40

Opstelbeoordeling met lijnen en getallen

Jan Binne Hoeksma

1. Inleiding

Opstellen beoordelen is meten. "Meten is het toekennen van getallen aan objecten of gebeurtenissen volgens regels" (Stevens 1951, in Kerlinger 1964). Het gebruik van beoordelingsschalen ter verbetering van de kwaliteit van opstelbeoordeling heeft direct betrekking op bovenstaande algemeen aanvaarde definitie van meten. Door gebruik te maken van beoordelingsschalen wordt geprobeerd de regels voor het toekennen van getallen eenduidig vast te leggen.

Een *beoordelingsschaal* bestaat uit een verzameling in kwaliteit oplopende opstellen die elk zijn voorzien van een schaalwaarde. De taak van de beoordelaar is de kwaliteit van een te beoordelen opstel te vergelijken met de kwaliteit van de opstellen in de beoordelingsschaal, en het opstel een waardering te geven overeenkomstig de schaalwaarde van het in kwaliteit best gelijkende opstel in de schaal. In vergelijking met de opstelbeoordeling waarbij van een schaal bestaande uit schoolcijfers gebruik wordt gemaakt heeft schaalbeoordeling onder meer het voordeel dat de schaalwaarden betekenisvol zijn: ieder schaalpunt representeert immers een opstel. Alleen al op grond van deze overweging mogen we verwachten dat schaalbeoordeling tot betrouwbaarder en valider beoordelingen leidt dan de traditionele beoordeling aan de hand van schoolcijfers (zie ook Wesdorp 1981).

Geen enkele meetprocedure is echter beter dan haar regels (Kerlinger 1974: 427). Bij schaalbeoordeling vormt de beoordelingsschaal met voorbeeldopstellen het hoofdbestanddeel van die regels. De schaal is de meetlat. Het is daarom zaak ruime aandacht en tijd te besteden aan de constructie van beoordelingsschalen. De *vergelijkingsmethode*, een schaalmethode afkomstig uit de psychofysica van het begin van deze eeuw, lijkt met het oog op de constructie van beoordelingsschalen veelbelovend. In een kleinschalig proefonderzoek hebben wij ervaring met de methode opgedaan. Daarvan doen wij hier verslag. Achtereenvolgens komen aan de orde: een concrete beschrijving van de methode, de vraagstelling en opzet van het proefonderzoek, alsmede de analyse en resultaten ervan.

W.K.B. Koning (red.), *Taalbeheersing in theorie en praktijk*. Dordrecht 1985: Foris Publications, 362–369.

2. Opstelbeoordeling met de vergelijkingsmethode

Voor de constructie van een beoordelingsschaal maken we gebruik van de *vergelijkingsmethode*. Deze schaalmethode werd aanvankelijk alleen gebruikt voor het schalen van percepties van fysische stimuli (bijvoorbeeld: toonhoogte of kleurintensiteit). Later is de methode met succes toegepast bij het schalen van sociale stimuli, bijvoorbeeld de status van beroepen (Neyens et al. 1981), opinies en attituden (Hamblin 1974). Toegepast op opstelbeoordeling ziet de methode er als volgt uit: beoordelaars krijgen een reeks opstellen voorgelegd samen met een zogenaamd standaardopstel; de opdracht voor de beoordelaar is zijn waardering voor een opstel te bepalen door het te vergelijken met het standaardopstel. Voor het aangeven van de waardering voor een opstel bestaan een aantal verschillende, voor de methode kenmerkende, uitdrukkingswijzen of antwoordmodaliteiten. Wij hebben er twee gebruikt:

- Bij de antwoordmodaliteit *lijnproduktie* is het standaardopstel voorzien van een lijnstuk met een bepaalde lengte. De beoordelaar drukt nu zijn waardering voor een opstel uit door het trekken van een lijn zoveel korter of langer dan de standaardlijn, in overeenstemming met hoeveel slechter of beter hij het opstel vindt dan het standaardopstel.
- Bij de tweede antwoordmodaliteit (*magnitude-estimation*) is het standaardopstel gewaardeerd met een standaardgetal, in ons geval het getal 100. De beoordelaar drukt zijn oordeel over een opstel eveneens in een getal uit. Vindt hij het opstel bijvoorbeeld tweemaal zo goed als het standaardopstel dan geeft hij het de waardering 200, bij driemaal zo slecht een waardering van rond de 33.

Het gebruik van andere antwoordmodaliteiten (bijvoorbeeld de manipulatie van geluidssterkte of lichtsterkte) behoort tot de mogelijkheden. Wij kozen voor lijnproduktie en magnitude-estimation vanwege hun eenvoud in het gebruik.

In het gebruik van meer antwoordmodaliteiten schuilt één van de voordelen van de vergelijkingsmethode. Door variatie in antwoordmodaliteiten is binnen een kort tijdsbestek herhaalde meting mogelijk zonder dat daarbij éénzelfde respons wordt herhaald. Een tweede voordeel van de vergelijkingsmethode is dat de betrouwbaarheid (nauwkeurigheid) van de metingen doorgaans zeer hoog is. Een derde voordeel is dat met de methode data op log-intervalniveau worden verkregen die na transformatie analyseerbaar zijn met (lineaire) intervaltechnieken (Saris et al. 1980).

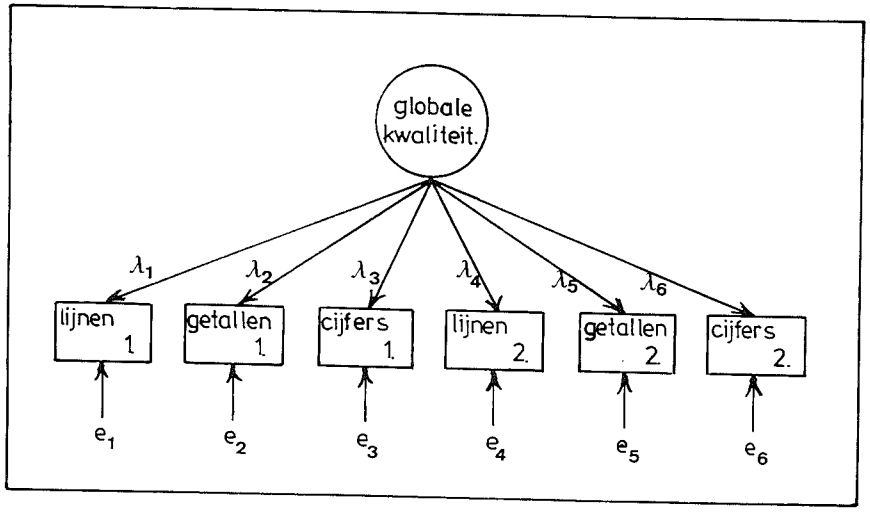
3. Het proefonderzoek

De vergelijkingsmethode is voor zover wij weten niet eerder toegepast bij de beoordeling van stelopdrachten. Daarom zijn wij in het proefonderzoek nagegaan of de methode bruikbaar is. Met name wilden wij de volgende vraag

beantwoorden: hoe verhoudt de vergelijkingsmethode zich tot de meer traditionele methode waarbij aan opstellen schoolcijfers (zgn. categorie-oordelen) worden toegekend? Toegespitst: (1) wordt met de vergelijkingsmethode hetzelfde gemeten als met de traditionele methode waarbij schoolcijfers worden toegekend en (2) komen beoordelaars met de vergelijkingsmethode tot betrouwbaarder oordelen dan met de traditionele methode?

In het proefonderzoek legden wij zes beoordelaars twintig opstellen voor met het verzoek deze (gespreid over twee zittingen) zes maal te beoordelen op *globale kwaliteit*. Op de eerste zitting werden de opstellen eerst beoordeeld volgens de vergelijkingsmethode met lijnproduktie, daarna met dezelfde methode waarbij het oordeel in getallen werd uitgedrukt en tenslotte werden de opstellen op de traditionele manier beoordeeld en gewaardeerd met schoolcijfers. Na vijf dagen volgde de tweede zitting waarop de drie beoordelingen werden herhaald. Verder zij opgemerkt dat de opstellen in a-selecte volgorde werden aangeboden en dat de opstellen tussen de twee zittingen in tweemaal zijn beoordeeld op 'inhoud en organisatie', op 'stijl' en op 'conventies'. Deze beoordelingen blijven hier verder buiten beschouwing.

De beschreven proefopzet levert per beoordelaar zes oordelen of responsen per opstel op. Saris et al. (1980) tonen aan dat de relaties tussen responsen verkregen met de vergelijkingsmethode na transformatie[1] geanalyseerd kunnen worden met LISREL, meer in het bijzonder met het zogenaamde *congenerieke testmodel* (Jöreskog 1971). Dit model, uitgewerkt voor de geschetste proefopzet, is in *figuur 1* weergegeven in de vorm van een pad-diagram.



Figuur 1: Eén-faktormodel

In het model worden de zes geobserveerde variabelen casu quo de zes oordelen per opstel weergegeven als een lineaire funktie van een gemeenschappelijke, niet geobserveerde variabele of factor *globale kwaliteit* en de storingstermen ε_i. De termen λ_i zijn regressie-gewichten die de invloed van de factor globale kwaliteit op de gegeven oordelen schalen.

Het model kan ook in meer psychometrische bewoordingen worden beschreven. In de klassieke testtheorie wordt een geobserveerde score – in ons geval een oordeel over een opstel – geacht te bestaan uit een ware score en een fout of scoringsterm (zie bijvoorbeeld Nunnaly 1974). Het model geeft weer dat alle zes de oordelen over een opstel dezelfde ware score hebben en slechts verschillen door toevallige fouten. De coëfficiënten λ_i zijn de correlaties tussen de geobserveerde score en de ware score waarvan de kwadraten geïnterpreteerd worden als betrouwbaarheidscoëfficiënten van de geobserveerde variabelen. De onbekende grootheden in het model (de regressiegewichten en de storingstermen) kunnen worden geschat met het computerprogramma LISREL (Jöreskog & Sörbom 1978), waarna met hetzelfde programma getoetst kan worden of het model de data adequaat beschrijft. De gebruikte toetsingsgrootheid is chi-kwadraat verdeeld. En het model wordt verworpen als de rechteroverschrijdingskans behorend bij de toetsingsgrootheid kleiner is dan .05.

4. De vergelijkingsmethode en de schoolcijfermethode vergeleken

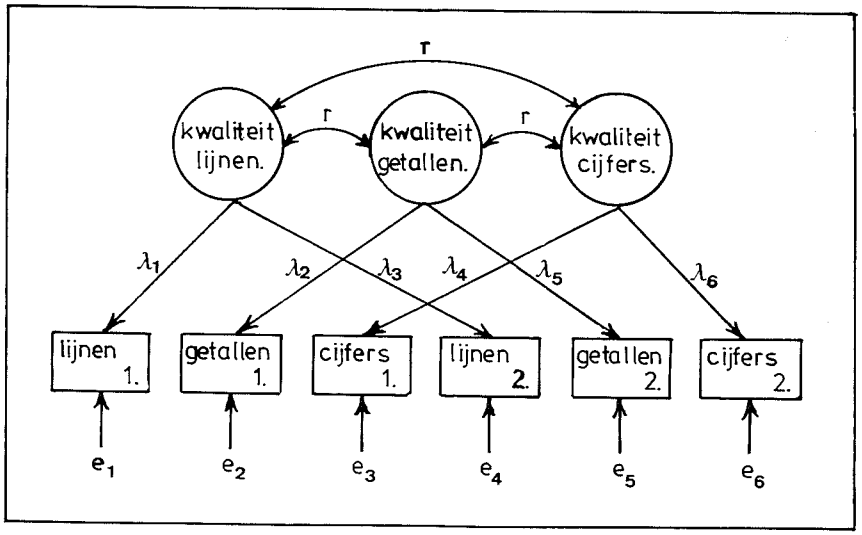
Het één-factormodel, zoals dat is weergegeven in *figuur 1*, vormt een representatie van de hypothese dat er naast de kwaliteit van de opstellen geen andere systematische invloeden zijn op de, op beide zittingen in drie modaliteiten, gegeven oordelen casu quo dat er met lijnen en getallen en schoolcijfers en op beide momenten hetzelfde worden gemeten. Dit model, deze hypothese hebben we voor de zes beoordelaars afzonderlijk getoetst. De toetsing (zie *tabel 1*) levert slechts voor twee van de zes beoordelaars een passend model op. Bij de overigen is er sprake van systematische variantie in de oordelen die *niet* verklaard kan worden door de factor *globale kwaliteit*. We mogen daarom niet concluderen dat met alle zes de oordelen hetzelfde wordt gemeten.

Beoordelaar	een-faktormodel (df=9)		twee-faktormodel (df=8)		drie-faktormodel (df=6)	
	χ^2	Pr	χ^2	Pr	χ^2	Pr
1	28.20	<.05*	6.62	>.05	19.64	<.05*
2	19.81	<.05*	11.46	>.05	14.09	<.05*
3	14.36	>.05	1.27	>.05	12.16	>.05
4	32.48	<.05*	11.34	>.05	30.22	<.05*
5	23.48	<0.5*	8.64	>.05	13.45	<.05*
6	15.42	>.05	9.91	>.05	13.29	<.05*

* Model wordt verworpen; (Pr: rechter overschrijdingskans).

Tabel 1: Toetsingsresultaten van het een-, twee- en drie-faktormodel

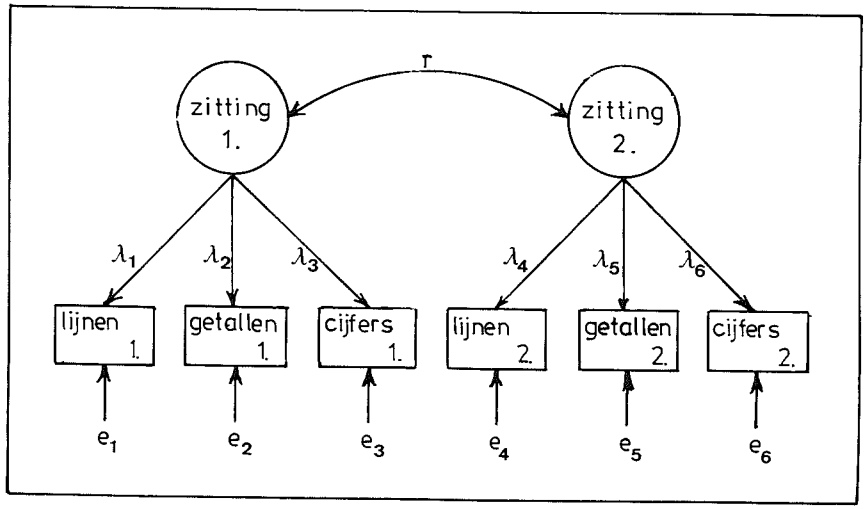
Er zijn twee min of meer plausibele bronnen voor deze onverklaarde variantie. Ten eerste kan worden verondersteld dat beoordelaars de kwaliteit van opstellen anders beoordelen als zij hun oordeel uitdrukken in een andere modaliteit (dat bijvoorbeeld het oordeel in lijnen verschilt van dat in schoolcijfers). Deze veronderstelling wordt gerepresenteerd door een drie-factormodel (*figuur 2*) bestaande uit een factor *kwaliteit uitgedrukt in lijnen*, een factor *kwaliteit uitgedrukt in getallen* en een factor *kwaliteit uitgedrukt in schoolcijfers*. Daarbij wordt aangenomen dat deze factoren onderling correleren, zij het niet perfect.



Figuur 2: Drie-faktormodel

Ten tweede kan worden verondersteld dat beoordelaars de opstellen op de twee zittingen verschillend beoordelen (dat het oordeel van de ene op de andere zitting is veranderd). Deze hypothese wordt gerepesenteerd door een twee-factormodel (*figuur 3*) bestaande uit een factor *kwaliteit op zitting één* en een factor *kwaliteit op zitting twee*. Daarbij wordt eveneens aangenomen dat beide factoren correleren.

Beide veronderstellingen zijn wederom getoetst[2] met LISREL. De toetsingsresultaten in *tabel 1* laten zien dat het drie-factormodel voor vijf van de zes beoordelaars wordt verworpen. Het twee-factormodel daarentegen beschrijft de beoordelingsdata voor alle beoordelaars adequaat. Zodat we mogen concluderen dat (1) de beoordelaars tussen beide zittingen hun oordeel of wijze van beoordelen hebben veranderd en (2) dat met de vergelijkingsmethode (lijnen en getallen) dezelfde *globale kwaliteit* wordt gemeten als met de traditionele methode waarbij schoolcijfers worden toegekend.



Figuur 3: Twee-faktormodel

Hiermee is één van de twee vragen van ons proefonderzoek beantwoord. Rest de vraag of beoordelaars met de vergelijkingsmethode tot even nauwkeurige of wellicht nauwkeuriger oordelen komen als met de traditionele methode. Om deze vraag te kunnen beantwoorden hebben we twee specifieke gevallen van het voorgaande twee-factormodel (dat voor alle beoordelaars paste) nader onderzocht. Te weten: *model 1* dat de hypothese representeert dat de zes beoordelingen (lijnen, getallen en schoolcijfers op twee zittingen) alle even betrouwbaar zijn, en *model 2* dat weergeeft dat de beoordelingen in de drie modaliteiten weliswaar even nauwkeurig zijn, maar over de twee zittingen variëren.

Beoordelaar	model 1 (df=18)		model 2 (df=16)		betrouwbaarheden (λ^1)	
	χ_1^2	Pr	χ_2^2	Pr	zitting 1	zitting 2
1	17.65	<.05	14.01	>.05	.78	.88
2	20.36	<.05	20.22	>.05	.60	.64
3	13.86	<.05	7.73	>.05**	.82	.92
4	31.33	>0.5*	22.97	>.05**	.81	.93
5	16.72	<.05	9.91	>.05**	.75	.89
6	27.07	<.05	14.42	>.05**	.77	.93

* model wordt verworpen.
** model 2 past significant beter dan model 1.

Tabel 2: Twee specifieke gevallen van het twee-faktormodel, model 1 en model 2, alsmede de betrouwbaarheidscoëfficienten onder model 2

De resultaten van beide toetsingen[3], weergegeven in *tabel 2*, laten zien dat *model 1* slechts voor één beoordelaar hoeft te worden verworpen en dat *model 2* voor allen mag worden geaccepteerd. Vergelijken we de passing van beide modellen onderling, dan constateren we dat *model 2* voor vier van de zes beoordelaars significant beter[4] past dan *model 1*. Op grond van deze bevindingen geven we de voorkeur aan *model 2* boven *model 1* en concluderen we dat met de vergelijkingsmethode even nauwkeurig wordt gemeten als met de traditionele methode waarbij schoolcijfers worden toegekend.

De betrouwbaarheidscoëfficiënten zoals die zijn berekend onder *model 2* voor de afzonderlijke beoordelaars in *tabel 2* bevestigen de in de inleiding gemaakte opmerking dat deze doorgaans hoog zijn. Geruststellend is dat de nauwkeurigheid van de oordelen is toegenomen. Kennelijk geldt dat beoordelaars na enige oefening trefzekerder lijnen trekken en getallen en schoolcijfers geven.

Resumerend concluderen we dat beoordelaars de vergelijkingsmethode goed kunnen toepassen. De globale kwaliteit zoals gemeten met de vergelijkingsmethode verschilt niet van de globale kwaliteit zoals die wordt gemeten met de traditionele methode. Er is geen sprake van methoden-variantie. Voorshands blijkt echter niet dat de vergelijkingsmethode in opstelbeoordeling tot nauwkeuriger oordelen leidt dan de traditionele methode waarbij schoolcijfers worden toegekend.

Noten

1. De lijnen en getallen worden getransformeerd volgens $y = \ln(y)$, waarin ln de natuurlijke logarithme is. De categorie-oordelen worden niet getransformeerd.
2. Naast de regressiegewichten (λ_i) en de fouten (ε_i) zijn ook de correlatie-coëfficiënten (r) geschat.
3. Bij de passing van *model 1* zijn de zes λ-coëfficiënten aan elkaar gelijkgesteld, terwijl bij passing van *model 2* de λ-coëfficiënten binnen de factoren zijn gelijkgesteld.
4. *Model 2* past significant beter dan *model 1* als het verschil van de chi-kwadraten behorend bij beide modellen groter is dan 5.99 (df=2).

Literatuur

Hamblin, R.L.
1973 'Social attitudes: magnitude measurement and theory'. In: H.M. Blalock Jr. (ed.), *Measurement in social sciences.* London: McMillan Press, 61–121.
Jöreskog, K.G.
1971 'Statistical analysis of sets of congeneric tests'. *Psychometrika 36*, 109–133.
Jöreskog, K.G. en D. Sörbom
1978 *LISREL IV: a general computer program for estimation of linear structural equation systems by maximum likelihood methods.* Chicago: International Education Services.
Kerlinger, F.N.
1964 *Foundations of behavioral research.* London: Holt.
Neijens, P., L. van Doorn en W.E. Saris
1981 *Mens en maatschappij 4*, 378–397.

Nunnaly, J.
1967 *Psychometric theory.* New York: McGraw-Hill.
Saris, W.E., P. Neijens en L. van Doorn
1980 'Scaling social science variables by multimodality matching'. *M.D.N. 2*, 3–21.
Wesdorp, H.
1981 *Evaluatie-technieken voor het moedertaalonderwijs.* Amsterdam/Den Haag: SCO/ Staatsuitgeverij.