.

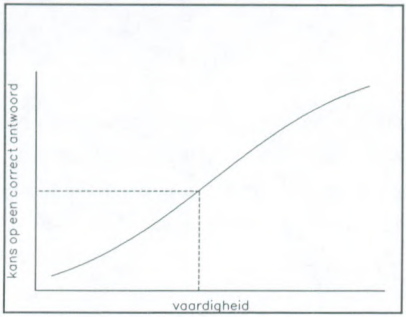
(i) Er kunnen statistische toetsen worden gebouwd die gevoelig zijn voor defecten van het model die het gevolg zijn van een misspecificatie van één of meer van de coëfficiënten. Toepassing van deze toetsen leidt tot het ontwikkelen van zogenaamde modificatie-indices die tamelijk gedetailleerd aangeven hoe een gegeven LLTM kan verbeterd worden door aanpassing van specifieke regressie-coëfficiënten. Hoewel deze manier van toetsen, bekend als Lagrange Multiplier toetsen, in de statistische literatuur reeds lang bekend is (Silvey,1959) en in de literatuur over lineaire structurele modellen extensief wordt gebruikt (Bollen, 1989), is het de verdienste van Glas (Glas & Verhelst, 1993) geweest de toepassingsmogelijkheden voor deze benadering in te zien voor gebruik binnen de IRT, en speciaal voor toepassingen met het LLTM. Hier ligt een potentieel heel vruchtbaar veld van onderzoek op ontginning te wachten.

(ii) De tweede manier bestaat er uit de coëfficiënten van de lineaire combinatie helemaal niet te specificeren, maar ze uit de data te schatten. Als we de constanten in (2) vervangen door onbekende grootheden, dan is onmiddellijk duidelijk dat het model niet meer lineair is, waardoor het schattingsprobleem, dat in het gewone LLTM een routinematige klus is die zelden problemen oplevert, behoorlijk ingewikkeld wordt. Doch daar is met een beetje inventiviteit wel overheen te komen. Een veel groter probleem is echter dat we niet alle coëfficiënten de status van onbekende parameter kunnen geven, want dan is het model niet meer geïdentificeerd. Er bestaan oneindig veel oplossingen voor de schattingsvergelijkingen, die wiskundig allemaal evenwaardig zijn, doch geen eenduidige interpretatie meer toelaten. Hoewel er reeds een paar successen met deze aanpak zijn geboekt (Butter, De Boeck & Verhelst, 1993; Butter, De Boeck & Stouthard, 1994) is het algemeen probleem van het formuleren van de precieze voorwaarden waaronder deze uitbreiding van het LLTM geïdentificeerd is nog niet helemaal opgelost.

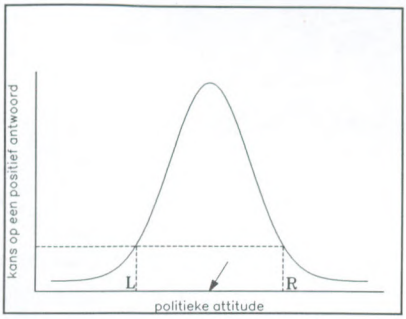
Adequate informatie

Hierboven heb ik er reeds op gewezen dat de gevoeligheid van de statistische toetsen in grote mate afhangt van de grootte van de steekproef. Dit houdt echter niet automatisch een pleidooi in om met grote steekproeven te gaan werken. Meer algemeen geldt dat hoe nauwkeuriger we uitspraken willen doen over het functioneren van het cognitieve systeem, des te meer informatie we via onze observaties dienen te verzamelen. De omvang van de steekproef laten toenemen is een manier om meer informatie te verzamelen, maar het is zeker niet de enige. Ik zal U hiervan een voorbeeld geven uit mijn eigen onderzoekswerk (Verhelst en Verstralen, 1993).

Bij het ontwerpen van modellen voor het meten van cognitieve vaardigheden maken we meestal gebruik van een zeer krachtig principe, namelijk de monotoniciteit, dat grofweg stelt dat hoe groter de vaardigheid is, des te groter de kans dat een item positief wordt beantwoord. Deze relatie wordt grafisch weergegeven door een curve die men de itemresponscurve pleegt te noemen (zie figuur 5). Deze relatie is een 1-1 relatie: uit de vaardigheid kan ondubbelzinnig de kans worden bepaald en uit de kans volgt ondubbelzinnig de vaardigheid. Wanneer men zich echter gaat bezighouden met het meten van attitudes en voorkeuren, kan men in veel gevallen geen gebruik meer maken van dit principe. Stel dat men in een onderzoek naar politieke attitude een aantal personen de volgende uitspraak voorlegt: "*Joop Den Uyl was een goede premier voor Nederland*", en laten we voorts veronderstellen dat het antwoord bepaald wordt door wat men gewoonlijk de politieke 'links-rechts'-oriëntatie noemt, dan is het redelijk aan te nemen dat de itemresponscurve een vorm aanneemt zoals weergegeven in figuur 6, waarbij het pijltje de politieke attitude aangeeft van iemand die het volstrekt eens is met bovenstaande uitspraak. Als de attitude van de ondervraagde persoon hiermee grote gelijkenis vertoont - d.w.z. als zijn politieke attitude kan worden afgebeeld als een punt op de lijn in de buurt van het pijltje - dan is de kans groot dat de uitspraak onderschreven wordt. Anders uitgedrukt: van twee personen die beide positief reageren op de uitspraak, is het redelijk te verwachten dat ze er ongeveer dezelfde politieke mening op nahouden. Helemaal anders echter is de situatie bij twee personen die het oneens zijn met de uitspraak. De persoon L verwerpt de uitspraak omdat hij Den Uyl te rechts vindt en persoon R verwerpt de uitspraak omdat hij Den Uyl te links vindt. Twee identieke doch negatieve keuzen kunnen dus geenszins opgevat worden als een indicatie voor gelijke of ongeveer gelijke politieke attitude: de negatieve keuzen bevatten duidelijk minder informatie over het te meten attribuut dan de positieve.



Figuur 5
Een monotone itemresponscurve



Figuur 6
Een niet-monotone itemresponscurve

De wijze waarop Verstralen en ik het antwoordgedrag op zo'n vragenlijst gemodelleerd hebben is in grote lijnen als volgt. Hoewel er bij elk item van de vragenlijst slechts twee antwoordcategorieën in aanmerking komen (positief of negatief) wordt verondersteld dat er eigenlijk drie antwoordcategorieën zijn: 'akkoord', 'niet akkoord want te links' en 'niet akkoord want te rechts'. Deze antwoordcategorieën, zo luidt de redenering, worden echter niet geobserveerd, en worden daarom latent genoemd. Wat we wel observeren is alleen een manifeste akkoordverklaring of verwerping. De relatie tussen latente en manifeste antwoordcategorieën is weergegeven in figuur 7.

latente antwoorden	manifeste antwoorden
akkoord	akkoord
niet akkoord, te links	niet akkoord
niet akkoord, te rechts	niet akkoord

Figuur 7. Relatie tussen latente en manifeste antwoorden

Het psychometrisch model, dit wil zeggen de algemene hypothese over het antwoordgedrag bij de vragenlijst bevat dus twee componenten: een die aangeeft wat de relatie is tussen de politieke attitude en het latente antwoord, en een die de relatie tussen latente en manifeste antwoorden modelleert. M.a.w. we hebben een behoorlijk ingewikkelde zwarte doos geconstrueerd.

Mathematisch is dit niet zo'n probleem. Voor de oplossing van de schattingsvergelijkingen beschikken we over een schitterend analyse-instrument, het EM-algoritme (Dempster, Laird en Rubin, 1977), dat de oplossing bijna zeker garandeert. Hoewel de toepassing van dit algoritme in dit voorbeeld zeker geen triviale kwestie is, zou ik ze toch willen categoriseren als een technisch vraagstuk. Het

creatieve moment bij de modellering is gelegen in het concept van het latente antwoord (zie ook Maris (1992) voor een buitengewoon vruchtbare toepassing van deze benadering), maar terzelfdertijd rijst de vraag of we het hierbij moeten laten. Hoe elegant deze benadering ook mag lijken, we blijven zitten met het feit dat we informatie-arme, want dubbelzinnige observaties hebben verzameld, en voor deze armoede moeten we de tol betalen. De statistische toetsen ter controle van de modelgeldigheid zijn niet erg krachtig, en de beslissing tussen dit en een concurrerend model zal noodzakelijkerwijze niet erg helder zijn. De uitweg uit deze weinig comfortabele positie lijkt voor de hand liggend: als we dan toch zo gecharmeerd zijn van het concept van het driewaardige latente antwoord, waarom proberen we dan niet dit antwoord manifest te maken, door aan de respondenten bijvoorbeeld naar de reden te vragen waarom ze een uitspraak verwerpen, en op die manier althans één laag van de zwarte doos aan het licht van de observaties bloot te stellen.

Deze oplossing lijkt misschien eerder karakteristiek voor het gezond verstand dan het resultaat van academisch intellect, en ik hoop van harte dat mij nog een beetje van dat eerste rest, maar het probleem is toch ingewikkelder dan op het eerste gezicht lijkt. Vergeten we niet dat de links-rechts oriëntatie als verklarende dimensie voor de politieke attitude geen gegeven is, maar alleen gehanteerd werd om het voorbeeld duidelijk te maken; ze kan dus hoogstens als hypothese in het onderzoek fungeren, niet als feit. Door de respondenten expliciet te vragen naar een reden voor verwerping die gesteld is in termen van deze dimensie, riskeren we die te induceren, zodat de uitkomst van het onderzoek eerder een artefact van de methode is dan een verklaring van politieke attitude en politiek gedrag. Wat we daarom moeten doen is op zoek gaan naar een methode van observatie of dataverzameling die terzelfdertijd neutraal (unobtrusive, zie Webb, Campbell, Schwartz en Sechrest, 1966) is en meer informatie oplevert dan de binaire antwoorden waaraan wij zo gewoon zijn. Men zou bijvoorbeeld de respondent na beantwoording van de vragenlijst kunnen vragen de uitspraken die hij heeft verworpen, op te delen in stapeltjes, zodat de uitspraken die tot hetzelfde stapeltje behoren om dezelfde of gelijkaardige redenen verworpen zijn. Het probleem is dan natuurlijk op welke wijze de twee soorten antwoorden, enerzijds de aanvaarding of verwerping, anderzijds de classificatie, in eenzelfde formeel psychometrisch model kunnen worden geïntegreerd. Voorwaar een uitdagend onderwerp voor een ondernemende AIO.

Het probleem dat door het voorgaande voorbeeld is gesignaleerd, heeft mijns inziens een veel bredere betekenis dan de loutere toepassing in de politicologie waarnaar het voorbeeld refereert. Het is niet toevallig dat vakgenoten-psychometrici in academische kringen meestal het -minder of meer formele - bijbaantje hebben van

methodoloog, en geacht worden toe te zien dat het meer inhoudelijk georiënteerde onderzoek van de collega's volgens de regels van de kunst verloopt, en waarvan soms verwacht wordt dat zij genoegen nemen met de rol van 'ancilla scientiae', de dienstmaagd der wetenschap, die met veel ijver, creativiteit en inventiviteit de data analyseert die hij toegeschoven krijgt door de 'echte' onderzoeker. Het is een grote verdienste van Coombs (1964), er op te wijzen dat observatiegegevens en data niet een en hetzelfde zijn: data zijn reeds ten dele geïnterpreteerde observaties, die beladen zijn met theorie. Het ruwe materiaal waarop de psychometricus zijn modellen loslaat is een zeer grote tabel die opgevuld is met enen en nullen. Of we een 1 in zo'n tabel nu moeten zien als een dominantierelatie (de leerling beschikt over meer vaardigheid dan nodig is om een item op te lossen), dan wel als een nabijheidsrelatie (de overeenkomst tussen de politieke uitspraak en de politieke attitude van de respondent is voldoende groot om een positieve reactie uit te lokken) valt niet aan de tabel af te zien. De methodoloog dient er op te wijzen dat zo'n tabel multi-interpretabel is, dat door de onderzoeker voor een interpretatie gekozen moet worden. Een dwingende interpretatie wordt niet door een slim computerprogramma opgelegd, ook niet als het aangestuurd wordt door een slimme methodoloog. Maar ik denk dat we nog een stap verder kunnen gaan: we dienen er onze cliënten op te wijzen dat sommige observaties zeer arm aan informatie zijn, dat de analyse op de data impliceert dat er een nogal complexe zwarte doos wordt geconstrueerd en dat de besluiten die uit de analyse volgen noodzakelijkerwijze met een brede marge van onzekerheid beladen zullen zijn. Dit is de passieve kant van ons werk.

Aan de actieve kant valt er echter ook een en ander te doen. Mijn grootste cliënt is het Cito, het instituut voor toetsontwikkeling, dat zich in zijn nu meer dan vijfentwintigjarig bestaan opgewerkt heeft tot de landelijke expert in het meten van leerresultaten, waarbij overvloedig gebruik wordt gemaakt van meerkeuzevragen, ik zou bijna zeggen op instigatie van De Groot en Van Naerssen (1966). En inderdaad, de voordelen van de meerkeuzevraag die deze auteurs aandragen zijn niet onbelangrijk: de scoring van de antwoorden is goedkoop en objectief, en leidt meestal tot meer betrouwbare toetsen dan het gebruik van open vragen. De vraag is of deze vergelijking wel helemaal eerlijk is. De kritiek op de open-vraagvorm is meestal dat er meetonnauwkeurigheid geïntroduceerd wordt door het subjectieve oordeel van de corrector. Dit zal wel waar zijn, doch dit betekent niet dat men deze onnauwkeurigheid niet kan reduceren. Uit onderzoek (Sanders, Hendrix en Luijten, 1984; Hendrix en Sanders, 1987) blijkt dat een analytische beoordelingsprocedure veel betrouwbaardere meetresultaten oplevert dan een globale procedure. Bovendien maak ik mij meer zorgen over de nadelen van de meerkeuzevraagvorm: een correct antwoord kan een

weerspiegeling zijn van kennis of vaardigheid doch het kan ook het resultaat zijn van een gelukkige gok. Dit wil zeggen dat, net als in het voorbeeld van de politieke attitude de antwoorden een zekere mate van dubbelzinnigheid hebben en daardoor minder informatie dragen over die ene variabele die ons interesseert, namelijk de vaardigheid. Hoewel we deze dubbelzinnigheid expliciet kunnen opnemen als onderdeel van de zwarte doos door een model te hanteren dat expliciet voorziet in de mogelijkheid tot correct raden, gebeurt dit op het Cito in de regel niet, waarbij de belangrijkste reden is dat het meest populaire model voor dit soort toetsen, het drie-parameter logistisch model of een variant daarvan, in de regel zeer instabiele resultaten oplevert (zie Westers, 1993).

De open-vraagvorm, al dan niet met gedetailleerd correctievoorschrift is echter niet het enige alternatief. Het zou naïef zijn te denken dat indien de leerling het juiste alternatief in een meerkeuzevraag niet met zekerheid herkent, dat hij dan blind gaat gokken tussen alle alternatieven. Partiële kennis zou bijvoorbeeld kunnen blijken uit het feit dat hij met zekerheid twee van de vier alternatieven kan elimineren, doch dat hij tussen de overblijvende twee niet kan beslissen en gaat raden. Verstralen (1994) heeft aangetoond dat, indien we de kandidaat zouden toestaan beide alternatieven aan te kruisen, en deze informatie op een gepaste manier weten te verwerken -bijvoorbeeld volgens het model dat Verstralen heeft ontwikkeld- een veel nauwkeuriger meetresultaat verkregen wordt dan bij de klassieke observatie van een enkelvoudige keuze per item. Het werk van Verstralen heeft mij uitermate gestimuleerd om mijn afkeer tegen de meerkeuze-vraagvorm wat te nuanceren: het moet mogelijk zijn om op een efficiënte manier - d.w.z. met objectieve en bij voorkeur machinaal te verwerken scoringsregels - meer eenduidige informatie uit de meerkeuze-vraagvorm te halen dan nu het geval is, althans als we het lef hebben enige sturing te geven aan de wijze waarop de data verzameld dienen te worden, en ons niet kritiekloos neerleggen bij het dataformaat dat we nu eenmaal gewoon zijn.

Meten met Correctie

De strategieën die ik tot hiertoe heb geschetst, hebben als gemeenschappelijk kenmerk dat ze er op gericht zijn - door meer geëigende statistische methodes van toetsen en schatten, dan wel door meer adequate observaties - de structuur van de zwarte doos inzichtelijker dan wel eenvoudiger te maken. Het is echter geen kwestie van alles of niets. Hoe goed onze modellen ook in elkaar steken, het blijven idealiseringen van de werkelijkheid. Toch is er een strategie mogelijk die ons moet toestaan in onze meetprocedures de defecten van ons geïdealiseerd model onder controle te houden.

Als voorbeeld grijp ik nog even terug naar de Lumsdenmachine. Bij de bespreking van dit meetinstrument zijn we ervan uit gegaan dat iedereen die aan de meetprocedure wordt onderwerpen gedurende de tijd dat de meting duur even lang blijft. We zien ook in dat de langste personen (gemiddeld) de meeste klappen zullen oplopen, en we weten uit ervaring dat een klap op onze schedel vaak als gevolg heeft dat daar een bult verschijnt, die ons tijdelijk wat langer maakt dan we eigenlijk zijn. Dit wil zeggen dat de meetprocedure zelf de eigenschap die gemeten wordt, verandert, en dat ze dit voor verschillende personen op ongelijke wijze kan doen, waarbij de ongelijkheid niet puur toevallig is, maar op een systematische wijze samenhangt met de te meten eigenschap. Dit soort effecten bestaat ook bij het meten van cognitieve vaardigheden in de vorm van vermoeidheids- of vervelingseffecten of in de vorm van leereffecten, waar de respondent als het ware bijleert gedurende het afleggen van de toets, waarbij de mate van leren dan weer afhankelijk kan zijn van de kwaliteit van zijn antwoord. Meer in het algemeen hebben dit soort effecten vaak als gevolg dat een belangrijke veronderstelling van het model, de lokale stochastische onafhankelijkheid, niet meer waar is, of althans niet meer bruikbaar is, waardoor toepassing van de meest gangbare psychometrische modellen niet meer gerechtvaardigd is.

Gelukkig zijn er psychometrische modellen ontwikkeld die dit soort effecten ook kunnen modelleren in zeer algemene zin zonder dat het daarbij noodzakelijk is dat hun werkingsmechanisme verklaard of begrepen wordt (Kelderman, 1984; Jannarone, 1986; Verhelst en Glas, 1993). Merkwaardig genoeg schijnen echter de auteurs van deze modellen er genoegen mee te nemen aan te tonen dat deze effecten bestaan en een schatting te maken van hun grootte. Nog nergens heb ik een toepassing gezien waarbij de individuele meetuitslag, de schatting van de individuele vaardigheid, voor dit soort effecten wordt gecorrigeerd, terwijl we het toch als heel normaal beschouwen bijvoorbeeld werkloosheidscijfers te horen die gecorrigeerd zijn voor seizoensfluctuaties, en die belangrijk kunnen afwijken van de feitelijke tellingen of schatting daarvan op een gegeven moment. Hoewel we natuurlijk voorzichtig moeten zijn met het toepassen van dergelijke procedures, die waarschijnlijk niet erg transparant zullen zijn, in gevallen waar belangrijke beslissingen worden genomen over individuen, denk ik dat het belangrijk is de nodige tijd aan dit probleem te besteden, omdat het de mogelijkheid biedt meetuitslagen van instrumentatie effecten te zuiveren, en aldus een bijdrage te leveren aan wetenschappelijk heldere concepten.

De psychometrie als zwarte doos

De afgelopen 30 jaar is behoorlijk veel over de moderne psychometrie gepubliceerd, waarbij een substantieel gedeelte van de publikaties lezenswaardig is. Ze is uitgegroeid tot een redelijk geïntegreerd pakket van theoretisch inzicht en technologisch kunnen dat eigenlijk zonder veel moeite kan ingezet worden in de meetpraktijk binnen de sociale wetenschappen. De psychometrische gemeenschap in Nederland heeft zich in dat opzicht niet onbetuigd gelaten. Veel creatieve en innovatieve bijdragen in de gespecialiseerde vakliteratuur zijn van Nederlandse hand. In dit licht is het wellicht een beetje bevreemdend dat van deze theorie (althans in Nederland) relatief weinig gebruik wordt gemaakt bij het construeren van meetinstrumenten. De zesde uitgave van 'Documentatie van Tests en Testresearch in Nederland' (Evers, Van Vliet-Mulder en Ter Laak, 1992) bevat een kwaliteitsbeoordeling van 376 tests die alle kunnen aangeduid worden als 'psychologische tests'. De beoordeling wordt verricht door de gezaghebbende Commissie Test Aangelegenheden Nederland (COTAN). Aan de moderne testtheorie worden in de verantwoording van de beoordelingsprocedure precies drie regels gewijd (p. 5) waarin wordt gezegd dat met enige soepelheid, het beoordelingssysteem dat op de klassieke testtheorie is gebaseerd, ook van toepassing is op tests die volgens de moderne testtheorie zijn geconstrueerd. Deze soepelheid is bovendien nauwelijks nodig, want referenties naar IRT-principes zijn in het meer dan 800 pagina's dikke boek nauwelijks te vinden.

Hoe komt het dat professionele instrumentenbouwers geen gebruik maken van een methodologie en een theorie die pretendeert, en mijns inziens terecht, superieur te zijn aan de klassieke testtheorie?

Ik heb reeds vele antwoorden op deze vraag gehoord, waarvan de leukste steevast refereren aan de tekortkomingen van anderen. Maar misschien is het passend in deze context ook eens de hand in eigen boezem te steken. Mijn indruk is dat studenten en collega's die van de psychometrie niet hun beroep maken, tegen ons vak een beetje aankijken als tegen een zwarte doos, maar dan in de betekenis van een mysterie dat slechts voor enkele ingewijden toegankelijk is. En dat de club van ingewijden dit eigenlijk wel leuk vindt. De sleutel om deze zwarte doos open te maken is gelegen in het onderwijs, vooral op elementair niveau, aan een zo breed mogelijk publiek van studenten in de sociale wetenschappen.

Referenties

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Butter, R., De Boeck, P., & Verhelst, N.D. (1993). *Item response model with internal restrictions on item difficulty*. (Research group on quantitative methods report 93-7). Leuven: Faculteit voor psychologie en pedagogische wetenschappen.

Butter, R., De Boeck, P., & Stouthard, M. (1993). *Item response model with internal restrictions on item difficulty applied to facet designs*. (Research group on quantitative methods report 94-2). Leuven: Faculteit voor psychologie en pedagogische wetenschappen.

Coombs, C.H. (1964). *A theory of data*. New York: Wiley.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B, 39*, 1-38.

Ebbinghaus, H. (1885). *Über das Gedächtnis*. Leipzig.

Evers, A., Vliet-Mulder, J.C. van, & Laak, J. ter (1992). *Documentatie van tests en testresearch in Nederland*. Amsterdam: Nederlands Instituut van Psychologen.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-373.

Fischer, G.H. (in voorbereiding). The linear logistic test model. In: G.H. Fischer, & I.W. Molenaar (red.). *Rasch models: their foundations, recent developments ans applications*.

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54*, 635-659.

Glas, C.A.W., & Verhelst, N.D. (1993). Een overzicht van itemresponsmodellen. In: T.J.H.M. Eggen en P.F. Sanders (red.). *Psychometrie in de praktijk* (pp. 179-238). Arnhem: Cito.

Glas, C.A.W., & Verhelst, N.D. (in voorbereiding). Testing the Rasch model. In: G.H. Fischer, & I.W. Molenaar (red.). *Rasch models: their foundations, recent developments ans applications*.

Groot, A.D. de, & Naerssen, R.F. van (1966). *Studietoetsen*. Den Haag: Mouton.

Hendrix, A.C., & Sanders, P.F. (1987). De beoordeling van de samenvatting Nederlands vwo: de analytische beoordelingsprocedure in de praktijk. *Tijdschrift voor taalbeheersing, 9*, 151-165.

Jannarone, R.J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Lumsden, J. (1976). Test Theory. *Annual review of psycholgy, 27*, 251-280.

Maris, E. (1992). *Psychometric models for psychological processes and structures.* Proefschrift, Unversiteit Leuven.

McCorquodale, K., & Meehl, P.E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological review, 55*, 95-107.

Pavlov, I.P. (1904). Dvadtzatiletni opit obektivnogo izucheniya vischei nervnoi deyatelnosti (povendeniya) zhivotnik. Moscow en Leningrad: State publishing house, 1923, 1925[3]. Derde editie vertaald door: W. Horsley Gantt, *Lectures on conditional reflexes*. New York: International Publishers, 1928, pp. 76-80.

Sanders, P.F., Hendrix, A.C., & Luijten, A.J.M. (1984). De beoordeling van de samenvatting Nederlands. *Tijdschrift voor taalbeheersing, 6*, 241-251.

Silvey, S.D. (1959). The Lagrangian multiplier test. *Annals of mathematical statistics, 30*, 389-407.

Spada, H., Fischer, G.H., & Heyner, W. (1973). Die Analyse von Denkoperationen und Lernprozessen bei der Lösung von Problemstellungen aus der Mechanik mittels des linearen logistischen Modells. In: H. Spada, P. Häussler, & W. Heyner (red.). *IPN-Arbeitsbericht: Denkoperationen und Lernprozesse als Grundlage für lernorientierter Unterricht*. Kiel: Institut für Pädagogik der Naturwissenschaften der Universität Kiel.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid: de ontwikkeling van een domeingericht meetinstrument*. Proefschrift. Arnhem: Cito.

Verhelst, N.D. (1992). *Het eenparameter logistisch model.* (OPD Memorandum 92-3). Arnhem:Cito.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek.* (PPON-rapport, nr. 4). Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395-415.

Verhelst, N.D., & Verstralen, H.H.F.M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve methoden, 14*, 73-92.

Verstralen, H.H.F.M (1994). *A logistic latent class model for multiple choice items.* (Measurement and Research Department Reports, 94-1). Arnhem: Cito.

Webb, E.J., Campbell, D.T., Schwartz, R.D., & Sechrest, L. (1966). *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.

Westers, P. (1993). *The solution-error response-error model: a method for the examination of test item bias.* Proefschrift. Enschede: Universiteit Twente.