TILBURG UNIVERSITY

Tilburg University

De invloed van het samenvoegen van geordende antwoordcategorieën op de H-coëfficiënt van het Mokken-model

van den Berg, P.T.; Sijtsma, K.; Feij, J.A.

Published in:
Kwantitatieve Methoden: Nieuwsbrief voor Toegepaste Statistiek en Operationele Research

Publication date:
1990

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
van den Berg, P. T., Sijtsma, K., & Feij, J. A. (1990). De invloed van het samenvoegen van geordende antwoordcategorieën op de H-coëfficiënt van het Mokken-model. *Kwantitatieve Methoden: Nieuwsbrief voor Toegepaste Statistiek en Operationele Research*, *11*(33), 41-56.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 06. Oct. 2022

KM 33(1990)
pag 41 - 56

DE INVLOED VAN HET SAMENVOEGEN VAN GEORDENDE ANTWOORD-CATEGORIEEN OP DE H-COEFFICIENT VAN HET MOKKEN-MODEL*

P.T. van den Berg, K. Sijtsma en J.A. Feij**

Samenvatting

In dit artikel wordt gedemonstreerd wat er met de H-coëfficiënt gebeurt, als geordende antwoordcategorieën van items worden samengevoegd. Dit wordt gedaan aan de hand van de scores van 441 sollicitanten op de vier schalen van de Spanningsbehoeftelijst-selectieversie (SBL-s). De items worden gescoord op zevenpuntschalen. Door het samenvoegen worden items verkregen met respectievelijk vijf, drie en twee antwoordcategorieën. De H-coëfficiënt blijkt hoger te zijn naarmate meer categorieën zijn samengevoegd. Ingegaan wordt op de oorzaken van deze gevonden tendens. Daarnaast blijkt dat hoe minder categorieën er zijn, hoe minder zekerheid men heeft dat H afwijkt van nul. Geconcludeerd wordt dat men dezelfde waarde van H bij verschillende aantallen categorieën verschillend moet interpreteren.

1 Inleiding

Het Mokken-model (Mokken, 1971; Mokken & Lewis, 1982; Sijtsma, 1988) is een lid van de familie der item-response modellen. In item-response modellen verklaart men het antwoordgedrag dat tot uiting komt in de itemscores uit zowel eigenschappen van de personen als eigenschappen van de items. Juist vanwege de zwakke theoretische aannamen van het Mokken-model zou een Mokken-schaalanalyse een goed alternatief kunnen zijn voor analyses volgens strengere item-response modellen als het Rasch-model. Tot voor kort kon een Mokken-schaalanalyse alleen worden toegepast op dichotome items. Molenaar (1986) heeft het model echter uitgebreid voor toepassing op polychotome items. Debets en Brouwer (1989) hebben voor dit model een computerprogram-

---

* Het onderzoek waarvan hier verslag wordt gedaan maakt deel uit van het onderzoeksproject Spanningsbehoefte en Selectie (Psychon; subsidie verstrekt door NWO). De auteurs danken het LTP te Amsterdam voor zijn medewerking.

** Vrije Universiteit, Vakgroep Arbeids- en Organisatiepsychologie
De Boelelaan 1081, 1081 HV Amsterdam, tel. 020-5485502.

ma geschreven: Mokken Scale analysis for Polychotomous items (MSP). Op basis van ervaring met Mokken-schaalanalyse van dichotome items zijn normen opgesteld voor de kwaliteit van een verzameling van items alsmede voor individuele items. Voor polychotome items ontbreekt die ervaring vooralsnog en derhalve is het onduidelijk of de normen in het geval van dichotome items hier kunnen worden toegepast.

Het doel van dit artikel is om op beperkte schaal de resultaten van Mokken-schaalanalyses van polychotome items te vergelijken met die van dichotome items. Hierbij beperken wij ons tot het model van monotone homogeniteit, een van de modellen uit de benadering volgens Mokken, en richten wij de aandacht op de H-coëfficiënt waarmee de kwaliteit van monotoon homogene verzamelingen van items kan worden beoordeeld. Het Mokken-model van dubbele monotonie wordt hier buiten beschouwing gelaten, omdat de H-coëfficiënt niet geschikt is om als kwaliteitsmaat voor dit model te fungeren (zie Mokken, Lewis & Sijtsma, 1986).

2 Het Mokken-model voor polychotome items: de H-coëfficiënt

We bespreken in deze paragraaf kort het Mokken-model voor polychotome items voor zover dit van belang is voor het te rapporteren onderzoek. Uitgebreide behandelingen kan men vinden bij Molenaar (1982; 1986) en Sijtsma, Debets en Molenaar (1990).

Het begrip 'itemstap' neemt in het model een centrale plaats in. Men kan zich daarbij het volgende voorstellen. Gesteld dat een item uit een psychologische test antwoorden toelaat in drie geordende antwoord-categorieën (bijvoorbeeld, 'mee oneens', 'neutraal' en 'mee eens', naar aanleiding van de stelling uit een attitude-vragenlijst: 'Euthanasie is toegestaan als de persoon de wens daartoe te kennen geeft') dan kan het antwoordgedrag worden beschreven met behulp van twee denkbeeldige itemstappen. Ten eerste bepaalt de respondent - althans, imaginair - of hij voldoende met de stelling instemt om de stap te kunnen nemen van 'mee oneens' naar 'neutraal'. Indien niet, dan is de stapscore gelijk aan nul; als de stap wel wordt genomen dan is de score gelijk aan één. In het laatste geval bepaalt de respondent tevens of hij voldoende instemt om de tweede stap van 'neutraal' naar 'mee eens' te nemen. Ook hier kan de itemstapscore weer de waarden nul en één aannemen. De totaalscore op het item is gelijk aan de som van de itemstapscores en in het voorbeeld is deze som dus gelijk aan nul, een of twee.

Aangezien een nul op de eerste stap een nul op de tweede stap impliceert

en een één op de tweede stap een één op de eerste stap, kunnen we de itemstappen niet opvatten als onafhankelijke dichotome 'pseudo-items'. De afhankelijkheid van stappen binnen hetzelfde item impliceert dat een speciaal model moet worden ontworpen voor de analyse van polychotome scores. Het eenvoudigweg toepassen van het bestaande Mokken-model voor dichotome items op dichotome itemstappen is niet mogelijk. We wijzen er terzijde op dat het hier beschreven idee van itemstappen ook is te vinden bij Andrich (1978), die het toepast binnen de klasse van logistische item-response modellen.

Naar analogie van de schaalbaarheidscoëfficiënt H volgens Mokken (1971), heeft Molenaar (1982) een dergelijke coëfficiënt voorgesteld voor polychotome itemscores. De definitie van deze laatste coëfficiënt is gebaseerd op de ordening naar opklimmende moeilijkheid van itemstappen behorend bij paren van items. We behandelen de H-coëfficiënt voor twee items waarna de uitbreiding naar een willekeurig aantal items gemakkelijk volgt.

We gaan uit van twee items, i en j, ieder met m+1 geordende antwoordcategorieën en derhalve m itemstappen. De itemscores geven we aan als X_i en X_j, waarbij X_i, X_j = 0, 1, ..., m. De itemstapscores geven we aan als Y_{gi} en Y_{hj}, met Y_{gi}, Y_{hj} = 1, 0 en g, h = 1, 2 ..., m.

Tussen itemscores en itemstapscores bestaat, bijvoorbeeld voor item i, de volgende relatie:

$$X_i = \sum_{g=1}^{m} Y_{gi} \ . \qquad (1)$$

Gegeven de definitie van de itemstap kan de moeilijkheid worden gedefinieerd als

$$\pi_{gi} = \text{Prob}(Y_{gi}=1) = \text{Prob}(X_i \geq g) \ . \qquad (2)$$

De moeilijkheid π_{gi} geeft de proportie van personen aan die stap g van item i heeft genomen. Deze personen hebben een itemscore die tenminste gelijk is aan g. In steekproeven wordt π_{gi} geschat door middel van

$$\hat{\pi}_{gi} = \left(\sum_{h=g}^{m} n_{hi} \right) / n \ , \qquad (3)$$

waarbij n de steekproefgrootte aangeeft, en n_{hi} het aantal personen dat een itemscore X_i=h heeft.

Voor twee items i en j worden 2m moeilijkheden geschat volgens (3). Deze moeilijkheden kunnen worden afgeleid uit een kruistabel van de orde (m+1)*(m+1), waarin per cel het aantal personen is gegeven met een specifieke scorecombinatie op items i en j. Langs de marginalen staan de cumulatieve frequenties waaruit de moeilijkheden volgens (3) direct kunnen worden geschat. Voorbeelden van deze kruistabellen staan in Tabel 2 en Tabel 3. We verwijzen voorts naar Molenaar (1982; 1986) en Debets en Brouwer (1988), waar eveneens voorbeelden van kruistabellen worden gegeven.

In navolging van de definitie van de H-coëfficiënt voor dichotome items gaan we er vanuit dat een scorepatroon waarbij een specifieke persoon een gemakkelijker stap op het ene item heeft gemist maar een moeilijker stap op het ander item wel neemt, niet is toegestaan. Dergelijke patronen worden aangeduid als fouten. De andere patronen zijn de toegestane scorepatronen. Binnen de genoemde kruistabel kan een pad worden aangegeven van cellen die corresponderen met toegestane patronen. In totaal zijn 2m + 1 patronen toegestaan en m^2 patronen niet.

Behalve dat de geobserveerde frequenties per cel in de kruistabel kunnen worden bepaald, kan ook met behulp van de marginalen, geschat volgens (3), worden vastgesteld wat de verwachte frequenties zouden zijn onder onafhankelijkheid van de itemscores. Laat O_{ij} het totale aantal geobserveerde fouten aangeven, gesommeerd over de foutencellen van de tabel behorend bij i en j en laat E_{ij} het totale aantal verwachte fouten aangeven onder marginale onafhankelijkheid. De schaalbaarheidscoëfficiënt H_{ij} is dan gedefinieerd als

$$H_{ij} = 1 - O_{ij}/E_{ij} \,. \qquad (4)$$

De H-coëfficiënt voor k items is gebaseerd op de kruistabellen van alle tweetallen van items. De definitie van de schaalbaarheidscoëfficiënt voor k items luidt

$$H = 1 - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} O_{ij} \; / \; \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} E_{ij} \,. \qquad (5)$$

De maxima van H_{ij} en H zijn gelijk aan één, een waarde die wordt behaald indien er geen foutenpatronen zijn. De waarde nul houdt in dat het aantal geobserveerde fouten gelijk is aan het verwachte aantal onder marginale

onafhankelijkheid. Het minimum ligt niet vast.

Voor dichotome items geeft Mokken (1971) vuistregels voor de interpretatie van H:

$H < 0.30$	; items niet schaalbaar
$0.30 \leq H < 0.40$	; items zwak schaalbaar
$0.40 \leq H < 0.50$	; items middelmatig schaalbaar
$H \geq 0.50$	; items sterk schaalbaar

Of deze vuistregels bruikbaar zijn bij polychotome items is nog maar de vraag (zie bijvoorbeeld, Sijtsma, 1990). Immers, bij dichotome items is de verhouding van mogelijke toegestane patronen en mogelijke foutenpatronen 3:1, bij items met vier geordende categorieën is deze verhouding 7:9 en bij zeven categorieën 13:36. Het aantal mogelijke foutenpatronen krijgt dus bij een toenemend aantal categorieën snel de overhand, hetgeen wellicht zou kunnen resulteren in lagere H-waarden.

Tenslotte vermelden we dat Molenaar (1982; 1983) de steekproeventheorie van H en H_{ij} heeft afgeleid. Hierdoor is het mogelijk om te toetsen of H en H_{ij} gelijk zijn aan nul, tegen het alternatief dat ze positief zijn. Uitgaande van vaste marginalen in de kruistabel voor i en j stelt Molenaar (1983) een benadering van de verdeling van het aantal fouten O_{ij} voor onder de nulhypothese van marginale onafhankelijkheid. Gebruiken we Φ voor de standaard normale verdelingsfunctie, $\mathcal{E}$ voor de verwachting en σ voor de standaarddeviatie, dan kan voor het gevonden aantal fouten o_{ij} de benadering

$$P(O_{ij} \leq o_{ij}) \approx \Phi(z) \ ,$$
$$z = (o_{ij} + 0.5 - \mathcal{E}O_{ij}) \ / \ \sigma(O_{ij}) \qquad (6)$$

worden bepaald. Met behulp van (6) kan worden nagegaan of het aantal fouten O_{ij} significant kleiner is dan het aantal verwacht onder marginale onafhankelijkheid. Dit komt overeen met het toetsen van de nulhypothese dat $H_{ij} = 0$ tegen het alternatief dat $H_{ij} > 0$. De waarde z in (6) wordt doorgaans aangegeven met de benaming Delta Star.

In dit artikel onderzoeken we de waarden van H en H_{ij} zoals die zijn berekend voor vier subtests van de Spanningsbehoeftelijst-selectieversie (Van den Berg & Feij, 1988). We gaan na wat er met H en H_{ij} gebeurt indien antwoordcategorieën van de oorspronkelijk, zeven-categoriële items worden samengevoegd. Het onderzoek heeft vooral heuristische waarde en leidt tot enige suggesties voor grondiger vervolgonderzoek.

3 De Spanningsbehoeftelijst-selectieversie

Spanningsbehoefte, een vertaling van het begrip 'sensation seeking tendency', is een persoonlijkheidstrek gedefinieerd als ".. the need for varied, novel and complex sensations and experiences and the willingness to take physical and social risks for the sake of such experience" (Zuckerman, 1979, p. 10). Zuckerman (1979) ontwikkelde vijf vragenlijsten, waarmee dit persoonlijkheidskenmerk gemeten kan worden. Hiervan stonden twee lijsten model voor de Nederlandse versie van de Spanningsbehoeftelijst, de SBL (Feij & Van Zuilen, 1984). Omdat de spanningsbehoefte van een persoon relevant lijkt voor de satisfactie en het functioneren in de werksituatie, werd een onderzoek opgezet om de SBL geschikt te maken voor gebruik in selectiesituaties ten behoeve van personeelsvoorziening. Met omissie van vragen over sex en drugs, die voor sollicitanten bedreigend zouden kunnen zijn, werd een speciale selectieversie ontwikkeld, de SBL-s (Van den Berg & Feij, 1988). De SBL-s bestaat uit 74 items met ieder zeven geordende antwoordcategorieën (zevenpuntschalen).

Uit onderzoek door Van den Berg en Feij (1988) bleek dat de SBL-s opgedeeld moet worden in vier subtests: Risicobereidheid, Ervaringsgerichtheid, Behoefte aan Verandering en Ontremming. De betrouwbaarheid van de scores op de subtests, geschat door middel van Cronbach's alpha, varieerde van 0.74 tot 0.86, hetgeen op het niveau van subtests bevredigend mag worden genoemd. Uit factoranalyse bleek dat de subtests mogen worden opgevat als unidimensioneel. Verder voldeden de subtests, na dichotomisering van de itemscores, aan de eisen van het Rasch model.

Vanuit psychometrisch gezichtspunt kan de SBL-s dus een goede test worden genoemd. Het onderzoek naar de predictieve validiteit is momenteel in gang (Van den Berg, in voorbereiding).

4 Werkwijze

Ten behoeve van het validiteitsonderzoek van de SBL-s werden gegevens verzameld bij 441 sollicitanten naar uiteenlopende functies. Deze gegevens werden gebruikt in het hier te rapporteren onderzoek. Zij hebben betrekking op de vier subtests van de SBL-s en zijn beschikbaar in de vorm van itemscores (X_i = 0, 1, ..., 6).

Om de invloed van het aantal antwoordcategorieën per item op de H-coëfficiënt te kunnen onderzoeken werden naburige categorieën samengevoegd

op een wijze die verderop in deze paragraaf wordt beschreven. Uitgaande van de oorspronkelijke zeven categorieën resulteerde dit achtereenvolgens in vijf, drie en twee antwoordcategorieën. Per geval werden de waarden van de H-coëfficiënt bestudeerd en werd gezocht naar oorzaken van eventuele afname of toename van H. Het achteraf door de onderzoekers samenvoegen van itemscores is in zoverre discutabel dat men niet weet of respondenten, wanneer zij daadwerkelijk met, bijvoorbeeld, vijfpuntschalen in plaats van zevenpuntschalen waren geconfronteerd, het veronderstelde antwoordgedrag hadden vertoond. Als voorbeeld nemen we de situatie dat iemand de score 5 heeft op de zevenpuntschaal, waarop 0 en 6 de extremen zijn. Na samenvoegen van de categorieën 5 en 6, alsmede 0 en 1, ontstaat er een vijfpuntschaal waarop de respondent nu in de extreme categorie scoort. De vraag is dan natuurlijk of deze respondent, ware hij geconfronteerd met de vijfpuntschaal, ook werkelijk de extreme categorie had gekozen. Uit de literatuur is bekend dat sommige personen de neiging hebben om extreme categorieën te ontlopen.

Onafhankelijk van het zojuist gesignaleerde probleem menen wij dat ons onderzoek in de huidige vorm zijn nut heeft. Regelmatig komt het immers voor dat onderzoekers antwoordcategorieën samenvoegen en daar berekeningen op baseren. Een reden daarvoor is dat respondenten soms niet alle antwoordcategorieën gebruiken, zodat het aantal waarnemingen per categorie erg klein is en het item de facto eigenlijk functioneert als, bijvoorbeeld, een vijfpuntitem in plaats van een zevenpuntitem. Een andere reden is dat achteraf kan blijken dat respondenten geen betrouwbaar onderscheid maken tussen een groot aantal antwoordcategorieën, zodat een kleiner aantal een betrouwbaarder beeld oplevert van het antwoordgedrag. Verder gaan wij er vanuit dat de huidige aanpak bruikbare informatie oplevert aangaande karakteristieken van de H-coëfficiënt, los van de vraag of het samenvoegen van antwoordcategorieën vanuit psychologisch standpunt verdedigbaar is.

De scores op enkele negatief geformuleerde items van de SBL-s zijn andersom gecodeerd. Hoge scores betekenen dus altijd dat de respondent de bedoelde eigenschappen in sterke mate bezit en lage scores dat hij of zij ze in geringe mate bezit.

De antwoordcategorieën van de items van de SBL-s zijn op drie manieren samengevoegd:

1. 0 en 1 samen en 5 en 6 samen: vijf categorieën.
2. 0 en 1 samen, 2, 3 en 4 samen en 5 en 6 samen: drie categorieën.
3. 0, 1, 2 en 3 samen en 4, 5, en 6 samen: twee categorieën.

Bij de derde manier van samenvoegen is de middenpositie (3) bij de laagste categorie gevoegd, omdat de mediane response voor de meeste items groter is dan 3.

De berekeningen in dit onderzoek werden uitgevoerd met het computerprogramma MSP (Debets & Brouwer, 1989).

5 Resultaten

Op de oorspronkelijke scores en de scores die ontstonden door het samenvoegen van categorieën werden Mokken-schaalanalyses uitgevoerd. De gevonden waarden van H en Delta Star staan in Tabel 1. De betrouwbaarheidsschattingen in de derde kolom komen in de Discussie aan de orde.

Op de eerste plaats valt in Tabel 1 op, dat de H-waarden, bepaald met de oorspronkelijke scores (zeven categorieën) bij alle subtests onder de kritische waarde van 0.30 voor dichotome items liggen. Volgens de vuistregels voor dichotome items zouden de subtests dus niet schaalbaar te noemen zijn.

Op de tweede plaats is duidelijk, dat de H-waarde in het algemeen toeneemt naarmate er meer categorieën zijn samengevoegd. Bij dichotomisering zijn twee van de subtests, Risicobereidheid en Ontremming, wel schaalbaar. Overigens merken we op dat bij een itemselectie, waarbij items met een lage schaalbaarheid worden weggelaten, Mokken-schaalanalyse op de dichotome scores vier kortere subtests oplevert met een H groter dan 0.30. Vergelijking van de in Tabel 1 vermelde H-waarden met die van tests, die zijn samengesteld op basis van itemselectie volgens Mokken, is daarom niet op zijn plaats.

Op de derde plaats is er een tendens te constateren, dat Delta Star lager is naarmate er minder categorieën zijn.

In de volgende paragraaf gaan we nader op deze resultaten in. Dit doen we aan de hand van kruistabellen, waarmee de schaalbaarheid van twee items ten opzichte van elkaar (H_{ij}) wordt bepaald. De bespreking van de resultaten voor Delta Star komt in paragraaf 5.2 aan de orde.

Tabel 1. Resultaten van Mokken-schaalanalyses op de subtests van de SBL-s bij verschillende aantallen categorieën.

	H	Delta Star	betrouwbaarheid*
Risicobereidheid (13 items)			
7 categorieën	.21	63	.86
5 categorieën	.22	54	.86
3 categorieën	.31	55	.85
2 categorieën	.34	50	.83
Ervaringsgerichtheid (11 items)			
7 categorieën	.14	31	.71
5 categorieën	.15	27	.71
3 categorieën	.21	28	.71
2 categorieën	.20	21	.65
Behoefte aan Verandering (20 items)			
7 categorieën	.14	56	.82
5 categorieën	.15	50	.81
3 categorieën	.20	45	.78
2 categorieën	.23	44	.78
Ontremming (8 items)			
7 categorieën	.20	32	.79
5 categorieën	.22	29	.78
3 categorieën	.34	30	.76
2 categorieën	.43	28	.77

* De betrouwbaarheid is geschat op basis van een methode beschreven door Molenaar en Sijtsma (1988).

5.1 De invloed van het samenvoegen van antwoordcategorieën op H_{ij}

Teneinde de invloed te onderzoeken, die het samenvoegen van antwoordcategorieën op H_{ij} kan hebben, bekijken we twee extreme gevallen. Eerst geven wij een voorbeeld waarbij H_{ij} afneemt door het samenvoegen van naburige antwoordcategorieën. Dit komt slechts weinig voor en alleen als zeven antwoordcategorieën worden samengevoegd tot vijf antwoordcategorieën. Het voorbeeld wordt gegeven, omdat het demonstreert dat H_{ij} niet noodzakelijk hoeft toe te nemen, maar zelfs af kan nemen bij samenvoeging. Als tweede voorbeeld nemen wij het geval, waarbij H_{ij} het sterkst toeneemt bij samenvoeging van antwoordcategorieën. Hierbij gaat het om een samenvoeging van zeven tot twee antwoordcategorieën. Aan het voorbeeld kan men duidelijk zien hoe het komt dat H_{ij} toeneemt.

Tabel 2 geeft de kruistabellen voor het eerste geval. De betreffende items luiden: 'Ik zou het liefst om de paar jaar willen verhuizen' en 'Als ik

lange tijd op dezelfde plaats moet blijven word ik onrustig'. In Tabel 2 is telkens per cel het geobserveerde aantal fouten aangegeven en daaronder het verwachte aantal fouten bij marginale onafhankelijkheid. Het verwachte aantal fouten per cel onder marginale onafhankelijkheid is het produkt van de marginalen gedeeld door het totaal aantal proefpersonen. De cellen van links boven naar rechts onder die door lijnen met elkaar zijn verbonden vormen de toegestane patronen van itemscores. De scores in de overige cellen worden als fouten aangemerkt. De categorieën 0 en 1 en de categorieën 5 en 6 in Tabel 2A zijn in Tabel 2B samengevoegd.

Zoals reeds vermeld neemt H_{ij} in het algemeen toe als er antwoordcategorieën worden samengevoegd. In het voorbeeld in Tabel 2 zien we echter dat H_{ij} afneemt van 0.22 naar 0.16. De vraag is hoe deze uitzondering op de algemene regel te verklaren is. Om te begrijpen op welke wijze het samenvoegen van categorieën tot een lagere H kan leiden bekijken we formule (4). Hieraan zien we dat H_{ij} afneemt, als O/E toeneemt. Veranderingen in O en E kunnen ontstaan, doordat foutencellen bij het samenvoegen in de toegestane patronen worden opgenomen. Dit is, zoals blijkt uit Tabel 2A, het geval bij de foutencellen behorend bij de scorepatronen (0,2), (1,0), (4,6) en (6,5). Van deze foutencellen zullen vooral de cellen met weinig geobserveerde fouten en veel verwachte fouten bij samenvoeging een toename in O/E en een afname in H_{ij} veroorzaken. Het grootste verschil tussen geobserveerd en verwacht aantal fouten zien we bij cel (0,2). Deze cel bevat slechts 7 geobserveerde fouten bij een verwacht aantal van 23.7. Het is vooral aan het gering aantal fouten in deze cel toe te schrijven, dat O/E toeneemt van 0.78 naar 0.84.

Als we kijken naar het verloop van het aantal fouten in de eerste rij van Tabel 2A, dan valt op dat het aantal geobserveerde fouten sterk daalt van cel (0,1) naar cel (0,2) en weer sterk toeneemt van cel (0,2) naar cel (0,3). Een dergelijk onregelmatig verloop is moeilijk anders dan door toeval te verklaren. Dit doet vermoeden dat in het algemeen een afname van H_{ij} bij het samenvoegen van categorieën aan toeval is toe te schrijven.

Uit het voorbeeld in Tabel 2 blijkt op de eerste plaats, dat H_{ij} niet noodzakelijk hoeft toe te nemen, maar zelfs kan afnemen bij samenvoeging. De toename van H_{ij} die in de meeste gevallen plaatsvindt, is dus geen noodzakelijk gevolg van het samenvoegen, maar een empirisch verschijnsel dat om een nadere verklaring vraagt. Op de tweede plaats blijkt de afname van H_{ij} het gevolg te zijn van het feit dat een van de foutencellen, die door het samenvoegen in de toegestane patronen worden opgenomen, weinig geobserveerde fouten bevat in verhouding tot het verwachte aantal fouten.

Tabel 2. Kruistabellen van de responsen op twee items uit de SBL-s bij zeven (Tabel 2A) en bij vijf antwoordcategorieën (Tabel 2B).
In de cellen: geobserveerde en verwachte aantallen responsen.

A. Kruistabel bij zeven categorieën

item		op dezelfde plaats blijven								
	cat.	0	1	2	3	4	5	6	marg.	cum.
verhuizen	0	32	22	7	27	7	10	5	110	441
		13.2	24.4	23.7	22.2	15.2	8.9	2.2		
	1	7	35	31	19	16	4	3	115	331
		13.8	25.6	24.8	23.2	15.9	9.4	2.4		
	2	6	18	41	17	14	8	0	104	216
		12.5	23.1	22.4	21.0	14.4	8.5	2.1		
	3	1	13	5	16	9	5	1	50	112
		6.0	11.1	10.8	10.1	6.9	4.1	1.0		
	4	2	7	6	7	11	3	0	36	62
		4.3	8.0	7.8	7.3	5.0	2.9	0.7		
	5	3	2	4	2	2	4	0	17	26
		2.0	3.8	3.7	3.4	2.4	1.4	0.4		
	6	2	1	1	1	2	2	0	9	9
		1.1	2.0	1.9	1.8	1.2	0.7	0.2		
marginalen		53	98	95	89	61	36	9		
cumul. marg.		441	388	290	195	106	45	9		

O = 220 E = 282.8 Sd(O) = 9.4 H = 0.22 Delta Star = 6.7

B. Kruistabel bij vijf categorieën

item		op dezelfde plaats blijven						
	cat.	0	1	2	3	4	marg.	cum.
verhuizen	0	96	38	46	23	22	225	441
		77.0	48.5	45.4	31.1	23.0		
	1	24	41	17	14	8	104	216
		35.6	22.4	21.0	14.4	10.6		
	2	14	5	16	9	6	50	115
		17.1	10.8	10.1	6.9	5.1		
	3	9	6	7	11	3	36	62
		12.3	7.8	7.3	5.0	3.7		
	4	8	5	3	4	6	26	26
		8.9	5.6	5.3	3.6	2.7		
marginalen		151	95	89	61	45		
cumul. marg.		441	290	195	106	45		

O = 204 E = 243.8 Sd(O) = 9.1 H = 0.16 Delta Star = 4.4

Tabel 3 geeft de kruistabellen van een itempaar, waarbij H_{ij} sterk toeneemt door het dichotomiseren van de itemscores. In Tabel 3A valt op dat het pad met toegestane patronen niet langs de diagonaal van links boven naar rechts onder loopt, maar afgebogen is naar de rechter bovenhoek. Dit komt doordat de proefpersonen meer instemmen met het item 'Ik houd er wel van om af en toe ontzettend veel gein te maken' dan met het item 'Ik houd van wilde feesten'. Door de kromlijnigheid van het pad zijn er foutencellen, die ver van het pad met toegestane patronen af liggen. De fouten uit deze cellen komen bij dichotomisering terecht in de nieuwe foutencellen. De foutencellen links onder, die wat verder van het pad af liggen, vormen samen de foutencel in Tabel 3B. Het lijkt er daarom op, dat de grote toename van H_{ij} samenhangt met het feit dat de foutencellen ver van het pad van toegestane patronen weinig fouten bevatten in verhouding tot het verwachte aantal en cellen dichtbij het pad van toegestane patronen veel fouten. We zullen nader bekijken wat er met deze cellen gebeurt bij het samenvoegen van antwoordcategorieën.

Aan de hand van Tabel 2 en Tabel 3 kunnen we zien, dat bij een geleidelijke samenvoeging van categorieën de foutencellen dichtbij het pad het eerst verdwijnen. Als we kunnen aangeven waarom in deze cellen O/E hoog is, is daarmee dus de toename van H_{ij} en van H verklaard bij samenvoeging van categorieën.

Sijtsma (1988, pag. 93, 94) demonstreert aan de hand van een voorbeeld, dat H_{ij} van twee dichotome items hoger is naarmate de moeilijkheidsgraad van de items sterker verschilt (zie ook Jansen, 1982). In het voorbeeld geldt dat de items voldoen aan de eisen van het Rasch-model, dat kan worden opgevat als een speciaal geval van het Mokken-model. In het algemeen geldt dat afhankelijk van het discriminerend vermogen bij dergelijke items met een sterk verschillende moeilijkheidsgraad relatief weinig fouten zullen optreden in verhouding tot het verwachte aantal fouten. Dit geldt evenzeer voor de itemstappen van twee polychotome items. Onder de aanname dat de items passen in het Mokken-model ontstaan dus bij itemstappen die sterk van elkaar in moeilijkheidsgraad verschillen weinig fouten in verhouding tot het verwachte aantal. De hiermee corresponderende cellen liggen ver van het pad met toegestane patronen (links onder en rechts boven in de kruistabellen). Dit zijn 'onwaarschijnlijke' fouten in de zin dat weinig respondenten een dergelijk patroon genereren. Fouten op paren van itemstappen met ongeveer gelijke moeilijkheidsgraad staan in de cellen dichtbij de toegestane patronen ('waarschijnlijke' fouten of fouten die door veel respondenten worden gegenereerd).

Tabel 3. Kruistabellen van de responsen op twee items uit de SBL-s bij zeven (Tabel 3A) en bij twee antwoordcategorieën (Tabel 3B).
In de cellen: geobserveerde en verwachte aantallen responsen.

A. Kruistabel bij zeven categorieën

item		gein maken								
	cat.	0	1	2	3	4	5	6	marg.	cum.
f	0	4	7	7	8	8	10	7	51	441
e		0.5	1.4	2.9	5.9	13.0	17.2	10.2		
e	1	0	3	8	13	21	25	8	78	390
s		0.7	2.1	4.4	9.0	19.8	26.4	15.6		
t	2	0	0	6	13	22	19	7	67	312
e		0.6	1.8	3.8	7.8	17.0	22.6	13.4		
n	3	0	1	3	9	26	27	21	87	245
		0.8	2.4	4.9	10.1	22.1	29.4	17.4		
	4	0	1	0	7	26	32	19	85	158
		0.77	2.3	4.8	9.8	21.6	28.7	17.0		
	5	0	0	1	0	9	30	19	59	73
		0.5	1.6	3.3	6.8	15.0	19.9	11.8		
	6	0	0	0	1	0	6	7	14	14
		0.1	0.4	0.8	1.6	3.6	4.7	2.8		
marginalen		4	12	25	51	112	149	88		
cumul. marg.		441	437	425	400	349	237	88		

0 = 229 E = 272.8 Sd(0) = 9.4 H = 0.16 Delta Star = 4.7

B. Kruistabel bij twee categorieën

item		gein maken			
f	cat.	0	1	marg.	cum.
e	0	82	201	283	441
e		59.0	224.0		
s	1	10	148	158	158
t		33.0	125.0		
e					
n					
marginalen		92	349		
cumul. marg.		441	349		

0 = 10 E = 33.0 Sd(0) = 4.1 H = 0.70 Delta Star = 5.6

Het feit dat H en H_{ij} lager zijn naarmate het aantal antwoordcategorieën groter is, is dus te verklaren doordat bij een groter aantal categorieën meer itemstappen met elkaar worden vergeleken die weinig verschillen in moeilijkheidsgraad. Hierdoor ontstaan specifieke foutenpatronen met grote waarschijnlijkheid. Omdat deze waarschijnlijke fouten bij de berekening van H en H_{ij} even zwaar meetellen als onwaarschijnlijke fouten zijn bij een groter aantal categorieën H en H_{ij} lager.

5.2 De invloed van het samenvoegen van antwoordcategorieën op Delta Star

In Tabel 1 zien we dat de toetsingsgrootheid Delta Star in tegenstelling tot H een tendens tot afname vertoont. Naarmate het aantal antwoordcategorieën afneemt, wordt Delta Star lager. Dit betekent bij deze gegevens dat als men categorieën samenvoegt H wel hoger wordt, maar dat men minder zekerheid heeft dat H in de populatie afwijkt van nul. Door het samenvoegen van categorieën neemt de 'power' van de toets af.

Het resultaat is in overeenstemming met in de literatuur vermelde gegevens over dichotome items. Zo geeft Sijtsma (1988, pag. 101) in geval van dichotome items aan dat Delta Star in tegenstelling tot H juist relatief klein is, als het ene item extreem gemakkelijk en het andere item extreem moeilijk is. Omdat bij meercategoriële items de opeenvolgende itemstappen weinig in moeilijkheidsgraad verschillen, kunnen er meer waarschijnlijke fouten optreden. Samenvoeging van antwoordcategorieën resulteert in grotere afstanden tussen de opeenvolgende stappen en daarmee, naar analogie van het geval van dichotome scores, tot een geringere Delta Star-waarde.

6 Discussie

De resultaten geven een eerste aanwijzing dat bij een Mokken-analyse voor polychotome items H toeneemt naarmate er per item meer antwoordcategorieën worden samengevoegd. Deze tendens is te verklaren, doordat bij een groter aantal categorieën de opeenvolgende itemstappen minder verschillen in moeilijkheidsgraad. Hierdoor worden meer waarschijnlijke schendingen tegen het model van monotone homogeniteit gemaakt. De tendens betekent dat bij een groter aantal categorieën meer fouten worden gemaakt in de ordening van de personen op de latente trek. Volgens de gegeven verklaring wordt dit niet veroorzaakt doordat men bij een groter aantal categorieën minder nauwkeurig zou meten, maar doordat de schaal waarop men meet sterker

gedifferentieerd is, zodat er meer fouten in de ordening van de personen kunnen ontstaan. Om deze reden moet men dezelfde waarde van H bij verschillende aantallen categorieën anders interpreteren. Men zou daarom ook verschillende normen voor de interpretatie van H en H_{ij} moeten gebruiken, die lager zijn naarmate er meer categorieën worden gebruikt. Om een alternatieve coëfficiënt te verkrijgen die minder afhankelijk is van het aantal categorieën zou men de fouten moeten wegen, zodat onwaarschijnlijke fouten (itemstapmoeilijkheden ver uiteen) zwaarder tellen dan waarschijnlijke fouten (itemstapmoeilijkheden dicht bijeen). Verschillende manieren van wegen zijn reeds ontwikkeld en zullen in een vervolgonderzoek aan de orde komen.

Tenslotte wijzen wij erop dat de betrouwbaarheid van de meting bij verschillende aantallen antwoordcategorieën niet dramatisch wordt aangetast; zie Tabel 1. Dat zou voor de SBL-s kunnen betekenen dat respondenten vaak niet in staat zijn om op de zevenpuntschalen een systematisch onderscheid te maken tussen naburige antwoordcategorieën. Vanuit dàt oogpunt bezien zou men voor de SBL-s en wellicht persoonlijkheidsvragenlijsten in het algemeen een kleiner aantal dan zeven categorieën kunnen gebruiken.

Literatuur

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Berg, P.T. van den & Feij, J.A. (1988). De ontwikkeling van een selectieversie van de Spanningsbehoeftelijst (SBL-s). *Nederlands Tijdschrift voor de Psychologie*, *43*, 328-334.

Berg, P.T. van den. Proefschrift in voorbereiding. Amsterdam, V.U.

Debets, P. & Brouwer, E. (1989). *User's manual MSP*. Groningen: iec ProGAMMA, Rijksuniversiteit Groningen.

Feij, J.A. & Zuilen, R.W. van (1984). *Handleiding bij de Spanningsbehoeftelijst (SBL)*. Lisse: Swets & Zeitlinger.

Jansen, P.G.W. (1982). Homogenitätsmessung mit hilfe des Koeffizienten H von Loevinger: Eine kritische Diskussion. *Psychologische Beiträge*, *24*, 96-105.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. 's-Gravenhage: Mouton.

Mokken, R.J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *10*, 279-285.

Mokken, R.J., Lewis, C. & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement*, *10*, 279-285.

Molenaar, I.W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden*, *3*, *no. 8*, 145-164.

Molenaar, I.W. (1983). Meten met Mokkenschalen, statistisch uitgediept. Samenvatting lezing Statistische Dag.

Molenaar, I.W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën. In: G.F. Pikkemaat & J.J.A. Moors (Eds.), *Liber Amicorum Jaap Muilwijk*. Groningen: Econometrisch Instituut.

Molenaar, I.W. & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, *9*, *nr. 28*, 115-126.

Sijtsma, K. (1988). *Contributions to Mokken's Nonparametric Item Response Theory*. Amsterdam: Free University Press.

Sijtsma, K. (1990). Betekenis en bruikbaarheid van het Mokken model. *Nederlands Tijdschrift voor de Psychologie* (geaccepteerd).

Sijtsma, K., Debets, P. & Molenaar, I.W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application. *Quality and Quantity* (in press).

Zuckerman, M. (1979). *Sensation Seeking. Beyond the optimal level of arousal*. New York: McGraw-Hill.

Ontvangen: 28-05-1989
Geaccepteerd: 26-10-1989