KM 5(1982)
pag 3-29

MENSEN DIE HET BETER METEN

Een inleiding tot enige latente trek modellen

Ivo W. Molenaar *)

Samenvatting

Op verzoek van de redactie heb ik een verkorte en enigszins bijbijgewerkte versie van Molenaar (1979) geschreven, met het doel om de lezers die van dit gebied niet op de hoogte zijn een overzicht te bieden van statistische modellen voor eendimensionele schaalconstructie in de sociale wetenschappen.

Na een korte inleiding over meetproblemen wordt de klassieke testtheorie in paragraaf 2 geschetst en bekritiseerd. De in paragraaf 3 algemeen ingeleide latente trek modellen proberen op die kritiek een antwoord te bieden. Op didactische en niet-historische gronden komen in paragraaf 4 eerst de modellen van Guttman en Mokken aan de orde. Het uiterst belangwekkende maar veeleisende Rasch-model wordt in paragraaf 5 ingeleid, gevolgd door enige varianten en uitbreidingen in paragraaf 6.

Paragraaf 7 bevat een persoonlijke evaluatie van enige voordelen en bezwaren van elk model. Mijn conclusie is dat er voorlopig geen dwingende redenen zijn in alle omstandigheden hetzelfde schaalmodel te gebruiken. De behandelde literatuur, waarvan alleen de in mijn ogen belangrijkste of meest toegankelijke werken zijn genoemd, geeft volgens mij wel een interessante aanzet tot verbeterde meetprocedures die voor de sociale wetenschappen van belang zijn.

Het hier gegeven overzicht is lang niet volledig, en ten dele door persoonlijke smaak bepaald. De statistische en inhoudelijke problemen bij schaalconstructie hebben ook nog geen (in alle opzichten) bevredigende oplossingen. Als deze bijdrage enige lezers zou stimuleren daaraan bij te dragen, en anderen een handzaam overzicht biedt, acht ik mijn doel al bereikt. Andere artikelen in deze en de volgende aflevering van Kwantitatieve Methoden gaan op enige specifieke problemen in.

---

*) Vakgroep Statistiek en Meettheorie, FSW, Oude Boteringestr. 23, 9712 GC Groningen.

1. Inleiding

"Measurement is the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals. Numbers are assigned to the individuals according to a carefully prescribed, repeatable procedure." Dit citaat uit Allen & Yen (1979) brengt mij bij de vraag waarom meten in de psychologie, of algemener in de sociale wetenschappen, tot extra problemen leidt. Zonder aanspraken op volledigheid noem ik ten eerste dat het voorwerp van meting een *individu* is, dat op het gemeten worden reageert, en waarbij herhaalde meting onder strikt identieke omstandigheden niet mogelijk is. Al is de procedure herhaalbaar, de uitkomst van een tweede meting kan door het feit van de eerste meting essentiëel anders zijn. Een tweede probleem is dat de te meten *eigenschap* in het algemeen nogal breed van aard is, en zich onder wisselende omstandigheden in verschillende gedragsvormen kan uiten.

De menselijke vrijheid, waaraan ik overigens geen strobreed in de weg zou willen leggen, maakt de toekenning van getallen aan personen dus tot een hachelijke onderneming. In de afgelopen decennia heeft de meettheorie getracht procedures en modellen te genereren die met deze specifieke problemen rekening houden. Onder namen zoals psychometrie of testtheorie hield men zich in het bijzonder bezig met het probleem van de *meetfouten*, en probeerde men de tests meer *betrouwbaar en valide* te maken (dat betekent in statistische termen ongeveer: een kleine variantie over meetherhaling resp. een zuivere schatting van het bedoelde begrip; validiteit is meer inhoudelijk dan wiskundig, zie par. 7).

De hier te bespreken modellen zijn zowel van toepassing op de meting van vaardigheden (zoals intelligentie of kennis van de biologie) als op de meting van houdingen of eigenschappen (zoals tolerantie t.o.v. abortus, emancipatiegezindheid of neuroticisme). In beide gevallen is eenmalige meting met één vraag notoir onbetrouwbaar. Omdat onafhankelijke herhaling zonder hersenspoeling doorgaans onmogelijk is, tracht men de betrouwbaarheid te verhogen door een test samen te stellen uit een aantal deelvragen die allen (iets van) dezelfde eigenschap meten.

Bij vaardigheidstests zijn dit verschillende opgaven, die elk met goed/fout worden beoordeeld; bij houdingstests zijn het meestal uitspraken, waarmee de proefpersoon het al dan niet eens is. Uit de antwoorden op de deelvragen, in de sociale wetenschappen "items" genoemd, construeert men dan de score op de gehele test, meestal door eenvoudig optellen. Het gaat er nu om de kwaliteit van zo'n meetconstructie te onderzoeken. Binnen een expliciet meetmodel is dat tot op zekere hoogte mogelijk; dat is een stap vooruit vergeleken met de "fiat-meting" waarbij de onderzoeker eenvoudigweg poneerde dat de totaalscore op de voorgelegde items de bedoelde eigenschap zou meten.

Nadat de *klassieke testtheorie* in haar bloeiperiode tot verbetering van de meetprocedures had geleid, zijn vervolgens de *latente trek modellen* opgekomen die weer andere voordelen bieden. Vooral het *Rasch-model* trekt daarbij sterk de aandacht. Het is de bedoeling hier kort na te gaan welke attracties en welke zwakheden deze modellen bieden voor de verantwoorde toekenning van getallen aan individuen ter meting van hun eigenschappen.

2. Klassieke testtheorie

De grofweg tussen 1930 en 1960 ontwikkelde klassieke testtheorie kan in statistische termen worden vertaald als een éénweg variantie-analyse met toevalseffecten (oneway random effects ANOVA). Het model wordt hier beknopt geschetst; een korte en duidelijke weergave staat in hoofdstuk 4 van Novick & Jackson (1974), een volledig werk is Lord & Novick (1968) en een inleiding vindt men bij Nunnally (1978) en Allen & Yen (1979). Fischer (1974) bespreekt de gehele theorie en voegt een zeer kritische evaluatie toe. De prijs voor de amusantste kritiek gaat naar Lumsden (1976).

Basis van de theorie is de hypothetische splitsing van elke geobserveerde score X van een persoon op een test in een ware score T en een meetfout E. In het dubbel stochastische model kiest men eerst aselect een persoon v uit de te meten populatie; deze heeft een ware score T_v. Per persoon v per replicatie k van de meting wordt daaraan een stochastische meetfout E_{vk} toegevoegd:

$$X_{vk} = T_v + E_{vk}.$$

Onder voor de hand liggende assumpties voor de verwachtingen, varianties en correlaties van de kansverdelingen kan men dan de betrouwbaarheid en validiteit van de test onderzoeken en waar nodig verbeteren. De al genoemde parallel met de variantie-analyse is overigens beperkt: daar wordt vooral de variantiecomponent σ_T^2 onderzocht, en met name getoetst of deze nul is. Die assumptie is bij de meting van individuen bij voorbaat zinloos. Het doel is veeleer tot betrouwbare uitspraken over elke individuele ware score T_v te komen (zie b.v. Novick, Jackson & Thayer, 1971).

Het begrippenapparaat van de klassieke testtheorie heeft tot een sterke verbetering van de meetprocedures geleid, vergeleken met de situatie waarin elke onderzoeker zo maar een stel uitspraken of opgaven verzon en zonder nadere adstructie beweerde dat de somscore een bepaalde eigenschap zou meten. Toch is in de loop der jaren ook veel kritiek geuit, die hier beknopt wordt samengevat.

Ten eerste heeft de splitsing in ware score en meetfout een sterk "fiat-karakter"; zij wordt gepostuleerd, maar omdat T en E niet afzonderlijk observeerbaar zijn is het nauwelijks mogelijk de assumpties over hun kansverdelingen empirisch te verifiëren.

Ten tweede krijgt het individu alleen een ware score op één bepaalde test, terwijl het meestal de bedoeling is iets te zeggen over de mate waarin de persoon de eigenschap heeft, los van de gebruikte test. Men kan in de praktijk moeilijk een "paralleltest" construeren waarop iedere persoon dezelfde ware score heeft als op de eerste. De invoering van een tweede factor (tests) in het variantie-analytisch model heeft tot interessante theorie over de zgn. generische en specifieke ware score geleid, maar ook hier is de praktische onmogelijkheid van onafhankelijke herhaalde meting een groot struikelblok gebleken: zonder replicaties weet men te weinig om de theorie zinvol te hanteren.

Ten derde is de betrouwbaarheid (dit is binnen het model de correlatie met de hypothetische paralleltests, tevens gelijk aan de verhouding van de varianties van ware en van geobserveerde scores) sterk afhankelijk van

de groep gemeten personen, en dit geldt ook voor de nauwkeurigheid van een uitspraak over de ware score van één persoon.

Ten vierde ligt het voor de hand dat men zich de te meten eigenschap als continu variërend op een onbeperkt interval voorstelt; de geobserveerde score is echter *discreet* en kent een *vloer* en *plafond*. Alleen hierom al zijn de assumpties (zoals homoscedasticiteit en onafhankelijkheid) over de verdelingen van T en E niet houdbaar.

3. Latente trek modellen

In de algemene inleiding op het begrip *latente trek* bij Lord & Novick (1968, par. 16.1) staat o.m.: "Any theory of latent traits supposes that an individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called *traits*, quantitatively estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations (...). Much of psychological theory is based on a trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behaves as if he were in possession of an amount of each of a number of relevant traits and that he behaves as if these amounts substantially determined his behavior. (...) The psychologist who wishes to use an examinee's responses to a mental test to make inferences about his psychological traits must have some knowledge of how psychological traits determine or are related to examinee's responses."

De trek wordt *latent* genoemd omdat zij niet direct observeerbaar (verborgen) is; het is een algemeen en Platonisch begrip dat zich alleen indirect manifesteert in de concrete reactie op specifieke items. In de meeste modellen over de relatie tussen trekken en items postuleert men *locale onafhankelijkheid*: in de deelverzameling van personen die op elke relevante trek dezelfde waarde hebben is de kansverdeling van de antwoorden tussen items stochastisch onafhankelijk; zij wordt alleen veroorzaakt door toevallige storingen. Die deelverzameling is, door het latente

karakter van de trekken,uiteraard nooit in concreto te bepalen. In de verzameling van alle personen is er uiteraard juist een *positieve afhankelijkheid* tussen de items: weet ik dat iemand al enige items positief heeft beantwoord dan zal hij/zij wel een gunstige positie op de relevante latente trek(ken) hebben en verwacht ik dus ook relatief veel positieve antwoorden op de overige items.

De assumptie dat een test door slechts één latente trek wordt bepaald (*unidimensionaliteit*) is nu nogal ingrijpend: geen enkele andere eigenschap van de te meten personen mag nog systematische invloed op de antwoorden uitoefenen. Toch streeft men naar een dergelijke eerlijke en ongecontamineerde meting, o.m. omdat het onethisch is als een test naar b.v. geslacht of ras zou discrimineren, of omdat het resultaat op een test voor rekenvaardigheid niet door taalkennis of leessnelheid zou moeten worden beinvloed. Daarmee is alleen bedoeld dat personen van *gelijke vaardigheid* (ongeacht hun ras, sexe, taalkennis of leessnelheid) ook dezelfde score behoren te krijgen; het is dus best mogelijk dat die vaardigheid ongelijk verdeeld is naar ras, sexe, enz., maar de itemkeuze mag niet ten nadele van zo'n groep werken door een specifieke interactie (zie ook Van der Flier, 1980).

In de volgende paragrafen worden een aantal modellen besproken voor tests bestaande uit *dichotome* items (twee antwoordcategorieën, b.v. goed/fout of eens/oneens) die één latente trek meten en daarin *cumulatief* zijn verondersteld (de kans op een positief antwoord is voor elk item een niet-dalende functie van de waarde op de latente trek).

Het ligt inderdaad voor de hand dat personen die meer van de te meten eigenschap bezitten, een even grote of grotere kans op het goede antwoord hebben dan personen die de eigenschap in mindere mate hebben. Soms neemt de onderzoeker ook uitspraken "in omgekeerde richting" op, waarbij dan juist "oneens" het goede antwoord is; dit biedt enige waarborgen tegen een neiging van sommige ondervraagden om uit luiheid of braafheid in twijfelgevallen steeds bevestigend te antwoorden. Zo kwam in een schaal ter meting van *politiek zelfvertrouwen* de bewering

voor: "er stemmen zo veel mensen bij de verkiezingen dat mijn stem er niet toe doet". Pas na omkering (oneens=goed) stijgt de kans op een goed antwoord met het zelfvertrouwen. De bewering "stemmen is de enige manier waarop mensen zoals ik iets te zeggen hebben over hoe de regering de zaken behartigt" bleek *niet cumulatief* te zijn (Mokken, 1971,p.234): personen zonder enig politiek zelfvertrouwen zeggen "oneens" omdat ze denken dat stemmen ook niet helpt, en personen met veel politiek zelfvertrouwen zeggen "oneens" omdat ze ook heil zien in acties, demonstraties of ledenvergaderingen van politieke partijen. Daartussen bevinden zich mensen die zich overwegend in de inhoud van de uitspraak kunnen vinden.

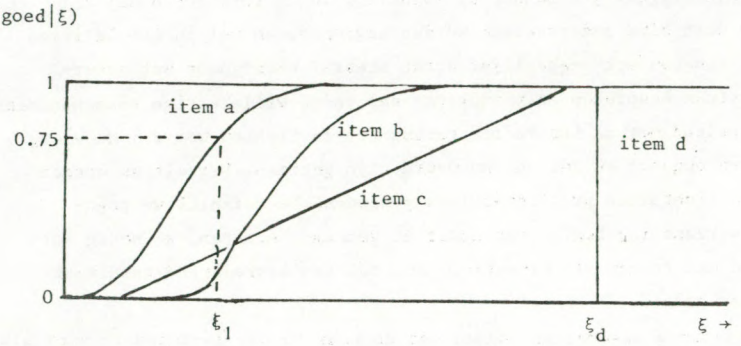
Het blijkt nog heel wat vakmanschap te vereisen de uitspraken van een vragenlijst zonder uitzondering zo te formuleren dat zij alle door elke ondervraagde worden begrepen, en wel in cumulatieve zin. Een concept-vragenlijst wordt meestal eerst door het onderzoeksteam besproken en vervolgens aan enige willekeurige respondenten voorgelegd met de instructie tevens toe te lichten hoe zij de vraag hebben opgevat en tot hun antwoord zijn gekomen. Vervolgens worden enige tientallen proefinterviews gehouden. De definitieve gegevensverzameling begint pas nadat de genoemde stappen, zo nodig herhaald met aangepaste formuleringen, tot een bevredigend resultaat hebben geleid.

Uit onze assumpties volgt, dat de kans op het ja-antwoord voor elk item i een niet-dalende functie $f_i(\xi)$ van de latente eigenschap is. Men noemt $f_i(\xi)$ de *item-karakteristieke curve* (Engels *trace line*). De diverse in de volgende paragrafen te bespreken latente trek modellen onderscheiden zich onderling door de assumpties die over $f_i(\xi)$ worden gemaakt. Deze zijn niet direct empirisch verifiëerbaar omdat ξ latent is, maar zij hebben wel observeerbare consequenties, die toelaten iets over hun aannemelijkheid te zeggen. In figuur 1 zijn enkele voorbeelden van karakteristieke curven geschetst die later aan de orde komen.

Binnen een specifiek latente trek model probeert men nu vragen te beantwoorden zoals:

1) Wat zijn de beste schattingen voor de parameters van de items en van de personen;
2) vormt een gegeven item verzameling een (goede)schaal in de zin van het model;
3) kan ik uit een gegeven itemverzameling door stapsgewijs opbouwen of stapsgewijs weglaten een goede schaal verkrijgen.

Recente korte inleidingen in de latente trek modellen komen voor in de dissertaties van Van der Flier (1980) en Van den Wollenberg (1979). In de paragrafen 4 t/m 6 worden speciale modellen besproken: evaluatie van hun voor- en nadelen volgt in paragraaf 7.



Figuur 1. Voorbeelden van item-karakteristieke curven. Elke persoon met waarde ξ_1 op de latente eigenschap heeft een kans van 0.75 om item a juist te beantwoorden. Voor items b, c, d is zijn/haar kans 0.10 resp. 0.20 resp. 0.Voor item c stijgt de kans op goed rechtlijnig met ξ, althans binnen de grenzen waarvoor deze kans tussen 0 en 1 ligt. Item d is een Guttman-item: het wordt met zekerheid fout beantwoord door elke respondent met een ξ-waarde kleiner dan ξ_d, en met zekerheid goed zodra $\xi \geq \xi_d$.

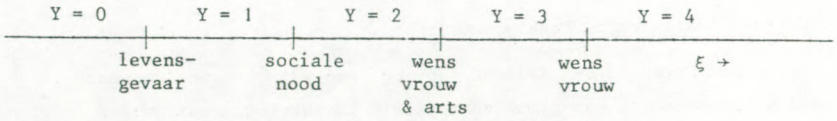
4. Guttman-scalogram en Mokken-model

Guttman begon tussen 1940 en 1950 met een vorm van meting waarbij items en personen op één gemeenschappelijk latent continuum worden *geordend*. Ik wil zijn procedure grofweg vergelijken met een sorteermachine waarbij een heterogene collectie aardappels over een serie zeven met toenemende maaswijdte wordt geleid: de kleinsten vallen al door de eerste zeef, de iets grotere door de tweede, enzovoorts, totdat de grootste aardappels zelfs op de laatste zeef blijven liggen. Maak van een zeef een item, van een aardappel een persoon, van blijven liggen bevestigend beantwoorden, en dan zijn er nog maar twee kleine veranderingen nodig om het Guttman-scalogram te verkrijgen: elke persoon krijgt elk item voorgelegd (een eenmaal gevallen aardappel komt niet op de grovere zeef) en de volgorde van de items is meestal niet die van opklimmende moeilijkheid.

Als een gestyleerd voorbeeld kan de volgende schaal gelden:

1. Als de vrouw in levensgevaar is vind ik abortus toelaatbaar
2. Bij ernstige sociale nood vind ik abortus toelaatbaar
3. Als de vrouw en de huisarts het beiden wensen vind ik abortus toelaatbaar
4. Als de vrouw het wenst vind ik abortus toelaatbaar

Wie op een vraag al met "nee" heeft geantwoord zal dat vrijwel zeker ook op de volgende vragen doen. Naar de antwoorden op alle vier vragen worden de personen dus in vijf geordende kategorieën gesorteerd: het aantal ja-antwoorden Y is een maat voor de tolerantie t.a.v. abortus. Als die latente eigenschap weer met de letter ξ wordt aangeduid, wordt de ξ-as dus in vijf intervallen verdeeld (fig. 2); uit de ordening tussen elke persoon en elk item (ja of nee) volgt tevens de ordening van de items (een item met veel ja-antwoorden heeft een lage ξ-waarde), en ook de ordening van de personen (met knopen als zij dezelfde score Y hebben).



Figuur 2. Ordening van respondenten door vier Guttman-items

In ons voorbeeld is het nogal inconsequent wanneer iemand het wel met items 1 en 3 maar niet met item 2 en 4 eens is. Geven we "eens" met 1 en "oneens" met 0 aan, dan heeft zo iemand het scorepatroon 1010. Van alle 2^4 = 16 theoretische mogelijke patronen mogen in het Guttman-scalogram voor vier items alleen 1111, 1110, 1100, 1000 en 0000 optreden: de elf overige patronen weerspreken het cumulatieve karakter van de schaal.

In de meeste schalen ter meting van eigenschappen blijken dergelijke "verboden patronen" echter af en toe op te treden. Dit gaf aanleiding tot de constructie van *reproduceerbaarheidscoëfficiënten*: bij een toelaatbaar scorepatroon is elk afzonderlijk antwoord reproduceerbaar uit de totale score Y, en bij een verboden patroon zoals 1010 wordt uit de informatie Y = 2 ten onrechte het patroon 1100 voorspeld. Hoewel eigenlijk elke groep personen met één of meer verboden patronen het strikt deterministische Guttman-model weerspreekt, ontwikkelde men in de jaren vijftig talrijke vuistregels waarbij een beperkt aantal schendingen nog toelaatbaar werd geacht. Een overzicht vindt men aan bij Mokken (1971, hoofdstuk 2).

Deze auteur ontwikkelde, uit onvrede met het ad hoc karakter van de Rep-coëfficiënten en de vuistregels, een probabilistisch model voor zo'n Guttman-schaal met een beperkt aantal schendingen. De kans op het ja-antwoord is bij Mokken een willekeurige stijgende functie van de latente parameter ξ, maar de item-karakteristieke curven mogen elkaar niet snijden. Deze eis (die in fig.1 voor items b en c geschonden wordt) zorgt er voor dat de ordening van de items naar moeilijkheid voor elke subgroep van personen dezelfde is.

De kans dat iemand met latente waarde ξ_1 item b juist beantwoordt is in fig. 1 gelijk aan $f_b(\xi_1) = 0.10$. In een populatie van te meten personen zal ξ niet steeds gelijk aan ξ_1 zijn, maar een kansdichtheid hebben die we met $g(\xi)$ aanduiden. Geven we met $Y_b = 0$ resp. 1 aan dat een willekeurig persoon uit deze populatie item b negatief resp. positief beantwoordt, dan volgt de marginale kans door integratie:

$$(1) \qquad \pi_b \stackrel{\text{def}}{=} P(Y_b = 1) = \int_{-\infty}^{\infty} f_b(\xi)\, g(\xi)\, d\xi.$$

Analoog worden Y_a en π_a gedefinieerd; omdat $f_a(\xi) > f_b(\xi)$ voor elke ξ zal $\pi_a > \pi_b$ zijn.

De kans dat iemand met latente waarde ξ_1 het item a onjuist en het moeilijker item b juist beantwoordt is in fig. 1 wegens de locale onafhankelijkheid gelijk aan

$$(2) \qquad \{1 - f_a(\xi_1)\}\, f_b(\xi_1) = 0.25 * 0.10 = 0.025.$$

In de gehele populatie is de kans op een antwoordpaar (0,1) dan, door integratie van (2),

$$(3) \qquad \pi_{ab}(0,1) \stackrel{\text{def}}{=} P(Y_a = 0, Y_b = 1) = \int_{-\infty}^{\infty} \{1 - f_a(\xi)\}\, f_b(\xi)\, g(\xi)\, d\xi.$$

Dit is dus de kans dat een persoon zich bij deze twee items niet volgens het Guttman-model gedraagt. Een kwaliteitsindex is dan de fractie modelconforme personen $M = 1 - \pi_{ab}(0,1)$.

Als de items in de gehele populatie statistisch onafhankelijk waren zou de kans op het antwoordpaar $Y_b = 1$, $Y_a = 0$ gelijk aan $\pi_b(1 - \pi_a)$ zijn. Mokken normeert nu, in navolging van Loevinger, de kwaliteitsindex tot $H_{ab} = \{M - E_0(M)\} / \{M_{max} - E_0(M)\}$:

$$(4) \qquad H_{ab} = \frac{\{1 - \pi_{ab}(0,1)\} - \{1 - \pi_b(1 - \pi_a)\}}{1 - \{1 - \pi_b(1 - \pi_a)\}} \qquad (\pi_a > \pi_b).$$

Eenvoudig rekenwerk in de 2x2 tabel van de vier mogelijke antwoordpatronen voor het itempaar leidt tot

$$(5) \qquad H_{ab} = \frac{\pi_{ab}(1,1) - \pi_a \pi_b}{\pi_b(1 - \pi_a)} = \frac{\phi_{ab}}{\phi_{ab(max)}} \qquad (\pi_a > \pi_b) ,$$

waar ϕ de product-momentcorrelatie tussen Y_a en Y_b is en $\phi_{ab(max)}$ de maximaal mogelijke correlatie bij de gegeven marginale kansen. Voor onafhankelijke items is $H_{ab} = 0$; alleen voor itemparen met een lege *foutencel* bereikt H_{ab} de maximale waarde van 1.

Wanneer een aselecte steekproef van n personen uit de genoemde populatie is getrokken, kan men de fractie personen daarin met item a goed, item b goed, en beide goed invullen in (5). Mokken berekent de asymptotische verdeling ($n \to \infty$) van deze schatter voor H_{ab} en baseert daarop procedures om de tweede en derde vraag aan het slot van paragraaf 3 te beantwoorden, en nog enige verwante problemen op te lossen. Merk op dat de eerste vraag hier niet terzake is, omdat hij een verdelingsvrij en geen parametrisch model hanteert, zowel voor de verdeling van de latente eigenschap als voor de vorm van de itemkarakteristieke curven. In een volgens de Mokken-procedures toelaatbare schaal zijn evenals bij Guttman personen en items geordend naar ξ. Het totaal aantal positieve antwoorden Y is een *ordinale* schaal ter meting van ξ; omdat het model niet-parametrisch is, laat het niet toe de eigenlijke ξ-waarden van personen te schatten.

5. Rasch-model

Om de zojuist genoemde metrische afbeelding van Y naar ξ mogelijk te maken, eiste Rasch (1960) voor dichotome cumulatieve items behalve unidimensionaliteit (slechts één latente trek) en locale onafhankelijkheid eveneens dat de totaalscore Y afdoende (sufficient) is voor ξ. Dan volgt (zie voor een bewijs Fischer (1974) pag. 195-199):

Zodra de ξ-as zodanig is getransformeerd dat de karakteristieke curve voor één item een logistische curve is, impliceert het afdoende zijn van Y voor ξ dat ook alle andere items logistische karakteristieke curven hebben, die door verschuiving uit de eerste ontstaan:

(6) $P(\text{persoon v maakt item i goed}) = P(Y_i = 1 | \xi_v, \sigma_i) =$
$= e^{\xi_v - \sigma_i}/(1 + e^{\xi_v - \sigma_i}) = \theta_v \varepsilon_i/(1 + \theta_v \varepsilon_i).$

Hier is σ_i de moeilijkheid van item i, dit is de ξ-waarde van een persoon die kans $\frac{1}{2}$ heeft om het item goed te doen. In diverse berekeningen werkt men sneller met de nieuwe parameters $\theta_v = \exp(\xi_v)$ en $\varepsilon_i = \exp(-\sigma_i)$, die we samenvoegen tot een vector ter lengte n (aantal personen) resp. k (aantal items). We merken nog op dat het model slechts toelaat de ε_i op een ratioschaal te schatten; men normeert bij conventie met $\prod_{i=1}^{k} \varepsilon_i = 1$ of door één vast item de parameterwaarde 1 toe te kennen. Zo kiest men één representant uit de klasse van equivalente parametriseringen, die uit elkaar ontstaan door alle ε_i met een constante te vermenigvuldigen en gelijktijdig alle θ_v door dezelfde constante te delen.

Alle antwoorden van n personen op k items vatten we samen tot een n x k matrix A, waar van het element a_{vi} de waarde 1 resp. 0 aanneemt als persoon v item i goed resp. fout maakt. Dan geldt dus

(7) $$P(A|\theta, \varepsilon) = \prod_{v=i}^{n} \prod_{i=1}^{k} \frac{(\theta_v \varepsilon_i)^{a_{vi}}}{1 + \theta_v \varepsilon_i} =$$

$$= \frac{\prod_v \theta_v^{a_{vo}} \prod_i \varepsilon_i^{a_{oi}}}{\prod_v \prod_i (1 + \theta_v \varepsilon_i)} .$$

Hier blijkt dat totaalscore $a_{vo} \overset{\text{def}}{=} \sum_{i=1}^{k} a_{vi}$, eerder met Y aangeduid, afdoende is voor θ_v (en voor ξ_v), terwijl $a_{oi} \overset{\text{def}}{=} \sum_{v=1}^{n} a_{vi}$ afdoende is voor ε_i (en voor σ_i). Elke matrix met dezelfde randtotalen a_{vo} (v=1,2,...,n) en a_{oi} (i=1,2,...,k) is blijkens (7) even waarschijnlijk.

De uitdrukking (7) is, na invulling van de geobserveerde persoons- en item totalen, een functie van de n+k onbekende parameters θ en ε ; voor de meest aannemelijke schattingen moet deze functie worden gemaximaliseerd. Het rekentechnisch probleem is iteratief oplosbaar, maar er is een fundamenteel statistisch probleem: de aldus verkregen item-parameterschattingen zijn niet zuiver en niet consistent: als het aantal proefpersonen onbeperkt toeneemt, brengt elke proefpersoon een nieuwe parameter θ_v mee, en deze storingsparameters beletten een bevredigende schatting van de itemparameters.

Een aantal Amerikaanse auteurs (zie b.v. Wright & Stone, 1979) lossen dit adhoc op door aan de meest aannemelijke schatter voor ε een correctiefactor (k-1)/k toe te voegen, die het grootste deel van de euvels wegneemt. In Europa kiest men meestal voor de door E.B.Andersen voorgestelde *conditionele* schatters, die $P(a_{oi}|a_{vo}, \varepsilon)$ maximaliseren. Deze uitdrukking blijkt inderdaad niet meer van θ af te hangen en wordt iteratief naar de k itemparameters gemaximaliseerd. Het daarbij gebruikte algorithme is o.m. door Gustafsson (1977, 1979) zo veel sneller en nauwkeuriger gemaakt dat het ook voor grote itemverzamelingen bevredigend werkt (zie ook Wainer, Morgan & Gustafsson, 1980).

Heeft men eenmaal goede schattingen voor ε verkregen dan worden de persoonsparameters doorgaans geschat door deze in (7) in te vullen; wegens de afdoendheid van a_{vo} krijgen telkens alle personen met dezelfde totaalscore ook dezelfde geschatte persoonsparameter. Wainer & Wright (1980) hebben interessante verbeteringen in de schatting van persoonsparameters voorgesteld.

Binnen het Rasch-model geldt voor elk paar items en voor elke subgroep van personen met dezelfde vaardigheidsparameter θ dat

$$(8) \qquad \frac{P(a_{vi}=1,\ a_{vj}=0|\theta)}{P(a_{vi}=0,\ a_{vj}=1|\theta)} = \frac{\{\theta\varepsilon_i/(1+\theta\varepsilon_i)\}.\{1/(1+\theta\varepsilon_j)\}}{\{1/(1+\theta\varepsilon_i).\{\theta\varepsilon_j/(1+\theta\varepsilon_j)\}} = \frac{\varepsilon_i}{\varepsilon_j} \quad .$$

Dit quotiënt hangt niet van θ af, en men kan de verhouding $\varepsilon_i/\varepsilon_j$ dus in principe volstrekt los van de gebruikte groep personen bepalen. Na spiegeling van personen en items is de verhouding van twee persoonsparameters eveneens onafhankelijk van de gebruikte items.

Deze zgn. *specifieke objectiviteit*, het ontbreken van interacties tussen personen en items, is vanuit meettheoretisch oogpunt een uiterst wenselijke eigenschap, zoals o.m. benadrukt wordt door Rasch (1960) en Fischer (1974). Deze eigenschap wordt in allerlei toepassingen uitgebuit, waarvan we er twee zullen noemen; zie verder Wright & Stone (1979), waarvan de inleiding het ideaal van specifieke objectiviteit welsprekend verdedigt.

Wanneer twee tests slechts een enkel item gemeenschappelijk hebben maar hun items wel aan het Rasch-model voldoen, kan men de ene test aan de ene groep respondenten voorzetten en de andere test aan de andere groep (zeg de parallelklas); wegens de specifieke objectiviteit kan men achteraf de scores van beide groepen tot vaardigheidsschattingen op een gemeenschappelijke θ-as herleiden. Daarmee wordt de eeuwige strijdvraag beslecht of de ene proefwerkversie nu moeilijker was dan wel de ene klas slechter voorbereid.

Een toepassing die een stap verder gaat wordt in het Engels met *tailored testing* aangeduid. Stel dat een onderzoeker over honderd items beschikt die volgens het Rasch-model eenzelfde eigenschap meten en waarvan hij de - uiteenlopende - moeilijkheidsgraad aan eerdere proefgroepen heeft geschat. Een nieuw te meten persoon krijgt nu eerst bijvoorbeeld tien items van doorsnee-moeilijkheid voorgeschoteld. Maakt hij er daarvan veel goed, dan wordt de volgende serie van tien uit de lastiger items van de voorraad gekozen, zijn er veel fout dan gaat de onderzoeker op makkelijker items over,en behaalt hij een middenscore dan wordt de test met items van matig niveau voortgezet. Desgewenst wordt na twintig items op grond van de tot op dat moment behaalde score nog een derde set van

tien items van passend niveau toegevoegd. De onderzochte persoon krijgt dus een test "naar maat" toegediend, vandaar de associatie met de maatkleermaker. Via de eigenschappen van het Rasch-model is het toch mogelijk achteraf de scores van alle personen op één schaal te herleiden, waarbij het verschil in moeilijkheid van de voorgelegde vragen verdisconteerd wordt. Het voordeel voor de proefpersoon is dat hij/zij niet verveeld wordt tijdens de test met te makkelijke vragen en evenmin gefrustreerd met te lastige opgaven. Het voordeel voor de testleider is dat juist de vragen van aangepast niveau een nauwkeurige schatting van de persoonsparameter mogelijk maken. De ingeving dat deze vragen een hoger "informatiegehalte" hebben dan te lastige of te eenvoudige items, blijkt in de hier niet gepresenteerde theorie over de in items bevatte informatie inderdaad juist te zijn.

6. Andere modellen en uitbreidingen

Onafhankelijk van Rasch bestudeerde Birnbaum (1968) de volgende klasse van itemkarakteristieke curven:

$$(9) \qquad f_i(\xi) = c_i + (1 - c_i) \exp [a_i(\xi - b_i)]/\{1 + \exp [a_i(\xi - b_i)]\},$$

waarin elk item gekenmerkt is door een giskans c_i, een moeilijkheid b_i en een discriminatieparameter a_i. Wanneer voor elk item $c_i = 0$ en $a_i = a$ geldt, is (9) equivalent met het Rasch-model; wanneer alleen $c_i = 0$ geldt blijkt de gewogen som $\sum_{i=1}^{k} a_i Y_i$ een afdoende grootheid voor ξ. Het uitgebreidere model is in veel toepassingen realistischer maar de specifieke objectiviteit gaat verloren. De problemen bij het schatten van alle itemparameters worden door sommige auteurs nagenoeg onoverkomelijk genoemd (Fischer, 1974, pag. 277-280), maar door anderen niet (Van der Flier, 1980).

Omdat de logistische curve met een geschikte schaalfactor nagenoeg samenvalt met de normale (zie b.v. Molenaar, 1974) is (9) vrijwel gelijkwaardig met

$$(10) \qquad f_i(\xi) = c_i + (1 - c_i)(2\pi d_i)^{-\frac{1}{2}} \int_{-\infty}^{\xi} \exp[-\tfrac{1}{2}(u - a_i)^2/d_i^2]du,$$

een model dat veelvuldig door Lord is bestudeerd (zie Lord & Novick, 1968).

Bij Lazarsfeld is de itemkarakteristieke curve een rechte lijn op het interval waar deze tussen 0 en 1 loopt (zoals item c in fig. 1). Dit model wordt weinig meer gebruikt. Van meer belang is het *latente klasse model* waarbij de latente parameter slechts een beperkt aantal waarden aanneemt zie Lazarsfeld & Henry (1968) of Fisher (1974, par. 10.2). Omdat de items binnen één klasse dan stochastisch onafhankelijk zijn, biedt dit mogelijkheden geobserveerde afhankelijkheid "weg te verklaren" en tot clusters van gelijkwaardige personen te komen (Mårdberg, 1974 ,1975).

Tenslotte worden nog drie interessante varianten van het Rasch-model genoemd. In het *lineair logistisch testmodel* (Fisher, 1974, hoofdstuk 17) neemt men aan dat de itemparameters σ_i een lineaire combinatie met bekende coëfficiënten zijn van een kleiner aantal basisparameters η_j:

$$(11) \qquad \sigma_i = \sum_{j=1}^{m} q_{ij}\eta_j \quad (i=1,2,\ldots,k).$$

Dit model blijkt bijvoorbeeld bij een klasse eenvoudige problemen uit de mechanica of bij aanvulling van getallenrijen mogelijkheden te bieden de moeilijkheid van elk item te verklaren uit een beperkt aantal facetten waarop de items onderling verschillen.

In het Rasch-model voor *meer dan twee antwoord kategorieën* (Fischer, 1974, hoofdstuk 20,21) laat men de dichotomie-eis vallen door extra kategorie-parameters in te voeren, met behoud van de specifieke objectiviteit. De plaatsruimte ontbreekt hier om dit nader uit te werken. Dat geldt eveneens voor de modellen van Spada, Scheiblechner, en Kempf waarin de locale onafhankelijkheid wordt verzacht tot een leermodel waarin het resultaat op voorafgaande items van invloed mag zijn op de kans een volgend item juist te beantwoorden (Fischer, 1974, pag. 18.3).

7. Evaluatie: conflicterende idealen

Met dit overzicht van diverse schaalmodellen, waaraan er makkelijk nog enkele kunnen worden toegevoegd, is nog niet gezegd welke modellen nu de voorkeur verdienen. Het is zelfs niet direct duidelijk wanneer het ene schaalmodel "beter" moet worden geacht dan het andere. Er zijn naar mijn mening verschillende,soms onderling strijdige, kwaliteits-criteria, en men kan ook uiteenlopende redenen hebben om schalen voor sociaalwetenschappelijk onderzoek te construeren, te gebruiken of te evalueren.

Deze afsluitende paragraaf kan daarom hoogstens de diverse modellen globaal toetsen aan een voorlopige verzameling van criteria. De uiteindelijke keuze blijft aan de gebruiker, evenals bij de vergelijkende onderzoeken van de Consumentenbond. Voorzover U hier een definitief eindoordeel had verwacht kan ik U misschien troosten: het loont de moeite de modelkeuze aan de hand van de criteria te overwegen,maar anderzijds zullen de uitkomsten verkregen uit diverse schaalmodellen dikwijls niet dramatisch van elkaar afwijken (zie b.v. de Vos, 1980 of Molenaar, 1983).

Mijn verzameling criteria, die een voorlopig karakter heeft, kan globaal worden gesplitst in twee rubrieken, die ik assumpties en prestaties heb genoemd.

ASSUMPTIES

A1. Welke assumpties bevat het schaalmodel?
A2 Is het wenselijk dat de data daar aan voldoen?
A3. Is het plausibel dat de data daar aan voldoen?
A4. Is empirisch onderzoek mogelijk of de data daar aan voldoen?
A5. Zijn de gevolgen voor de gebruiker bekend van schending van een assumptie?

PRESTATIES

P1. Is het duidelijk hoe de relevante itemkenmerken worden bepaald?
P2. Is de nauwkeurigheid van die bepaling te achterhalen?

P3. Is het duidelijk hoe de score van een persoon wordt bepaald?
P4. Is de nauwkeurigheid van die bepaling te achterhalen?
P5. Is voorspelbaar welke items een persoon met een gegeven score positief zal beantwoorden?
P6. Krijgen de personen een redelijke verdeling van scores?
P7. Is er een methode om de schaal te verbeteren door eliminatie van items (of van personen)?
P8. Is er een methode om de schaal te verbeteren door toevoeging van items?
P9. Is er een bescherming tegen kanskapitalisatie?
P10. In hoeverre zijn de resultaten generaliseerbaar naar andere items en naar andere personen?

Voorzover een antwoord op deze vragen los staat van de aard van het onderzoek zal ik in tabel 1 globaal mijn antwoord trachten te geven. Dit voorbehoud betreft niet alleen de specifieke eigenschappen van bepaalde datamatrices, maar ook de omstandigheid dat de statistische generalisering ("wat gebeurt er als ik het overnieuw doe") soms naar andere items, soms naar andere personen, en soms naar een replicatie met dezelfde personen en items gericht is.

De tabel is een subjectieve evaluatie, waarin door de beknoptheid ook veel verloren gaat. Zo kan "ja?" zowel betekenen: " Het kan maar het is lastig" als ook "Ik twijfel of het wel echt kan". Persoonlijk is mij geen Birnbaum- of Mokkenmodel voor polychotome items bekend, maar het zou misschien op te stellen zijn. Het assumptie-onderzoek bij de laatste drie kolommen is voor sommige assumpties formeel uitgewerkt en voor andere alleen via een inspectie met het blote oog mogelijk. Het Birnbaum-model leidt, zoals op pag. 15 al opgemerkt, soms tot schattingsproblemen, die mogelijk met singulariteiten van de aannemelijkheidsfunctie te maken hebben (Wainer, pers. communicatie). Het Guttman-model is in zijn strikte en deterministische vorm perfect, en perfect toetsbaar, maar de pogingen om schendingen via Rep-coëfficiënten weg te praten stellen m.i. niet veel voor.

Tabel 1. Enkele assumpties en kwaliteitsvragen voor schaalmodellen.

					latente trek modellen			
	Thurstone Chaves	Likert	Factor anal.	Kl. test th.	Guttman	Mokken	Birnbaum	Rasch
DICHOTOOM	ja	ja?	ja?	ja?	ja	ja	ja	ja
POLYCH. GEORDEND	nee	ja	ja	ja	nee	nee	nee	ja
UNIDIM.	ja	ja	ja	ja?	ja	ja	ja	ja
LOC. ONAFH.	nvt	nvt	ja	ja	ja	ja	ja	ja
MONOTOON	nee	ja	ja	ja	ja	ja	ja	ja
ΣX_i AFDOENDE	nvt	nee	nee	ja	orde	orde	nee	ja
MEETMODEL	nee	nee	nee?	nee?	ja	ja	ja	ja
ONDZ. ASSUMPTIES	vaag	vaag	vaag	vaag	ja?	ja?	nee?	ja?
ELIMINATIE	ja	ja	ja	ja	ja	ja?	nee?	ja??
TOEVOEGING	kan	kan	kan	kan	kan	ja!	kan	kan
ITEM & NAUWK.	ja?	ja?	ja?	ja?	orde	orde	ja(?)	ja
SCORE & NAUWK.	ja?	nee	ja?	ja?	orde	orde	ja(?)	ja
PATROON VOORSP.	ja?	nee?	nee?	nee?	ja!	ja?	prob	prob
SPECIF. OBJECTIEF	nee	nee	nee	nee	nee	orde	nee	ja
KRITISCH	6	3	4	6	10	8	8	9
KRACHTIG	5	5	6	7	7	7	8	9

Afkortingen: nvt = niet van toepassing; prob = probabilistisch. Een hoog cijfer voor "kritisch" betekent: moeilijk aan te voldoen; een hoog cijfer voor "krachtig" betekent: veel dingen zijn mogelijk.

Er is veel voor te zeggen bij de huidige stand van zaken, om zich niet vast te leggen op één schaalmodel, en de keuze mede door het gebruiksdoel en de omstandigheden te laten bepalen. Wel heb ik, vanuit een groter vertrouwen in de assumpties, een voorkeur voor latente trek modellen, en daarbinnen voor het Rasch-model zodra het mogelijk lijkt een valide schaal te vormen die ongeveer aan de stringente assumpties voldoet. Ongetwijfeld is Birnbaum realistischer, met de gisparameter en het verschil in helling, zie formule (9). Er zijn echter enige aanwijzingen dat het Rasch-model redelijk werkt ook als de data aan Birnbaum voldoen, zie Lumsden (1976), Hambleton & Traub (1971) en Forsyth et al.(1981).

Alle latente trek modellen sluiten aan bij de gedachte dat de te meten eigenschap een continue kansverdeling heeft en los van de gebruikte items bestaat. Daarmee wordt een deel van de in paragraaf 2 genoemde bezwaren tegen de klassieke testtheorie ondervangen. De kwaliteitseisen van cumulativiteit, unidimensionaliteit en locale onafhankelijkheid zijn weliswaar alleen indirect te toetsen, omdat de latente trek even Platonisch is als de ware score, maar zij sluiten beter aan bij de grondslagen van het meten dan de in de klassieke testtheorie gemaakte assumpties over varianties en correlaties. Het wordt uiteraard wel steeds lastiger om aan de bij zo'n meting gestelde eisen te voldoen.

Er is geen sprake van een ordening, waarbij alle besproken modellen van laag tot hoog uit elkaar ontstaan door extra eisen toe te voegen. Veeleer verschillen de modellen wezenlijk van elkaar in de zin dat telkens andere eigenschappen van de schaal centraal worden gesteld in het kwaliteitsoordeel. De gebruiker dient zelf na te gaan welk ideaalbeeld voor elke specifieke toepassing de meeste voordelen lijkt te bieden.

Het ideaal van de klassieke testtheorie is een kleine meetfout, dus een hoge betrouwbaarheid en hoge item-totaalcorrelaties. Uit een gegeven itemverzameling worden daartoe bij voorkeur items met een populariteit van omstreeks ½ geselecteerd: items met een lage of hoge populariteit zullen immers door het verdelings-effect minder hoog correleren. Personen in de (0,1) cel èn in de (1,0)cel van een item paar belemmeren allen een optimale correlatie, en zijn dus ongewenst.

Het ideaal van Mokken is een schaal die zo weinig mogelijk afwijkt van een perfect Guttman-scalogram: uit de totaalscore Y van een persoon moet, afgezien van een beperkt aantal fouten, te voorspellen zijn dat de Y gemakkelijkste items goed en de k-Y overige items onjuist zijn beantwoord. De H-coëfficiënt meet per itempaar of het aantal personen in de "foutencel" (0,1) duidelijk kleiner is dan bij items zonder enig verband mag worden verwacht. Uit een gegeven itemverzameling worden items met tamelijk gespreide populariteiten in de schaal geselecteerd, die een ordinaal karakter krijgt.

Het ideaal van Rasch is de specifiek objectieve meting, zonder interactie van personen en items. Uit de eis dat de totale score afdoende is voor de latente parameter volgt dat de items door evenwijdige logistische curven worden gekenmerkt. Uit een gegeven itemverzameling worden items geselecteerd met tamelijk gespreide populariteiten maar met een ongeveer even sterke relatie tot de te meten eigenschap.

Wanneer het mogelijk is een aantal goed discriminerende Rasch-items te vinden, die het bedoelde begrip betrouwbaar en valide meten, heeft het gebruik van deze schaal een aantal theoretische en practische voordelen die al in paragraaf 5 zijn geschetst. Vooral bij attitudemeting en bij meting van vrij algemene vaardigheden moet men er echter rekening mee houden dat de eisen van unidimensionaliteit, onafhankelijkheid en afdoendheid erg streng zijn. Selectie van deelverzamelingen van items die aan deze eisen voldoen is niet alleen moeizaam (Van den Wollenberg, 1979; Molenaar, 1980; Formann, 1981), maar kan er ook snel toe leiden dat een schaal over blijft waarin maar één aspect van het bedoelde begrip wordt gemeten. Deze bandbreedte-vernauwing is overigens een gevaar bij elke eendimensionele

schaalconstructie die te veel op de wiskundige betrouwbaarheid en te weinig op de inhoud let: een stel items waarin nagenoeg dezelfde vraag in telkens iets andere bewoordingen wordt herhaald is doorgaans erg betrouwbaar maar weinig valide!

Op de weg van fiat-meting via factoranalyse en klassieke testtheorie naar de latente trek modellen koopt de gebruiker steeds meer kennis met steeds meer assumpties. Er valt, in termen van Hofstee (1980), steeds meer te wedden. In toenemende mate zijn de modellen falsifiëerbaar; als de data steeds meer beproevingen doorstaan, heeft de resulterende meting ook steeds mooiere eigenschappen, met als voorlopig eindpunt de item-onafhankelijke ratio-schaal voor persoonsparameters (en vice versa) binnen het Rasch-model.

De falsifiëring van een latente trek model is overigens niet zonder problemen, omdat de assumpties over latente parameters alleen indirect via hun observeerbare consequenties kunnen worden getoetst. De dreigende onvolledigheid van zo'n toetsing is een gevaar, zie o.m. Mokken (1971); Van den Wollenberg (1979, 1981abc); Molenaar (1980) en de discussie tussen Jansen (1981ab) en Molenaar (1981). Dat de eigenschappen niet kunnen worden geextrapoleerd naar latente waarden die niet of nauwelijks bij de geobserveerde personen optreden is niet zo erg, zoals Jansen (1981a) voor de cumulativiteitsassumptie laat zien. Dat bijvoorbeeld de zeer gangbare toets van Andersen voor het Rasch-model tegen diverse alternatieven niet onderscheidend blijkt te zijn is ernstiger (Van den Wollenberg, 1979).

Het ligt voor de hand dat alle beschouwde schaalmethoden tot steeds slechtere resultaten zullen leiden naarmate de te meten groep personen een geringere spreiding heeft in de latente eigenschap c.q. de ware score. In de limiet treedt er nog maar één ware waarde op en is alle geobserveerde spreiding aan meetfout toe te schrijven, wegens de locale onafhankelijkheid. Terwijl de klassieke testtheorie en het Mokken-model hiertegen openlijk waarschuwen, via lage correlaties c.q. lage H-waarden, blijft een Rasch-schaal een Rasch-schaal: zo kwam Wood (1978) tot de conclusie dat toevalsgetallen aan het Rasch-model voldoen.

Onder formule (8) is niet toevallig neergeschreven dat de verhouding tussen de item-parameters *in principe* los van de gebruikte groep personen kan worden bepaald. Als de te meten personen in overgrote meerderheid een latente waarde ver onder de itemwaarden hebben,krijgen zij vrijwel allen de (0,0) score op elk itempaar; eveneens bij "ver boven" en (1,1). De *nauwkeurigheid* van de schatting berust dan ook op het optreden van redelijke aantallen (0,1) *en* (1,0) antwoorden, al is de *verwachte waarde* bij elke personengroep dezelfde. Het is bizar dat een perfect Guttman-scalogram zich niet volgens Rasch laat analyseren, omdat personen of items met een nulscore of een perfecte score voorafgaand aan een Rasch-analyse worden verwijderd.

De keuze voor een schaalmethode moet volgens mij in sterke mate worden bepaald door de aard van de te meten eigenschap en van de te meten groep personen, en door het doel waarvoor de onderzoeker zijn / haar metingen wil gebruiken. Bij elke methode zijn er onbevredigende aspecten en onopgeloste problemen. Toch bieden zij met elkaar een aantal belangwekkende mogelijkheden om tot meer inzicht en beter meten in de sociale wetenschappen te geraken.

Literatuur

Allen, M.J. & Yen, W.M., (1979) Introduction to Measurement Theory. Monterey: Brooks/Cole.

Birnbaum, A. (1968), zie hoofdstuk 17 t/m 20 van Lord & Novick (1968).

Fischer, G.H. (1974) Einführung in die Theorie Psychologischer Tests. Bern: Huber.

Flier, H. van der (1980) Vergelijkbaarheid van individuele testprestaties (dissertatie V.U. Amsterdam).Lisse: Swets & Zeitlinger.

Formann, A.K. (1981) Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch, Zeitschrift für experimentelle und angewandte Psychologie, Band XXVIII, Heft, 2,S.

Forsyth, R., Saisangjan, U. & Gilmer, R. (1981) Some empirical results related to the robustness of the Rasch-model, Applied Psychological Measurement, 5, 175-186.

Gustafsson, J.E. (1977) The Rasch model for dichotomous items: Theory applications and a computer program. Reports from the Institute of Education, University of Göteborg, no. 63.

Gustafsson, J.E. (1979) PML: A computer program for conditional estimation and testing in the Rasch-model for dichotomous items, Reports from the Institute of Education, University of Göteborg, no. 85.

Gustafsson, J.E. (1980) Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

Hambleton, R.K. & Traub, R.E. (1971) Information curves and efficiency of three logistic test models, British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hofstee, W.K.B., (1980) De empirische discussie, Meppel: Boom.

Jansen, P.G.W. (1981a) Spezifisch objective Messung im Falle monotoner Einstellungsitems, Zeitschrift für Sozialpsychologie, 12, 24-41.

Jansen, P.G.W. (1981b) De onbruikbaarheid van Mokkenschaalanalyse, Tijdschrift voor Onderwijsresearch, 6.

Lazarsfeld, P.F. & Henry, N.W. (1968) Latent Structure Analysis. Boston: Houghton Mifflin.

Lewis, C. (1981) Estimating abilities: inference for random variables, Kwantitatieve Methoden, nr. 3, 2, Rotterdam: Vereniging voor Statistiek.

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

Lumsden, J. (1976) Test theory, Annual Review of Psychology, 1976, 251-280.

Mårdberg, B. (1974/1975) Latent profile analysis, Reports from the Institute of Psychology, University of Bergen, Norway, no.10 en no. 4 (1975).

Mokken, R.J. (1971) A theory and procedure of scale analysis, Den Haag: Mouton.

Molenaar, W. (1974) De logistische en de normale kromme, Nederlands Tijdschrift voor de Psychologie, 29, 415-420.

Molenaar, I.W. (1979) Meten in de sociale wetenschappen, Voordracht Vakantiecursus voor leraren, Stichting Mathematisch Centrum augustus 1979, Rapport nr. VC 33/79, p. 19-49.

Molenaar, I.W. (1980) Some improved diagnostics for failure of the Rasch model, Heymans Bulletin Psychologische Instituten R.U. Groningen, HB-80-482-EX (to be published).

Molenaar, I.W. (1981) Beperkte bruikbaarheid van Jansen's kritiek, Tijdschrift voor Onderwijsresearch, 6.

Molenaar, I.W. (1983) De latente trek modellen van Rasch en Mokken: voordelen en problemen. To be published.

Novick, M.R., Jackson, P.H. & Thayer, D.T. (1971) Bayesian inference and the classical reliability model: reliability and true scores. Psychometrika, 36, 261-288.

Novick, M.R. & Jackson, P.H. (1974) Statistical methods for educational and psychological research. New York: McGraw-Hill.

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

Vos, H.H. de (1980) Het meten van werkorientaties, dissertatie, R.U.Groningen.

Wainer, H., Morgan, A. & Gustafsson, J.E. (1980) A review of estimation procedures for the Rasch model with an eye toward longish tests, Journal of Educational Statistics, 5, 35-64.

Wainer, H. & Wright, B.D. (1980) Robust estimation of ability in the Rasch model, Psychometrika, 45, 3, 373-391.

Wollenberg, A. van den (1979) The Rasch model and time-limit tests, (dissertatie K.U.Nijmegen), Nijmegen: Studentenpers.

Wollenberg, A.L. van den (1981a) Two new test statistics for the Rasch model, University of Nijmegen, to be published.

Wollenberg, A.L. van den (1981b) A simple and effective method to test the dimensionality axiom of the Rasch model, University of Nijmegen, to be published.

Wollenberg, A.L. van den (1981c) On the Wright-Panchapakesan goodness of fit test for the Rasch model, University of Nijmegen, to be published.

Wood,R. (1978) Fitting the Rasch model - a heady tale, British Journal of Mathematical and Statistical Psychology, 31, 27-32.

Wright, B.D. & Stone, M.H. (1979) Best test design, Rasch measurement, Chicago: Mesa Press, 5835 Kinbarg Avenue.