Rasch-homogeniteit empirisch onderzocht

Sijtsma, K.

Published in:
Tijdschrift voor Onderwijsresearch

Publication date:
1983

Rasch-Homogeniteit Empirisch Onderzocht

K. Sijtsma

Research Instituut voor het Onderwijs in het Noorden, Haren (Groningen)

Rasch homogeneity empirically examined

A few of the many methods for testing the Rasch model are combined in order to select homogeneous subsets of items from a larger set. Andersens conditional likelihood ratio test (1973) is used for globally testing the model assumptions of monotonicity and sufficiency on the one hand and onedimensionality and local stochastic independence on the other hand. In case the model is globally rejected, depending on the kind of model violation a test proposed by Molenaar (1980) and a graphical method (see e.g. Fischer, 1974) are used for detecting deviating items.
For a data set of 40 arithmetic items it is shown how the proposed combination of the methods leads to homogeneous tests in the sense of the Rasch model.

De laatste jaren is er een verscheidenheid aan statistische toetsen en exploratieve methoden beschikbaar gekomen waarmee kan worden onderzocht of een verzameling van items voldoet aan de assumpties van het dichotoom logistisch model van Rasch (zie bijv. Fischer, 1974, p. 281 e.v.; Van den Wollenberg, 1979, p. 31 e.v., p. 107 e.v.; Gustafsson, 1980, p. 211 e.v.; Wright & Stone, 1979, p. 66 e.v.; Molenaar, 1980). De assumpties zijn: eendimensionaliteit van de meting, monotoon niet-dalende item karakteristieke curven, lokaal stochastische onafhankelijkheid van de item-antwoorden van een vaste persoon en afdoendheid van de ruwe testscore voor de schatting van de latente persoonsparameter ξ . Deze vier assumpties zijn nodig en voldoende voor het Rasch model:

$$P(X_{vi} = 1|\xi_v, \sigma_i) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)}$$

De formule geeft de kans dat persoon v met parameterwaarde ξ_v op de latente trek ξ item i met moeilijkheid σ_i goed beantwoordt. Uitgebreide verhandelingen over het Rasch model zijn o.a. te vinden bij Fischer (1974) en Wright & Stone (1979).
Tevens is de laatste jaren duidelijk geworden dat geen van de statistische toetsen en exploratieve methoden kan worden beschouwd als een globale toets van het Rasch model voor alle assumpties tegelijk (zie bijv. Van den Wollenberg, 1979, p. 94 e.v.; Stelzl, 1979; Gustafsson, 1980; Molenaar, 1980; Formann, 1981).
In dit artikel wordt een poging gedaan om enkele toetsende en exploratieve methoden zodanig te combineren in empirisch onderzoek, dat het mogelijk wordt om via een gestandaardiseerde cyclus schalen te construeren, die aan alle assumpties van het Rasch model voldoen. Daarbij

---

Met dank aan P.H. Been en I.W. Molenaar voor hun kritische commentaar op eerdere versies van dit artikel.
Adres auteur: RION, Postbus 132, 9751 SZ Haren Gn.

worden alle assumpties van het model expliciet, dan wel impliciet onderzocht op empirische gegevens. Dat de cyclus 'gestandaardiseerd' is wil overigens niet zeggen dat de onderzoeker geen subjectieve beslissingen meer hoeft te nemen bij de schaalconstructie.

Toetsen voor het Rasch model

De belangrijkste eigenschap van het Rasch model is de populatieonafhankelijkheid van de parameterschattingen: de itemparameters worden onafhankelijk van de verdeling van de persoonsparameters geschat en de persoonsparameters worden onafhankelijk van de verdeling van de itemparameters geschat. Een gevolg van deze eigenschap is bijvoorbeeld dat personen dezelfde parameterwaarden krijgen toegekend, onafhankelijk van de test die ze hebben gemaakt. Uiteraard moeten de tests wel uit hetzelfde Rasch-homogene itemdomein komen. Andersen (1973) heeft een toetsingsmethode gepresenteerd die gebaseerd is op de eigenschap van populatie-onafhankelijkheid van de itemparmeters.

De procedure voor deze toets is dat men de groep van proefpersonen indeelt in selecte deelgroepen. Dit kunnen scoregroepen zijn, maar andere indelingscriteria zoals geslacht of leeftijd zijn ook toegestaan. De keuze van indelingscriteria is afhankelijk van de bedoelingen van de onderzoeker. In de wederzijds uitsluitende en uitputtende deelgroepen en de gehele groep worden de itemparameters geschat en er wordt vervolgens getoetst in hoeverre de schattingen in de deelgroepen van elkaar verschillen. Wanneer de gegevens voldoen aan het Rasch model zullen die schattingen vanwege hun populatie-onafhankelijkheid niet significant van elkaar verschillen. De toetsingsgrootheid is asymptotisch χ^2 verdeeld met (g-1)(k-1) vrijheidsgraden, waarbij g het aantal deelgroepen is en k het aantal items. Als de nulhypothese van gelijke itemparameters in alle deelgroepen wordt verworpen, valt met de globale toets van Andersen niet zonder meer vast te stellen welke assumptie(-s) van het Rasch model zijn geschonden. Een aantal auteurs (Van den Wollenberg, 1979; Stelzl, 1979; Gustafsson, 1980) heeft met behulp van gesimuleerde gegevens onderzocht voor welke modelschendingen de toets van Andersen gevoelig is. Van den Wollenberg (1979, p. 84-95) heeft aangetoond dat de toets van Andersen bij deelgroepen op basis van de ruwe score gevoelig is voor schendingen van de assumpties van monotonie en afdoendheid. De toets blijkt echter in hoge mate ongevoelig te zijn voor schendingen van eendimensionaliteit, en daarmee voor schendingen van lokaal stochastische onafhankelijkheid (Molenaar, 1980, p. 18). Deze ongevoeligheid voor meerdimensionaliteit van de gegevens hangt samen met het feit dat de criteria voor de vorming van deelgroepen, zoals de ruwe score op de test, meestal in dezelfde mate samenhangen met de latente trekken die ten grondslag liggen aan de testprestatie. Van den Wollenberg (1979, p. 108) heeft laten zien dat de toets van Andersen wel geschikt is voor het toetsen van eendimensionaliteit wanneer deelgroepen worden gevormd met behulp van 'splitteritems' (zie ook Molenaar, 1980, p. 4). Dat zijn items die in hoge mate een beroep doen op één latente trek, maar vrijwel ongerelateerd zijn aan andere latente trekken die bij sommige items een rol spelen. De proefpersonengroep wordt gesplitst in de groep die het splitteritem goed en de groep die het splitteritem fout heeft beantwoord. Stel nu dat het splitteritem en een aantal andere items overwegend een meting van trek 1 zijn en de overige items overwegend een meting zijn van trek 2. De veronderstelling van Van den Wollenberg (1979, p. 109) is dat de gemiddelde persoonsparameter op trek 1 in de eerste groep relatief hoog is en in de tweede groep relatief laag. Op trek 2 zullen de parameterwaarden voor beide groepen rond het totale gemiddelde liggen. Wanneer nu twee items i en j worden vergeleken, waarbij i overwegend trek 1 meet en j overwegend trek 2, dan zullen de schattingen van de itemparameters in beide

groepen systematisch verschillen. De eigenschap van populatie-onafhankelijke schattingen geldt dus niet. Wanneer men de schattingen in beide groepen grafisch tegen elkaar afzet, kan men de items die sterk met het splitteritem samenhangen herkennen doordat ze in de 'foute' groep moeilijker zijn dan in de 'goede' groep.

De toets van Andersen kan dus enerzijds gebruikt worden als globale toets voor de assumpties van monotonie en afdoendheid, en anderzijds als globale toets voor de assumpties van eendimensionaliteit en lokaal stochastische onafhankelijkheid. In het laatste geval kan met behulp van een grafiek tevens op itemniveau worden nagegaan welke items verantwoordelijk zijn voor modelschendingen. Molenaar (1980) heeft een toets gepresenteerd waarmee op itemniveau kan worden onderzocht welke items de assumpties van monotonie en afdoendheid, of alleen de laatste, schenden. Schendingen komen tot uiting in item karakteristieke curven die te vlak of te steil zijn in vergelijking met die van de overige items, hetgeen een schending van afdoendheid is, of item karakteristieke curven die niet monotoon stijgend zijn. Bij de toets van Molenaar wordt de frequentie goede antwoorden op een vast item i per ruwe scoregroep vergeleken met de verwachte frequentie. De gestandaardiseerde verschillen tussen beide frequenties worden voor een vast item i over de ruwe scoregroepen uit het laagste en het hoogste kwartiel gecombineerd tot een toetsingsgrootheid U_i. Deze toetsingsgrootheid is bij benadering standaardnormaal verdeeld wanneer het aantal waarnemingen in iedere ruwe scoregroep groot is. Grote positieve waarden van U_i geven aan dat de item karakteristieke curve van item i relatief te vlak is, grote negatieve waarden dat de item karakteristieke curve relatief te steil is. Bij de interpretatie van U_i lijkt het verstandig om tevens de geobserveerde en verwachte frequenties per scoregroep te vergelijken. Op die manier kan worden nagegaan of U_i niet ten gevolge van lokale uitschieters erg hoog of laag is dan wel of een afwijkende trend van een item karakteristieke curve over het gehele bereik van de latente trek vóórkomt. Verder lijkt het bij de verwijdering van items op basis van hun U_i-waarden het beste om eerst items met te vlakke karakteristieke curven weg te laten (Molenaar, 1980, p. 14), omdat dit de items zijn met een relatief zwak discriminerend vermogen. Tenslotte kan het vóórkomen dat groepen van items te grote of te kleine U_i-waarden hebben. Dit kan een aanwijzing zijn voor meerdimensionaliteit van de itemverzameling (Molenaar, 1980, p. 14; Gustafsson, 1980, p. 208).

Met behulp van de tot nu toe behandelde methoden kunnen alle assumpties van het Rasch model voor empirische gegevens worden onderzocht. De assumptie van lokaal stochastische onafhankelijkheid wordt hier niet expliciet onderzocht: het streven is gericht op eendimensionele metingen en wanneer hier aan voldaan is volgt uit eendimensionaliteit lokaal stochastische onafhankelijkheid (Gustafsson, 1980, p. 207).

Een overzicht van de behandelde toetsen en exploratieve methoden is te vinden in Tabel 1. In de tabel staat door middel van plustekens aangegeven welke assumpties per toets expliciet worden onderzocht. In het geval van bijvoorbeeld U_i zijn dat monotonie en afdoendheid, maar deze toetsingsmethode kan de onderzoeker tevens op het spoor zetten van meerdimensionaliteit. Met behulp van de toetsen en exploratieve methoden uit tabel 1 wordt in de volgende paragraaf een onderzoekscyclus geconstrueerd waarmee schaalconstructie volgens het Rasch model kan plaatsvinden. Hierbij wordt voor empirische gegevens systematisch onderzocht of assumpties geschonden zijn en welke items daarvoor verantwoordelijk zijn.

Een onderzoekscyclus voor itemanalyse

Wanneer men een verzameling van items wil analyseren volgens de genoemde onderzoeks-

Tabel 1.
Toetsen en exploratieve methoden voor het onderzoeken van Rasch-homogeniteit. In de tabel wordt aangegeven welke assumpties expliciet worden onderzocht.

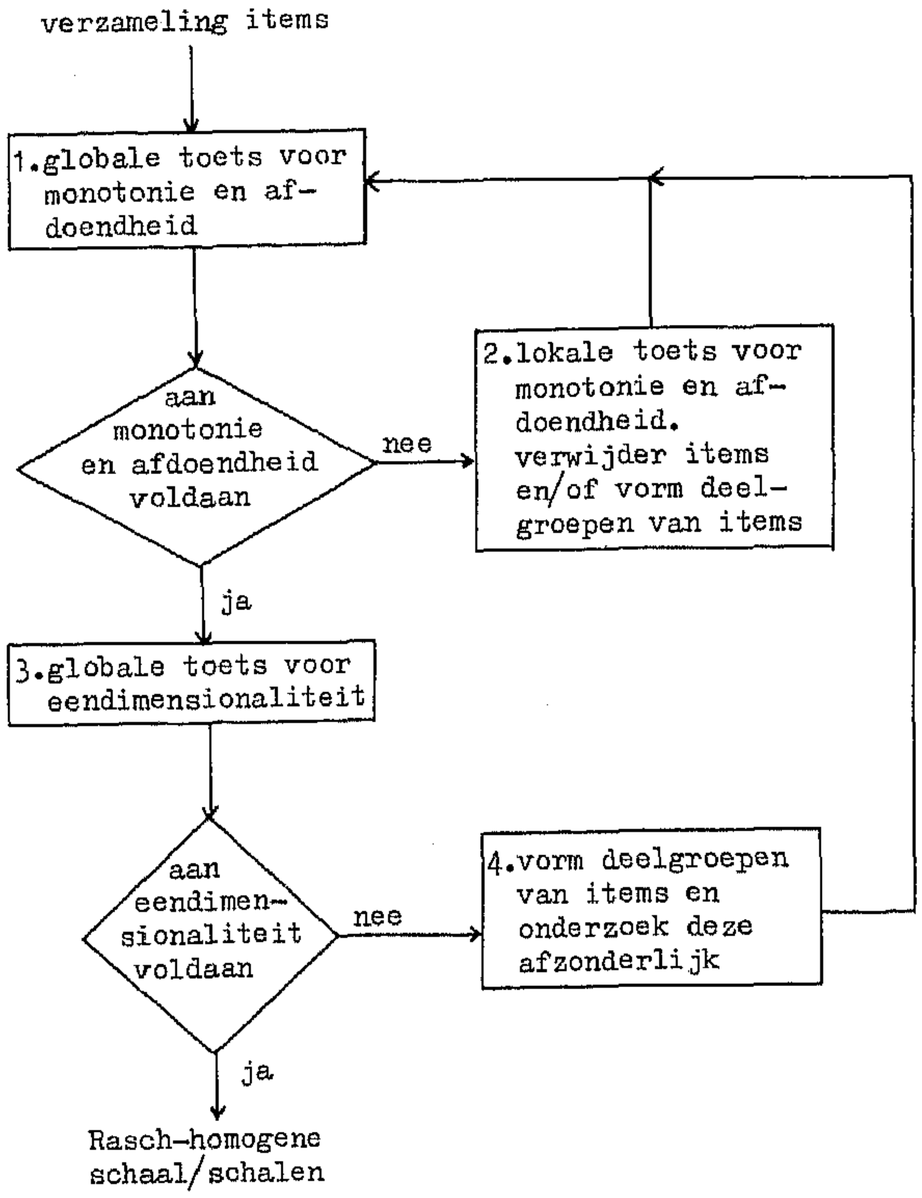
niveau	toets/exploratieve methode	assumpties eendimensionaliteit	monotonie	afdoendheid
globaal	Andersen-toets (scoregroepen)		+	+
	Andersen-toets (splitteritems)	+		
lokaal	U_i van Molenaar		+	+
	Grafische controle n.a.v. Andersen-toets	+		

methoden, dan biedt de volgende onderzoekscyclus een aanpak die tot één of meer Rasch-homogene schalen leidt, wanneer deze althans in de gegevens aanwezig zijn.

1. Voer de globale toets van Andersen voor monotonie en afdoendheid uit. Wanneer de nulhypothese niet wordt verworpen, ga verder met 3; wordt de nulhypothese wel verworpen, ga verder met 2.
2. Bereken voor ieder item (lokaal) de U_i-toetsingsgrootheid en beschouw het empirische verloop van de item karakteristieke curven. Verwijder bij voorkeur items op basis van zowel formele als inhoudelijke gronden en/of verdeel de verzameling van items in inhoudelijke betekenisvolle deelgroepen. Ga voor iedere groep items weer naar 1.
3. Wanneer het vermoeden bestaat dat de assumptie van eendimensionaliteit geschonden is, voer dan de toets van Andersen volgens scores op splitteritems uit. Wanneer de nulhypothese niet wordt verworpen, dan is de onderzoekscyclus ten einde. Wordt de nulhypothese wel verworpen, ga dan naar 4.
4. Wanneer uit de grafische modelcontrole, waarin de schattingen van de itemparameters in deelgroepen tegen elkaar afgezet worden, een inhoudelijk zinvol interpreteerbare vorm van meerdimensionaliteit blijkt, ga dan voor iedere deelschaal terug naar 1 en doorloop de cyclus opnieuw.

De onderzoekscyclus wordt in Figuur 1 in de vorm van een stroomdiagram weergegeven.

Uiteraard is de onderzoekscyclus slechts een ruwe schets voor schaalconstructie volgens het Rasch model. Ook via andere methoden en combinaties van methoden lijkt het mogelijk om tot Rasch-homogene schalen te komen. Zo zou men kunnen beginnen met het onderzoeken van de assymptie van eendimensionaliteit. Een voordeel van de hier geschetste cyclus lijkt echter dat men met behulp van de U_i-waarden alvast informatie kan verkrijgen over mogelijke onderverdelingen van de items. Tevens kan men items herkennen die item-specifieke factoren meten. Dat laatste is bij een minder nauwkeurige grafische modelcontrole naar aanleiding van de splittermethode lastiger. Van de alternatieve methoden moeten de Q_1- en Q_2-toets van Van



Figuur 1: Onderzoekscyclus voor itemanalyse volgens het Rasch model.

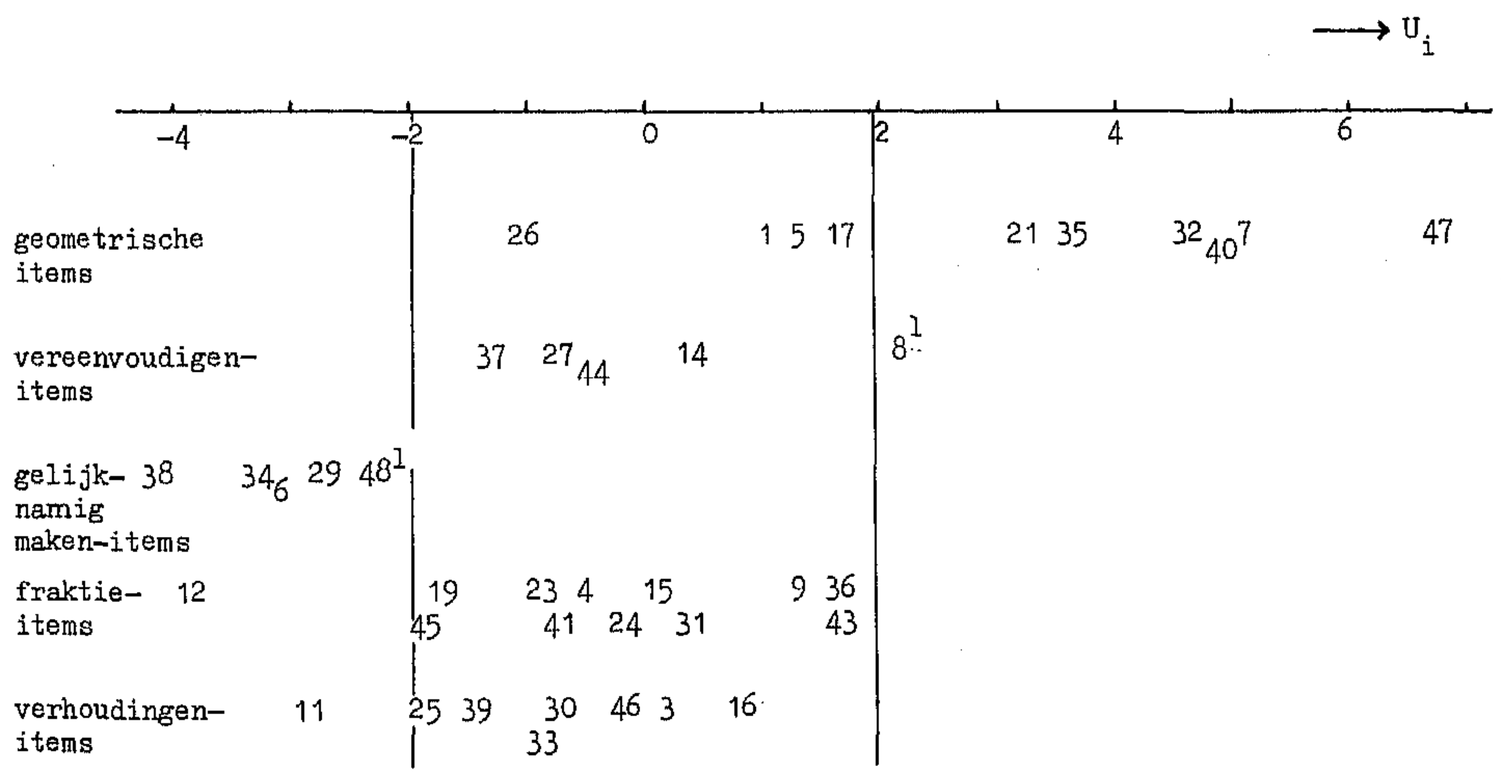
den Wollenberg (1982a) worden genoemd. Q_1 is een globale toets voor monotonie en afdoendheid, en Q_2 is een globale toets voor eendimensionaliteit en lokaal stochastische onafhankelijkheid. Met behulp van Q_1 kan tevens informatie op itemniveau worden verkregen (Van den Wollenberg, 1982a, p. 135). Q_2 heeft een paar technische onvolkomenheden (Van den Wollenberg, 1982a, p. 139). Een nadeel voor de toepassing van Q_2 is dat de itemparameters in alle scoregroepen apart moeten worden geschat, hetgeen in iedere scoregroep veel waarnemingen vereist. Samenvoeging van scoregroepen lijkt dit praktische probleem te ondervangen (Van den Wollenberg, 1982b, p. 49).

EEN EMPIRISCH VOORBEELD VAN SCHAALCONSTRUCTIE VOLGENS DE ONDERZOEKSCYCLUS

De empirische bruikbaarheid van de hier voorgestelde cyclus werd onderzocht met behulp van een verzameling van 40 breukrekenitems. De verzameling bevat uiteenlopende soorten van items (bijlage 1) waarvoor waarschijnlijk verschillende vaardigheden zijn vereist. Zo wordt bij de geometrische items bijvoorbeeld expliciet een beroep gedaan op het geometrisch inzicht

van leerlingen. Bij de andere items lijkt een goed geometrisch inzicht in mindere mate vereist. Een kanttekening hierbij is dat sommige leerlingen een numeriek gepresenteerde breuk misschien cognitief representeren als een geometrische figuur (zie bijvoorbeeld Greeno, 1976). Vanwege de veronderstelde meerdimensionaliteit van de vaardigheid in breukrekenen, valt te verwachten dat de verzameling van 40 items als geheel niet Rasch-homogeen is.
De items werden door 360 leerlingen gemaakt. De leerlingen waren afkomstig uit de eerste klas van een LEAO-school en brugklassen op MAVO/HAVO- en HAVO/Atheneum-niveau. De items werden in open-vraag-vorm gepresenteerd, zodat gissen zoveel mogelijk werd voorkomen. Verder waren items van hetzelfde type over de test verspreid om leereffecten zoveel mogelijk te vermijden. Om dezelfde reden werden groepjes van steeds zes à zeven items in aselecte volgorde aan de leerlingen voorgelegd.
In eerste instantie is onderzocht of de 40 items een schaal vormen in de zin van het Rasch model. Daarbij is gebruik gemaakt van de methoden en de cyclus die in de voorgaande twee paragrafen werden besproken. De analyses zijn uitgevoerd met behulp van het computerprogramma PML (Gustafsson, 1979).

Analyse van 40 breukrekenitems
De toets van Andersen op twee groepen (stap 1 in de cyclus) – één met ruwe scores boven de mediaan en één met ruwe scores onder de mediaan – leidt tot verwerping van de nulhypothese van gelijke itemparameters: $\chi^2 = 234.6$, df = 39 en $p < .001$.
Hieruit blijkt dat de assumpties van monotonie en afdoendheid, of alleen de laatste, voor de verzameling van items als geheel zijn geschonden. In eerste instantie blijken vooral de items met een geometrische presentatievorm bij te dragen aan de schending van de beide assumpties (stap 2 in de cyclus). Over het algemeen hebben deze items een relatief te vlakke item karakteristieke curve, hetgeen tot uitdrukking komt in hoge U_i-waarden (Figuur 2).



Figuur 2: U_i-waarden voor de verzameling van 40 items. Bij een item met $|U_i| > 1.96$ betekent 'l' dat de schending 'lokaal' is; waar geen l wordt vermeld is de gehele curve te steil ($U_i < 0$) of te vlak ($U_i > 0$).

Een interpretatie van dit verschijnsel is dat de items met geometrische presentatievorm in hoge mate een beroep doen op één of meer andere latente trekken dan de overige items. Verder valt op dat de gelijknamig maken – items relatief te steile item karakteristieke curven hebben. Gezien de grote negatieve U_i-waarden lijkt het echter verstandiger deze voorlopig te handhaven omdat ze relatief sterk discrimineren tussen personen.
Op grond van de resultaten werd besloten de 10 items met geometrische presentatievorm (in het vervolg: geometrische items) en de overige 30 items met numerieke presentatievorm (in het vervolg: numerieke items) apart te onderzoeken.

Analyse van de geometrische items
Monotonie en afdoendheid. Ofschoon de globale toets van Andersen (stap 1) geen duidelijke verwerping van de assumpties van monotonie en afdoendheid oplevert ($\chi = 18.9$, df = 9 en p = .03), kan de waarde van de toetsingsgrootheid aanzienlijk worden verkleind door item 5 en 17 uit de schaal verwijderen. Deze twee items hebben binnen de verzameling van geometrische items een relatief te steile item karakteristieke curve (stap 2), hetgeen tot uiting komt in grote negatieve U_i-waarden (tabel 2). Een inhoudelijk a posteriori argument voor de verwijdering van item 5 en 17 is dat bij beide het antwoord kan worden gevonden door het aantal gearceerde en het totale aantal vierkanten te tellen. Behalve een 'tel'-operatie is bij de overige items vereist dat men de geometrische figuur in congruente deelfiguren indeelt voordat men het antwoord kan geven. Item 5 en 17 worden op grond van dit inhoudelijke argument uit de schaal verwijderd. De geometrische items zullen in vervolgonderzoek met behulp van het lineair logistisch model (zie bijvoorbeeld Fischer, 1974) worden onderzocht. Bij deze toepassing is het gewenst dat trekken zo zuiver mogelijk worden gemeten. Op grond van alleen psychometrische argumenten zouden de relatief sterk discriminerende items 5 en 17 hier niet worden verwijderd.
De globale toets van Andersen (stap 1) leidt tot de conclusie dat de assumpties van monotonie en afdoendheid voor de resterende acht items niet worden geschonden ($\chi^2 = 5.3$, df = 7 en p = .63). De U_i-waarden (Tabel 2) vertonen geen belangrijke afwijkingen.

Tabel 2.
U_i-waarden binnen de verzameling van alle geometrische items, en na weglating van item 5 en 17. Een 'l' en een 'g' geven aan dat een schending 'lokaal' respectievelijk 'globaal' is op het onderzochte interval van de latente trek.

itemnummer	U_i	U_i
7	2.1^l	1.0
21	.4	–1.1
32	2.1^l	.7
40	1.4	1.7
5	-3.5^g	
17	-3.1^g	
26	–1.9	-2.2^l
47	1.2	– .3
1	2.3^l	.6
35	– .2	– .5

Tabel 3.
Resultaten van de toets van Andersen voor vier splitter-items, gekozen uit de acht resterende geometrische items van Tabel 2.

splitteritem	1	47	7	32
populariteit	.68	.66	.50	.44
χ^2	14.5	7.8	6.3	8.8
df	6	6	6	6
p	.02	.26	.39	.19

Eendimensionaliteit. De keuze van splitteritems voor de toets van Andersen (stap 3) is bij voorkeur gebaseerd op een hypothese over een indeling van de items in inhoudelijk homogene deelgroepen. De deelgroepen bevatten hier respectievelijk item 7, 21, 32 en 40 (arceer een deel van de figuur), item 26 en 47 (bepaal welk deel van de figuur gestreept is) en item 1 en 35 (bepaal de verhouding van twee gestreepte delen). Verder mag de populariteit van splitteritems niet extreem groot of klein zijn, opdat de nauwkeurigheid van de schattingen van de itemparameters in beide groepen bevredigend is.
Voor de toets van Andersen werden twee groepen proefpersonen gevormd met respectievelijk $X_i = 0$ (splitteritem i fout) en $X_i = 1$ (splitteritem i goed). De resultaten van de toetsen volgens vier verschillende splitteritems – de helft van de items uit ieder groepje – staan in Tabel 3 en de conclusie is dat de assumptie van eendimensionaliteit en daarmee lokaal stochastische onafhankelijkheid niet ernstig is geschonden.
De algemene conclusie is dat de gegevens van de resterende acht geometrische items geen ernstige schendingen van de assumpties van het Rasch model vertonen en derhalve als Rasch-homogeen mogen worden opgevat.

Analyse van de numerieke items
Monotonie en afdoendheid. Uit de toets van Andersen (stap 1) op een lage en een hoge scoregroep blijkt dat de assumpties van monotonie en afdoendheid of alleen de laatste voor de verzameling van 30 numerieke items zijn geschonden: $\chi^2 = 74.5$ df $= 29$ en $p < .001$. Op grond van de lokale toets voor monotonie en afdoendheid (stap 2) kan men zien dat de gelijknamig maken-items evenals bij de analyses op 40 items als afwijkende groep naar voren komen (Tabel 4), hetgeen waarschijnlijk kan worden opgevat als een indicatie van meerdimensionaliteit.
Van de fractie-items hebben o.a. item 9 en 43 (lokaal) te vlakke item karakteristieke curven. Uit retrospectieve analyses van gegevens van 19 leerlingen kan worden afgeleid dat de door hen gebruikte oplossingsprocessen verschillen voor item 9, 36 en 43 enerzijds en item 4, 15 en 24 anderzijds. Naar aanleiding hiervan worden item 9, 36 en 43 uit de verzameling van items verwijderd, hoewel items 36 naar U_i-waarde wel in de verzameling past. Van de overige items met hoge of lage U_i-waarden komt item 8 het eerst voor verwijdering in aanmerking, ofschoon er hier geen inhoudelijke redenen voor kunnen worden gegeven.
Na verwijdering van item 8, 9, 36 en 43 voldoen de overige 26 numerieke items globaal gezien (stap 1) aan de assumpties van monotonie en afdoendheid: $\chi^2 = 38.1$, df $= 25$ en $p = .05$, ofschoon enkele items nog wel afwijken (stap 2) van de overige (Tabel 4). Om kanskapitalisatie zoveel mogelijk te voorkomen, worden verder geen items meer op grond van hun U_i-waarden verwijderd.

Tabel 4.
U_i-waarden binnen de verzameling van 30 numerieke items, en na weglating van item, 8, 9, 36 en 43. De betekenis van 'l' en 'g' is dezelfde als in Tabel 3.

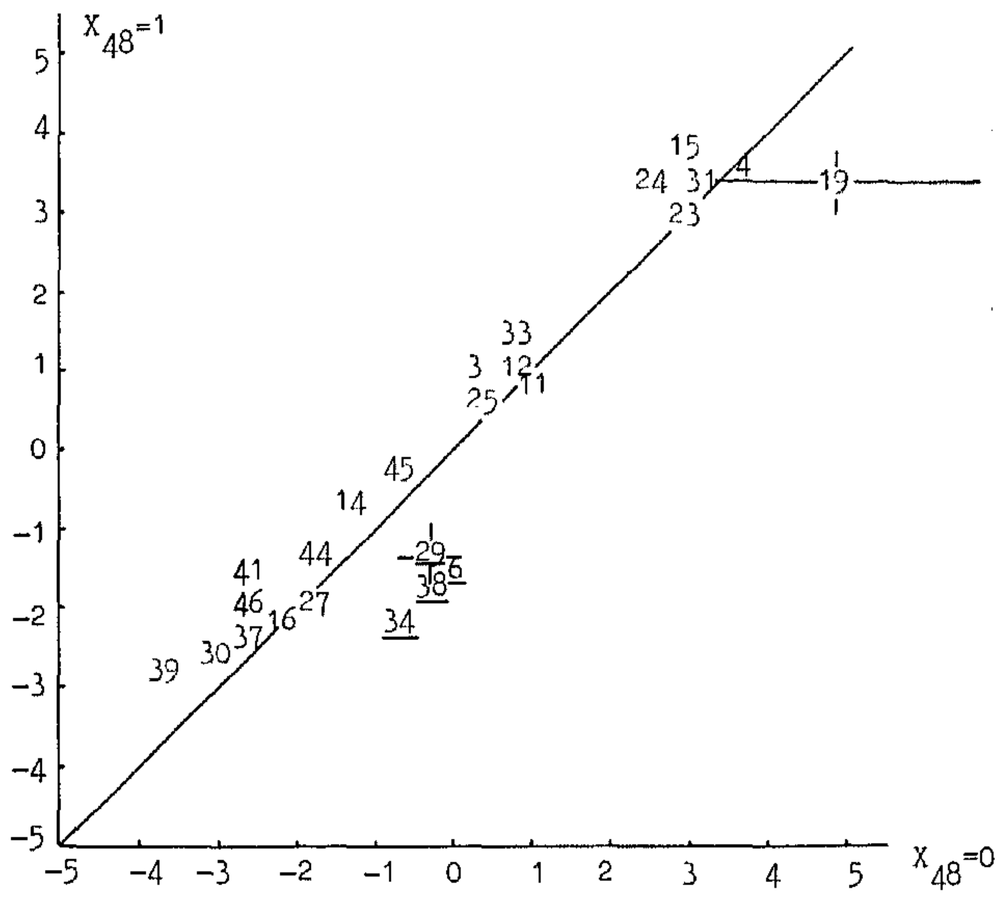
	itemnummer	U_i	U_i
vereenvoudigen-items	8	5.4[g]	
	14	1.9	3.8[g]
	27	− .1	.4
	37	− .7	− .7
	44	− .3	− .4
gelijknamig maken-items	6	−2.5[g]	−2.2[l]
	29	−2.3[g]	−2.1[l]
	34	−2.3[l]	− .3
	38	−2.8[g]	−2.6[g]
	48	−1.2	−1.5
fractie-items	4	.1	− .1
	9	3.0[g]	
	15	.5	1.3
	24	− .1	− .0
	36	1.0	
	43	2.6[l]	
	12	−2.7[g]	− .8
	19	− .9	− .5
	23	− .5	.9
	31	6.1[l]	1.0
	41	.9	2.1[g]
	45	.2	.7
verhoudingen-items	3	3.1[l]	4.8[l]
	11	−1.6	−1.6
	16	1.3	.8
	25	1.2	.8
	30	− .1	− .1
	33	2.3[l]	3.2[l]
	39	.2	.1
	46	.3	1.4

Eendimensionaliteit. Bij de gelijknamig maken-items werden al eerder (Figuur 2 en Tabel 4) aanwijzingen gevonden dat ze een aparte latente dimensie meten. Voor het onderzoek naar eendimensionaliteit wordt behalve van dit itemtype ook van de andere itemtypen steeds één representant als splitteritem genomen (stap 3). De resultaten van de toetsen staan in Tabel 5. De nulhypothese wordt voor de indeling op item 14 (vereenvoudigen), item 48 (gelijknamig maken) en item 25 (verhoudingen) duidelijk verworpen. De indeling op item 12 (fracties) levert een minder duidelijke verwerping van de nulhypothese op dan bij de andere drie splitteritems.
In Figuur 3 zijn de schattingen van de itemparameters in de groepen met item 48 fout en item 48 goed tegen elkaar afgezet (stap 4). De vier gelijknamig maken-items zijn in de groep met item 48 fout aanzienlijk moeilijker dan in de andere groep, hetgeen betekent dat deze items relatief sterk met het splitteritem samenhangen. De moeilijkheid van item 19 (fracties) is in de groep met $X_{48} = 0$ erg onnauwkeurig geschat, zodat aan de positie van dit item niet teveel waarde moet worden gehecht.

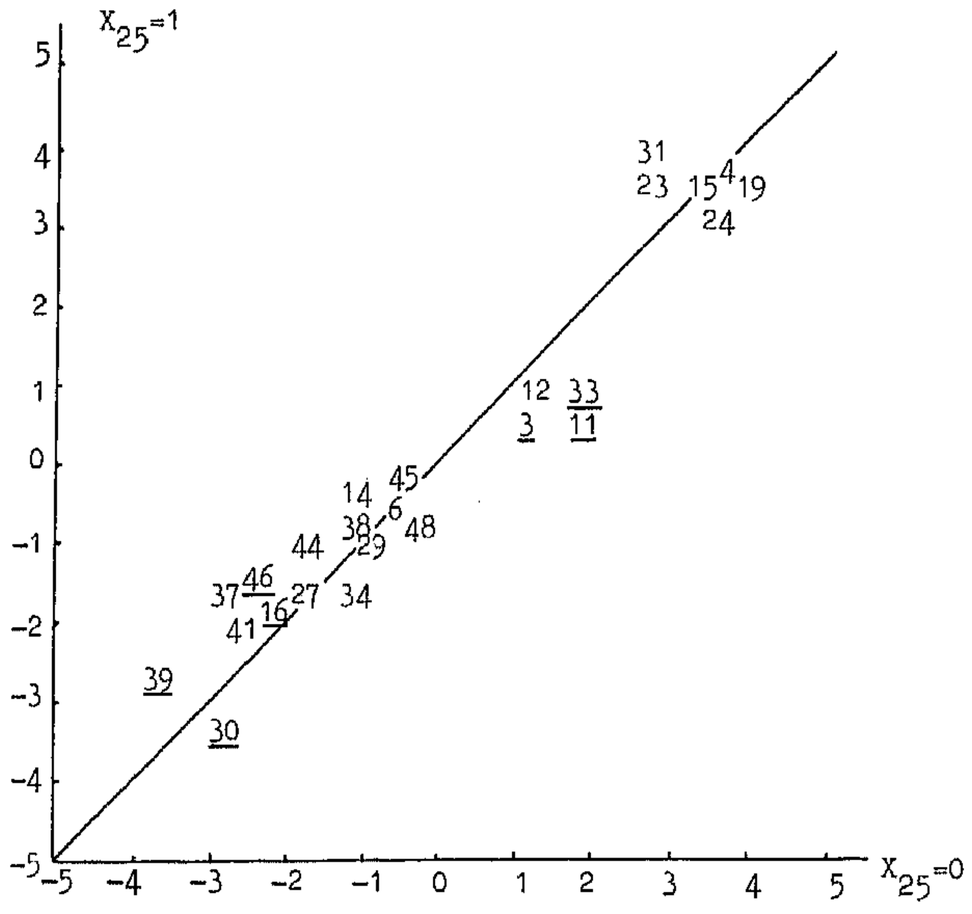
Tabel 5.
Resultaten van de toets van Andersen voor splitteritems van ieder item-type.

splitteritem	14	48	12	25
populariteit	.66	.59	.39	.46
χ^2	63.0	124.2	45.1	76.2
df	24	24	24	24
p	.00	.00	.01	.00

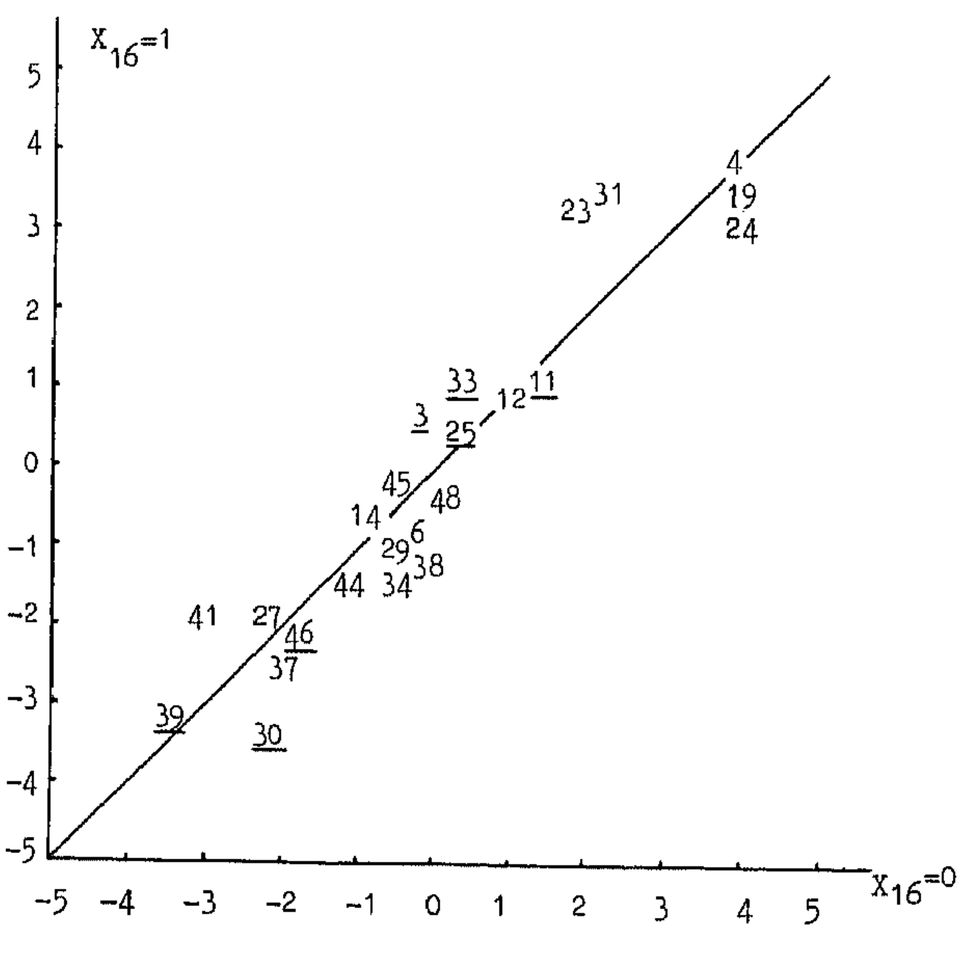
Voor de overige drie splitteritems zijn eveneens de schattingen van de itemparameters in de deelgroepen in een grafiek tegen elkaar afgezet (stap 4). Bij de indeling op item 14 (vereenvoudigen) komen de vereenvoudigen-items als inhoudelijk homogene deelgroep naar voren. Bij de indeling op item 12 (fracties: a/b deel van c is . . .) kunnen de fractie-items met dezelfde presentatievorm als homogene deelgroep worden onderscheiden van de overige items. De fractie-items met de vorm 'a/b is het . . .deel van c/d' lijken relatief minder goed in deze deelgroep te passen.
Voor de indeling op item 25 (verhoudingen) is het beeld minder duidelijk. In Figuur 4 zijn de schattingen van de itemparameters in de groepen met respectievelijk item 25 fout en item 25 goed tegen elkaar afgezet (stap 4). De verhoudingen-items 3, 11 en 33 hangen het sterkst met item 25 samen; de geschatte moeilijkheid van item 30 heeft een relatief grote schattingsfout en hoort derhalve niet bij dit groepje. Item 3, 11, 25 en 33 hebben gemeen dat de deling van de



Figuur 3: Grafiek van σ_i voor de groepen met item 48 fout ($X_{48} = 0$) en item 48 goed ($X_{48} = 1$). Voor item 19 en 29 zijn de asymptotische 95% betrouwbaarheidsintervallen aangegeven. De nummers van de items van het type gelijknamig maken zijn onderstreept.



Figuur 4: Grafiek van σ_i voor de groepen met item 25 fout ($X_{25} = 0$) en item 25 goed ($X_{25} = 1$). De nummers van de items van het type verhoudingen zijn onderstreept.



Figuur 5: Grafiek van σ_i voor de groepen met item 16 fout ($X_{16} = 0$) en item 16 goed ($X_{16} = 1$). Item 15 ontbreekt vanwege een minimale itempopulariteit in de groep met item 16 fout. De nummers van de items van het type verhoudingen zijn onderstreept.

tellers (item 11 en 25) of de noemers (item 3 en 33) een gebroken getal als uitkomst heeft. Bij de andere verhoudingen-items is deze uitkomst een positief geheel getal. Dit verschil tussen beide groepen items zou van invloed kunnen zijn op het door de leerlingen gevolgde oplossingsproces: Bij de delingen met een gebroken getal als uitkomst zijn er aanwijzingen uit retrospectieve analyses dat leerlingen de bekende breuk vereenvoudigen en dan opnieuw proberen het probleem op te lossen. Uit Figuur 4 blijkt dat item 16, 30, 39 en 46 – deling van de tellers of de noemers levert een positief geheel getal – niet sterker met het splitteritem samenhangen dan de overige items. Van deze vier items diende item 16 als splitteritem voor de toets van Andersen. De nulhypothese van eendimensionaliteit werd in dit geval verworpen ($\chi^2 = 54.6$, df = 23 en $p < .001$). In Figuur 5 zijn de schattingen van de itemparameters in de deelgroepen tegen elkaar afgezet (stap 4). Van de verhoudingen-items hangen item 30 en 46 het sterkst samen met het splitteritem. Deling van tellers of noemers van deze drie items heeft een positief geheel getal als uitkomst. Item 30 en 46 hangen echter niet sterker met het splitteritem samen dan een aanzienlijk aantal andere items, zie Figuur 5. De conclusie is dat de verhoudingen-items zich niet duidelijk als homogene deelgroep van de overige items onderscheiden.

De resultaten van het onderzoek naar eendimensionaliteit van de numerieke items zijn de volgende: De gelijknamig maken-, vereenvoudigen- en fractie-items met de vorm 'a/b deel van c is . . .' schenden de assumptie van eendimensionaliteit niet in ernstige mate. De overige drie fractie-items zijn om inhoudelijke redenen aan de schaal toegevoegd, hoewel daarmee een schending van de assumptie van eendimensionaliteit wordt veroorzaakt. De fractieschaal zal in vervolgonderzoek echter niet nader worden onderzocht met behulp van het lineair logistisch model, omdat de items voor de onderzochte populatie te moeilijk zijn (Bijlage 1). De verhoudingen-items zijn waarschijnlijk niet eendimensioneel, maar worden wel als groep verder geanalyseerd aangezien er geen inhoudelijk zinvolle onderverdeling van deze items in eendimensionele metingen mogelijk is.

Analyse van de deelgroepen van numerieke items

Monotonie en afdoendheid. Voor iedere deelgroep van numerieke items werd de toets van Andersen op een hoge en een lage scoregroep berekend (stap 1). De resultaten staan in Tabel 6 en men kan er uit afleiden dat de assumpties van monotonie en afdoendheid niet ernstig zijn geschonden voor de groepen van items als geheel.

De algemene conclusie van de analyses volgens het Rasch model en de onderzoekscyclus is dat de geometrische items en de vereenvoudigen- en gelijknamig maken-items kunnen worden opgevat als Rasch-homogene tests. Bij de fractie- en verhoudingen-items is de assumptie van eendimensionaliteit in geringe mate geschonden. Behalve voor de fractie-items, zal voor de overige itemtypen in vervolgonderzoek worden nagegaan welke deelvaardigheden nodig zijn

Tabel 6.
Resultaten van de toets van Andersen voor de vier deelgroepen van numerieke items.

	χ^2	df	p
vereenvoudigen	6.5	3	.09
gelijknamig maken	3.9	4	.42
fracties	20.8	8	.01
verhoudingen	12.8	7	.08

bij het maken van de items. Dit onderzoek zal plaatsvinden met behulp van het lineair logistisch model. Voor de gelijknamig maken-, de vereenvoudigen- en de verhoudingen-items is hiermee reeds een begin gemaakt (Sijtsma, 1982).

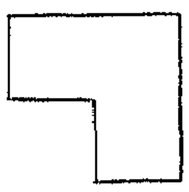
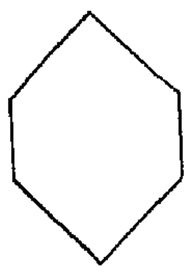
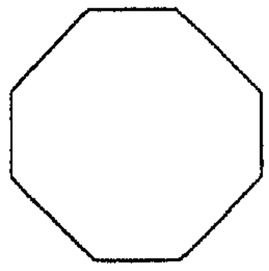
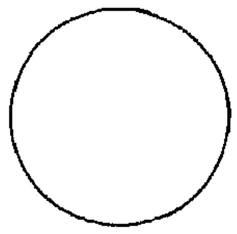
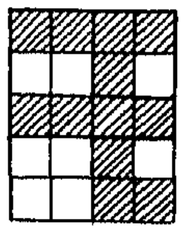
DISCUSSIE

De methoden die in dit onderzoek zijn gebruikt om te onderzoeken of empirische gegevens voldoen aan het Rasch model, lijken gecombineerd goed te voldoen aan het doel van een itemanalyse, het vinden van Rasch-homogene groepen van items. In dit artikel is de onderzoekscyclus toegepast op een verzameling van breukrekenitems. De bruikbaarheid van de cyclus is ook gebleken bij andere soorten van items. Kok[1] (1982) construeerde met behulp van de cyclus een aantal Rasch-homogene schalen uit items van een persoonlijkheidsvragenlijst. Het is echter geenszins de bedoeling om te suggereren dat itemanalyse volgens het Rasch model altijd op de hier voorgestelde manier moet plaatsvinden. De bedoeling van het artikel is vooral om een illustratie te geven van een itemanalyse volgens het Rasch model, zodanig dat ook werkelijk alle assumpties van het model op empirische gegevens worden onderzocht.
Wanneer men Rasch-homogene groepen van items heeft gevonden, beschikt men echter nog niet in alle gevallen over bruikbare tests. Wanneer men bijvoorbeeld tests zou willen gebruiken voor het nemen van individuele beslissingen in selectiesituaties, is het een vereiste dat de test goed discrimineert rond de aftestgrens. Hiervoor biedt een Rasch-homogene test echter geen garanties (Wood, 1978; Molenaar, 1982a, p. 24, 25; 1982b, p. 177). Het lijkt daarom verstandig om naast het onderzoek naar de assumpties van het Rasch model, eveneens populatie-afhankelijke indices in beschouwing te nemen, zoals item-rest correlaties en betrouwbaarheidschattingen. In dit onderzoek hebben bijvoorbeeld de geometrische items ($k = 8$) een coëfficiënt alpha van .47 en de gelijknamig maken-items ($k = 5$) een coëfficiënt alpha van .89. Wanneer men de test zodanig zou willen verlengen dat ze voor selectie-onderzoek een betrouwbaarheid van .95 verkrijgen, dan moet de geometrische test volgens de Spearman-Brown formule uit 172 items bestaan en de gelijknamig maken-test uit 12 items. De eerste test is vanwege de lengte ongeschikt om in de praktijk te gebruiken. De tweede test is na verlenging op psychometrische gronden wel geschikt, maar men kan zich afvragen of de inhoud van de test niet te eenzijdig is voor praktische toepassingen. Voor theoretisch onderzoek, zoals bijvoorbeeld naar deelvaardigheden die leerlingen toepassen bij het maken van breukrekenitems, lijken de gevonden schalen, eventueel na verlenging, wel geschikt. De betrouwbaarheidseisen zijn hier minder streng dan bij praktische toepassingen.
In het vorige hoofdstuk bleek al dat itemselectie volgens het Rasch model, zoals hier wordt voorgesteld, niet enkel en alleen volgens formele criteria moet geschieden. Inhoudelijke argumenten dienen steeds een rol te spelen naast formele argumenten. Verder lijkt het nodig om met het oog op praktische toepassingen de samenhang tussen de items over personen te beschouwen (zie bijvoorbeeld ook Molenaar, 1982b, p. 178). Deze laatste eis impliceert het gebruik van populatie-afhankelijke indices naast het populatie-onafhankelijke Rasch model.

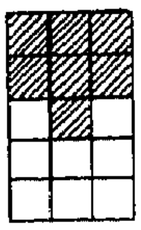
---

1. Interne notitie. Publicatie volgt in 1983.

Bijlage 1: De items die in dit onderzoek werden gebruikt en de itempopulariteiten.

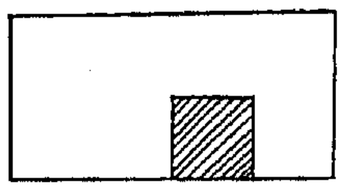
	itempopulariteit
GEOMETRISCHE ITEMS	
7. Maak met behulp van een potlood $\frac{1}{4}$ deel van deze figuur zwart. 	.50
21. Maak met behulp van een potlood $\frac{1}{8}$ deel van deze figuur zwart. 	.73
32. Maak met behulp van een potlood $\frac{1}{7}$ deel van deze figuur zwart. 	.44
40. Maak met behulp van een potlood $\frac{5}{8}$ deel van deze figuur zwart. 	.74
5. Deze figuur is in gelijke stukken verdeeld. Welk deel van deze figuur is gestreept?  Antwoord:	.73

17. Deze figuur is in gelijke stukken verdeeld. Welk deel van deze figuur is gestreept? .78



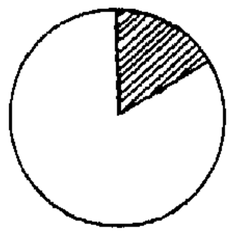
Antwoord: .....

26. Welk deel van deze figuur is gestreept? .84



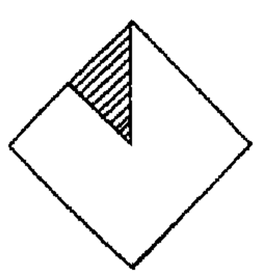
Antwoord: .....

47. Welk deel van deze figuur is gestreept? .66

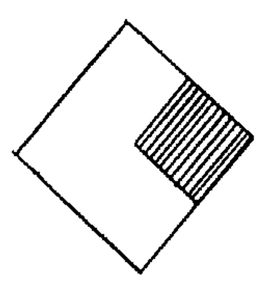


Antwoord: .....

1. Het gestreepte deel van figuur B is ... maal zo groot als het gestreepte deel van figuur A. .68

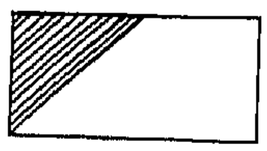


A

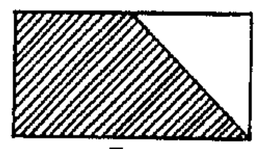


B

35. Het gestreepte deel van figuur B is ... maal zo groot als het gestreepte deel van figuur A. .37



A



B

VEREENVOUDIGEN - ITEMS

Vereenvoudig de volgende breuk zoveel mogelijk.

8.	$\frac{18}{45}$ =	.61
14.	$\frac{30}{54}$ =	.66
27.	$\frac{6}{36}$ =	.80
37.	$\frac{12}{15}$ =	.84
44.	$\frac{42}{48}$ =	.74

GELIJKNAMIG MAKEN - ITEMS

Maak de volgende twee breuken gelijknamig.

6.	$\frac{5}{6}$ en $\frac{3}{10}$ Antwoord: ..... en	.66
29.	$\frac{3}{4}$ en $\frac{3}{7}$	.66
34	$\frac{11}{14}$ en $\frac{5}{7}$	.72
38.	$\frac{5}{8}$ en $\frac{2}{5}$	.62
48.	$\frac{4}{9}$ en $\frac{5}{12}$	.59

FRAKTIE - ITEMS

Vul op de puntjes een <u>breuk</u> in en vereenvoudig het antwoord <u>zoveel</u> mogelijk.

4.	$\frac{1}{4}$ is het ... deel van $\frac{3}{5}$	.08
9.	4 is het ... deel van 10	.30
15.	$\frac{2}{3}$ is het ... deel van 3	.09
24.	2 is het ... deel van $3\frac{1}{2}$	.10

36. $\frac{2}{5}$ is het ... deel van 4	.18
43. $\frac{1}{2}$ is het ... deel van 2	.47

Vul op de puntjes het antwoord in. Wanneer het antwoord een breuk bevat, vereenvoudig dan het antwoord zoveel mogelijk.

12. $\frac{2}{6}$ deel van 3 is ...	.39
19. $\frac{5}{8}$ deel van 5 is ...	.09
23. $\frac{2}{7}$ deel van 9 is ...	.12
31. $\frac{2}{9}$ deel van 5 is ...	.10
41. $\frac{1}{4}$ deel van 20 is ...	.82
45. $\frac{2}{3}$ deel van 15 is ...	.57

VERHOUDINGEN - ITEMS

Vul op de puntjes het ontbrekende getal in.

3. $\frac{..}{6} = \frac{10}{15}$	.42
11. $\frac{4}{..} = \frac{6}{9}$	.39
16. $\frac{..}{28} = \frac{4}{7}$	.81
25. $\frac{6}{12} = \frac{4}{..}$	.46
30. $\frac{4}{9} = \frac{..}{27}$	.87
33. $\frac{8}{12} = \frac{..}{9}$	.37
39. $\frac{2}{5} = \frac{4}{..}$	.90
46. $\frac{15}{..} = \frac{5}{6}$	.82.

LITERATUUR

Andersen, E.B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140.

Fischer, G.H. *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen.* Bern: Hans Huber, 1974.

Formann, A.K. Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch. *Zeitschrift für experimentelle und angewandte Psychologie*, 1981; *28*, 541-560.

Greeno, J.G. Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In: D. Klahr (Red.), *Cognition and instruction*. New York: John Wiley & Sons, Inc., 1976.

Gustafsson, J.E. 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*, nr. 85, 1979.

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

Molenaar, I.W. Some improved diagnostics for failure of the Rasch model. *HB-80-482-EX*, Vakgroep statistiek en meettheorie, FSW, Rijksuniversiteit Groningen, 1980. (Te verschijnen in Psychometrika, 1983).

Molenaar, I.W. Mensen die het beter meten. *Kwantitatieve Methoden*, 1982a, *3*, nr. 5, 3-29.

Molenaar, I.W. Een tweede weging van de Mokkenschaal. *Tijdschrift voor Onderwijsresearch*, 1982b, *7*, 172-181.

Sijtsma, K. Een lineair logistisch model ter verklaring van de moeilijkheidsparameters van breukrekenitems. In: F.G.L.C. Lodewijks & P.R.F. Simons (Red.), *Strategieën in leren en ontwikkeling*. Lisse: Swets & Zeitlinger, 1982.

Stelzl, I. Ist der Modelltest des Rasch-Modells geeignet Homogenitätshypothesen zu prüfen? Ein Bericht über Simulationsstudien mit inhomogenen Daten. *Zeitschrift für experimentelle und angewandte Psychologie*, 1979, *26*, 652-672.

Wollenberg, A.L. van den. *The Rasch model and time-limit tests. An application and some theoretical contributions*. (dissertatie). Nijmegen: Stichting Studentenpers Nijmegen, 1979.

Wollenberg, A.L. van den. Two new test statistics for the Rasch model. *Psychometrika*, 1982a, *47*, 123-140.

Wollenberg, A.L. van den. On the applicability of the Q_2 test for the Rasch model. *Kwantitatieve Methoden*, 1982b, *3*, nr. 5, 30-55.

Wood, R. Fitting the Rasch model – A heady tale. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 27-32.

Wright, B.D. & Stone, M.H. *Best test design. Rasch measurement*. Chicago: MESA Press, 1979.

Manuscript ontvangen 26-5-1982
Definitieve versie ontvangen 7-1-1983