Maastricht University

Het gebruik van einddoeltoetsen bij aanvang van de studie

Citation for published version (APA):

Imbos, T. (1989). *Het gebruik van einddoeltoetsen bij aanvang van de studie*. [Doctoral Thesis, Maastricht University]. Rijksuniversiteit Limburg. https://doi.org/10.26481/dis.19890101ti

Document status and date:
Published: 01/01/1989

DOI:
10.26481/dis.19890101ti

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

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.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 18 Apr. 2024

HET GEBRUIK VAN EINDDOELTOETSEN BIJ AANVANG VAN DE STUDIE

HET GEBRUIK VAN EINDDOELTOETSEN BIJ AANVANG VAN DE STUDIE

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de

Rector Magnificus, Prof.Dr. F.I.M. Bonke,

volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op
vrijdag, 17 februari 1989 om 16.00 uur

door

Tjaart Imbos

geboren te Sappemeer in 1946

Promotor:	Prof.Dr. W.H.F.W. Wijnen
Co-promotor:	Dr. R.J.M.M. Does
Beoordelingscommissie:	Prof.Dr. M.J. Drop, RL (voorzitter)
	Prof.Dr. C.P.A. van Boven, RL
	Prof.Dr. W.K.B. Hofstee, RUG
	Prof.Dr. G.J. Mellenbergh, UvA
	Prof.Dr. H.G. Schmidt, RL

Colofon:
Omslag: De alternerende boom van Pythagoras als model voor de ontwikkeling van Medische Kennis.
Lay-out: Michel Janssen.
Druk: Drukkerij Alberts/Druko, Gulpen.
ISBN: 90-9002697-5

Establishing construct validity is an ongoing process. Based on current theory regarding the trait being measured, the test developer makes predictions about how the test scores should behave in various situations.

M.J. Allen en W.M. Yen, 1979.

HET GEBRUIK VAN EINDOELTOETSEN BIJ AANVANG VAN DE STUDIE

Ter nagedachtenis van:
H.W. Imbos (1907-1975)
E. Imbos-Kraai (1913-1987)
T. Kraai (1907-1987)

-ze hadden er zo graag bij willen zijn-

Voor:
Nelly, Nienke, Sanne en Jelle

Omdat het ook jullie tijd was!

INHOUDSOPGAVE

	Voorwoord	1
1.	**Einddoeltoetsen in het eerstejaar? Probleemanalyse**	3
1.1	Inleiding	4
1.2	De rol van theorie in empirisch onderzoek: een keuze	4
1.3	Begripsvalidatie van toetsen	7
1.4	Het begrip-zoals-bedoeld	8
1.4.1	Medische kennis	8
1.4.2	De ontwikkeling van medische kennis	9
1.4.3	Implicaties voor het toetsen van medische kennis	11
1.5	Het begrip-zoals-bepaald; de voortgangstoets en de bloktoets	11
1.6	De voortgangstoets in het eerstejaar	13
1.7	Vraagstellingen van het onderzoek	14
1.8	Werkwijze, items, respondenten	16
1.9	Verdere empirische afbakening	16
1.10	Hypothesen voor het empirische domein	18
1.10.1	Hypothesen over verschillen in interne structuur	18
1.10.2	Hypothesen over verschillen in externe structuur	22
1.10.3	Psychometrische modellen als hypothese	24
1.11	Werkwijze bij het toetsen van de hypothesen	26
1.12	Samenvatting	28
2.	**Selectie van eerstejaarsitems**	29
2.1	Inleiding	30
2.2	Selectie van Eerstejaarsitems (Eerstejaarstoetsen)	31
2.2.1	Procedure	31
2.2.2	Resultaten beoordelingsprocedure	32
2.3	Selectieprocedure van controle items (Algemene Kennistoetsen)	35
2.4	Beschrijving Eerstejaarstoetsen	36
2.4.1	Inhoudelijke beschrijving	36
2.4.2	Relevantie van de Eerstejaarstoetsen	42
2.4.3	Verdeling juist- en onjuistvragen in beide toetsen	43
2.5	Inhoudelijke itemselectie vergeleken met empirische itemselectie	44
2.5.1	Vergelijking naar aantal items	44
2.5.2	Vergelijking naar moeilijkheidsgraad van de items	45
2.6	Samenvatting en conclusies	49

3.	**Vergelijking Eerstejaarstoetsen met Algemene kennistoetsen (interne analyses)**	53
3.1	Inleiding	54
3.2	Resultaten interne analyses	55
3.2.1	Verdelingen van itembeantwoording (hypothese 1)	55
3.2.2	Verdelingen van p-waarden (hypothese 2)	58
3.2.3	De variantie in p-waarden (hypothese 3)	60
3.2.4	Verdelingen van itemrestcorrelaties (hypothese 4)	62
3.2.5	Betrouwbaarheid van de toetsen (hypothese 5)	64
3.2.6	Verdelingen van toetsscores (hypothese 6)	66
3.2.7	Variantie in toetsscores (hypothese 7)	69
3.2.8	Itemvariantie en personenvariantie (hypothese 8)	71
3.2.9	Onderwijseffecten in Eerstejaarstoetsen (hypothese 9)	73
3.3	Samenvatting en conclusie	76
4.	**Vergelijking Eerstejaarstoetsen met Algemene Kennistoetsen (externe analyses)**	79
4.1	Inleiding	80
4.2	Het toetsen van hypothesen over correlaties (hypothesen 10a en 10b)	81
4.3	Statistische modellen	81
4.3.1	Simplex modellen	81
4.4	Resultaten	85
4.4.1	Correlaties tussen Eerstejaarstoetsen en Algemene Kennistoetsen	85
4.4.2	Twee toetsen, één simplex voor de latente variabelen	87
4.4.3	Conclusie	88
4.4.4	Parameterschattingen in de simplex modellen	89
4.4.5	Kennis- en toetsmethode variantie	93
4.4.6	Samenvatting en conclusies (hypothesen 10a en 10b)	94
4.5	Voorspellen van studieprestaties (hypothese 11)	95
4.5.1	Inleiding	95
4.5.2	Methode	95
4.5.3	Resultaten	96
4.5.4	Samenvatting en conclusies (hypothese 11)	99
4.6	Medische kennis in het eerstejaar (hypothese 12)	100
4.6.1	Inleiding	100
4.6.2	Resultaten	100
4.6.2.1	Correlaties tussen bloktoetsen, Eerstejaarstoetsen en Algemene Kennistoetsen	100

4.6.3	Samenvatting en conclusies (hypothese 12)	102
4.7	Samenvatting en conclusies (externe analyses)	102
5.	**Modellen als hulptheorie voor medische kennis**	105
5.1	Inleiding	106
5.2	De interactie tussen personen en items	107
5.3	Modellen voor het beantwoorden van toetsitems	110
5.3.1	De Guttmanschaal als basismodel	110
5.3.2	Het VGT-model (hypothese 13)	113
5.3.3	Het Raschmodel (hypothesen 14a, 14b en 14c)	115
5.3.4	Het Mokkenmodel (hypothesen 15a en 15b)	117
5.4	Procedure	119
5.4.1	De personen	119
5.4.2	De items	119
5.4.3	De statistische procedures	120
5.4.3.1	Procedures voor het VGT-model	120
5.4.3.2	Procedures voor het **Ras**chmodel	121
5.4.3.3	Procedures voor het Mokkenmodel	124
5.5	Samenvatting	125
6.	**Resultaten van het toetsen van de modellen**	127
6.1	Inleiding	128
6.2	Toetsing van het VGT-model (hypothese 13)	128
6.3	Samenvatting	135
6.4	Toetsing van het Raschmodel (hypothesen 14a, 14b en 14c)	136
6.5	Samenvatting	147
6.6	Toetsing van het Mokkenmodel (hypothesen 15a en 15b)	148
6.7	Samenvatting	156
6.8	Correlaties tussen Raschitems, Mokkenitems en niet-schaalbare items	156
6.9	Samenvatting	160
6.10	Constructrepresentatie van Medische Kennis	160
6.10.1	Inleiding	160
6.10.2	Grafische representatie van de variabele	161
6.10.3	De relatie tussen inhoud en moeilijkheid van de toetsitems	164
6.11	Samenvatting constructrepresentatie	171
6.12	Samenvatting en conclusies	172

Hoofdstuk 5: Modellen

7.	**Algemene discussie**	175
7.1	Inleiding	176
7.2	Samenvatting van de onderzoeksresultaten	176
7.3	Het gebruik van einddoeltoetsen in het eerstejaar: bespreking van de onderzoeksresultaten	179
7.4	Bespreking onderzoeksresultaten vanuit het cognitief psychologisch perspectief	183
7.5	Bespreking onderzoeksresultaten vanuit het psychometrisch perspectief	185
7.6	Bespreking onderzoeksresultaten vanuit het data-analytisch perspectief	188
7.7	Slotbeschouwing	189
	Bijlage 1	191
	Bijlage 2	193
	Referentielijst	215
	Summary	229
	Dankwoord	233
	Curriculum vitae	235

VOORWOORD

Dit proefschrift gaat over de Maastrichtse Voortgangstoets, met name over het gebruik ervan in het begin van de studie. De voortgangstoets is een instrument om het cognitieve einddoel van de medische studie te meten. De mate waarin de studenten het einddoel beheersen wordt gemeten door vier keer per jaar een voortgangstoets af te nemen. Tijdens deze afnames worden alle studenten getoetst: eerstejaars tot en met zesdejaars. Vanwege de gerichtheid van de voortgangstoets op het einddoel, kan het instrument ook getypeerd worden als een einddoeltoets.
De centrale vraagstelling, waarop in de hier gerapporteerde studie een antwoord wordt gezocht luidt: Is het zinvol om een einddoeltoets bij het begin van de studie te gebruiken?
Dit hier in globale termen geformuleerde probleem is in verschillende bewoordingen en in wisselende gremia naar voren gekomen. Dit kan worden gezien als een illustratie van het belang dat de facultaire gemeenschap hecht aan het beantwoorden van deze vraag.
De mogelijkheid om de vraagstelling aan te pakken deed zich voor toen een aantal docenten met specifieke verantwoordelijkheden voor het onderwijsprogramma in het eerstejaar bereid bleken te zijn om aan het onderzoek mee te werken. Van hun bereidheid om mee te werken is dankbaar gebruik gemaakt. Hun deskundigheid maakte het mogelijk om een eerste deelvraagstelling te beantwoorden namelijk de vraag of de voortgangstoets items bevat, die door eerstejaars studenten, gegeven het onderwijsprogramma te beantwoorden zijn.
Nadat deze vraag in positieve zin beantwoord bleek te kunnen worden was het mogelijk om de vraagstelling verder uit te diepen en toe te spitsen op hypothesen met betrekking tot verschillen en overeenkomsten tussen een groep items, welke door de docenten waren aangewezen en een aselect getrokken steekproef uit de overige items.
De hypothesen werden onderzocht in vier voortgangstoetsen, die in het academisch jaar 1983-1984 waren afgenomen. Afhankelijk van het soort hypothese werden de antwoorden van de deelnemers aan alle vier toetsen gebruikt, dan wel van een aantal groepen (jaarklassen en afgestudeerden) uit de deelnemers.
In de studie komen achtereenvolgens aan de orde:

- hoofdstuk 1: de vraagstelling en deelvraagstellingen, de hypothesen en de gekozen werkwijze.
- hoofdstuk 2: de resultaten van de itembeoordeling door de docenten. De door docenten geselecteerde items (Eerstejaarsitems) worden inhoudelijk vergeleken met de aselect gekozen items (Algemene Kennis items).

- hoofdstuk 3: de resultaten van de hypothesetoetsing met betrekking tot verschillen en overeenkomsten tussen de twee itemgroepen in beantwoording, goede antwoorden, discriminerend vermogen, betrouwbaarheid, toetsscores en onderwijseffecten.
- hoofdstuk 4: de resultaten van de hypothesetoetsing met betrekking tot verschillen tussen beide itemgroepen in correlatiegedrag.
- hoofdstuk 5: drie psychometrische modellen, die in theorie de structuur van de antwoorden van de respondenten kunnen verklaren.
- hoofdstuk 6: de resultaten van analyses met behulp van deze modellen.
- hoofdstuk 7: discussie over de resultaten tegen de achtergrond van de (deel)vraagstelling(en). Nieuw opgekomen vragen worden gesignaleerd en er worden suggesties voor vervolg onderzoek gedaan.

Lezers die een speciale interesse hebben in de wijze waarop de voortgangstoets in de Faculteit der Geneeskunde wordt gebruikt, worden verwezen naar bijlage 2. In deze bijlage komen verschillende aspecten van de voortgangstoets aan de orde: het idee dat aan de ontwikkeling ten grondslag lag, het begrip voortgangstoets, de blauwdruk van de toets, de wijze van samenstellen, de rapportage van de resultaten, de opslag van gegevens en itembanking, de mogelijkheden met voortgangstoetsen, de toepasbaarheid van voortgangstoetsen en tenslotte de kwaliteit ervan als meetinstrument.
Met deze studie hoop ik een bijdrage te hebben geleverd aan een beter begrip van de voortgangstoets, en deze een plaats te hebben gegeven in de onderwijskundige en psychometrische literatuur.
Een belangrijke bron van inspiratie bij het uitvoeren van het onderzoek was de oproep van Hofstee (1971) om studietoetsen niet alleen op inhoudsvaliditeit te beoordelen. Dit is "maar het begin van een stuk werk dat gevoeglijk met construct- of begripsvalidatie kan worden omschreven. Want wie weet is de operationalisering niet gelukt" (blz. 494).
Als dit boek een bijdrage levert aan de begripsvaliditeit van de voortgangstoets, in de door Hofstee bedoelde betekenis, is de opzet geslaagd. De Maastrichtse Voortgangstoets verdient zo'n poging zeer zeker.

1. EINDDOELTOETSEN IN HET EERSTEJAAR? PROBLEEMANALYSE

Inhoud

1.1 Inleiding
1.2 De rol van theorie in empirisch onderzoek: een keuze
1.3 Begripsvalidatie van toetsen
1.4 Het begrip-zoals-bedoeld
1.4.1 Medische kennis
1.4.2 De ontwikkeling van medische kennis
1.4.3 Implicaties voor het toetsen van medische kennis
1.5 Het begrip-zoals-bepaald: de voortgangstoets en de bloktoets
1.6 De voortgangstoets in het eerstejaar
1.7 Vraagstellingen van het onderzoek
1.8 Werkwijze, items, respondenten
1.9 Verdere empirische afbakening van het onderzoek
1.10 Hypothesen over het empirische domein
1.10.1 Hypothesen over verschillen in interne structuur
1.10.2 Hypothesen over verschillen in externe structuur
1.10.3 Psychometrische modellen als hypothese
1.11 Werkwijze bij het toetsen van de hypothesen
1.12 Samenvatting

Hoofdstuk 1: Probleemanalyse

1.1 Inleiding

De voortgangstoets is een einddoeltoets met betrekking tot de kennis waarover een basisarts zou moeten beschikken.
In dit hoofdstuk wordt de centrale probleemstelling van het onderhavige onderzoek uitgewerkt. Deze probleemstelling houdt rechtstreeks verband met het einddoelkarakter van de voortgangstoetsen. De vraag is namelijk of einddoeltoetsen reeds bij de aanvang van de studie informatief zijn. Dat wil zeggen, dat een redelijk deel van de items beantwoordbaar is voor eerstejaars studenten.
Toetsen zijn echter nooit een doel op zich. Ze zijn slechts een middel om informatie te verwerven over iets dat zelf niet observeerbaar is. Uiteindelijk gaat de interesse uit naar deze niet observeerbare theoretische grootheid, of het begrip-zoals-bedoeld. (De Groot, 1968).
Bij de uitwerking van de probleemstelling worden dan ook de relatie tussen het theoretische begrip en het middel om dat begrip meetbaar te maken aan een nadere beschouwing onderworpen.
Pas als deze relatie duidelijk is gemaakt, is het mogelijk om de vraagstellingen en de daarbij behorende hypothesen te ontwikkelen. Hiermede is al een eerste min of meer arbitraire keuze gemaakt namelijk, dat het zinvol is om in het kader van empirisch onderzoek uit te gaan van theoretische begrippen. Niet iedereen is die mening toegedaan. Men raadplege daarvoor Roskam (1982) en Dessens en Jansen (1987).
Onnodig te zeggen dat de gemaakte keuze een nadere argumentatie verdient. Dit gebeurt in de eerste paragraaf, in de tweede paragraaf wordt duidelijk gemaakt dat de keuze leidt tot de opvatting dat studietoetsen aan begripsvalidatie onderworpen dienen te worden (Hofstee, 1971). Daarop volgt een paragraaf waarin het in deze studie beoogde theoretische begrip wordt beschreven, waarna in de volgende paragraaf het meetinstrument voor dat begrip aan de orde komt. Daarna worden de vraagstellingen en de daarbij behorende hypothesen beschreven. Tenslotte volgt de beschrijving van de gekozen werkwijze.

1.2 De rol van theorie in empirisch onderzoek: een keuze

Een probleem, dat momenteel binnen de sociale wetenschappen in de aandacht staat is de relatie tussen theoretische begrippen en empirische begrippen. Hox (1986) en Dessens en Jansen (1987) besteden in hun studies, ieder vanuit hun eigen optiek, aandacht aan het probleem van het operationaliseren van kernbegrippen uit sociaal wetenschappelijke theorieën. Voor een uitvoerige verkenning van dit probleemgebied zij verwezen naar genoemde auteurs. Hier wordt volstaan met het kort weergeven van de voornaamste standpunten terzake.
Globaal genomen kan men met betrekking tot het operationaliseringsprobleem twee

standpunten onderscheiden. Het ene standpunt, door Dessens en Jansen aangeduid als de operationaliseringstraditie, gaat uit van een zorgvuldige analyse van het theoretische begrip (het begrip-zoals-bedoeld, De Groot, 1968) en vertaalt dat in een meetinstrument (de variabele-zoals-bepaald, De Groot, 1968). De tweede opvatting, door Dessens en Jansen omschreven als de modeltraditie, gaat daarentegen uit van de empirie, een nauwkeurig omschreven domein van gegevens, waarover men uitspraken wil doen. Voorstander van de eerste opvatting is De Groot. Voorstander van de modeltraditie is Roskam.
Volgens Roskam (1982) is de taak van de onderzoeker om de verschijnselen te ordenen en te trachten er enige samenhang in te ontdekken. Dit gebeurt door het "denkbeeld" van de onderzoeker omtrent de te verwachten structuur in de data op houdbaarheid te onderzoeken. En toetsing van de structuur in de data dient in de opvatting van Roskam te gebeuren met behulp van een model. Dit model heeft de status van een minitheorie, die de structuur in de data verklaart. De parameters in het model zijn de grootheden die inhoudelijk kunnen worden geïnterpreteerd. Van validiteit in de gebruikelijke zin van het woord is in dit geval geen sprake. In tegenstelling tot de gebruikelijke validiteit, die een externe referentie heeft, berust de validiteit in de modeltraditie op een interne referentie; van een eventuele geldigheid is alleen sprake binnen het omschreven datadomein. Door het koppelen van het verklarende begrip-als-bedoeld aan zijn operationele bepaling door middel van het model is er in de opvatting van Roskam geen verschil tussen beide begrippen. "Theorieën en theoretische begrippen gaan over data, en geven uitdrukking aan de structuur daarvan" (blz. 351). De ideeën die aan een en ander ten grondslag liggen, het begrip-zoals-bedoeld, bij De Groot een wezenlijk onderdeel van wetenschappelijke activiteit, behoren in de opvatting van Roskam tot de voorwetenschappelijke fase van het onderzoek.
Zoals bekend wil De Groot (1968) werken vanuit de theorie naar de meting, waarbij de theorievorming uitdrukkelijk tot de taak van de wetenschap wordt gerekend. De omgekeerde weg die Roskam voorstaat leidt volgens De Groot (1982) tot één grote modelmakerij, waar men zich niet bekommert om de opvattingen van het publiek en men geen belangstelling heeft voor "grote theorieën" en om het bereiken van forumconvergentie. Het leidt tot een "cyclisch modellenspel met grote omloopsnelheid. Niet alleen voor methodologen maar ook voor (vrijgestelde) onderzoekers is zo'n spel ongetwijfeld aantrekkelijk; maar aan "samenhang in de theorieproduktie en aan consistente (en begrijpelijke, volgbare) begripsverheldering voor het derde publiek wordt weinig of niets gedaan" (De Groot, 1982, blz. 260).
Het is De Groot met name te doen om theorieën over belangrijke onderwerpen als "leren (en leerervaringen en leereffecten), intelligentie en welzijn" (De Groot, 1982, blz. 238). Over dit soort onderwerpen wil De Groot "akkoordtheorieën". "Daaronder versta ik voorstellen tot theoretische aanzetten - analyses en definities van basisbegrippen, postulaten, modellen, werktheorieën - die in eerste instantie bedoeld

zijn om het eens te worden: "theories of agreement", waarover de deelnemers, plus eventueel de medegebruikers van het derde publiek zich akkoord verklaren omdat er goed mee te werken valt" (blz. 239 e.v.).
De tegenstelling tussen beide opvattingen lijkt kleiner dan door de voor- en tegenstanders wordt gesuggereerd. In de opvatting van De Groot behoren inhoudelijke theorie en toetsing daarvan in de empirie tot de wetenschappelijke activiteit. Roscam verklaart de eerste als wetenschappelijk niet relevant. Alleen de empirie en de modeleerbaarheid daarvan is van belang. De opvatting van de Groot laat zich zeer wel combineren met de strakke empirische werkwijze uit de modeltraditie. Ook Roskam lijkt de indruk te wekken dat voortdurend allerlei opvattingen die de onderzoeker in z'n hoofd heeft toch een rol spelen. En bovendien hoe kiest een voorstander van de modeltraditie het model? Zijn de overwegingen die daarbij een rol spelen niet theoretisch van aard? En hoe kan een onderzoeker empirische gegevens hanteren zonder een theoretisch begrip? (Mellenbergh, persoonlijke mededeling, november 1988).
Er blijven dus allerlei bezwaren kleven aan de opvatting binnen de modeltraditie. Bezwaren die moeilijk te ondervangen zijn.
De bezwaren die vanuit de modeltraditie tegen de methodologie van De Groot naar voren zijn gebracht kunnen daarentegen wel worden ondervangen. Men kan van een wetenschappelijke theorie of begrip-zoals-bedoeld eisen dat daarbij wordt aangegeven op welke empirische gegevens ze betrekking heeft, waarbij gebruik gemaakt kan worden van niet-triviale modellen die de waarnemingen kunnen verklaren. Daartoe kan er gebruik worden gemaakt van hulptheorie in één of andere vorm (Blalock, 1968, 1979, 1982, Borgatta en Bohrnstedt, 1972, Mellenbergh, 1980 en Hox, 1986).
Mellenbergh (1980) lijkt de tegenstelling te overbruggen. Hij onderscheidt theorieën op drie niveaus: een inhoudelijk niveau, een psychometrisch niveau en een data-analytisch niveau. De inhoudelijke theorie biedt een verklaring voor het gedrag dat men observeert, de psychometrische theorie biedt een verklaring voor de reacties van de personen op de aangeboden stimuli (bijvoorbeeld items uit een studietoets) en de data-analytische theorie biedt een kader voor het beschrijven en verklaren van relaties tussen de gemeten variabelen. In elk empirisch sociaal wetenschappelijk onderzoek moeten de drie niveaus goed op elkaar worden afgestemd (Mellenbergh, 1980).
In het onderzoek dat in deze studie wordt gerapporteerd is gewerkt vanuit de opvatting dat beide hierboven beschreven opvattingen elkaar eerder aanvullen dan uitsluiten. Onderzoekers die werken vanuit de operationaliseringstraditie verontachtzamen nog al eens de inmiddels zeer goed ontwikkelde psychometrische- en data-analytische theorie. Omgekeerd loopt een onderzoeker die uitsluitend werkt vanuit de modeltraditie het risico niet verder te komen dan het model. In een empirische wetenschap kan dat nooit de bedoeling zijn. Het onderhavige onderzoek

steunt op drie pijlers:

1. de inhoudelijke theorie met name de inzichten uit cognitieve psychologie en de in de loop van de jaren impliciet en expliciet verworven inzichten van de onderzoeker en z'n collega's met betrekking tot het onderwijs en de evaluatie van studieresultaten,
2. de psychometrische theorie.
3. de analytische theorie, met name de inzichten op het gebied van de longitudinale data-analyse met behulp van lineaire structurele modellen.

1.3 Begripsvalidatie van toetsen

De bruikbaarheid van tests en toetsen moet blijken. Deze algemene stelling is eenvoudig af te leiden uit standaardwerken op het gebied van de toets- en testresearch (Nunnally, 1967, Lord en Novick, 1968, Drenth, 1975, Visser, van Vliet, Evers en Ter Laak, 1982 en Levy en Goldstein, 1984).
Het empirisch onderzoek naar tests en toetsen richt zich doorgaans op het beantwoorden van vragen met betrekking tot de betrouwbaarheid en de validiteit van de meetinstrumenten. Informatie omtrent deze aspecten is nodig om de wetenschappelijke waarde van sociaal wetenschappelijke meetinstrumenten vast te kunnen stellen.
Drenth (1975) benadrukt dat het bij het belichten van de kwaliteit van dit type meetinstrument vooral moet gaan om de begripsvaliditeit. In dit verband wijst hij op de sterke overeenkomst tussen een experimenteel onderzoeker en een testconstructeur. Beiden zijn er idealiter op uit om hun verwachtingen in de vorm van toetsbare hypothesen te onderzoeken. Wat voor de experimentator de wetenschappelijke theorie is, die de richting van het onderzoek bepaalt, is voor de testconstructeur het met de test gemeten begrip.
De redenering van Drenth is ook toepasbaar op studietoetsen.
Het is Hofstee (1971) geweest die een pleidooi heeft gehouden voor begripsvalidatie van studietoetsen. Hij denkt bij begripsvalidatie vooral aan toetsing van expliciete en impliciete hypothesen. Als voorbeelden van dergelijke implicaties noemt Hofstee (1971) "dat bestudering van de leerstof in verder onderwijs of praktijk zijn vruchten zal afwerpen; dat chemie-studenten niet zonder meer een voldoende op de propedeuse psychologie moeten kunnen halen" (blz. 494).
Hofstee (1971) eindigt z'n pleidooi met een opsomming van methoden, die voor de begripsvalidatie van toetsen kunnen worden aangewend.
Welke impliciete en expliciete hypothesen getoetst moeten worden en hoe deze met betrekking tot een bepaalde toets moeten worden afgeleid, laat Hofstee (1971) in het midden.
Op dit punt kunnen de in paragraaf 1.2 genoemde inhoudelijke-, psychometrische-, en data-analytische theorieën een nuttige bijdrage leveren. Bovendien kunnen

dergelijke theorieën worden gebruikt om te bepalen welke methoden het meest in aanmerking komen voor het toetsen van de geformuleerde hypothesen.
Met andere woorden, in het kader van het onderhavige onderzoek worden deze theorieën gebruikt als hulpmiddelen bij het expliciteren van de te onderzoeken hypothesen. Genoemde theorieën kunnen met Hox (1986) derhalve hulptheorieën worden genoemd; het zijn hulpmiddelen bij het formuleren en toetsen van hypothesen. Hulpmiddelen derhalve bij de begripsvalidatie van voortgangstoetsen.

1.4 Het Begrip-zoals-bedoeld

1.4.1 Medische kennis

Hoewel in het spraakgebruik geldt dat onderwijs wordt genoten, vereist het doel-middel denken in de onderwijskunde (De Groot, (1986)) dat er iets wordt bereikt. Iemand die een opleiding heeft gevolgd, heeft na afloop daarvan iets waarover hij eerder niet de beschikking had. De Groot onderscheidt opleidingen nadrukkelijk van andere vormen van onderwijsaanbod. In zijn visie (De Groot, 1986, blz. 109 e.v.) is er pas van een opleiding sprake indien er vier hoofdkenmerken te onderscheiden zijn:

- een substantiële omvang
- samenhang in het pakket
- duidelijke doelgerichtheid
- een afsluitend examen in de een of andere vorm als evaluatie.

Het hoeft geen betoog dat een medische opleiding er een is in de door De Groot bedoelde betekenis. Eén van de doelstellingen van een medische opleiding is die kennis aan te brengen, die het adequaat funktioneren in de gezondheidszorg, met de daar voorkomende problemen mogelijk maakt. Hoe men de opleiding ook inricht de afgestudeerde moet over "voldoende" kennis beschikken. In principe moet het qua kennisniveau niet uitmaken of men nu in Groningen of in Maastricht gestudeerd heeft. (Zie voor het optreden van verschillen tussen de faculteiten bij een gelijk niveau bijlage II paragraaf 8).
De beoogde kennis maakt men zich gedurende de opleiding eigen, door te werken aan de leertaken, de verschillende werkvormen, door de terugkoppeling via de evaluaties, door gesprekken met docenten en studenten over het vakgebied, enzovoort.
Het curriculum is de didactische operationalisatie van de opleiding (De Groot, 1986, blz. 107)) en het evaluatiesyteem, de evaluatieve operationalisatie (De Groot, ibid) van de opleiding. Voor de Maastrichtse opleiding is het eerste beschreven door Snellen (1982), het tweede door Verwijnen e.a. (1982).
In deze studie gaat de interesse vooral uit naar de ontwikkeling van de medische kennis door dit curriculum heen. De interesse gaat uit naar de ontwikkeling van de

medische kennis van de leek (de beginnende eerstejaars student) naar die van de arts, de afgestudeerde.
Hoewel deze studie vooral is gericht op de kennis van eerstejaars studenten en de problemen die deze met zich meebrengt bij het beantwoorden van einddoeltoetsen, moet dit ontwikkelingsaspect van de medische kennis worden benadrukt (zie ook Messick, 1984). Verontachtzaming van dit aspect zou te veel de nadruk leggen op de vraag: wat een eerstejaars wel en niet weet. Het gaat er veeleer om wat weet een eerstejaars van hetgeen ook bijvoorbeeld zesdejaars en afgestudeerden weten. Medische kennis, in deze context, wordt opgevat als een continue variabele met verschillende gradaties in competentie. Deze gradatie in de medische kennis ontwikkelt zich binnen en tussen individuen onder invloed van verschillende factoren: intelligentie, motivatie en iemands persoonlijke omstandigheden zijn daarvan slechts enkele voorbeelden. Voornaamste factor die van invloed is op het ontwikkelen van medische kennis is het curriculum als de didactische operationalisatie van de medische opleiding.
Met medische kennis wordt hier niet de inhoud ervan bedoeld, niet wat iemand weet. Veeleer gaat de interesse uit naar medische kennis als theoretisch begrip, dat als verklaring kan dienen voor de verschillen die men kan waarnemen in het gedrag van artsen, leken en alle aspirant-deskundigen daartussen in. Dergelijke abstracties met betrekking tot de kennis zijn het onderwerp van de cognitieve psychologie. Met name gaat het daarbij om opvattingen die enig houvast kunnen bieden omtrent de ontwikkeling van medische kennis.

1.4.2 De ontwikkeling van medische kennis

Aanknopingspunten daarvoor zijn te vinden bij Boshuizen en Schmidt (1986). Uitgangpunt voor deze auteurs is het lagenmodel van Feltovich en Barrows (1984). Dit model is een expert model voor medische kennis. Op het hoogste niveau is de kennis van de expert georganiseerd rondom "ziektebeschrijvingen" ("illness scripts"). Deze bevatten drie componenten: de oorzaken, de stoornis en de gevolgen die tot uiting komen in de symptomen. Het idee is dat bij de ontwikkeling van een ziektebeschrijving, de basiskennis (bedoeld wordt niet alleen kennis van basisvakken, ook basiskennis uit de klinische vakken valt onder het hier gebruikte begrip basiskennis) een integrerende rol speelt. Het zou met name deze basiskennis zijn die de onderdelen van een "ziektebeschrijving" structuur geeft en bundelt. De aanwezigheid van een "ziektebeschrijving" stelt volgens deze opvatting de arts in staat om een causale redenering met betrekking tot het verklaren van klachten en symptomen op te stellen. Op grond van zo'n causale redenering wordt de diagnose gesteld op het moment dat de klachten en symptomen van de patiënt in verband zijn gebracht met de stoornis.
In het onderzoek van Boshuizen en Schmidt wordt aangetoond dat het model van

Feltovich en Barrows een adequate beschrijving geeft van de kennis van de expert. Zijn kennisorganisatie stelt hem in staat om goed om te gaan met twee ingewikkelde ziektebeschrijvingen. Op grond daarvan komt de expert tot de goede bevindingen en de juiste diagnose. Een tweedejaars student blijkt over te weinig kennis te beschikken om het probleem te kunnen hanteren. Twee studenten van een tussenliggend kennisniveau blijken de problemen wel te begrijpen maar kunnen ze toch niet goed hanteren. De auteurs concluderen dat de ontwikkeling van kennis een aantal stadia doormaakt; van een totaal gebrek aan kennis, via een tussenstadium waarin er gebruik wordt gemaakt van kennis van de basisvakken, naar een kennisniveau dat iemand in staat stelt om de verschillende mechanismen van het ziektebeeld te begrijpen. In het tussenstadium is er wel sprake van adequate kennis met betrekking tot het probleem. De kennis wordt echter gekarakteriseerd als inflexibel (Boshuizen en Schmidt, 1986).
Later onderzoek van Boshuizen (1987) laat zien dat toenemende expertise gepaard gaat met een meer uitgewerkt pathofysiologisch semantisch netwerk. Het cognitief schema van de "ziektebeschrijving" dat wordt toegepast door de expert, lijkt meer geïntegreerd met pathofysiololgische kennis. De ziektebeschrijving stelt de expert in staat om relevante basiskennis vast te houden en toe te passen.
Het theoretische model voor de ontwikkeling van medische kennis wordt gekenmerkt door een aantal stadia die hiërarchisch lijken te zijn. Vanuit algemene alledaagse medische kennis, via allerlei weinig samenhangende concepten, ontstaan er geleidelijk zeer algemene concepten omschreven als "ziektebeschrijvingen", die de gebruiker in staat stellen om relevante kennis te integreren, te onthouden en uit te breiden.
Bovenstaande opvatting waarin het leerproces een aantal fasen zou doorlopen berust grotendeels op de theorie van Anderson (1982). Deze theorie onderscheidt in het menselijk leren drie fasen:

- een eerste fase, aangeduid als de declaratieve fase. Dit is de fase waarin een beginner zich bevindt.De beginner heeft weinig kennis en de kennis die er is, is onsamenhangend. De beginner "redt" zich met algemene cognitieve vaardigheden, die weinig specifiek voor het betreffende domein zijn. Deze vaardigheden zijn ontwikkeld op basis van eerdere ervaring: schoolse en niet schoolse.
- een overgangsfase, waarin declaratieve kennis door middel van het vergaren van veel domein specifieke kennis (knowledge compilation) overgaat in de laatste fase
- de procedurele fase, waarin de kennis gemakkelijk en vloeiend kan worden toegepast in situaties waarin deze geldig is. De kennis komt als het ware "vanzelf" beschikbaar

In de opleiding tot medicus ontstaan aldus geleidelijk cognitieve structuren, die men ook kan omschrijven als het "geestelijke bezit" van de studenten (De Groot, 1986, blz. 69). Wanneer er gesproken wordt over een van de onderwijsdoelstellingen van een medische opleiding, dan is het begrip-zoals-bedoeld niet de toets waarmee men de examens regelt, maar veeleer de "disposities tot gedrag" (De Groot, ibid, blz. 69), mentale programma's die de studenten bewust kunnen aanwenden wanneer ze geconfronteerd worden met medische problemen. Dat kunnen zijn de problemen in een spreekkamer, of de problemen in een toets, geconcretiseerd in een toetsitem. Het begrip-zoals-bedoeld, waar het in de onderhavige studie omgaat, kan worden omschreven als het kennisbestand of de structuur in de cognities, welke studenten zich eigen maken gedurende hun medische studie of welke reeds eerder in een adequate vooropleiding tot ontwikkeling is gebracht.

1.4.3 Implicaties voor het toetsen van medische kennis

Wat is nu de betekenis van dit alles voor het toetsen van medische kennis? Er zijn ten minste twee mogelijkheden:

- In de eerste plaats kan de beschreven theorie over de ontwikkeling van kennis aanknopingspunten opleveren voor de constructie van toetsen, waarmee niet alleen kan worden vastgesteld hoeveel iemand weet, maar ook hoe de kennis in elkaar zit. Dit is de invalshoek die Messick (1984) kiest waar deze auteur stelt dat "Hence educational achievement testing should not focus just on the retrieval and processing of information but also on the nature and utility of the schemas invoked in information processing" (blz. 218).
- In de tweede plaats biedt deze theorie een kader voor het voorspellen van toetsgedrag van respondenten van verschillende cognitieve niveau's. Daarnaast biedt de theorie aanknopingspunten met betrekking tot het voorspellen van itemmoeilijkheden en met betrekking tot de empirische relaties tussen items en de kennis van de persoon. Met andere woorden de besproken theorie geeft richtlijnen voor het vernieuwen van bestaande toetsen en voor het doen van voorspellingen over empirische gegevens verzameld met deze toetsen. Welke concrete voorspellingen uit de theorie afgeleid kunnen worden, zal verderop worden besproken.

1.5 Het begrip-zoals-bepaald: de voortgangstoets en de bloktoets

Of studenten (en afgestudeerden) beschikken over een kennisbestand, dat hun in staat stelt om medische problemen op een adequate wijze op te lossen, kan slechts indirect worden vastgesteld. Om de kwaliteiten en de modaliteiten van een

kennisbestand te kunnen beoordelen, moet men studenten (artsen) problemen voorleggen en ze vragen deze op te lossen. In cognitief psychologisch onderzoek maakt men, om zicht te krijgen op het oplossingsproces, gebruik van de methode van het "hard-op-denken" (zie bijvoorbeeld Boshuizen en Schmidt, 1986). In de traditie van het toetsen van studieresultaten, beschouwt men de toetsscore als een indicator van de aanwezige kennis, zonder dat men zich om het proces als zodanig bekommert.
Het is in Nederland Elshout (1976) geweest die beide tradities bij elkaar heeft gebracht door items uit intelligentietests die psychometrisch homogeen (in factoranalytische betekenis) waren gebleken, via de hard-op-denk methode procesmatig te analyseren. Dit is een begaanbare weg gebleken: "Individuele verschillen in prestatie zijn manifestaties van individuele verschillen in een denkvaardigheid" (...) "en het probleemoplossingsproces is deze vaardigheid aan het werk" (blz. 457).
Voor zover bekend, is de werkwijze van Elshout nooit toegepast op studietoetsen of op items daar uit. Impliciet is een suggestie daar voor te vinden bij Hofstee (1971) in de aanbeveling om in het kader van begripsvalidatie van een toets experimenteel onderzoek te doen (blz. 498).
In het kader van de onderhavige studie wordt de werkwijze van Elshout als een algemeen denkkader bij het interpreteren van individuele verschillen van studenten op toetsen beschouwd. Er worden geen denkprocessen geanalyseerd. In deze studie gaat de aandacht uit naar de vraag of items beantwoordbaar zijn, gegeven de cognitieve structuren van de student, en zo ja welke items dan wel beantwoordbaar zijn.
Toetsitems zullen in deze studie worden opgevat als moeilijkheden voor de cognitieve structuur (of de kennis) van de student. Bij het al dan niet overwinnen van die moeilijkheden manifesteert zich de cognitieve structuur van de student. Toetsitems zijn in deze opvatting remplacanten voor andere problemen uit het medische domein (open vragen over patiëntproblemen, een gestructureerd mondeling examen). Van belang is slechts de demonstreerbaarheid van de cognitieve structuur aan de hand van problemen. Het begrip-zoals-bedoeld in deze studie, de zich ontwikkelende cognitieve structuur aangeduid als medische kennis, kan op zeer verschillende wijzen gedemonstreerd worden. Dat is ook de reden waarom in de titel van deze paragraaf twee realisaties van het begrip-zoals-bedoeld worden genoemd: de voortgangstoets en de bloktoets.

Een voortgangstoets kan worden gedefinieerd als:

> Een schriftelijke toets, bestaande uit een gestratificeerde aselecte steekproef van ca. 250 juist-onjuist vragen uit het gehele gebied van de medische wetenschap, dat door een basisarts beheerst dient te worden.

Deze toets wordt vier keer per jaar afgenomen en deelname aan de toets is verplicht voor studenten uit de studiejaren één tot en met zes.
Een bloktoets kan worden gedefinieerd als:

> Een schriftelijke toets met een variabele lengte bestaande uit juist-onjuist vragen handelend over de thematiek van het onderwijsblok.

De toets wordt aan het einde van het blok afgenomen. Deelname aan de bloktoetsen is verplicht voor studenten die het onderwijsblok (= een periode van 6 weken thematisch onderwijs) hebben gelopen.
Het verschil tussen beide toetsen is gradueel en kan wellicht het beste getypeerd worden aan de hand van een beschrijving van de bedoeling van de toetsen. De bedoeling van de voortgangstoets is de student en de faculteit inzage te geven in de gemaakte vorderingen in het cognitieve domein van de medische opleiding. De gemaakte vorderingen kunnen op grond van geheel verschillende leeractiviteiten tot stand komen. De bedoeling van de toets is daarbij zowel "vooruitziend" als "terugziend" te zijn (Drenth, 1975).
De bloktoets daarentegen is alleen "terugziend". De toets geeft informatie over de kennis verworven in dit ene specifieke blok. De toets behoort zo specifiek te zijn dat leeractiviteiten buiten het blok niet tot een hoge score kunnen leiden.
Het hier beschreven verschil tussen beide toetsen houdt overigens niet in dat items uit de bloktoets onbruikbaar zouden zijn voor opname in de voortgangstoets. In principe kunnen ook bloktoetsitems waarvan is vastgesteld dat een basisarts het item moet kunnen beantwoorden, voor opname in de voortgangstoets in aanmerking komen. Ook een item uit de voortgangstoets kan indien het past in de blauwdruk in de bloktoets, worden opgenomen.

1.6 De voortgangstoets in het eerstejaar

De vergelijking van beide toetsen kan worden gebruikt om de vraagstelling van dit onderzoek te verduidelijken. Het essentiële verschil tussen beide is niet zozeer dat de items in de twee toetsen van elkaar verschillen. In principe kunnen bloktoetsvragen in de voortgangstoets worden opgenomen (mits ze passen in de blauwdruk, die de inhoudelijke rubrieken van de toets definiëert) en omgekeerd.
Het verschil bestaat hierin dat de items in de bloktoets passen bij de verworven cognitieve structuur van de student op een bepaald moment waardoor de prestatie op de items in de bloktoets beter kan worden gestuurd. Er kan speciaal voor worden getraind waardoor het kennisbestand (tijdelijk) groter is dan vóór de training (c.q. het blok) het geval zou zijn.
Het verschijnsel waar hier op wordt gedoeld kent iedereen uit eigen ervaring. Een goede voorbereiding op het tentamen over het boek "Methodologie" van De Groot

kan de kandidaat gedurende enige dagen in de waan laten over de cognitieve structuur van de auteur van dit boek te beschikken. Wordt de kandidaat een aantal weken later gevraagd zelf een methodologische verhandeling te schrijven over een bepaald onderwerp, dan zal waarschijnlijk blijken dat de alledaagse cognitieve structuur een andere is. De "retrieval paden" die er waren op het moment van het tentamen zijn weer buiten gebruik geraakt (Schmidt, persoonlijke mededeling, juni 1988). Soortgelijke ervaringen kennen docenten, die een college of werkgroep voorbereiden. Natuurlijk wordt er in dit soort onnatuurlijke omstandigheden wel iets geleerd en zal er ook wel een blijvende verandering in de kennisstructuur optreden. Dat is het nuttige effect van tentamens. Dat is in Maastricht ook het nuttige effect van de bloktoetsen.
Waar het hier om gaat is duidelijk te maken dat in het geval van gewone studietoetsen (zoals bloktoetsen) er een betrekkelijk probleemloze relatie is tussen de toetsvragen en de kennisstructuur van de studenten. Beide zijn voor het toetsmoment zo gemaakt dat een optimale "matching" gegarandeerd is. De bloktoets past op de verworven cognitieve structuren. Vandaar ook de goede prestaties op dit soort toetsen; de procentuele score van eerstejaars op bloktoetsen is net zo hoog als die van zesdejaars op voortgangstoetsen. Bij voortgangstoetsen is deze "matching" voor jongerejaars wel degelijk problematisch. De meeste toetsvragen passen nog niet of niet meer op hun verworven cognitieve structuren.
Op dit probleem richt deze studie zich. Jongerejaars zijn in deze studie, de volstrekte beginners; de doorsnee eerstejaars. Hun cognitieve structuur is nog zo rudimentair dat vele problemen uit de voortgangstoetsen te moeilijk zijn. De vraag rijst dan ook of eerstejaars wel op deze wijze getoetst moeten worden.

1.7 Vraagstellingen van het onderzoek

In de voorgaande paragrafen is op een aantal plaatsen de vraagstelling van het onderzoek globaal aangeduid. In deze paragraaf wordt de vraagstelling geëxpliciteerd en in een aantal deelvragen opgesplitst.
In de eerste plaats richt het onderzoek zich op de volgende vragen:

1 Bevatten voortgangstoetsen items die door beginnelingen in het medisch domein beantwoord kunnen worden?

   a. Uit welke categorieën van de toetsblauwdruk en uit welke vakgebieden zijn deze items afkomstig?
   b. Op welke onderwijsaktiviteiten heeft de inhoud van deze items betrekking?
   c. Zijn er verschillen in relevantie tussen deze items en andere items uit de voortgangstoetsen?

d. Is er verschil tussen deze items en andere items in het aantal juist en onjuist gesleutelde items?

Bovenstaande vraagstelling I en de deelvraagstellingen Ia tot en met Id hebben betrekking op het eerste deel van het onderzoek (zie hoofdstuk 2). Dit deel is voornamelijk beschrijvend. Het tweede deel heeft een hypothesetoetsend karakter (zie hoofdstukken 3 tot en met 6). Uitgaande van een bevestigend antwoord op vraagstelling I, heeft het tweede deel van het onderzoek betrekking op de volgende vraagstelling:

II Zijn er voorspelbare verschillen en overeenkomsten in het "gedrag" van de in vraagstelling I bedoelde items en andere items uit de voortgangstoets?

Evenals vraagstelling I kan deze vraagstelling aan de hand van een nadere specificatie van de term "gedrag" in een aantal deelvragen worden opgesplitst. Met "gedrag" wordt hier bedoeld het psychometrische "gedrag" van items en toetsen. Dit kan worden bestudeerd in termen van de mate waarin de items door respondenten worden beantwoord, het aantal goede antwoorden, het discriminerend vermogen, de betrouwbaarheid, de correlaties tussen de scores op de items en de scores op andere items, de reproduceerbaarheid van itemscores uit de somscores, etc.
Vraagstelling II wordt in dit onderzoek toegespitst op de volgende deel-vraagstellingen:

Zijn er voorspelbare verschillen tussen de in I bedoelde items en andere items met betrekking tot:
a. de beantwoording?
b. de correcte beantwoording?
c. het discriminerend vermogen?
d. de betrouwbaarheid?
e. de toetsscores?
f. het correlatiegedrag?
g. het voorspellend vermogen?
h. het voldoen aan de eisen van probalistische modellen?

De vraagstellingen I en II worden in deze studie beantwoord op de wijze zoals beschreven in paragraaf 1.8.

1.8 Werkwijze, items en respondenten

Om de vraagstellingen te beantwoorden werden als onderzoeksmateriaal de vier voortgangstoetsen uit het academisch jaar 1983- 1984 gekozen. Deze toetsen bevatten in totaal 1029 items. Deze items werden door 8 docenten beoordeeld. Van deze docenten waren 6 coördinator van een onderwijsblok in het eerste studiejaar. Beide andere docenten waren de coördinator van het eerstejaar en zijn voorganger. Aan de docenten werd gevraagd de items te selecteren, die op grond van het programma van het eerstejaar beantwoord moesten kunnen worden. Bij onenigheid tussen de coördinatoren besliste de jaarcoördinator.
De aldus géselecteerde items werden Eerstejaarsitems genoemd. Deze items werden nader onderscheiden naar het tijdstip van afname gedurende het academisch jaar 1983-1984. Een groep Eerstejaarsitems die op hetzelfde tijdstip waren afgenomen werd Eerstejaarstoets genoemd. Op deze wijze werden er vier Eerstejaarstoetsen gevormd. Als controleitems werd er op elk tijdstip uit de resterende voortgangstoetsitems aselect een groep items gekozen van dezelfde omvang als de Eerstejaarsitems. Deze items werden Algemene Kennisitems genoemd. Op alle vier tijdstippen waren er derhalve twee toetsen: één Eerstejaarstoets en een Algemene Kennistoets. Met andere woorden er werden in totaal 4 x 2 = 8 toetsen onderzocht.
De respondenten in dit onderzoek waren de studenten die gedurende het genoemde academisch jaar aan de toetsen deelnamen. Voor het toetsen van sommige hypothesen werd daarnaast gebruik gemaakt van de antwoorden van een groep afgestudeerde basisartsen in opleiding voor huisarts. Niet in alle gevallen kon over de antwoorden van deze groep worden beschikt. Dit zal verderop uit de context blijken.
Afhankelijk van de hypothese werden de antwoorden van alle studenten gebruikt, dan wel de antwoorden van een paar van de (jaar)groepen, met name eerstejaars, zesdejaars en artsen. De antwoorden van de artsen konden alleen worden gebruikt, voor zover het ging om reeds gearchiveerde data. Daar waar nieuwe berekeningen nodig waren kon niet over deze gegevens worden beschikt.
Meer specifieke informatie betreffende de gevolgde werkwijze bij het toetsen van de verschillende hypothesen wordt in de betreffende hoofdstukken gegeven waar de resultaten van de toetsing worden beschreven.

1.9 Verdere empirische afbakening van het onderzoek

In vraagstelling I is sprake van de voortgangstoetsen. In de loop van de jaren zijn er zeer veel voortgangstoetsen afgenomen. Het is uiteraard ondoenlijk deze alle te bestuderen met betrekking tot de geformuleerde vraagstelling. Een eerste empirische afbakening betreft met andere woorden de vraag welke voortgangstoetsen bestudeerd worden.

In dit onderzoek werden de vier voortgangstoetsen bestudeerd, die werden afgenomen in het academisch jaar 1983-1984. De keuze voor dit jaar was vrij willekeurig. De discussie over de vraag naar de bruikbaarheid van de voortgangstoets in het eerste studiejaar ontstond opnieuw in het najaar van 1984. Toen het besluit eenmaal was genomen om het onderzoek te starten, was de keuze voor de laatst afgenomen toetsen erg voor de hand liggend.
Hoewel er dus sprake is van een gelegenheidssteekproef (Hofstee, 1980) zijn er geen redenen om aan te nemen dat de keuze voor een andere serie voortgangstoetsen tot geheel andere conclusies zou hebben geleid. Met andere woorden, aangenomen wordt dat de hier bestudeerde toetsen representatief zijn voor de voortgangstoetsen. Het voornaamste argument daarvoor, naast de reeds genoemde keuze, is dat de voortgangstoetsen volgens dezelfde blauwdruk worden samengesteld.
Een tweede afbakening betreft het begrip beantwoordbaar en beginnelingen in het medisch domein.
Beantwoordbaarheid van toetsvragen voor beginnelingen zou kunnen worden afgeleid uit de p- waarden (= % goed) in het eerstejaar. In deze redenering is een vraag beter beantwoordbaar naarmate de p-waarde hoger is. Deze op zichzelf bruikbare redenering heeft als bezwaar dat niet duidelijk wordt hoeveel hogere p- waarden er verwacht mogen worden in het eerste studiejaar.
Om aan dit bezwaar tegemoet te komen is getracht de relatie tussen de inhoud van de items en de inhoud van het eerste onderwijsprogramma in het eerstejaar te leggen. In deze studie is de relatie gelegd door gebruik te maken van het oordeel van docenten.
Als docenten zeggen dat er een correspondentie bestaat tussen het item en het onderwijsprogramma in het eerstejaar dan mag worden verwacht dat het item voor een beginneling beter beantwoordbaar is dan een willekeurig ander item. Met andere woorden in deze studie is het begrip beantwoordbaar voor beginnelingen een relatief begrip, dat wordt gekoppeld aan de beantwoordbaarheid van andere - willekeurig - gekozen items.
De in paragraaf 1.4 beschreven theorie over de ontwikkeling van (medische) kennis claimt dat kennis hiërarchisch is. Niet voor alle door middel van toetsitems afvraagbare kennis is dit het geval. Voortgangstoetsen zijn echter bedoeld om de ontwikkeling in medische kennis vast te stellen. Dit kan alleen als alle items verschillen tussen de verschillende kennisniveaus te zien geven. Om deze reden worden in deze studie niet alleen de antwoorden van beginnelingen bestudeerd, maar ook die van meer gevorderden en afgestudeerden. Hierbij wordt er van uit gegaan dat de kennisniveaus in ieder geval globaal overeenkomen met de indeling in de (jaar)groepen: eerstejaars tot en met zesdejaars en afgestudeerden. Gemiddeld genomen mag worden verwacht dat de laagste kennisniveaus onder de eerstejaars zitten en de hoogste onder de zesdejaars en de afgestudeerden.

Met het bovenstaande kan het empirische domein van deze studie worden omschreven: de antwoorden van studenten en afgestudeerden op items afkomstig van twee deelverzamelingen uit elk van de vier voortgangstoetsen van het academisch jaar 1983-1984 worden bestudeerd. Eén deelverzameling is door docenten aangewezen als zijnde beantwoordbaar door eerstejaars, de tweede deelverzameling is aselect gekozen uit de overblijvende items van elk van de vier voortgangstoetsen. Het bereik van de antwoorden die bestudeerd worden kan worden geordend van beantwoord tot niet beantwoord, en van goed beantwoord tot niet goed beantwoord. Aan de items wordt slechts één facet onderscheiden, de "moeilijkheid van de items", dat twee niveau's bevat:

1. corresponderend met het eerstejaarsprogramma (minder moeilijk), en
2. niet corresponderend met het eerstejaarsprogramma (moeilijk).

1.10 Hypothesen over het empirische domein

Uitgaande van de omschrijving van het begrip-zoals-bedoeld en het begrip-zoals-bepaald is het mogelijk om een aantal hypothesen op te stellen, dat als leidraad bij de analyse van de verschillen tussen de itemdomeinen aangeduid als Eerstejaarstoets en Algemene Kennistoets kan dienen. De hypothesen kunnen in drie groepen worden onderverdeeld:

1. Hypothesen met betrekking tot verschillen in de interne structuur van de toetsen.
2. Hypothesen met betrekking tot verschillen in de externe structuur van de toetsen.
3. Psychometrische modellen als hypothese.

1.10.1 Hypothesen over verschillen in interne structuur

Met de interne structuur van een toets wordt hier bedoeld (zie ook Drenth, 1975) de kenmerken van de verdelingen van de gegeven antwoorden, de goede antwoorden, de interne consistentie, de item rest correlaties enzovoort.
Met betrekking tot de interne structuur van de toetsen kunnen de volgende hypothesen worden geformuleerd:

HYPOTHESE 1

a. In het eerste studiejaar is het gemiddelde beantwoordingspercentage van de Eerstejaarsitems hoger dan het gemiddelde beantwoordingspercentage van de Algemene Kennisitems.
b. In het zesdejaar zijn deze gemiddelden in beide itemgroepen gelijk.
c. In de groep afgestudeerden zijn deze gemiddelden ook gelijk.

Rationale: Als eerstejaarsitems beter passen bij de kennisstructuur van de eerstejaars, dan mag worden verwacht dat deze vragen door meer studenten zullen worden beantwoord. Zesdejaars en afgestudeerden hebben een kennisstructuur, die het mogelijk maakt om de problemen uit beide domeinen adequaat op te lossen. De resultaten met betrekking tot deze hypothese zijn te vinden in paragraaf 3.2.1.

HYPOTHESE 2

a. In het eerstejaar is het gemiddelde percentage goede antwoorden (de p-waarden) van de Eerstejaarsitems hoger dan het gemiddelde percentage goede antwoorden van de Algemene Kennisitems.
b. In het zesdejaar zijn deze gemiddelden in beide itemgroepen gelijk.
c. In de groep afgestudeerden zijn deze gemiddelden in beide itemgroepen gelijk.

HYPOTHESE 3

a. In het eerstejaar is de variantie van het percentage goede antwoorden (de p-waarden) van de Eerstejaarsitems groter dan de variantie van het percentage goede antwoorden van de Algemene Kennisitems.
b. In het zesdejaar zijn deze varianties in beide itemgroepen gelijk.
c. In de groep afgestudeerden zijn deze varianties in beide itemgroepen gelijk.

Rationale: Als de eerstejaarsitems beter passen bij de kennisstructuur van de eerstejaars dan mag verwacht worden dat deze items, ook beter worden beantwoord. Beide itemgroepen leveren naar verwachting voor de kennisstructuur van gevorderden dezelfde problemen op zodat een gelijk gemiddelde in goede antwoorden wordt verwacht. Wat betreft de schaal van de verdeling is de geformuleerde verwachting gebaseerd op de overweging dat onderwijseffecten zich manifesteren in de verschillen tussen items (Wijnen, 1971). Voor eerstejaars zullen onderwijseffecten zich vooral voordoen in de Eerstejaarstoetsen. Vandaar dat verwacht mag worden dat in het eerstejaar de verschillen tussen de items binnen de Eerstejaarstoets groter zijn dan de verschillen tussen items binnen de Algemene Kennistoets. Op grond van dezelfde overweging worden er geen verschillen bij zesdejaars en artsen verwacht. De resultaten met betrekking tot deze hypothesen worden beschreven in de paragrafen 3.2.2 en 3.2.3.

HYPOTHESE 4

a. In het eerstejaar is de gemiddelde item-restcorrelatie van de Eerstejaarsitems lager dan de gemiddelde item-restcorrelatie van de Algemene Kennisitems.
b. In het zesdejaar zijn deze gemiddelden in beide itemgroepen gelijk.

Rationale: De ervaring heeft geleerd dat de Maastrichtse eerstejaars studenten met betrekking tot kennis van het medische domein zeer heterogeen van samenstelling zijn. Deze observatie leidt tot de verwachting dat de variantie van toetsscores op een Algemene (Medische) Kennistoets relatief groot is. Groter dan de variantie van toetsscores van een meer op het niveau van een eerstejaars toegesneden toets. Naar verwacht mag worden vormen de Eerstejaarsitems een toets in de laatstgenoemde betekenis. Derhalve mag worden verwacht dat de variantie van de toetsscores op de Eerstejaarstoets kleiner zal zijn dan die van de toetsscores op de Algemene Kennistoets. Dit verschil leidt op grond van de klassieke test theorie tot de voorspelling dat de item-restcorrelaties van de Eerstejaarsitems gemiddeld lager zijn dan van de Algemene Kennisitems. In het zesdejaar en in de artsengroep zal er geen verschil in varianties van de toetsscores zijn. Derhalve wordt daar geen verschil voorspeld. De resultaten met betrekking tot deze hypothese worden beschreven in paragraaf 3.2.4.

HYPOTHESE 5

a. In het eerstejaar is de betrouwbaarheid (Cronbach's alfa) voor Eerstejaarsitems lager dan voor Algemene Kennisitems.
b. In het zesdejaar is de betrouwbaarheid in beide itemgroepen gelijk.

Rationale: Op grond van de onderwijseffect hypothese (zie rationale voor hypothese 3 en 4) moet worden verwacht dat bij eerstejaars de interne consistentie van de Eerstejaars toets lager zal zijn dan die van de Algemene Kennis toets. Bij zesdejaars en afgestudeerden moet op grond van dezelfde redenering worden verwacht dat de interne consistentie van beide toetsen gelijk zal zijn. De resultaten met betrekking tot deze hypothese worden beschreven in paragraaf 3.2.5.

HYPOTHESE 6

a. In het eerstejaar is het gemiddelde van de toetsscores op de Eerstejaarsitems hoger dan het gemiddelde van de toetsscores op de Algemene Kennisitems.
b. In het zesdejaar zijn deze gemiddelden in beide itemgroepen gelijk.

HYPOTHESE 7

a. In het eerstejaar is de variantie van de toetsscores op de Eerstejaarsitems groter dan de variantie van de toetsscores op de Algemene Kennisitems.
b. In het zesdejaar zijn deze varianties in beide itemgroepen gelijk.

Rationale: Deze is bij 3 en 4 al gegeven. Op grond van deze redenering, moet in het eerstejaar op de Eerstejaarstoets een hoger gemiddelde en een kleinere spreiding in de verdeling van toetsscores worden verwacht. De resultaten met betrekking tot deze hypothese worden beschreven in paragraaf 3.2.6 en 3.2.7.

HYPOTHESE 8

a. In de scorematrix (items x personen) van Eerstejaarsitems en eerstejaars studenten is de itemvariantie groter dan de personenvariantie. In de scorematrix van Algemene Kennisitems en eerstejaars studenten is de personenvariantie groter dan de itemvariantie.
b. In het zesdejaar is in beide scorematrices de itemvariantie groter dan de personenvariantie.

Rationale: Deze hypothese is gebaseerd op een door Wijnen (1971) voorgestelde typologie van studietoetsen. In de toets zijn volgens deze beschrijving tenminste twee variantiebronnen aan te wijzen: een effect dat is toe te schrijven aan verschillen tussen personen en een effect dat is toe te schrijven aan verschillen tussen items.
Onderwijseffecten komen in de toetsresultaten tot uiting doordat de verschillen tussen items relatief groot zijn. Persoonseffecten komen in de toetsresultaten tot uiting doordat de verschillen tussen personen relatief groot zijn.
In het eerstejaar zijn Eerstejaarsitems op te vatten als toetsen die vooral onderwijseffecten meten. Algemene Kennisitems zijn op te vatten als toetsen die vooral reeds bestaande verschillen in kennis tussen personen meten. Dit leidt tot de in de hypothese geformuleerde verwachting.
In het zesdejaar zijn beide toetsen in dit opzicht als equivalent te beschouwen, leidend tot de in hypothese 8b geformuleerde verwachting. De resultaten van de toetsing van deze hypothesen worden vermeld in paragraaf 3.2.8.

HYPOTHESE 9

a. In het eerstejaar is het gemiddelde percentage goede antwoorden van Eerstejaarsitems die een natoets vormen hoger dan het gemiddelde percentage goede antwoorden van Eerstejaarsitems die een voortoets vormen.
b. In het zesdejaar zijn deze gemiddelden gelijk.

Rationale: Onderwijs en andere ervaringen opgedaan in een bepaald domein beïnvloeden de medische kennis. Eerstejaarsitems zijn items die volgens docenten een relatie hebben met het onderwijsprogramma in het eerste studiejaar. Het derde tijdstip van toetsing was maart 1984. Op dat moment was een deel van het programma door de studenten doorlopen, terwijl een ander deel nog moest komen. Items die betrekking hebben op het reeds doorlopen deel kunnen worden beschouwd als een natoets. Items die betrekking hebben op het deel dat nog moet komen kunnen worden opgevat als een voortoets.
Indien de verwachte hogere kans op een goed antwoord op de Eerstejaarsitems toe te schrijven is aan een onderwijseffect dan luidt de verwachting met betrekking tot het verschil zoals verwoord in hypothese 9a.
In het zesdejaar mag er geen verschil meer zijn, beide itemgroepen zijn voor deze groep als een natoets te beschouwen. De resultaten met betrekking tot deze hypothesen worden beschreven in paragraaf 3.2.9.

1.10.2 Hypothesen over verschillen in externe structuur

Onder externe structuur wordt hier verstaan het geheel aan empirische relaties tussen de bestudeerde toets en andere relevante externe variabelen (zie ook Drenth, 1975). Relevant in het kader van het achterhalen van de begripsvaliditeit van een studietoets zijn in principe alle variabelen die bijdragen aan die begripsvaliditeit. De verzameling daarvoor in aanmerking komende variabelen is onuitputtelijk, zodat het antwoord op de begripsvaliditeit nooit definitief te geven is. De beperkingen die de onderzoeker zich in deze moet opleggen zijn dan ook niet van principiële aard, doch veeleer gerelateerd aan practische overwegingen, zoals de variabelen waarover men relatief eenvoudig kan beschikken: bijvoorbeeld omdat ze reeds gemeten zijn.
In dit geval waren dat van het cohort 1983 de bloktoetsen uit het eerste-, het tweede-, het derde studiejaar. Verder werden de voortgangstoetsen uit het tweede- en derde studiejaar gebruikt.
Met betrekking tot de relaties tussen de Eerstejaarstoetsen en deze externe variabelen kunnen de volgende hypothesen worden geformuleerd:

HYPOTHESE 10

a. In het eerstejaar en in het zesdejaar worden op ieder tijdstip de correlaties tussen de toetsscores van Eerstejaarsitems en de toetsscores van Algemene Kennisitems door één latente variabele verklaard, die over de tijdstippen heen met behulp van een simplex model kan worden beschreven (zie o.a. Jöreskog, 1970).
b. Indien het model dat besloten ligt in hypothese 12a, de data niet goed beschrijft, dan kan een "passend" model worden gevonden door twee latente variabelen toe te voegen: één specifiek voor de Algemene Kennistoetsen en één specifiek voor de Eerstejaarstoetsen.

Rationale: De verwachte sterke correlatie tussen beide toetsen komt voort uit de overweging die al eerder werd genoemd n.l. dat de itemdomeinen waar het hier om gaat voornamelijk verschillen in moeilijkheid. Alle items doen een beroep op eenzelfde latente variabele, medische kennis. Hoe beter deze is ontwikkeld des te hoger de prestatie op beide toetsen. Vandaar dat er een sterke correlatie wordt verwacht, zo sterk dat beide toetsen eigenlijk naar dezelfde latente variabele wijzen. Op ieder tijdstip een latente variabele, die over tijdstippen heen in het eerstejaar minder stabiel zal zijn dan in het zesdejaar. Dit op grond van de overweging dat er in het eerstejaar met de cognitieve structuur van de studenten het meest gebeurt. In het eerstejaar zal er behalve variatie, die aan de niet observeerbare trek kan worden toegeschreven, ook nog toets-specifieke methode variatie zijn, waarin de "eigenheid" van de Eerstejaarstoets en de Algemene Kennistoets zich manisfesteert. In het zesdejaar is deze "eigenheid" van de toetsen verdwenen en is de latente trek voldoende om de variatie te verklaren. De resultaten met betrekking tot deze hypothesen worden beschreven in paragraaf 4.4.2.

HYPOTHESE 11

Er is geen verschil tussen de toetsscores van Eerstejaarsitems en Algemene Kennisitems met betrekking tot het voorspellend vermogen van later studiesucces.

Rationale: Beide toetsen zullen voldoende verschillen tussen de studenten registreren om als voorspeller te kunnen fungeren. Op grond van de te verwachten grotere spreiding in de toetsscores op de Algemene Kennistoets, moet deze als voorspeller enigszins hoger worden aangeslagen. Het voorspellend vermogen op zich, moet worden verwacht op grond van het feit dat studiesucces in latere jaren ten dele is gerelateerd aan de cognitieve structuur in het eerste studiejaar. De

wijze waarop deze zich ontwikkelt in het eerstejaar zal daarom ook (hoewel in afnemende mate) betekenis hebben voor de verdere ontwikkeling in de latere studiefases. De resultaten met betrekking tot het voorspellend vermogen van beide toetsen worden beschreven in de paragraaf 4.5.3.

HYPOTHESE 12

> Alle correlaties tussen toetsscores van Eerstejaarsitems en toetsscores van Algemene Kennisitems met toetsscores van bloktoetsen in het eerstejaar zijn niet negatief.

Rationale: Van eenzelfde verklarende latente variabele, i.c. medische kennis mag worden verwacht dat goede prestaties op het ene instrument samengaan met goede prestaties op het andere instrument. Uitzonderingen hierop zijn in deze redenering anomalieën van de veronderstelde wetmatigheid, die naar analogie van Guttman's (1981) eerste wet van antwoorden in het empirische domein van de intelligentieitems, kan worden geformuleerd. De eerste wet voor kennis zou kunnen luiden: als twee willekeurige items uit het domein van kennisitems worden geselecteerd dan zullen deze items in een adequate populatie personen niet negatief correleren.
Als deze wetmatigheid geldt dan zullen toetsscores van itemgroepen uit het betreffende domein eveneens niet negatief correleren.
De resultaten met betrekking tot dit punt worden beschreven in paragraaf 4.6.3.

1.10.3 Psychometrische modellen als hypothese

HYPOTHESE 13

> Beide itemgroepen voldoen aan de assumpties van het VGT-model.

Rationale: In deze hypothese wordt de redenering die aan de voorgaande verwachtingen ten grondslag ligt doorgetrokken. Als de verschillen, zoals reeds besproken, worden veroorzaakt door het verschil in moeilijkheid van de items in relatie tot de kennisstructuur, dan moet worden verwacht dat de kans op een goed antwoord alleen afhangt van de moeilijkheid van het item en van de ontwikkeling van de cognitieve structuur. Hoe meer deze is ontwikkeld des te hoger is de kans op een goed antwoord. En voor de kennisstructuur op een bepaald niveau, geldt dat de kans op een goed antwoord kleiner wordt naarmate de items moeilijker zijn. Eerstejaars zullen op grond van deze redenering op een willekeurig item uit de Eerstejaarstoets een grotere kans op een goed antwoord hebben dan op een willekeurig item uit de Algemene Kennistoets.

De uitwerking van deze redenering wordt beschreven in hoofdstuk 5. De toetsing ervan is te vinden in hoofdstuk 6.

HYPOTHESE 14

a. Een deel van de Eerstejaarsitems voldoet aan de assumpties van het Raschmodel.
b. Een deel van de Algemene Kennisitems voldoet aan de assumpties van het Raschmodel.
c. Gezamenlijk vormen deze Eerstejaars- en Algemene Kennisitems één Raschhomogene itemgroep.

Een nadere beschrijving van het Raschmodel is te vinden in hoofdstuk 5. De resultaten van het toetsen van de assumpties van het model zijn te vinden in hoofdstuk 6.

Rationale: De voortgangstoetsitems zijn nog niet eerder met behulp van item response modellen onderzocht. Op grond van de literatuur over deze modellen (o.a. Wright & Stone, 1979) moet worden verwacht dat niet alle items aan de assumpties van dit model voldoen. Hoe groot het Raschhomogene deel zal zijn is op voorhand moeilijk te zeggen. Vandaar dat dit in de hypothese niet verder is uitgewerkt. Wel kan worden opgemerkt dat hier geldt hoe meer items aan de assumpties voldoen des te beter de itemgroep het construct meet. Eerstejaarsitems en Algemene Kennisitems, die wel aan het Raschmodel voldoen, zijn uit hetzelfde medische domein afkomstig. Dit betekent dat de beantwoordbaarheid van de items op beide itemgroepen in theorie afhangt van eenzelfde latente variabele. Deze variabele werd in paragraaf 1.4 omschreven als medische kennis. De beantwoording van Raschhomogene Eerstejaarsitems kan worden verklaard door één latente variabele. De beantwoording van Raschhomogene Algemene Kennisitems kan eveneens worden verklaard door één latente variabele. Op grond van de theorie over medische kennis moet worden verwacht dat aan de beantwoording van Eerstejaarsitems en Algemene Kennisitems, die afzonderlijk Raschhomogeen zijn gebleken, één latente variabele ten grondslag ligt. Deze redenering leidt tot de verwachting, die in hypothese 10c is geformuleerd. De resultaten met betrekking tot deze hypothesen worden beschreven in hoofdstuk 6.

HYPOTHESE 15

a. Een deel van de niet-Raschhomogene Eerstejaarsitems voldoet wel aan de eisen van het niet-parametrische Mokkenmodel.

b. Een deel van de niet-Raschhomogene Algemene Kennisitems voldoet wel aan de eisen van het niet-parametrische Mokkenmodel.

Rationale: De items die niet Raschhomogeen zijn, kunnen zowel inhoudelijk als psychometrisch nog wel goede items zijn. Voortgangstoetsitems moeten over de verschillende kennisniveau's van de gemeten variabele monotoon niet-dalende item karakteristieke curves hebben. Dat wil zeggen, een hogere score op de toets als geheel moet op itemniveau een hogere kans op een goed antwoord te zien geven dan een lagere toetsscore. Hoe groot het aantal Mokkenitems zal zijn is evenmin aan te geven (zie rationale bij hypothese 14).
Dit soort uitgangspunten zijn redelijk als het gaat om het meten van het kennisbestand van studenten. Immers, iemand met een kennisbestand van een "arts" zal veel problemen uit de geneeskunde kunnen oplossen. Dat wil zeggen, van zo iemand wordt verwacht dat hij een hoge succeskans heeft op een willekeurig probleem en een hogere succeskans op zo'n probleem dan iemand met een kleiner kennisbestand van het betreffende gebied. Op itemniveau leidt deze redenering tot het hiervoor beschreven monotoniciteitsprincipe en kan worden getoetst met behulp van het Mokkenmodel. Het Mokkenmodel wordt beschreven in hoofdstuk 5. De resultaten van de toetsing zijn te vinden in hoofdstuk 6.

1.11 Werkwijze bij het toetsen van de hypothesen

In deze laatste paragraaf wordt globaal beschreven op welke wijze de hiervoor genoemde hypothesen werden getoetst. Details met betrekking tot de toetsingsprocedures zijn te vinden in de paragrafen waar de resultaten van de betreffende hypothesetoetsing beschreven worden.
Alle percentages gescoorde variabelen werden met behulp van de logittransformatie omgezet in logitscores. Dit werd gedaan omdat deze transformatie meestal leidt tot min of meer normale verdelingen. De hypothesen werden getoetst in het eerste studiejaar en in het zesde studiejaar. In enkele gevallen werd er ook een toets uitgevoerd in een groep afgestudeerden.
De meeste hypothesen werden per tijdstip getoetst. Daar waar simultaan over de tijdstippen heen werd getoetst, zal dit steeds bij de betreffende resultaten vermeld worden.
Een aantal hypothesen kon niet op "harde" wijze worden getoetst (hypothesen 5 en 8). In deze gevallen werd de uitgevoerde analyse opgevat als een beschrijving, waarvan werd vastgesteld of deze al dan niet in overeenstemming was met de verwachting.
Andere hypothesen hebben betrekking op het voldoen van de data aan een psychometrisch model (hypothesen 13, 14 en 15)). In deze gevallen werd een modeltoets uitgevoerd om na te gaan in hoeverre het betreffende model de

datastructuur benaderde. Deze hypothesen werden onderzocht door gebruik te maken van de antwoorden van alle jaargroepen. De reden wordt in hoofdstuk 5 gegeven. Het zal duidelijk zijn dat sommige hypothesen onderling afhankelijk zijn. Het streven om het empirische domein van dit onderzoek zo volledig mogelijk te beschrijven bracht met zich mee dat de gegevens vanuit verschillende invalshoeken bestudeerd moesten worden.
Tenslotte nog enkele woorden over het feit dat in dit geval er een hele reeks hypothesen moet worden getoetst om de indicaties met betrekking tot de begripsvaliditeit naar waarde te kunnen schatten (zie paragraaf 1.3). Als er veel hypothesen worden getoetst dan zijn er altijd wel een paar significant, zonder dat daaraan echt betekenis mag worden gehecht.
Meer algemeen kan gesteld worden dat wanneer alle n nulhypothesen juist zijn en men toetst op alfa-niveau, dan zullen er gemiddeld genomen alfa * n significante toetsingsresultaten zijn. Dit verwachte aantal significante resultaten zal dus moeten worden vergeleken met het aantal feitelijk gevonden significante resultaten. Met andere woorden pas als dit feitelijk gevonden aantal redelijk veel groter is dan alfa * n kan men waarde hechten aan de gevonden toetsingsresultaten.
Dat betekent dat er pas van ondersteuning van de begripsvaliditeit van de Eerstejaarstoetsen kan worden gesproken als het aantal feitelijk gevonden significante resultaten tenminste groter is dan alfa * n.
Een ander probleem is het constant houden van de kans op een fout van de eerste soort indien een allesomvattende nulhypothese opgesplitst wordt in een aantal deelhypothesen die samen de nulhypothese omvatten. Bij herhaald toetsen van een (deel)hypothese wordt de kans op het ten onrechte verwerpen van de nulhypothese veel groter dan het door de onderzoeker gekozen individuele significantie-niveau van bijvoorbeeld .05. Bij tien keer toetsen van een deelhypothese op een significantie-niveau van .05, is de kans op het ten onrechte verwerpen van tenminste één nulhypothese niet gelijk aan de beoogde .05, maar ongeveer gelijk aan .40. In zo'n geval is het wenselijk om correcties op de overschrijdingskans toe te passen. Een veel gebruikte correctiemethode is die welke is voorgesteld door Bonferroni (Miller, 1981). Deze methode corrigeert op eenvoudige wijze door het gekozen significantie-niveau te delen door het aantal keren dat men toetst.
Bij het toetsen van de afzonderlijke nulhypothesen is steeds een aantal paarsgewijze vergelijkingen in het geding. Immers binnen elk van de onderzochte jaargroepen worden de geschatte parameters van de Eerstejaarstoetsen vergeleken met die van de Algemene Kennistoetsen. In dit geval wordt het significantie-niveau constant gehouden door alfa te delen door het aantal paren. Dit betekent in het geval van bijvoorbeeld drie onderzochte groepen dat per paar een significantieniveau geldt van .016 om zodoende de type 1 fout te beperken tot .05.

1.12 Samenvatting

In dit hoofdstuk werden de vraagstelling en de hypothesen beschreven. In paragraaf 1.2 werd een keuze gemaakt uit een aantal mogelijke opvattingen betreffende de rol van theoretische begrippen bij empirisch onderzoek. Gekozen werd voor de operationaliseringstraditie aangevuld met de "harde" modelbenadering van de "modeltraditie". Deze keuze leidde min of meer vanzelfsprekend tot de opvatting die Hofstee (1971) bepleitte, namelijk het doen van onderzoek naar de begripsvaliditeit van studietoetsen (paragraaf 1.3).
In paragraaf 1.4 werd de inhoudelijke theorie, het begrip-zoals-bedoeld beschreven vanuit de cognitief psychologische benadering. Benadrukt werd dat aan de ontwikkeling van medische kennis een aantal stadia te onderscheiden zijn (paragraaf 1.4.2). De consequenties van deze opvatting voor het toetsen van medische kennis werd besproken in paragraaf 1.4.3. Met name het hiërarchische karakter van het begrip medische kennis leek daarbij van belang te zijn.
Vervolgens werden in paragraaf 1.5 twee instrumenten om kennis te meten beschreven: de voortgangstoets en de bloktoets. De verschillen tussen deze twee werden beschreven en gebruikt om de vraagstellingen van het onderzoek te verduidelijken (paragraaf 1.6).
In paragraaf 1.7 werden deze vraagstellingen geëxpliciteerd. Vraagstelling I luidde:

I Bevatten voortgangstoetsen items die door beginnelingen in het medisch domein beantwoord kunnen worden?

en vraagstelling II luidde:

II Zijn er voorspelbare verschillen en overeenkomsten in het "gedrag" van de in vraagstelling I bedoelde items en andere items uit de voortgangstoets?

Hierna werden in paragraaf 1.8 en 1.9 de gekozen werkwijze en de empirische afbakening van het onderzoek beschreven.
De hypothesen voor het omschreven domein werden geformuleerd en naar drie aspecten onderverdeeld:

- hypothesen over verschillen in interne structuur (paragraaf 1.10.1)
- hypothesen over verschillen in externe structuur (paragraaf 1.10.2)
- hypothesen over de houdbaarheid van psychometrische modellen (paragraaf 1.10.3).

Tenslotte werd in paragraaf 1.11 de gevolgde procedure bij het toetsen van de hypothesen beschreven.

2. SELECTIE VAN EERSTEJAARSITEMS

Inhoud

2.1 Inleiding
2.2 Selectie van Eerstejaarsitems (Eerstejaarstoetsen)
2.2.1 Procedure
2.2.2 Resultaten beoordelingsprocedure
2.3 Selectieprocedure van controle items (Algemene Kennistoetsen)
2.4 Beschrijving Eerstejaarstoetsen
2.4.1 Inhoudelijke beschrijving
2.4.2 Relevantie van de Eerstejaarstoetsen
2.4.3 Verdeling juist- en onjuistvragen in beide toetsen
2.5 Inhoudelijke itemselectie vergeleken met empirische itemselectie
2.5.1 Vergelijking qua aantal geselecteerde items
2.5.2 Vergelijking van moeilijkheidsgraad van de items
2.6 Samenvatting en conclusies

Hoofdstuk 2: Selectie Eerstejaarsitems

2.1 Inleiding

Vanaf de invoering van voortgangstoetsen in het curriculum van de Faculteit der Geneeskunde, bestond er twijfel over de zinvolheid van deze vorm van toetsen in het eerste studiejaar. Immers, voortgangstoetsen representeren de kennis waarover een arts aan het einde van de opleiding dient te beschikken. Zo'n omschrijving impliceert, dat de toetsen overwegend kennis meten, waarover een eerstejaars nog niet behoeft te beschikken. Twijfels over het nut van voortgangstoetsen in het eerstejaar zijn dan ook begrijpelijk en wellicht gerechtvaardigd. Onderzoek naar de bruikbaarheid van voortgangstoetsen in het eerstejaar van de geneeskunde studie leek dan ook geïndiceerd.
Eerder al onderzochten Wijnen (1984) en Imbos (1982) de betekenis van voortgangstoetsen voor het eerste studiejaar. Wijnen construeerde een eerstejaars subtoets op basis van de p-waarden (= percentage goede antwoorden) van de toetsvragen. Een vraag met een relatief hoge positie in de verdeling van p-waarden in het eerstejaar werd door hem beschouwd als een vraag die in een eerstejaars-subtoets thuis zou horen. De grens werd daarbij gelegd bij .40, zodat alle vragen met een p-waarde hoger dan .40 in de subtoets werden opgenomen.
Imbos onderzocht de voorspellende waarde van voortgangstoetsen in het eerstejaar ten opzichte van latere fases in de studie. De voorspelbaarheid van de voortgangstoetsen werd vergeleken met die van de bloktoetsen.Beide studies werden in het voorjaar van 1983 opnieuw in de discussie gebracht door het JOC-1 (= vergadering van de coördinatoren die de verantwoordelijkheid voor het eerstejaar hebben). De aanleiding voor de discussie was het invoeren van een nieuw examenreglement. Daarin werd het mogelijk gemaakt om ook bloktoetsen te betrekken bij het beoordelen van de studenten. Dit leidde tot protesten van studenten en sommige docenten. Het alternatief voor hetgeen in het concept-reglement werd voorgesteld was, dat de beoordeling van de kennis van de studenten in het eerstejaar alleen op basis van de resultaten op de voortgangstoetsen zou plaats vinden.
De uiteindelijke conclusie van deze discussie was dat het zinvol geacht werd om een gedetailleerde studie te doen naar de vraag die in hoofdstuk 1 (paragraaf 1.7) geformuleerd werd als:

I Bevatten voortgangstoetsen items die door beginnelingen in het medische domein beantwoord kunnen worden?

2.2 Selectie van Eerstejaarsitems

2.2.1 Procedure

Alle vier in het studiejaar 1983-1984 afgenomen voortgangstoetsen werden door de coördinatoren van de zes eerstejaars blokken (JOC-1) beoordeeld op de vraag, welke toetsvragen betrekking hadden op de inhoud van hun blok. De coördinatoren beoordeelden de vragen onafhankelijk van elkaar. Daarnaast werden alle vier voortgangstoetsen, met dezelfde opdracht, door de jaarcoördinator beoordeeld. De toetsvragen die zowel door de blokcoördinator als door de jaarcoördinator als een eerstejaars-vraag werden aangemerkt, werden zonder verdere discussie beschouwd als vragen die deel zouden moeten uitmaken van de Eerstejaarstoets. Vragen waarover men van mening verschilde werden door de jaarcoördinator en zijn voorganger besproken. In onderling overleg bepaalde men vervolgens of een dergelijke vraag in de Eerstejaarstoets opgenomen zou moeten worden. Deze selectie van items werd aan de onderzoeker doorgegeven. Informatie over de door individuele coördinatoren geselecteerde items was voor de onderzoeker derhalve niet beschikbaar.
Ter bepaling van de interbeoordelaars-overeenstemming van deze itemselectie-procedure, werden de vier toetsen ook nog eens beoordeeld door iemand die als een deskundige van het Maastrichts curriculum bij uitstek beschouwd mag worden, n.l. de auteur van het rapport "Curriculumoverzicht Faculteit der Geneeskunde" (Snellen,1982). Deze, ten opzichte van het JOC-1 als extern te beschouwen beoordelaar, gaf bovendien voor de door haar geselecteerde vragen nog aan op welk eerstejaars-blok de vragen betrekking hadden. Hierdoor werd het mogelijk om aan te geven of een bepaalde vraag, op het moment van afname van de toets, al dan niet in het onderwijsprogramma aan de orde was geweest.
Centraal in deze beoordelingsprocedures stond de vraag of de inhoud van een toetsitem redelijkerwijs bestudeerd zou kunnen worden door een eerstejaars student, die zich bij zijn studie laat leiden door het curriculum, zoals dat terug te vinden is in de blokboeken. De opdracht van de beoordelaars was dus met andere woorden, om na te gaan of de inhoud van de toetsitems in de voortgangstoetsen al dan niet binnen de thematiek van de eerstejaars blokken viel. In deze beoordelingsprocedure moest voor elk item worden aangegeven, of het al dan niet past binnen de doelstellingen van de eerstejaars blokken, die voor ieder blok globaal zijn omschreven. De taak van de beoordelaars kwam dus neer op het vergelijken van de inhoud van de vraag met de omschreven blokdoelstellingen en vervolgens te beslissen of een vraag wel of niet in de omschreven doelstellingen past.
Van Berkel (1984) omschrijft een dergelijke beoordelingstaak als het bepalen van

de relevantie van een toetsvraag. "De relevantie van een vraag is de mate waarin de vraag meet wat deze bedoelt te meten; de mate waarin de toetsvraag een of meerdere doelstellingen, of deeldoelstellingen dekt." (blz. 33). De relevantie van een toetsvraag, opgevat in de hierboven vermelde betekenis, is één van de uitwendige kenmerken van een toetsvraag die voor het samenstellen van toetsen, met behulp van inhoudsdeskundigen zou kunnen worden gebruikt. Het is bovendien één van de weinige uitwendige kenmerken van toetsvragen die door docenten correct gehanteerd blijken te kunnen worden. Docenten blijken niet in staat te zijn andere uitwendige kenmerken van toetsvragen, zoals de gemakkelijkheid en het discriminerend vermogen, goed te kunnen beoordelen (Van Berkel, 1984).
Zowel uit het onderzoek van Van Berkel als dat van Rovinelli en Hambleton (1977) kan men de conclusie trekken dat inhoudsdeskundige beoordelaars in staat moeten worden geacht om uit een voorraad toetsvragen, de vragen te selecteren, die bepaalde onderwijsdoelstellingen dekken. Dit laatste nu, was precies wat in de onderhavige studie van de beoordelaars werd gevraagd. De gebruikte beoordelingsprocedure kan dus op basis van de aangehaalde literatuur verdedigd worden als een procedure, waarvan verwacht mag worden dat ze tot het beoogde resultaat zal leiden, namelijk het samenstellen van voortgangstoetsen die relevant zijn voor de onderwijsdoelstellingen in het eerste studiejaar van de Faculteit der Geneeskunde van de Rijksuniversiteit Limburg.

2.2.2 Resultaten beoordelingsprocedure

Met behulp van de hiervoor beschreven beoordelingsprocedure werden de Eerstejaarsitems geselecteerd. In totaal werden door de beoordelaars 1029 toetsvragen beoordeeld. Van deze 1029 toetsvragen werden 226 (22%) geschikt gevonden om opgenomen te worden in een Eerstejaarstoets. De externe beoordelaar achtte in totaal 189 (18%) toetsvragen voor het gestelde doel geschikt. De resutaten in termen van aantallen al dan niet door beide beoordelende instanties geselecteerde toetsvragen worden vermeld in tabel 2.1.
Uit de resultaten in deze tabel blijkt dat de beide beoordelende instanties het oneens waren over in totaal slechts 71 van de 1029 beoordeelde toetsvragen. In percentages is dat 7% van de vragen waarover men het oneens was. De op basis van deze getallen berekende, voor toeval gecorrigeerde, overeenstemming leidde tot een kappa van .78 (Schouten, 1985). Op de verschillen tussen de twee beoordelende instanties zal bij de bespreking van de inhoudelijke kenmerken van de geselecteerde subtoetsen nader worden ingegaan. Hier kan worden volstaan met de constatering dat ook in dit geval is gebleken dat docenten het redelijk eens blijken te zijn wanneer het gaat om de vraag of een toetsvraag wel of niet de doelstellingen van het onderwijs dekt.

TABEL 2.1
Overeenstemming bij indeling van toetsvragen in: Eerstejaarsitems en Algemene Kennisitems

		JOC 1		
		Eerste-jaars items	Algemene Kennis items	Totaal
Externe beoordelaar	Eerstejaars items	172	17	189
	Algemene kennis items	54	786	840
	Totaal	226	803	1029

In het vervolg van het hier te beschrijven onderzoek werd verder gewerkt met de door het JOC-1 aangewezen vragen. De resultaten van de externe beoordelaar dienden als controle op de gevolgde beoordelingsprocedure. Gezien de grote mate van overeenkomst in de beoordelingsuitkomsten lijkt dit verdedigbaar.
Zoals hierboven werd beschreven, werden door het JOC-1 in totaal 226 toetsvragen, als Eerstejaarsitems aangemerkt. De verdeling van deze vragen over de vier voortgangstoetsen, wordt vermeld in tabel 2.2.

TABEL 2.2
Overzicht Eerstejaarsvragen in de vier voortgangstoetsen

	sept.'83	dec.'83	mrt.'84	mei '84	Totaal
Algemene Kennis-vragen	224	200	190	189	803
Eerstejaars-vragen	51 (19%)	53 (21%)	61 (24%)	61 (24%)	226
Totaal	275	253	251	250	1029

Uit tabel 2.2 blijkt dat tussen 19% en 24% van de gebruikte vragen in een toets door de beoordelaars als Eerstejaarsitems werd aangemerkt.
Gezien het feit dat het hier om voortgangstoetsen gaat, die op het einddoel van de opleiding tot arts zijn afgestemd, is dit een verrassende constatering. Immers voortgangstoetsen worden geconstrueerd zonder rekening te houden met specifieke onderdelen van het curriculum. Dit is ook de reden waarom voortgangstoetsen in verschillende curricula toepasbaar zijn gebleken (Bender, e.a., 1984). Bovendien zijn de hier vermelde percentages hoger dan op grond van een door de jaren heen gelijkmatige groei van het aantal beantwoorde items mag worden verwacht. Uitgaande van het gemiddeld door afgestudeerden beantwoorde items van 90% (70% goed + 20% fout) kan bij een gelijkmatige toename op de vier tijdstippen in het eerstejaar een percentage beantwoorde items van 4%, 7%, 10% en 14% worden verwacht.
Hier kan worden volstaan met de constatering dat het voor vier voortgangstoetsen in het academisch jaar 1983-1984 mogelijk is gebleken om subtoetsen samen te stellen, die volgens docenten passen in de doelstellingen van het eerstejaar in Maastricht. Uitgaande van een totale toetslengte van 250 items, kan de lengte van de door docenten geselecteerde Eerstejaarstoets redelijk worden genoemd. Ongeveer 1/5 deel van alle vragen hebben volgens docenten een relatie met het onderwijsprogramma in het eerste studiejaar. Of het samenstellen van eerstejaarssubtoetsen, ook in andere academische jaren mogelijk zal zijn, kan uiteraard op basis van deze studie alleen niet worden gezegd. Daartoe zou een herhaling van het onderzoek in een ander studiejaar nodig zijn. Hoewel ook dan geen sluitend bewijs geleverd kan worden.
Wel lijkt het verantwoord om te verwachten, dat een herhaling van het onderzoek dat hier beschreven wordt, tot het samenstellen van Eerstejaarstoetsen zou leiden, die qua aantallen vragen vergelijkbaar zijn met de subtoetsen uit het studiejaar 1983-1984. Immers de constructieregels voor voortgangstoetsen zijn voor ieder studiejaar gelijk. Ook de omvang van het vragenbestand is aan weinig fluctuaties onderhevig. Wel zijn er schommelingen in de inhoudelijke samenstelling van het vragenbestand. De te verwachten fluctuaties in een Eerstejaarstoets zouden dan ook vooral -evenals dat trouwens bij de totale voortgangstoets op dit moment nog steeds het geval is- inhoudelijk van aard zijn. Hoe de samenstelling van de Eerstejaarstoetsen in het onderhavige studiejaar was zal in paragraaf 2.4 aan de orde komen.
Deze resultaten samenvattend kan worden gesteld dat vraagstelling I door docenten in positieve zin wordt beantwoord: Voortgangstoetsen bevatten items die door eerstejaarsstudenten te beantwoorden zijn op grond van de relatie die de docenten constateerden tussen de inhoud van de items en het onderwijs in het eerstejaar.

De vraagstellingen Ia tot en met Id komen verderop in het hoofdstuk aan de orde. Eerst wordt beschreven op welke wijze de controle items, aangeduid als Algemene Kennisitems, werden geselecteerd.

2.3 Selectieprocedure van items (Algemene Kennistoetsen)

Dit onderzoek werd uitgevoerd om zicht te krijgen op een eventueel verschil in resultaten tussen gewone voortgangstoetsen en een subtoets daaruit, die met name voor eerstejaars geschikt zou zijn. Om vergelijkingen van welke aard dan ook te kunnen maken, moesten er voor elke Eerstejaarstoets een of meer controletoetsen worden geconstrueerd. De controletoets moest daarbij in ieder geval voldoen aan het criterium dat het afname tijdstip voor beide toetsen identiek zou zijn.
Uitgaande van deze eis waren er een aantal mogelijkheden. Het meest voor de hand liggend was om de totale voortgangstoets als controletoets te nemen. Het voordeel daarvan was dat de resultaten konden worden vergeleken met een criterium dat voor iedereen in de Faculteit der Geneeskunde duidelijk was. Het probleem was in dat geval dat sterk correlerende toetsen daarvan het resultaat zouden zijn. Een beter alternatief leek het te zijn om de rest van de voortgangstoets als controletoets te nemen. Het voordeel daarvan zou zijn: een controletoets die bijna perfect zou correleren met de totale voortgangstoets. Immers op vijftig tot zestig vragen na is het dezelfde toets als de totale voortgangstoets. Een nadeel van deze procedure is dat de controletoets vele malen (ongeveer vijf keer) langer zou worden dan de te onderzoeken Eerstejaarstoets. Voor een controletoets leek dit een minder geschikte eigenschap. Een andere mogelijkheid was om als controletoets een gestratificeerde aselecte steekproef te trekken, met als strata de deelgebieden uit de voortgangstoets. Deze mogelijkheid werd verworpen op grond van de overweging dat de samenstelling van de oorspronkelijke strata was verstoord door de selectie van Eerstejaarsitems. Een laatste mogelijkheid om een controletoets te maken, was het trekken van een aselecte steekproef, met eenzelfde aantal vragen als de Eerstejaarstoets, uit de rest van de voortgangstoets. Dit leek de best mogelijke controletoets op te leveren.
In dit onderzoek werden dus als de controletoets, waarmee de resultaten op de Eerstejaarstoets worden vergeleken, vier aselect getrokken steekproeven genomen van 51, 53, 61 en 61 vragen uit de resterende vragen van de voortgangstoetsen van september 1983, december 1983, maart 1984 en mei 1984. Deze toetsen zullen in het vervolg van dit onderzoek Algemene Kennistoetsen worden genoemd. De resultaten op de twee andere zojuist beschreven "controletoetsen" werden eveneens uitgerekend om, indien nodig, ook de vergelijkingen met deze twee subtoetsen uit te kunnen voeren. In ieder geval kon op deze wijze het effekt van de aselecte steekproeftrekking worden bestudeerd.

2.4 Beschrijving van Eerstejaarstoetsen

In deze paragraaf zal een aantal kenmerken van de Eerstejaarstoetsen beschreven worden, die een antwoord geven op de deelvragen Ia en Ib. De beschrijvingen van de Eerstejaarstoetsen zullen in de volgende paragrafen worden vergeleken met die van de totale voortgangstoetsen en met de aselect getrokken steekproeven, die aan het eind van de vorige paragraaf werden beschreven. Deze beschrijvingen zijn vooral gericht op de uiterlijke kenmerken van de toetsen. Uiterlijke kenmerken evenwel, die voor het interpreteren van de toetsresultaten, onontbeerlijk zijn.

2.4.1 Inhoudelijke beschrijving

Vragen in de voortgangstoetsen worden inhoudelijk op twee manieren beschreven in termen van de coderingen die standaard aan de toetsvragen worden toegekend. Elke vraag krijgt een code die aangeeft uit welke vakgroep de vraag afkomstig is. Deze vakgroepscode bestaat uit een letteraanduiding van maximaal drie posities voor de herkomst van de vraag en een cijferaanduiding van eveneens drie posities die het volgnummer van de vraag binnen de betreffende deelverzameling aangeeft.
De tweede codering geeft aan op welke categorie (orgaansystemen) uit de voortgangstoets de vraag betrekking heeft. Ook hier bestaat de codering uit twee delen, een aanduiding voor de categorie, aangevuld met een volgnummer binnen die categorie. Gebruik makend van deze twee coderingen is getracht een -zij het globale- beschrijving van de Eerstejaarstoetsen te geven. Tabel 2.3 geeft een overzicht van de verdeling van de vragen uit de vier Eerstejaarstoetsen naar herkomst van de verschillende vakgroepen en naar de categorieën uit de voortgangstoets. Tabel 2.3 biedt tevens de mogelijkheid om de Eerstejaarstoets in termen van deze verdelingen te vergelijken met de Algemene Kennistoets en de totale Voortgangstoets.
De in tabel 2.3 vermelde getallen zijn percentages van het totaal aantal vragen in de drie reeds genoemde toetsen. Het meest opvallende in de tabel is dat in alle vier Eerstejaarstoetsen de basisvakken relatief het sterkst vertegenwoordigd waren. In de vier toetsen waren deze percentages achtereenvolgens 65%, 45%, 43% en 44%. Deze percentages waren voor de vragen afkomstig uit de klinische vakken achtereenvolgens 12%, 25%, 31% en 36%. De gedragswetenschappelijke vakken tenslotte waren in de eerstejaarstoetsen vertegenwoordigd met respectievelijk 23%, 30%, 26% en 20% van de toetsvragen.

TABEL 2.3
Procentuele verdeling* van vragen over vakgroepen en categorieën in Eerstejaarstoets (Ej), Algemene Kennistoets (Ak) en de totale Voortgangstoets (VT)

TOETSEN	1			2			3			4		
	Ej	Ak	VT	Ej	Ak	VT	Ej	Ak	VT	Ej	Ak	VT
BASISVAKKEN:												
AE	4	6	4	2	8	10	5	10	8	10	7	7
BC	5	-	4	11	13	6	10	7	5	10	5	5
BF	6	-	2	8	-	2	5	-	2	2	3	1
FAR	8	6	7	8	4	7	8	2	5	-	8	3
FYS	6	-	1	6	4	5	2	-	0	3	3	3
GC	-	-	-	-	-	-	-	-	-	-	-	-
IM	12	2	3	6	8	5	8	3	4	8	3	5
MM	18	6	6	4	6	4	3	3	1	5	5	3
PA	6	-	2	2	6	3	2	3	2	7	3	6
Totaal	65	18	27	45	47	41	43	28	28	44	38	32
KLINISCHE VAKKEN:												
AG	-	-	-	-	-	-	-	-	-	-	-	-
AH	2	6	5	4	6	8	12	8	9	15	8	11
AN	-	-	-	8	-	2	2	-	1	2	-	0
CAR	4	2	3	4	-	4	3	-	2	2	5	4
DER	2	4	3	-	4	3	-	2	2	2	2	3
GO	-	2	5	-	2	1	-	5	3	-	-	2
HAG	-	8	4	-	-	3	-	2	5	-	3	3
IG	-	10	13	4	2	3	3	7	8	10	5	9
KG	2	8	9	-	11	6	-	13	7	-	8	8
KNO	-	-	2	-	-	-	-	2	1	-	-	-
NEU	-	6	3	-	-	1	3	5	5	2	3	4
OHK	-	6	3	-	4	2	-	-	2	-	-	-
ORT	-	8	4	-	2	2	-	2	3	2	7	4
PUL	2	2	2	6	-	4	3	2	2	3	3	3
REV	-	2	3	-	4	3	5	10	7	-	2	1
Totaal	12	62	56	25	34	42	31	56	54	36	46	50

TABEL 2.3 (vervolg)

GEDRAGSWET. VAKKEN:												
EGZ	6	2	3	6	4	3	-	2	2	-	2	1
EP	4	-	1	4	-	1	12	2	3	7	-	2
GZR	2	-	0	2	-	1	7	3	3	3	2	3
KP	-	6	3	-	2	1	2	2	2	7	2	4
ME	-	-	-	2	2	3	-	2	1	-	3	3
MP	6	2	3	9	-	3	7	5	5	2	5	3
MS	6	4	4	6	4	3	-	2	0	2	2	1
SP	-	2	2	2	6	3	-	-	2	-	2	2
Totaal	23	19	17	30	19	17	26	16	18	20	16	18
CATEGORIEEN:												
1	10	6	12	15	4	10	8	11	11	10	10	10
2	14	4	7	13	6	7	15	3	6	16	2	6
3	4	10	10	8	8	8	3	8	9	3	8	8
4	2	10	7	11	6	6	8	7	7	7	8	7
5	-	6	7	2	8	6	2	5	5	5	7	7
6	10	10	9	19	4	12	21	5	11	21	7	11
7	4	6	6	6	8	7	8	8	7	3	7	7
8	4	4	4	2	4	4	3	5	6	7	3	5
9	12	2	5	8	4	4	5	5	5	5	3	5
10	6	8	9	-	13	9	-	7	8	3	7	9
11	-	8	7	4	8	6	5	11	6	2	13	7
12	4	12	8	-	11	10	3	11	12	2	13	9
13	16	6	6	4	6	5	8	3	3	12	5	5
14	6	4	3	6	2	3	7	3	3	5	-	2
15	6	2	3	4	6	3	2	2	2	2	3	2

* Ten gevolge van afrondingsfouten sommeren niet alle percentages tot 100%.

AE = Anatomie/Embryologie
AG = Arbeidsgeneeskunde
AH = Algemene Heelkunde
AN = Anesthesiologie
BC = Biochemie
BF = Biofysica
IG = Inwendige Geneeskunde
IM = Immunologie
KG = Kindergeneeskunde
KNO = Keel-, Neus- en Oorheelkunde
KP = Klinische Psychiatrie

TABEL 2.3 (vervolg)

CAR	= Cardiologie	ME	= Gezondheidsethiek en Wijsbegeerte
DER	= Dermatologie	MM	= Medische Microbiologie
EGZ	= Economie van de Gezondheidszorg	MP	= Medische Psychologie
EP	= Epidemiologie	MS	= Medische Sociologie
FAR	= Farmacologie	NEU	= Neurologie
FYS	= Fysiologie	OHK	= Oogheelkunde
GC	= Genetica/Celbiologie	ORT	= Orthopedie
GO	= Gynaecologie/Obstetrie	PA	= Pathologie
GZR	= Gezondheidsrecht	PUL	= Pulmonologie
HAG	= Huisartsgeneeskunde	REV	= Revalidatie
		SP	= Sociale Psychiatrie

Categorie	1	= ademhalingsstelsel
Categorie	2	= hematologisch en lymfatisch stelsel
Categorie	3	= spier en skeletstelsel
Categorie	4	= geestelijke gezondheidszorg
Categorie	5	= voortplantingsstelsel
Categorie	6	= hart en vaatstelsel
Categorie	7	= hormoonstelsel en metabolisme
Categorie	8	= huid en bindweefsel
Categorie	9	= maatschappelijke factoren t.a.v. ziekte en gezondheid(szorg)
Categorie	10	= spijsverteringsstelsel
Categorie	11	= nieren en urinewegen
Categorie	12	= zenuwen en zintuigen
Categorie	13	= rest
Categorie	14	= wetenschapsleer en methodologie
Categorie	15	= persoonsgebonden factoren t.a.v. ziekte en gezondheid(szorg)

In vergelijking met de totale voortgangstoets en de Algemene Kennistoets zijn deze percentages opvallend, omdat in de twee laatstgenoemde toetsen de verhouding tussen basisvakken en klinische vakken -toets 2 uitgezonderd- precies andersom was. Met andere woorden, in de vier Eerstejaarstoetsen waren de basisvakken sterker vertegenwoordigd dan in de vier voortgangstoetsen, waaruit ze een steekproef vormden. Ook de gedragswetenschappelijke vakken waren in de Eerstejaarstoetsen in vergelijking met de Voortgangstoetsen enigszins oververtegenwoordigd.
De Algemene Kennistoets en de voortgangstoets waren qua gedrags-wetenschappelijke vakken goed vergelijkbaar. In de laatste twee toetsen waren deze percentages in de Algemene Kennistoets en de Voortgangstoets weer vergelijkbaar.
Wat betreft de verdeling van de vragen over de categorieën, waren de categorieën 2 (hematologisch en lymfatisch stelsel) en 6 (hart- en vaatstelsel) in de Eerstejaarstoets in vergelijking tot de controletoetsen sterk vertegenwoordigd. Dat was in mindere mate het geval met de categorieën 1 (ademhalingsstelsel), 9 (maatschappelijke faktoren t.a.v. ziekte en gezondheidszorg) en 13 (een niet nader omschreven restcategorie).
Uit nadere analyses van de vragen waarover de beoordelaars van het JOC-1 en de externe beoordelaar met elkaar van mening verschilden, kwam vast te staan dat deze vragen in termen van de hier gehanteerde indelingscriteria basisvakken, klinische vakken, gedragswetenschappelijke vakken en de orgaansystemen noch over- noch ondervertegenwoordigd waren. Met andere woorden geen van de twee beoordelende instanties had systematisch vragen uit bepaalde categorieën in de Eerstejaarstoets geweerd. De vragen waarover men van mening verschilde waren in ongeveer gelijke mate uit de genoemde indelingen afkomstig.
Tenslotte kunnen de Eerstejaarstoetsen worden beschreven in termen van de blokken waarop de vragen, volgens de externe deskundige, betrekking hadden. In tabel 2.4 wordt daarvan een samenvatting gegeven.
Uit tabel 2.4 blijkt dat niet alle onderwijsblokken in dezelfde mate in de Eerstejaarstoetsen vertegenwoordigd waren. Wanneer blok 1.1 buiten beschouwing wordt gelaten omdat dit blok bedoeld is als een kennismaking met het Maastrichtse Onderwijssyteem, dan zou men gezien de gelijke lengte van ieder blok (6 weken in totaal) een uniforme verdeling van de Eerstejaarsvragen over de blokken 1.2 tot en met 1.6 verwachten. Tegen deze achtergrond kan worden gesteld dat de blokken 1.2 en 1.4 in de Eerstejaarstoets binnen de voortgangstoets ondervertegenwoordigd waren.

TABEL 2.4
Procentuele verdeling van Eerstejaarsvragen over de onderwijsactiviteiten in het eerste jaar

Onderrwijs-activiteit	toets 1	2	3	4	tot
blok 1.1	2	-	-	1	1
blok 1.2	-	7	11	9	7
blok 1.3	20	9	11	22	16
blok 1.4	11	11	5	6	8
blok 1.5	13	24	20	10	16
blok 1.6	18	6	15	22	16
Trainingsweek	4	4	3	4	4
Computerpracticum	7	-	3	-	2
Orientatie eerste lijn	-	2	-	1	1
Voorkennis	-	2	9	4	4
Skillslab	2	2	5	4	3
Niet als eerstejaarsvraag herkend door externe beoordelaar	24	37	17	15	22
Totaal aantal vragen	55	54	66	68	243

Het best vertegenwoordigd waren de blokken 1.3, 1.5 en 1.6. Tenslotte blijkt uit de tabel 2.4 dat relatief weinig vragen volgens de externe beoordelaar op grond van middelbare schoolkennis te beantwoorden waren. Slechts 4% van het totaal aantal vragen in de Eerstejaarstoetsen werd als zodanig beoordeeld.
In deze paragraaf werd op grond van de vakgebieden, de voortgangstoetscategorieën en de onderwijsaktiviteiten in het eerste studiejaar, een inhoudelijke beschrijving van de Eerstejaarstoetsen gegeven. Hiermede zijn de vraagstellingen Ia en Ib in beschrijvende zin beantwoord.
De categorieën 2 en 6 waren in de Eerstejaarstoetsen relatief sterk vertegenwoordigd. Ook de categorieën 1, 9 en 13 waren, zij het in mindere mate, redelijk vertegenwoordigd.
Ten aanzien van de verdeling naar vakgebieden bleek dat vooral de basisvakken sterk vertegenwoordigd waren.

Van de onderwijsaktiviteiten waren de blokken 1.3, 1.5 en 1.6 het best vertegenwoordigd in de Eerstejaarstoetsen.
In de volgende paragraaf zullen de Eerstejaarstoetsen en van de Algemene Kennisstoetsen worden vergeleken met betrekking tot hun relevantie voor de basisarts. De beschrijving daarvan is relevant in het kader van deelvraagstelling Ic.

2.4.2 Relevantie van de Eerstejaarstoetsen

Jarenlang werden de voortgangstoetsen door een referentiegroep bestaande uit ± 60 artsen in opleiding voor huisarts beantwoord en op relevantie voor de basisarts beoordeeld. De reden daarvoor was het ontbreken van enig inzicht in wat voor afgestudeerden op dit soort toetsen een haalbare score zou zijn. Het beoordelen van de relevantie gebeurde door elke toetsvraag op een driepuntsschaal te scoren in één van de volgende categorieën: relevant, neutraal, irrelevant.
Van deze gegevens werd hier gebruik gemaakt om de relevantie van de Eerstejaarstoetsen te beoordelen. Voor iedere vraag uit de vier voortgangstoetsen, uit het onderhavige studiejaar, werd als relevantiescore genomen het percentage personen uit de referentiegroep dat de betreffende vraag als relevant beoordeelde. Deze relevantiescores van Eerstejaarstoetsen en Algemene Kennistoetsen werden vervolgens vergeleken. De resultaten daarvan worden vermeld in tabel 2.5.

TABEL 2.5
Relevantiebeoordeling van de vragen in Eerstejaarstoetsen (Ej) en Algemene Kennistoetsen (Ak)

	1		2		3		4	
	Ej	Ak	Ej	Ak	Ej	Ak	Ej	Ak
Gemiddelde	.59	.67	.63	.62	.58	.61	.80	.79
Std. dev.	.23	.20	.18	.15	.17	.16	.16	.17
Maximum	.94	.91	.89	.82	.79	.82	1.0	1.0
Minimum	.03	.12	.29	.25	.21	.29	.32	.39
t-waarde		-1.82 NS		.30 NS		-1.42 NS		.47 NS
Aantal vragen		51		53		61		61

NS = niet significant; significantieniveau = .013, tweezijdig getoetst.

Uit de tabel blijkt dat in geen van de vier gevallen de relevantie van de Eerstejaarstoetsen significant van die van de Algemene Kennistoetsen verschilde. Bij de eerste toets waren de verschillen het grootst, hoewel ook in dit geval niet significant zoals uit de resultaten van de t-toets onderaan de tabel blijkt. In alle gevallen vertoonde de verdeling van relevantiescores een grote spreiding. Deze spreiding was voor de Eerstejaarstoets en voor de Algemene Kennistoets ongeveer gelijk. Bij beide toetsen kwamen met andere woorden nogal extreme relevantiescores voor. De totale voortgangstoetsen hebben normaal een relevantiescore variërend van 60% tot 78% (Verwijnen en Fröberg, 1986). Tegen deze achtergrond, vallen de schommelingen in de relevantiescores van de hier besproken Eerstejaarstoetsen en Algemene Kennistoetsen binnen de normale variatiebreedte.
Uit deze gegevens blijkt dat vraagstelling Ic negatief moet worden beantwoord: er waren geen verschillen in relevantie tussen Eerstejaarsitems en Algemene Kennisitems.

2.4.3 Verdeling van juist en onjuist vragen in beide toetsen

Een beschrijving van de toetsen is niet compleet zonder dat aandacht is besteed aan de verdeling van de juist en onjuist vragen (vraagstelling Id). Dit is nodig omdat uit verschillende onderzoekingen (Ebel, 1972, Grosse en Wright, 1985 en Verwijnen, Imbos, Van Hessen en Wijnen, 1987) bekend is dat vragen waarop het bedoelde antwoord juist is, vaker goed worden beantwoord dan vragen met onjuist als bedoeld antwoord. In deze studie gaat het om het bestuderen van het antwoordgedrag van eerstejaarstudenten op toetsvragen die qua inhoud betrekking hebben op het eerste studiejaar. Dit antwoordgedrag wordt bestudeerd in vergelijking met toetsvragen die geen correspondentie zouden hebben met het eerste studiejaar. Vandaar dat de verdeling van de juist- en onjuist-vragen over de Eerstejaarstoetsen en de Algemene Kennistoetsen van direct belang is voor deze vergelijking. Immers een eventueel verschil in verdeling van juist- en onjuist-vragen zou een vergelijking al bij voorbaat in de een of de andere richting beïnvloeden.
De verdeling van de twee vraagtypes over de verschillende toetsen is weergegeven in tabel 2.6. Uit deze tabel blijkt dat de verdeling van de juist- en onjuist-vragen over de vier toetsen nagenoeg identiek was. Alleen de derde toets had een afwijkende verdeling. Hier was het aantal onjuist vragen in de Eerstejaarstoets groter dan in de Algemene Kennistoets. Dit verschil was niet significant, de chi-kwadraat voor de subtabel behorend bij de derde toets was 4.73. De chi-kwadraat waarden voor de eerste, tweede en vierde toetsen waren respectievelijk: 0.35, 0.94, en 0.00. Geen van deze waarden is significant op het, voor het aantal uitgevoerde toetsen, gecorrigeerde signficantieniveau van .013.

TABEL 2.6
Verdeling van juist en onjuist vragen over de twee toetsen op vier verschillende tijdstippen

Tijdstip	1		2		3		4	
	Ej	Ak	Ej	Ak	Ej	Ak	Ej	Ak
onjuist	28	25	25	30	38	26	30	30
juist	23	26	28	23	23	35	31	31
	NS		NS		NS		NS	
totaal	51		53		61		61	

NS = niet significant; significantieniveau = .013.

Met het bovenstaande werd een uitvoerige beschrijving gegeven van de wijze waarop de Eerstejaarstoetsen werden samengesteld. Bovendien werd uitvoerig stil gestaan bij een aantal inhoudelijke kenmerken van de Eerstejaarstoetsen, die daarbij werden vergeleken met de vermelde controletoetsen.
De vraagstellingen I en Ia tot en met Id zijn daarmee beschrijvenderwijs beantwoord: de voortgangstoetsen bevatten Eerstejaarsitems, die qua relevantie en qua aantal juist en onjuist vragen niet verschillen van de overige vragen.
Deze beschrijving van de constructiemethode zou echter niet volledig zijn als er geen vergelijking met een andere voor de hand liggende item-selectiemethode, namelijk die gebaseerd op p-waarden, zou plaats vinden. Deze vergelijking van een inhoudelijke toetsconstructie en een op de empirie gebaseerde methode zal in de volgende paragraaf aan de orde worden gesteld.

2.5 Inhoudelijke itemselectie vergeleken met empirische itemselectie

2.5.1 Vergelijking qua aantal geselecteerde items

Veel handboeken op het gebied van toetsconstructie ten behoeve van het meten van studieresultaten (Nunnally, 1967, De Groot en Van Naerssen, 1969, Thorndike, 1982 en Ebel, 1972) benadrukken het belang van een inhoudelijke toetsconstructiemethode. Een dergelijke methode gaat er van uit dat de toets en de inhoud van het onderwijsprogramma goed op elkaar zijn afgestemd. Een eerste stap in de toetsconstructie volgens deze methode is dan ook altijd een grondige analyse van het onderwijsprogramma.
Voortgangstoetsen worden in feite volgens dit principe samengesteld. De gehanteerde blauwdruk moet in dit verband worden gezien als het resultaat van een

dergelijke analyse, zij het in dit geval een analyse van het totale medische domein waarop de onderwijsprogramma's voor de basisarts zich richten. Eerstejaarstoetsen zijn eveneens samengesteld op basis van een inhoudelijke correspondentie met het eerstejaars programma van de Faculteit der Geneeskunde Maastricht. Dit leverde de inmiddels besproken selectie van toetsvragen in de vier Eerstejaarstoetsen op.

2.5.2 Vergelijking van moeilijkheidsgraad van de geselecteerde items

De vraag kan nu gesteld worden in hoeverre een selectiemethode die gebaseerd op de feitelijke antwoordpercentages van de studenten, en die in het vervolg de empirische methode genoemd zal worden, tot een van de inhoudelijke methode sterk afwijkende itemselectie geleid zou hebben. Op deze vraag wordt hier nader ingegaan.
Het criterium dat werd gebruikt is, dat het aantal geselecteerde vragen met de hoogste percentages goede antwoorden gelijk moest zijn aan het aantal vragen dat door de coördinatoren was geselecteerd. Dit komt er op neer dat de n- best beantwoorde vragen worden geselecteerd, waarbij n gelijk is aan het aantal vragen uit de inhoudelijke selectie. Het criterium werd toegepast op eerstejaars- tweedejaars- en zesdejaarsgegevens. De tweedejaars gegevens werden onderzocht omdat tweedejaars, qua kennisniveau nog het meest lijken op de eerstejaars en bovendien schattingen van de itemmoeilijkheden opleveren die onafhankelijk zijn van de schattingen op basis van het eerstejaar. De zesdejaars gegevens werden geanalyseerd om na te kunnen gaan of de docenten die de Eerstejaarstoetsen hadden geselecteerd dit niet voornamelijk op grond van de itemmoeilijkheden hadden gedaan, aannemende dat docenten de moeilijkheid van toetsvragen het best kunnen inschatten voor een groep die qua kennisniveau niet te ver van het eigen kennisniveau afstaat.
Aldus ontstonden er voor ieder tijdstip drie, langs empirische weg geselecteerde Eerstejaarstoetsen, die elk werden vergeleken met de op inhoudelijke gronden geselecteerde toetsen. Het resultaat van deze vergelijking is samengevat in tabel 2.7.
Om aan het vereiste aantal vragen te komen, moesten er in de drie jaargroepen op achtereenvolgende tijdstippen, de volgende grenzen voor de p-waarden worden gesteld: in jaar 1: .13, .14, .18 en .23; in jaar 2: .33, .40, .42 en .47; in jaar 6: .78, .76, .75 en .80.
Uit de tabel blijkt dat de overeenstemming tussen beide methoden in het eerstejaar en in het tweedejaar meer dan toevallig was. In het eerstejaar vertoonden de waarden van kappa een stijgende tendens. In het tweedejaar waren de kappa's op de eerste tijdstippen het hoogst.

TABEL 2.7
Overeenstemming tussen inhoudelijke selectie en een selectie op p-waarden bij gelijke toetslengte

		tijdstip 1			tijdstip 2			tijdstip 3			tijdstip 4		
		p-waarden eerste jaar											
		+	-	t	+	-	t	+	-	t	+	-	t
	+	19	32	51	23	30	53	38	23	61	37	24	61
inhoud	-	32	192	224	30	170	200	23	167	190	24	165	189
	t	51	224	275	53	200	253	61	190	251	61	189	250
		kappa = .23			kappa = .27			kappa = .51			kappa = .49		
		p-waarden tweede jaar											
		+	-	t	+	-	t	+	-	t	+	-	t
	+	24	27	51	28	25	53	33	28	61	28	33	61
inhoud	-	27	197	224	25	175	200	28	162	190	33	156	189
	t	51	224	275	53	200	253	61	190	251	61	189	250
		kappa = .35			kappa = .39			kappa = .40			kappa = .27		
		p-waarden zesde jaar											
		+	-	t	+	-	t	+	-	t	+	-	t
	+	9	42	51	19	34	53	15	46	61	14	47	61
inhoud	-	42	182	224	34	166	200	46	144	190	47	142	189
	t	51	224	275	53	200	253	61	190	251	61	189	250
		kappa < .00			kappa = .18			kappa < .00			kappa < .00		

+: geselecteerd door de methode;-: niet geselecteerd door methode

Dit is te begrijpen gezien het feit dat beginnende tweedejaars qua kennisniveau het meest lijken op eerstejaars. Hoe verder in het tweedejaar, des te meer het onderwijsprogramma van het tweedejaar doorwerkt in de studieresultaten, hetgeen tot grotere verschillen zal leiden.
De overeenstemming tussen de inhoudelijke samenstelling en de empirische in het zesdejaar, was in drie van de vier gevallen niet groter dan men op grond van het toeval mag verwachten (drie kappa's hadden een waarde < 0).
Tenslotte kan de vraag gesteld worden welke relatie er bestaat tussen de itemselectie-strategie en de moeilijkheidgraad van de toetsitems in het eerstejaar.

De vraag is met andere woorden of er door het hanteren van een bepaalde strategie of combinatie van twee strategieën relatief moeilijke of relatief gemakkelijke vragen worden geselecteerd. Het effect van de empirische methode, gebaseerd op p-waarden in het eerstejaar, is uiteraard voorspelbaar. Alle volgens dit criterium geselecteerde items hebben noodzakelijkerwijs een relatief hoge p-waarde. Bovendien moet de gemiddelde p-waarde van deze items hoger zijn dan die van de items die niet geselecteerd werden of alleen door de inhoudelijke methode geselecteerd werden. Dit is niet het geval wanneer er langs empirische weg items worden geselekteerd in het tweede studiejaar.
Over de verschillen in gemiddelde p-waarde in het eerstejaar van de items die alleen door de empirische methode werden geselecteerd en de items die door beide methodes werden geselecteerd is echter op voorhand, zonder inspectie van de data niets te zeggen. De empirie kan zowel op de hand van de combinatie inhoud en empirie zijn als op de hand van alleen de empirische methode zijn.
Om de relatie tussen itemselectie-strategie en de p-waarden van de items in het eerstejaar te bestuderen werden drie itemgroepen met elkaar vergeleken:

1. items alléén door de inhoudelijke methode geselecteerd,
2. items alléén door de empirische methode geselecteerd, en
3. items door beide methodes geselecteerd.

De resultaten van de vergelijkingen staan in tabel 2.8.
Uit tabel 2.8 blijkt dat de drie itemgroepen qua gemiddelde p-waarde in het eerstejaar tot dezelfde rangorde leidde, ongeacht het studiejaar waarop de empirische methode werd toegepast. Het moeilijkste waren de vragen die alléén door de inhoudelijke methode werden geselecteerd. De vragen die alléén door de empirische methode waren geselecteerd vormden een tussen-categorie, terwijl de vragen die op grond van beide methoden waren geselecteerd in alle gevallen voor eerstejaars het gemakkelijkst te beantwoorden waren. Hoewel de verschillen niet allen significant waren, was deze trend duidelijk aanwezig. Hierbij werden twee empirische methoden gebruikt: één gebaseerd op eerstejaars p- waarden en één gebaseerd op tweedejaars p-waarden.

TABEL 2.8
Relatie tussen selectiestrategie en gemiddelde p-waarden in het eerstejaar

		jaargroep waarop empirische selectie gebaseerd was							
		1				2			
tijdstip		1	2	3	4	1	2	3	4
selectie-strategie									
	N	32	30	23	24	27	25	28	33
A	M	4	7	8	14	4	10	14	21
	sd	2	3	6	6	4	7	15	14
	N	32	30	23	24	27	25	28	33
B	M	24	25	30	40	22	18	24	31
	sd	12	11	11	16	15	15	14	21
*	t	9.34	9.14	8.21	7.45	6.02	2.42	2.58	2.27
	p	+	+	+	+	+	0.08	0.006	0.024
	N	19	23	38	37	24	28	33	28
C	M	36	40	40	49	29	31	40	49
	sd	16	20	19	18	19	24	20	19
**	t	3.06	3.49	2.31	1.99	1.47	2.33	3.56	3.49
	p	+	+	0.012	0.02	0.08	0.016	+	+

A = alleen op inhoud geselecteerd
B = alleen op p-waarden geselecteerd
C = geselecteerd op inhoud en p-waarden
N = aantal items
M = gemiddelde p-waarde
* = methode 1 en 2 vergeleken
** = methode 2 en 3 vergeleken
\+ = significant op 0.005-niveau

Tenslotte werden de twee inhoudelijke itemselectie-strategieën vergeleken, d.w.z. de items die alleen door de externe deskundige en de items die alleen door het JOC-I werden geselecteerd. De resultaten van deze vergelijking staan in tabel 2.9.

TABEL 2.9
Verschillen in gemiddelde p-waarde in het eerstejaar tussen twee inhoudelijke selectiestrategieën.

tijd-stip	Externe deskundige			Jaarcommissie eerstejaar				
	N	M	sd	N	M	sd	t	p
1	4	2	1	13	20	16	2.20	0.02
2	1	21	–	20	23	17	niet bepaald	
3	5	20	16	11	27	21	0.68	0.25
4	7	42	27	10	22	15	1.96	0.03

Uit de gegevens die in tabel 2.9 worden weergegeven blijkt niets van een voorkeur van één van de twee beoordelende instanties ten opzichte van gemakkelijke of moeilijke items.
Van belang voor het onderhavige onderzoek is dat uit de verschillende vergelijkingen de conclusie kan worden getrokken dat het gebruiken van een inhoudelijke methode niet leidde tot het systematisch negeren van gemakkelijke of moeilijke items. Het bleek dat de items die alléén door de empirische methode zouden zijn geselecteerd, qua gemiddelde moeilijkheidsgraad voor eerstejaars precies lagen tussen de itemgroepen die door beide methoden en die alléén door de inhoudelijke methode zouden zijn geselecteerd. Gezien de spreiding in moeilijkheidsgraad van de laatste twee itemgroepen kan worden geconcludeerd dat de inhoudelijke methode óók dit midden-gebied bestrijkt.

2.6 Samenvatting en conclusies

In de voorgaande paragrafen werd beschreven op welke wijze er binnen vier bestaande voortgangstoetsen werd gezocht naar Eerstejaarstoetsen (paragraaf 2.2.1). Het bleek mogelijk om een panel van deskundigen vragen uit bestaande voortgangstoetsen te laten selecteren, die correspondeerden met het onderwijsprogramma in het eerstejaar. De overeenstemming met een andere deskundige bleek zeer hoog te zijn. Op een totaal van 1029 beoordeelde toetsvragen bleek men het eens te zijn over 975 vragen met betrekking tot de vraag of het toetsitem wel of geen inhoudelijke correspondentie had met het eerstejaars programma (paragraaf 2.2.2).
In de vier onderzochte voortgangstoetsen bleken tussen de 19% en 24% van het totaal aantal toetsvragen gerelateerd te zijn aan de eerstejaars doelstellingen. De

door het panel geselecteerde toetsvragen in elk van de vier voortgangstoetsen werden beschouwd als de Eerstejaarstoetsen. In vergelijking tot de reguliere voortgangstoetsen bleken de Eerstejaarstoetsen relatief gezien meer basisvak-vragen dan klinische vragen te bevatten. Bij de reguliere voortgangstoetsen is deze verhouding precies andersom. De Eerstejaarstoetsen bleken eveneens meer gedrags-wetenschappelijke items te bevatten dan de reguliere voortgangstoetsen. Wat betreft de relatie tussen de diverse onderdelen van het eerstejaars programma en de toetsvragen bleken de blokken 1.3, 1.5 en 1.6 sterk vertegenwoordigd te zijn in de Eerstejaarstoetsen (paragraaf 2.4.1).
De relevantie van de Eerstejaarstoetsen bleek niet te verschillen van de relevantie van een evenlange controle toets, die ter onderscheiding van de reguliere voortgangstoets, Algemene Kennistoets werd genoemd (paragraaf 2.4.2). Deze twee verkorte toetsversies hadden eenzelfde verdeling van het aantal juist en onjuist gesleutelde vragen. (paragraaf 2.4.3).
Tenslotte werd een vergelijking gemaakt tussen de in dit onderzoek gehanteerde inhoudelijke itemselectie-methode, waarbij gebruik wordt gemaakt van het oordeel van deskundigen en een empirische, op de p-waarden van de toetsvragen gebaseerde, selectie-methode (paragraaf 2.5.1).
Het bleek dat de toetsvragen die met behulp van de empirische methode geselecteerd zouden worden als zijnde eerstejaarsvragen, bij toepassing van de methode op eerstejaars- en tweedejaars data, een meer dan toevallige kans hadden om eveneens door de inhoudelijke methode geïdentificeerd te worden. De inhoudelijke methode bleek bovendien geen systematische onzuiverheid te vertonen voor relatief moeilijke of relatief gemakkelijke vragen in het eerstejaar (paragraaf 2.5.2).
Met betrekking tot de vraagstelling I kan op grond van de hier beschreven resultaten geconcludeerd worden dat naar de mening van docenten de voortgangstoetsen items bevatten, die corresponderen met het onderwijsprogramma in het eerste studiejaar.Op grond van deze correspondentie mag worden verwacht dat ook beginnelingen (in deze studie eerstejaars studenten) op deze items een redelijke kans op een goed antwoord hebben. Hoe groot die kans is werd nog niet besproken. Wel bleek dat een deel van de door de docenten geselecteerde items hoge p-waarden hadden. Deze resultaten zijn in overeenstemming met de veronderstelde hogere kans op een goed antwoord op deze items.
Niet alle door de docenten geselecteerde items bleken tot de groep met de hoogste p- waarden te behoren (paragraaf 2.5.2). De docenten selecteerden items over een breed bereik van p-waarden. Een verklaring daarvoor zou kunnen zijn dat docenten correspondentie tussen items en onderwijsprogramma toch niet goed kunnen beoordelen. Beoordelingsfouten kunnen niet worden uitgesloten. Het ligt echter meer voor de hand om aan te nemen dat de bedoelde items met lage p-waarden corresponderen met programmaonderdelen, die op het tijdstip van de toetsing nog

niet aan de orde waren geweest (zie ook hoofdstuk 3, paragraaf 3.2.9).
In deze studie wordt als werkhypothese aangenomen dat de door de docenten geselecteerde items door beginnelingen beantwoord kunnen worden. Als deze aanname klopt dan moeten er een aantal verschillen en overeenkomsten zijn tussen Eerstejaarsitems en willekeurige items (Algemene Kennisitems). Op deze verschillen en overeenkomsten richt vraagstelling II zich, nader uitgewerkt in de deelvragen. Het antwoord op vraagstelling II en de daarbij behorende deelvragen wordt in de volgende hoofdstukken gezocht door het toetsen van de in paragraaf 1.10 beschreven hypothesen.

Hoofdstuk 2: Selectie Eerstejaarsitems

3. VERGELIJKING EERSTEJAARSTOETSEN MET ALGEMENE KENNISTOETSEN (INTERNE ANALYSES)

Inhoud

3.1 Inleiding
3.2 Resultaten interne analyses
3.2.1 Verdelingen van itembeantwoording (hypothese 1)
3.2.2 Verdelingen van p-waarden (hypothese 2)
3.2.3 De variantie in p-waarden (hypothese 3)
3.2.4 Verdelingen van itemrestcorrelaties (hypothese 4)
3.2.5 Betrouwbaarheid van de toetsen (hypothese 5)
3.2.6 Verdelingen van toetsscores (hypothese 6)
3.2.7 Variantie in toetsscores (hypothese 7)
3.2.8 Itemvariantie en personenvariantie (hypothese 8)
3.2.9 Onderwijseffecten in Eerstejaarstoetsen (hypothese 9)
3.3 Samenvatting en conclusie

Hoofdstuk 3: Interne analyses

3.1 Inleiding

In dit hoofdstuk worden de hypothesen getoetst die betrekking hebben op de verschillen in interne structuur tussen Eerstejaarstoetsen en Algemene Kennistoetsen (hypothesen 1 tot en met 9).
Onder interne analyses wordt hier met Hofstee (1971) verstaan, de beschrijving van de verdelingen van itemindices, de interne consistentie en de verdeling van toetsscores van de hier bestudeerde toetsen. De resultaten in het eerstejaar zullen steeds worden vergeleken met die in het zesdejaar en in bepaalde gevallen met de resultaten van een groep afgestudeerden.
De vergelijking van de eerstejaars resultaten met die van ouderejaars en eventueel afgestudeerden is nodig om de verwachte kenmerken van de Eerstejaarstoetsen te kunnen beschrijven en te verklaren. Immers de verwachte kenmerken zouden zich alleen in het eerstejaar moeten voordoen en niet in de latere studiejaren en bij afgestudeerden. Voor de laatste groepen zou er geen verschil moeten zijn tussen resultaten op de Eerstejaarstoetsen en die op de Algemene Kennistoetsen (zie voor nadere argumentatie hoofdstuk 1, paragraaf 1.9).
Allereerst worden de verdelingen van de itemindices bestudeerd:

- het percentage van de respondenten dat het item beantwoordt. Deze index wordt itembeantwoording genoemd en is de som het percentage goede antwoorden en het percentage foute antwoorden. Het complement hiervan is het percentage dat het vraagtekenalternatief heeft gekozen.
  De verwachte verschillen werden geformuleerd in de hypothesen 1a, 1b en 1c, die op houdbaarheid worden onderzocht in paragraaf 3.2.1.

- het percentage respondenten dat het goede antwoord op het item heeft gekozen (de p-waarden). Het complement hiervan is het percentage foute antwoorden plus het percentage vraagtekens. De verwachting met betrekking tot deze resultaten werden geformuleerd in de hypothesen 2a, 2b en 2c. Deze hypothesen worden getoetst in paragraaf 3.2.2. De verwachte verschillen met betrekking tot de variantie in p-waarden, geformuleerd in de hypothesen 3a, 3b en 3c worden getoetst in paragraaf 3.2.3.

- de item-restcorrelatie van de items. Dit is een maat voor de associatie tussen het dichotome item (goed vs niet goed) en de totaalscores op de overige items. In de hypothesen 4a en 4b werden de verwachtingen met betrekking tot dit itemkenmerk geformuleerd. De resultaten van de toetsing van de hypothesen komen in paragraaf 3.2.4 aan de orde.

Na deze itemindices worden achtereenvolgens de resultaten vermeld van:

- de betrouwbaarheid van de Eerstejaarstoets en de Algemene Kennistoets. De verwachte verschillen werden in de hypothesen 5a en 5b geformuleerd. Deze hypothesen werden getoetst in paragraaf 3.2.5.

- de toetsscores, het percentage goede antwoorden per student. De verwachte verschillen in toetsscores werden geformuleerd in de hypothesen 6a en 6b. Deze worden getoetst in paragraaf 3.2.6.

- de variantie in de toetsscores. De verwachte verschillen die getoetst wordenin paragraaf 3.2.7, werden geformuleerd in de hypothesen 7a en 7b.

- de itemvarianties en de personenvarianties, waarover in de hypothesen 8a en 8b voorspellingen werden gedaan, die in paragraaf 3.2.8 getoetst worden.

- de onderwijseffecten, tot uiting komend in verwachte verschillen in p-waarden van items uit een voortoets en items uit een natoets van van Eerstejaarsitems. Hypothesen 9a en 9b geven de verwachte verschillen weer. Deze verwachtingen worden getoetst in paragraaf 3.2.9.

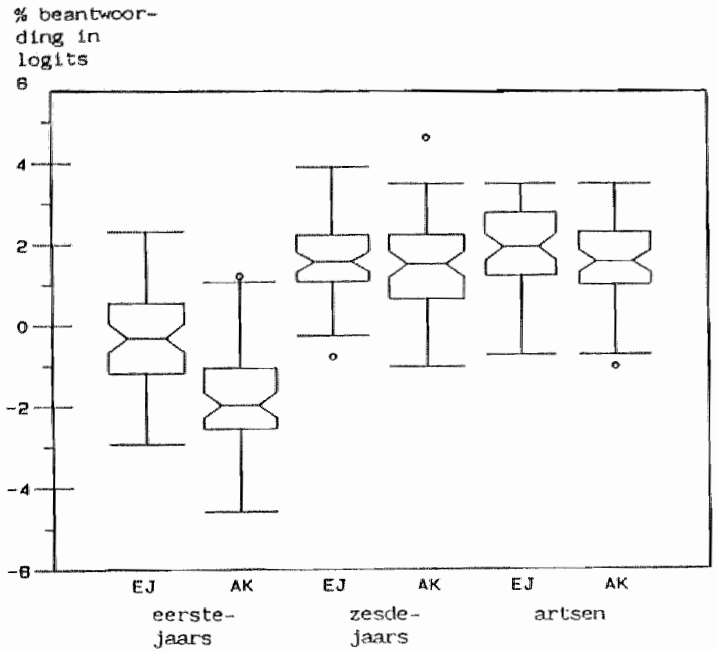
De hierboven genoemde item- en persoonsscores uitgedrukt in een percentage werden, voordat er mee werd gewerkt, getransformeerd met behulp van de logit transformatie (Mosteller and Tukey, 1977). Dit is een veel gebruikte transformatie voor proporties of percentages en leidt veelal tot min of meer normale verdelingen. De transformatie die werd toegepast is log (p/(100-p)), waarbij p een percentage is en 100-p het complement van dat percentage. Deze transformatie leidt tot positieve, negatieve of nulscores, afhankelijk van de hoogte van p. Als p groter is dan 50% is de logitscore positief, is p kleiner dan 50% dan is de logitscore negatief. Als p gelijk is aan 50% dan is de logitscore gelijk aan nul.
Hieronder zal uit de context steeds blijken welk percentage p wordt bedoeld.

3.2 Resultaten interne analyses

3.2.1 Verdelingen van itembeantwoording (hypothese 1)

Zoals reeds vermeld, werd voor alle toetsvragen een score voor itembeantwoording bepaald. In dit geval is p het percentage respondenten dat de vraag goed of fout beantwoordt en 100-p het percentage dat de vraag niet beantwoordt. De verdelingen van deze itemscores waren nagenoeg identiek op de vier tijdstippen. Om het verschil in locatie tussen de verdelingen te illustreren werd in figuur 3.1

derhalve alleen tijdstip 4 genomen. Een uitleg van deze grafische weergave is te vinden in Bijlage 1.
De verschuiving in de locatie van de verdelingen bij de drie onderzochte groepen respondenten werd getoetst met behulp van de t-toets voor twee onafhankelijke steekproeven. De resultaten van deze toetsingen zijn te vinden in tabel 3.1.



Figuur 3.1

Verdelingen van beantwoordingspercentage op tijdstip 4 (in logits) door drie groepen respondenten (Ej = Eerstejaarstoets, Ak = Algemene Kennistoets).

Uit de figuur en de tabel blijkt dat in het eerstejaar de Eerstejaarstoetsen naar verwachting meer werden beantwoord dan de Algemene Kennistoetsen. Hoewel dezelfde tendens ook bij de zesdejaars kon worden geconstateerd werd slechts één keer een significant resultaat gevonden. Bij de afgestudeerden was deze tendens ook aanwezig. Gemeten aan het exacte Bonferroni criterium werd hier echter geen enkel significant resultaat gevonden.

TABEL 3.1
Resultaten t-toetsen* voor onafhankelijk steekproeven met betrekking tot verschil in beantwoording

JAAR 1

Tijdstip	1	2	3	4
M Ej	-1.52	-.90	-.64	-.26
M Ak	-2.54	-2.25	-2.48	-1.98
t-waarde	4.72	5.76	7.48	7.86
p	.00	.00	.00	.00
aantal items	51	53	61	61

JAAR 6

Tijdstip	1	2	3	4
M Ej	1.33	1.75	1.46	1.70
M Ak	1.04	.90	1.12	1.43
t-waarde	1.53	4.17	1.67	1.37
p	.13	.00	.10	.17
aantal items	51	53	61	61

ARTSEN

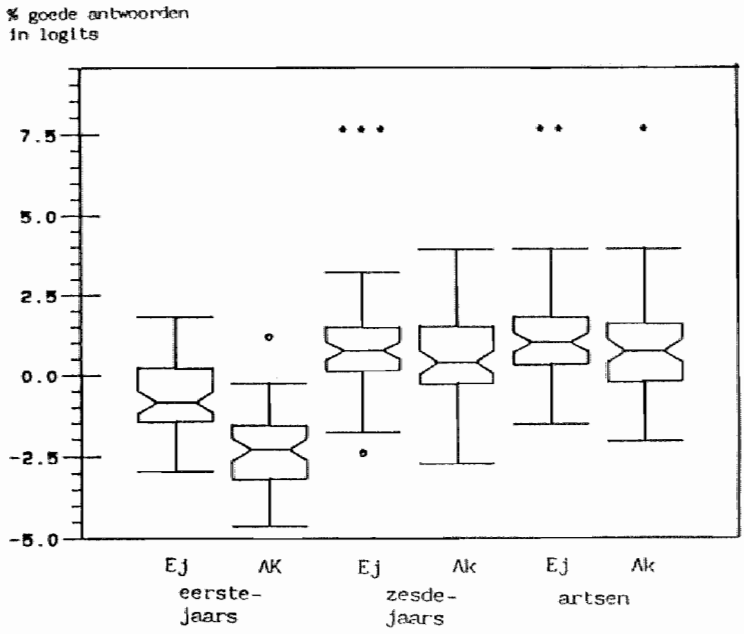
Tijdstip	1	2	3	4
M Ej	1.57	1.56	1.86	2.05
M Ak	1.61	1.13	1.63	1.70
t-waarde	-.18	2.35	1.03	1.67
p	.86	.02	.31	.10
aantal items	51	53	61	61

* jaar 1 eenzijdig ;jaar 6 en artsen tweezijdig getoetst.
Voor ieder tijdstip wordt een gecorrigeerde alfa van .05/3 = .016 (afgerond) gehanteerd.

In de eerstejaars groep werden in totaal vier toetsingen uitgevoerd. Er werden vier significante resultaten gevonden. In de controlegroepen werden in totaal acht toetsingen uitgevoerd. Er werd in totaal één significant resultaat gevonden. Voor de hypothese met betrekking tot het beantwoorden van de items komt dit neer op 11 successen en 1 mislukking. Immers in het licht van de hypothese geldt een significant resultaat in het eerstejaar en een niet-significant resultaat in de controlegroepen als succes. Een mislukking is een niet-significant resultaat in het eerstejaar en een significant resultaat in de controlegroepen. De resultaten in de eerstejaarsgroep en in de twee controlegroepen kunnen op grond van deze bevindingen dan ook worden opgevat als een ondersteuning van de theoretische overwegingen die met betrekking tot de begripsvaliditeit van de Eerstejaarstoetsen werden geformuleerd.
Aangetoond werd dat in het eerste studiejaar de Eerstejaarstoetsen gemiddeld meer werden beantwoord dan de controletoetsen. Op de vraag of het vaker beantwoorden van een item uit de Eerstejaarstoetsen ook terecht gebeurd, wordt in de volgende paragraaf ingegaan. Daar zullen de verdelingen van goede antwoorden nader worden onderzocht.

3.2.2 Verdelingen van p-waarden (hypothese 2)

In deze paragraaf zullen de verdelingen van goede antwoorden op de toetsitems worden onderzocht. Ten behoeve daarvan werden weer de logits (p/(100-p)) uitgerekend, waarbij p in dit geval het % goede antwoorden was en 100-p het % foute antwoorden + het % vraagtekens, dus de niet-correcte antwoorden. In de grafische weergave van de verdelingen wordt de y-as dus gevormd door de logits van de goede antwoorden. Ook in dit geval was de verschuiving in locatie van de verdelingen van Eerstejaarsitems en Algemene Kennisitems op de vier tijdstippen nagenoeg identiek.
De verdelingen van goede antwoorden op tijdstip 4 worden afgebeeld in figuur 3.2. De resultaten van de toetsingen zijn te vinden in tabel 3.2.
Uit deze figuur en tabel blijkt dat in het eerstejaar alle vier de toetsen een significant resultaat opleverden.
Bij de controlegroepen werden in totaal acht toetsen uitgevoerd, waarvan er twee met een significant resultaat waren en dus als mislukking aangemerkt dienen te worden. In totaal betekent dit voor de hypothese met betrekking tot de verschillen in goede antwoorden 10 geslaagde en 2 mislukte voorspellingen. Samenvattend kan dan ook worden gesteld dat net als bij het beantwoorden van de toetsvragen, ook de resultaten van de goede antwoorden gezien mogen worden als een ondersteuning van de begripsvaliditeit van de voortgangstoetsen.



Figuur 3.2

Verdelingen van p-waarden op tijdstip 4 (in logits) in drie groepen respondenten (Ej = Eerstejaarstoets, Ak = Algemene Kennistoets).

TABEL 3.2
Resultaten t-toetsen* voor onafhankelijke steekproeven met betrekking tot verschil in goede antwoorden

	JAAR 1			
Tijdstip	1	2	3	4
M Ej	-2.18	-1.63	-1.29	-.74
M Ak	-3.22	-2.90	-2.92	-2.48
t-waarde	4.52	5.99	6.59	8.02
p	.00	.00	.00	.00
aantal items	51	53	61	61

TABEL 3.2 (vervolg)

JAAR 6				
Tijdstip	1	2	3	4
M Ej	.26	.66	.45	.73
M Ak	-.05	-.37	.18	.38
t-waarde	1.20	4.66	1.16	1.47
p	.23	.00	.25	.14
aantal items	51	53	61	61

ARTSEN				
Tijdstip	1	2	3	4
M Ej	.39	.67	.76	1.05
M Ak	.38	-.07	.43	.63
t-waarde	.07	3.29	1.40	1.77
p	.94	.00	.16	.08
aantal items	51	53	61	61

* in jaar 1 eenzijdig getoetst; andere groepen tweezijdig getoetst: .05/3 = .016 op ieder tijdstip

3.2.3 De variantie in p-waarden (hypothese 3)

Eerder in hoofdstuk 1, werd in de rationale voor de hypothese over de goede antwoorden gesteld dat de verdelingen van goede antwoorden bij de Eerstejaarstoetsen en de Algemene Kennistoetsen niet alleen qua locatie, maar ook qua spreiding van elkaar zouden verschillen. Op de spreiding van de verdelingen zal in deze paragraaf nader worden ingegaan.
Voor het onderzoeken van de spreiding in de verdelingen van goede antwoorden op de items zullen de oorspronkelijke percentages worden genomen. Door de logit transformatie die in de vorige paragrafen werd besproken, wordt een eventueel verschil in spreiding van de verdelingen vrijwel geheel teniet gedaan.
Als maat voor de spreiding in goede antwoorden werd de variantie van de goede antwoorden op de items binnen elke toets genomen. De verhouding tussen de variantie van goede antwoorden op de Eerstejaarstoets en die op de Algemene Kennistoets werd uitgerekend. Voor de drie groepen personen wordt deze

verhouding vermeld in tabel 3.3. De verwachte verschillen in varianties werden getoetst met behulp van een t-toets voor onafhankelijke steekproeven. Daartoe werd allereerst de logaritme van de varianties genomen. Dit werd gedaan om het probleem, dat kleeft aan het direct toetsen van de varianties, nl. de grote gevoeligheid voor afwijkingen van de normaliteitsassumptie, te ondervangen (Box, Hunter en Hunter, 1978). Als grenswaarde voor een significant resultaat werd weer de Bonferroni kans .05/3 = .016 gehanteerd. De resultaten van deze berekeningen staan in tabel 3.3.

TABEL 3.3
Vergelijking van de varianties van de goedscores (%) op de items in Eerstejaarstoets (Ej) en Algemene kennistoets (Ak)

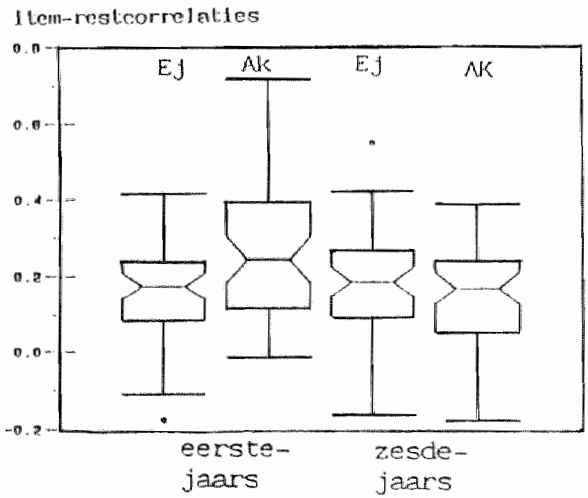
Varianties van de percentages goede antwoorden									
	Eerstejaars			Zesdejaars			Artsen		
Tijd stip	Ej	Ak	Ej/Ak	Ej	Ak	Ej/Ak	Ej	Ak	Ej/Ak
1	289	25	11.56	625	576	1.09	576	625	0.92
2	441	64	6.89	484	484	1.00	484	484	1.00
3	484	196	2.47	576	625	0.92	484	676	0.72
4	529	169	3.13	529	676	0.78	484	625	0.77
	logaritme uit de varianties								
1	2.46	1.40		2.80	2.76		2.76	2.80	
2	2.64	1.81		2.68	2.68		2.68	2.68	
3	2.68	2.29		2.76	2.80		2.65	2.82	
4	2.72	2.23		2.72	2.83		2.65	2.80	
t-waarde	3.22			-.66			-2.18		
p	.02			.54			.08		

* eenzijdig in het eerste jaar; tweezijdig in de controle groepen.

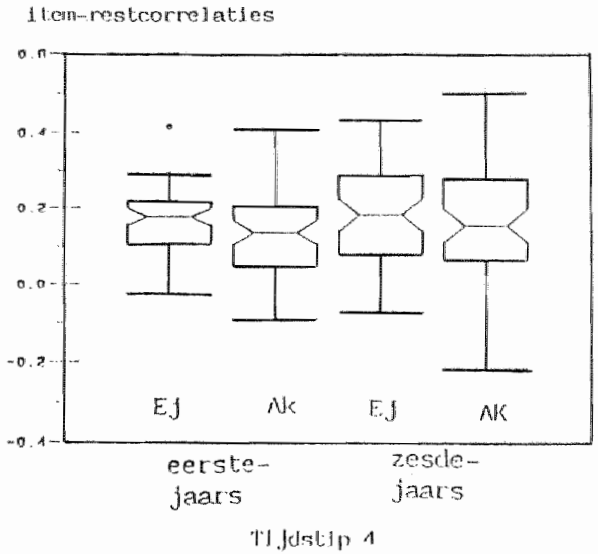
Uit tabel 3.3 blijkt dat de varianties van de goedscore verdelingen van de eerstejaars op de Eerstejaarstoets inderdaad groter waren dan die van de Algemene Kennistoets. Bij de zesdejaars en de afgestudeerden waren de twee spreidingsmaten nagenoeg gelijk. Het toetsen van de verschillen leverde naar verwachting dan ook alleen in het eerstejaar een significant resultaat op. Wanneer dit resultaat opgevat wordt als 4 successen in het eerstejaar en acht successen bij de zesdejaars en de afgestudeerden, dan kan dit worden opgevat als een ondersteuning van de begripsvaliteit.

3.2.4 Verdelingen van item-restcorrelaties (hypothese 4)

In deze paragraaf zal onderzocht worden of de verdelingen van de item-restcorrelaties in de Eerstejaarstoetsen en in de Algemene Kennistoetsen van elkaar verschillen. Item-restcorrelaties zijn zogenaamde "point biserial" correlaties die de associatie aangeven tussen het dichotome item en de min of meer continue somscore-variabele. De somscore is in dit geval de totaalscore op de items minus het betreffende item zelf. Alvorens de verschillen te toetsen werden de correlaties getransformeerd tot z-scores (Fisher's z-transformatie).
De verdelingen van deze indices op tijdstip 2 worden afgebeeld in figuur 3.3. De afgebeelde tijdstippen waren nagenoeg identiek aan de beide overige tijdstippen. Tijdstip 2 was illustratief voor tijdstip 1 en tijdstip 4 was illustratief voor tijdstip 3.



Tijdstip 2



Figuur 3.3 Verdelingen van item-restcorrelaties op de tijdstippen 2 en 4 in twee groepen respondenten (Ej = Eerstejaarstoets, Ak = Algemene Kennistoets).

TABEL 3.4
Resultaten t-toetsen (tweezijdig) voor onafhankelijke steekproeven met betrekking tot verschillen in itemrest-correlaties van de items.

	JAAR 1			
Tijdstip	1	2	3	4
M Ej	.28	.17	.18	.17
M Ak	.36	.26	.14	.14
t-waarde	-2.84	-2.96	2.25	1.40
p	.001	.002	.99	.92
Ej-items	51	53	61	61
Ak-items	48	53	61	61

TABEL 3.4 (vervolg)

	JAAR 6			
Tijdstip	1	2	3	4
M Ej	.15	.18	.15	.18
M Ak	.20	.14	.15	.15
t-waarde	-1.06	1.33	.16	1.01
p	.15	.19	.87	.32
Ej-items	51	53	61	61
Ak-items	51	53	61	61

In figuur 3.3 en tabel 3.4 worden de verdelingen van de itemrestcorrelaties uit de Eerstejaarstoetsen en uit de Algemene Kennistoetsen samengevat en op eventuele verschillen getoetst. Alle verdelingen in het zesde studiejaar bleken zich overeenkomstig de verwachting te gedragen. Dat wil zeggen in geen van de gevallen kon de nulhypothese dat de locatieparameters van de verdeling van de Eerstejaarstoets niet verschillen van die van de Algemene Kennistoets worden verworpen.
In het eerstejaar waren alleen de verschillen op de eerste twee tijdstippen in de verwachte richting (gemiddelde correlatie van Eerstejaarsitems kleiner dan van Algemene Kennisitems). Op de laatste twee tijdstippen kon de nulhypothese (geen verschil) niet meer worden verworpen.
In totaal was er sprake van 6 succesvolle en 2 mislukte voorspellingen.
In het eerstejaar werd hiermede onvoldoende ondersteuning gevonden voor de geformuleerde hypothese. In het zesdejaar werd de hypothese wel voldoende ondersteund.

3.2.5 Betrouwbaarheid van de toetsen (hypothese 5)

In deze paragraaf worden de resultaten van de betrouwbaarheidsanalyse van de Eerstejaarstoetsen en de Algemene Kennistoetsen onderzocht. De betrouwbaarheid werd op de klassieke manier bepaald, namelijk door de Cronbach's Alfa voor dichotome items uit te rekenen. Deze versie van de Cronbach's Alfa wordt de KR-20 (Lord en Novick, 1968) genoemd en is een maat voor de interne consistentie van de toetsen.
In hoofdstuk 1 werd gesteld dat in het eerstejaar de betrouwbaarheid van de Eerstejaarstoetsen lager (vanwege de grotere homogeniteit in de toetsscores) zou

zijn dan die van de Algemene Kennistoetsen. In de twee controlegroepen zou er geen verschil zijn tussen de betrouwbaarheden van de twee toetsen.
Vrijwel altijd rapporteert men betrouwbaarheden van toetsen of tests in termen van het al dan niet voldoen aan criteria die in de praktijk van het toets- en testgebruik ontstaan zijn. Deze criteria worden ondermeer beschreven door Visser, e.a. (1982). Zij hanteren de volgende - wel zeer strenge - criteria:

- een betrouwbaarheid van .90 of hoger wordt gekwalificeerd als goed.
- een betrouwbaarheid tussen .80 en .90 geldt als matig; onder de grens van .80 spreken deze auteurs van een lage betrouwbaarheid.
- in theoretisch onderzoek kunnen lagere normen voor de betrouwbaarheid worden geaccepteerd (blz. 14).

Ook in dit onderzoek zullen de criteria van Visser e.a. (1982), daar waar het gaat om de relatieve hoogte van de betrouwbaarheids-coëfficiënten te interpreteren, worden gebruikt. Om toch ook iets te kunnen zeggen over de grootte orde van de verschillen, zal hier volgens de volgende procedure worden gewerkt:

> Een verschil in betrouwbaarheden van .10 of meer wordt beschouwd als een relevant verschil in de betrouwbaarheid.
>
> In het eerstejaar wordt een verschil van .10 of meer, afhankelijk van de richting van het verschil, opgevat als een positief of een negatief resultaat. Een verschil ten gunste van de Eerstejaarstoetsen wordt beschouwd als een negatief resultaat. Een verschil ten gunste van de Algemene Kennistoetsen wordt beschouwd als een positief resultaat. Een verschil dat ligt binnen de hier afgesproken grens van .10 wordt beschouwd als een negatief resultaat.
> In het zesdejaar en in de groep afgestudeerden wordt een afwijking van .10 of meer, ongeacht de richting opgevat als een negatief resultaat. Een afwijking binnen de marge van .10 wordt gezien als een positief resultaat.

De berekende betrouwbaarheden zijn vermeld in tabel 3.5.
Wat in deze tabel allereerst opvalt is dat de vermelde betrouwbaarheden in alle gevallen redelijk hoog waren. Dit geldt zeker, wanneer men in aanmerking neemt dat het in dit geval om relatief korte toetsen ging. De maximale toetslengte was 61 items.

TABEL 3.5
Betrouwbaarheden van Eerstejaarstoetsen (Ej) en Algemene kennistoetsen (Ak)

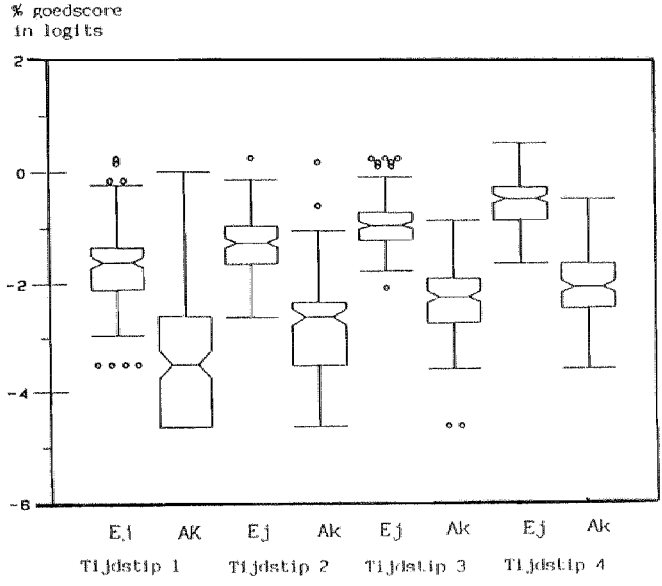
jaargroep	1		6		
tijdstip	Ej	Ak	Ej	Ak	aantal vragen
1	.80	.86	.67	.74	51
2	.64	.77	.71	.64	53
3	.74	.66	.68	.67	61
4	.71	.66	.73	.67	61

In het eerstejaar werd er drie keer een negatief resultaat geconstateerd. In drie gevallen waren de verschillen binnen de gedefiniëerde marges en één keer was het verschil, zoals verwacht, ten gunste van de Algemene Kennistoets.
In het zesdejaar daarentegen was geen van de verschillen groter dan de marge van .10. Met andere woorden in alle vier vergelijkingen werd de verwachting bevestigd. In totaal was er met betrekking tot de hypothese over de verschillen in betrouwbaarheid sprake van 5 successen en 3 mislukkingen. Op grond van deze resultaten kan geconcludeerd worden dat deze hypothese slechts gedeeltelijk door de onderzoeksresultaten werd ondersteund.

3.2.6 Verdelingen van toetsscores (hypothese 6)

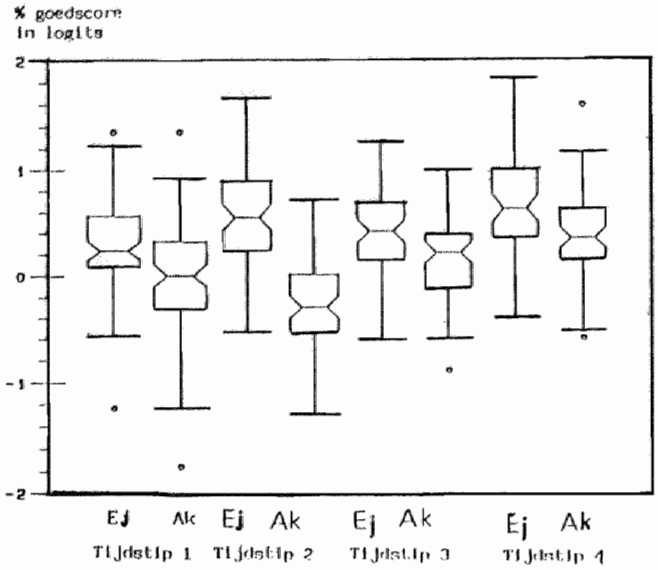
In de vorige paragrafen werd uitvoerig stilgestaan bij de verschillende kenmerken van de items uit de Eerstejaarstoets en de Algemene Kennistoets. Wil men iets zeggen over de soort toets dat men aan het bestuderen is, dan is het beschrijven van de empirische kenmerken van de toetsitems onvoldoende. Een zorgvuldige beschrijving van de toetsen vereist dat naast de kenmerken van de items ook de kenmerken van de toetsscores worden beschreven (Wijnen, 1971, Hofstee, 1971). In deze paragraaf zullen de verdelingen van de goedscores, omgezet in logits (= log(%goed/(100-%goed)) worden beschreven. De interesse gaat daarbij vooral uit naar de vraag of de verdelingen qua locatie al dan niet in hetzelfde gebied liggen en naar de vraag of de spreiding van de verdelingen van goedscores in de Eerstejaarstoetsen, in het eerstejaar, groter is dan die in de Algemene Kennistoets. Het toetsen van een eventueel verschil in locatieparameter van de verdelingen werd gedaan met behulp van de t-toets voor afhankelijke steekproeven. Omdat het ook in dit geval ging om het uitvoeren van meerdere toetsen in een keer werden de overschrijdingskansen op de reeds eerder besproken wijze

gecorrigeerd voor het aantal vergelijkingen, dat werd gemaakt. In het zesdejaar geven de resultaten voldoende ondersteuning aan de hypothese, in het eerstejaar echter niet.
In het eerstejaar zou volgens hypothese 6 aan de verwachtingen worden voldaan in het geval de verdelingen significant van elkaar verschillen. In het zesdejaar daarentegen zouden de verwachtingen worden bevestigd indien de locaties van de verdelingen niet significant van elkaar verschilden.
De grafische weergave van de verdelingen van de goedscores (in logits) wordt voor de eerstejaars weergegeven in figuur 3.4 en voor zesdejaars in figuur 3.5. De resultaten van de uitgevoerde statistische toetsing in tabel 3.6.



Figuur 3.4

Verdelingen van goedscores (in logits) in jaar 1 (Ej = Eerstejaarstoets, Ak = Algemene Kennistoets).



Figuur 3.5

Verdelingen van goedscores (in logits) in jaar 6 (Ej = Eerstejaarstoets Ak = Algemene Kennistoets).

TABEL 3.6
Resultaten van de t-toetsen* voor gepaarde waarnemingen met betrekking tot verschillen in toetsscores

tijdstip	1		2		3		4	
jaar	1	6	1	6	1	6	1	6
N	151	82	157	68	157	63	158	61
$\bar{x}(Ej)-\bar{x}(Ak)$	1.63	0.22	1.51	0.83	1.37	0.24	1.50	0.25
t-waarde	22.99	5.39	28.53	16.49	29.94	5.67	42.30	5.48
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* in jaar 1 eenzijdig; in jaar 6 tweezijdig.

Uit de hierboven staande figuren blijkt dat op alle tijdstippen de locatie van de twee verdelingen verschillend was. De resultaten van de toetsingen waren alle

significant. In alle acht verdelingen lag de locatie van de verdeling van de scores op de Eerstejaarstoets boven die van de verdeling van scores op de Algemene Kennistoets.
Hoewel de verschillen in locatie tussen de verdelingen van toetsscores zowel in het eerste- als in het zesdejaar significant waren, was er toch een duidelijk verschil in de resultaten van de twee jaargroepen.
De verdelingen van toetsscores op de Eerstejaarstoets en die op de Algemene Kennistoets vertoonden in het eerstejaar een overlap van ten hoogste 25%. Hetgeen betekent dat alleen de hoogste 25% van de toetsscores op de Algemene Kennistoets in het gebied van de Eerstejaarstoets lagen.
In het zesdejaar daarentegen lag op alle vier tijdstippen tenminste 75% van de toetsscores op de Algemene Kennistoets in het gebied dat door de Eerstejaarstoets werd bestreken. Deze grote overlap duidt er op dat voor het merendeel van de studenten in het zesdejaar de toetsscores enigszins vergelijkbaar waren.
In het licht van de eerder beschreven verwachte verschillen tussen de verdelingen van de toetsscores is dit een belangrijke constatering omdat het aangeeft dat de trend van de verschillen tussen de verdelingen in ieder geval in overeenstemming is met hetgeen op grond van de theoretische overwegingen mocht worden verwacht.
Samenvattend betekenen de resultaten dat de nulhypothese in beide jaargroepen op alle vier tijdstippen moest worden verworpen. Dit houdt in dat de verwachting in vier gevallen werd bevestigd en in vier gevallen werden ontkracht. De bevestigingen hadden betrekking op het eerste studiejaar (hypothese 6a) en de ontkrachtingen op het zesde studiejaar (hypothese 6b). De conclusie moet dan ook luiden dat de onderzoeksresultaten de hypothese slechts gedeeltelijk ondersteunen. In de discussie wordt op deze observatie en de betekenis ervan in het kader van deze studie nader ingegaan.

3.2.7 Varianties in toetsscores (hypothese 7)

Ook het bestuderen van de spreiding van de verdeling van de toetsscores kan inzicht geven in het soort toets waarmee men te maken heeft.
Tabel 3.7 vermeldt de varianties van de toetsscore-verdelingen van beide toetsen op de vier verschillende tijdstippen. Onder de aanname dat er geen verschil bestaat werden de verschillen in de logaritmen van de varianties getoetst met behulp van de t-toets.

TABEL 3.7
Verhouding van de varianties van logit scores op Eerstejaarstoetsen (Ej) en Algemene kennistoetsen (Ak)

	eerstejaars			zesdejaars		
tijd-stip	Ej	Ak	Ej/Ak	Ej	Ak	Ej/Ak
1	0.48	1.17	0.41	0.19	0.26	0.73
2	0.23	0.74	0.31	0.24	0.18	1.33
3	0.22	0.45	0.49	0.17	0.16	1.06
4	0.20	0.34	0.59	0.21	0.16	1.31
	Logaritme uit de varianties * 100					
1	1.68	2.07		1.28	1.41	
2	1.36	1.87		1.38	1.26	
3	1.34	1.65		1.23	1.20	
4	1.30	1.53		1.32	1.20	
t-waarde*		-6.02			0.59	
p		0.00			0.59	

* eenzijdig in het eerste jaar; tweezijdig in het zesdejaar.

In het eerste studiejaar bleek op alle vier tijdstippen de variantie van de Algemene Kennistoets groter dan die van de Eerstejaarstoets. Op drie van de vier tijdstippen was de variantie van Algemene Kennistoets ruim twee keer zo groot als die van de Eerstejaarstoets. De toetsing van dit verschil leidde in het eerstejaar dan ook tot een significant resultaat. In het zesde studiejaar waren de verschillen in varianties zeer gering. De geringe verschillen waren niet significant.
Het bovenstaande samenvattend, kan worden geconcludeerd dat de Algemene Kennistoetsen in het eerste studiejaar inderdaad een grotere spreiding hadden dan de Eerstejaarstoetsen. Verder kan de conclusie zijn dat dit effect in het zesde studiejaar is verdwenen. In dat studiejaar waren de spreidingen van beide toetsen in sterke mate vergelijkbaar.
Een en ander betekent dat in het eerstejaar de nulhypothese, naar verwachting, verworpen moest worden. In het zesdejaar kon de nulhypothese, eveneens conform de verwachting, niet verworpen worden. Beide resultaten kunnen worden opgevat als een ondersteuning van de validiteit van de voortgangstoets.

3.2.8 Itemvariantie en personenvariantie (hypothese 8)

Wijnen (1971) karakteriseert studietoetsen door middel van een variantieanalytische beschrijving. Tenminste drie effecten moeten in een dergelijke beschrijving zichtbaar worden gemaakt:

een effect als gevolg van verschillen tussen personen,
een effect als gevolg van verschillen tussen items en
een effect als gevolg van (nog) niet te interpreteren restvariantie.

Aan de hand van het aandeel van de itemvariantie en de persoonsvariantie met betrekking tot de totale variantie maakt Wijnen onderscheid tussen selectietoetsen en voortgangstoetsen (in de terminologie van Wijnen heeft dit begrip nog de betekenis van een gewone studietoets). Selectietoetsen zijn erop gericht om onderscheid te maken tussen personen, die op grond van hun toetsprestatie al dan niet in de groep thuis horen. Voortgangstoetsen, daarentegen zijn er vooral op gericht om onderscheid te maken tussen onderwerpen, die in het onderwijs aan bod zijn gekomen en onderwerpen die nog aan bod moeten komen. Selectietoetsen worden onder meer gekenmerkt door een grote bijdrage van de persoonsvariantie en een kleine bijdrage van de itemvariantie aan de totale variantie.
Bij voortgangstoetsen is dit precies andersom; daar is de relatieve bijdrage van de itemvariantie aan de totale variantie het grootst. Voortgangstoetsen zijn in de visie van Wijnen bij uitstek geschikt voor het opsporen van zaken die in het onderwijs onvoldoende, of niet op de juiste manier, aan de orde zijn geweest. De relatieve bijdrage van de itemvariantie aan de totale variantie wordt door Wijnen dan ook gekozen als operationalisatie van onderwijseffecten. Hoe groter deze bijdrage des te groter is, in deze redenering, het onderwijseffect. Selectietoetsen zijn vooral geschikt voor het opsporen van studenten die zich duidelijk van de rest van de groep onderscheiden. Kenmerken van deze toetsen worden door Wijnen dan ook beschouwd als een middel om persoons-effecten te operationaliseren. Derhalve wordt door Wijnen als operationalisatie van de persoonseffecten gekozen voor de grootte van de bijdrage van de persoonsvariantie aan de totale variantie.
Uit de studie van Wijnen bleek dat de door hem onderzochte studietoetsen zonder uitzondering te karakteriseren waren als voortgangstoetsen. In alle onderzochte toetsen bleek het aandeel van de itemvariantie vele malen groter dan de personenvariantie.
Wanneer de observatie van Wijnen een wetmatig karakter heeft, hetgeen alleszins aannemelijk lijkt, dan betekent dit voor de onderhavige studie dat aangetoond zou moeten worden dat in de Eerstejaarstoets afgenomen bij eerstejaars de itemvariantie groter is dan de personenvariantie. In de Algemene Kennistoets afgenomen in het eerstejaar zou daarentegen de personenvariantie groter moeten zijn dan de itemvariantie. In het zesdejaar zou in beide toetsen de itemvariantie het grootst moeten zijn.

Om de Eerstejaarstoetsen en de Algemene Kennistoetsen in de typologie van Wijnen (1971) te kunnen indelen werden per tijdstip de personenvariantie, de itemvariantie en de residuele variantie geschat. De verhouding tussen item- en personenvariantie in elk van de twee toetsen werd vervolgens vergeleken. In hoofdstuk 1 werd al opgemerkt dat de resultaten van de variantieanalyse alleen in beschrijvende zin gebruikt werden. Een formele toetsing van de grootte van de verschillende componenten werd niet gedaan. De resultaten van de vergelijking staan vermeld in tabel 3.8.

TABEL 3.8
Verhouding tussen itemvariantie en personenvariantie in Eerstejaarstoetsen en Algemene Kennistoetsen

tijd-stip	Eerstejaarstoets		Algemene Kennistoets		
	1	6	< jaar >	1	6
1	3.7	6.0		0.05	4.4
2	6.5	4.0		1.25	4.8
3	4.8	5.5		5.50	6.0
4	5.5	4.0		3.70	6.5

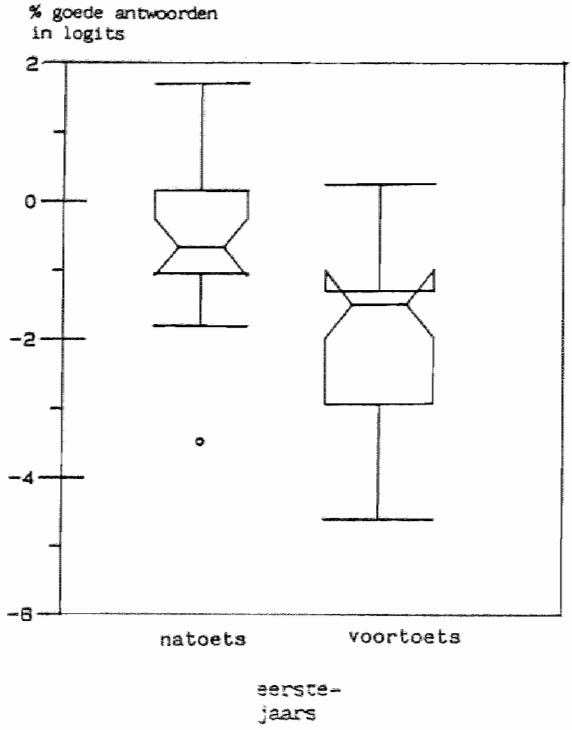
Uit de tabel blijkt, dat de verwachting in grote lijnen wordt bevestigd door de resultaten van de variantieanalyse. Bij de Eerstejaarstoetsen bleek dat in het eerstejaar de itemvariantie op alle vier tijdstippen groter was dan de personenvariantie.
Op twee van de vier tijdstippen was dit ook het geval bij de Algemene Kennistoetsen. Op het eerste tijdstip was de personenvariantie groter dan de itemvariantie. Op het tweede tijdstip waren beide variantiecomponenten nagenoeg even groot.
In het zesdejaar was in beide toetsen in alle vier gevallen de itemvariantie groter dan de personenvariantie.
Samenvattend kan worden geconcludeerd dat uit deze beschrijvende analyse bleek dat zes van de acht toetsen in het eerstejaar getypeerd moeten worden als "voortgangstoetsen", één als een "selectietoets" en één als niet de classificeren. In het zesdejaar bleken alle acht toetsen in de terminologie van Wijnen (1971) getypeerd te kunnen worden als "voortgangstoetsen".
Deze resultaten geven slechts een gedeeltelijke ondersteuning aan de hypothese voor wat betreft de typering van de toetsen in het eerstejaar. De resultaten waren

daarentegen volledig in overeenstemming met de verwachting die voor het zesdejaar werd geformuleerd.

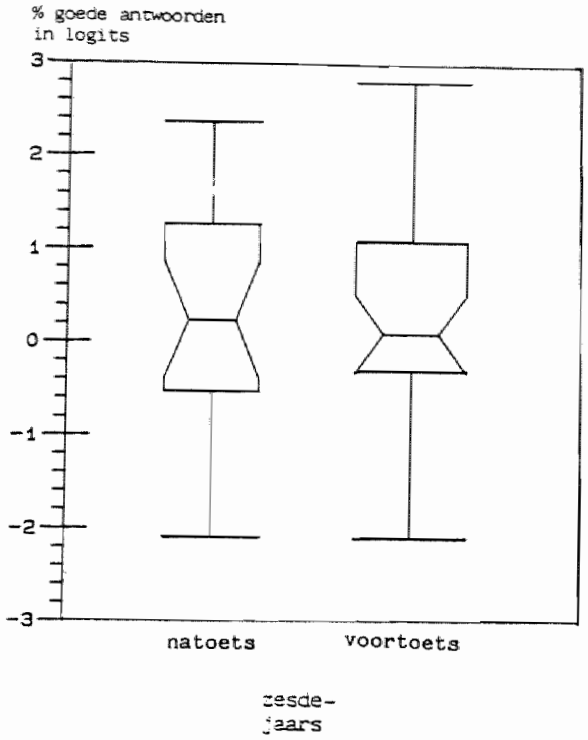
3.2.9 Onderwijseffecten in Eerstejaarstoetsen (hypothese 9)

In het voorafgaande werd aangetoond dat de Eerstejaarstoetsen zich, vooral in het eerstejaar van de medische opleiding, anders gedragen dan de in deze studie als controle gebruikte Algemene Kennistoetsen.
Centraal staat uiteraard de vraag of dit andere toetsgedrag uitgelegd mag worden als een gevolg van het onderwijsprogramma in het eerste studiejaar. Een alternatieve verklaring voor een aantal van de gerapporteerde bevindingen zou immmers gevonden kunnen worden in de constatering dat de Eerstejaarstoetsen vragen bevatten die gewoon gemakkelijker te beantwoorden zijn, ongeacht het eerstejaars programma. Immers ook zesdejaars scoorden hoger op de Eerstejaarstoetsen, ook al waren de verschillen geringer dan in het eerstejaar.
Het lijkt alleszins gerechtvaardigd nog eens gedetailleerd te kijken naar de Eerstejaarstoetsen met name met het oog op de vraag of de verschillen die geconstateerd werden, geïnterpreteerd mogen worden als onderwijseffecten.
Onderwijseffecten geoperationaliseerd als een relatief grote bijdrage van de itemvariantie aan de totale variantie, waren in het eerstejaar vooral in de Eerstejaarstoets aantoonbaar. In het zesdejaar waren deze effecten in beide toetsen in even sterke mate aanwezig.
Wanneer nu de in het eerstejaar gesignaleerde onderwijseffecten inderdaad toegeschreven mogen worden aan het programma in het eerste studiejaar, dan moet een opsplitsing van een Eerstejaarstoets in twee subtoetsen, waarvan er één bestaat uit vragen die betrekking hebben op zaken die aan de orde zijn geweest vóór het tijdstip waarop de toets werd afgenomen, en één met vragen die betrekking hebben op zaken die pas aan de orde kwamen nàdat de toets werd afgenomen, alleen in het eerstejaar een significant verschil te zien geven. Indien de redenering in overeenstemming is met de gegevens, mag er in het zesdejaar geen verschil zijn.
Om praktische redenen kwam niet elk van de vier toetsen in aanmerking voor een dergelijke opsplitsing. Op het moment dat de eerste toets plaatsvond, was er nog nauwelijks sprake geweest van onderwijs. Ten tijde van de laatste toets had bijna al het onderwijs in het eerste studiejaar plaatsgevonden.
Op grond van praktische overwegingen werd de derde toets als het meest geschikt voor een opsplitsing in subtoetsen beoordeeld.
Van de 61 toetsvragen uit de derde toets bleken er 47 naar herkomst van een bepaald onderwijsblok te kunnen worden ingedeeld. Van deze 47 toetsvragen hadden er 20 betrekking op onderwerpen die vóór de afname van de derde toets (maart 1984) aan de orde waren geweest. De overige 27 vragen hadden betrekking op

onderwerpen die ná maart aan de orde kwamen.
De eerste subtoets kan worden beschouwd als een natoets. De tweede subtoets kan worden beschouwd als een voortoets.
Uiteraard kunnen deze toetsen alleen in het eerstejaar als voor- en natoets worden beschouwd. In het zesdejaar zijn beide toetsen te beschouwen als natoetsen.
De frequentieverdelingen van de itemscores in logits van de voortoets en de natoets in het eerste en het zesdejaar worden in figuur 3.6 grafisch weergegeven.



Figuur 3.6 Verdelingen van de p-waarden van een natoets en een voortoets in jaar 1 en jaar 6.



TABEL 3.9
Resultaten t-toetsen voor onafhankelijke steekproeven op voortoets en natoets

	jaar 1	jaar 6
gem. natoets	-0.47	0.25
gem. voortoets	-1.99	0.29
t-waarde	4.22	-0.10
p	0.00	0.92
aantal items natoets	27	27
aantal items voortoets	20	20

Uit de figuur blijkt dat in het eerstejaar de Eerstejaarstoets van maart gesplitst kan worden in twee subtoetsen die qua locatie significant van elkaar verschillen. De natoets had gemiddeld een hogere p-waarde dan de voortoets.
Het tweede deel van de bovenstaande figuur geeft de verdelingen in het zesdejaar weer. Uit dat deel blijkt dat alle verdelingen de vorm hadden die verwacht mag worden van toetsen die op onderwijs betrekking hebben, n.l. toetsen met verdelingen van itemscores die zich over een groot gebied uitstrekken. In het zesdejaar was er, zoals verwacht, geen verschil tussen voor- en natoets. De twee verdelingen van de itemscores overlapten elkaar volledig.
Het bovenstaande samenvattend kan worden geconcludeerd dat het mogelijk is om op grond van zaken die al dan niet in het onderwijs in het eerste studiejaar aan de orde zijn geweest subtoetsen samen te stellen die zich gedragen zoals verwacht mag worden. Een natoets wordt beter gemaakt dan een voortoets.
In het zesdejaar kan een voor- en een natoets, geconstrueerd aan de hand van het onderwijs in het eerstejaar niet van elkaar worden onderscheiden. Ook dat is in overeenstemming met hetgeen werd verwacht. Voor zesdejaars behoren beide toetsen natoetsen te zijn. Met deze resultaten worden hypothesen 9a en 9b ondersteund.

3.3 Samenvatting en conclusies

In de voorafgaande paragrafen werden de hypothesen met betrekking tot de interne structuur van de Eerstejaarstoetsen en Algemene Kennistoetsen op houdbaarheid onderzocht. Een samenvatting van het resultaat van de uitgevoerde toetsingen is opgenomen in tabel 3.10.
Uit die tabel kunnen de volgende conclusies worden getrokken:

- hypothese 1.
  Deze hypothese werd in de afzonderlijke groepen voldoende ondersteund, zodat geconcludeerd kan worden dat Eerstejaarsitems alleen in het eerstejaar meer beantwoord werden dan de Algemene Kennisitems.

- hypothese 2.
  Ook deze hypothese werd in de afzonderlijke groepen in voldoende mate ondersteund. De conclusie kan derhalve zijn dat Eerstejaarsitems alleen in het eerstejaar beter beantwoord werden.

TABEL 3.10
Samenvatting van het toetsen van de hypothesen met betrekking tot verschillen in overeenkomsten in interne structuur.

tijdstip	1			2			3			4		
jaargr	1	6	7	1	6	7	1	6	7	1	6	7
1	+	+	+	+	=	+	+	+	+	+	+	+
2	+	+	+	+	=	=	+	+	+	+	+	+
3	+	+	+	+	+	+	+	+	+	+	+	+
4	+	+	0	+	+	0	=	+	0	=	+	0
5	=?	+?	0	+?	+?	0	=?	+?	0	=?	+	0
6	+	=	0	+	=	0	+	=	0	+	=	0
7	+	+	0	+	+	0	+	+	0	+	+	0
8	+?	+?	0	+?	+?	0	+?	+?	0	+?	+?	0
9	+	+	0	+	+	0	+	+	0	+	+	0

\+ geslaagde voorspelling
= mislukte voorspelling
+? beschrijving tendeert in de voorspelde richting
=? beschrijving tendeert niet in de voorspelde richting
0 niet getoetst
1 eerstejaars
6 zesdejaars
7 afgestudeerden

- hypothese 3.
  Werd eveneens voldoende ondersteund. De conclusie kon dan ook luiden dat er alleen in het eerstejaar grotere verschillen in de variantie van de p-waarden waren.

- hypothese 4.
  Werd in het eerstejaar alleen op de eerste twee tijdstippen ondersteund. Derhalve was er onvoldoende steun voor de voorspelling dat de gemiddelde item-restcorrelatie in de Eerstejaarstoetsen lager zou zijn.
  In het zesdejaar werd de hypothese voldoende ondersteund: hier bleek geen verschil te zijn in de gemiddelde item-restcorrelatie van beide toetsen.

- hypothese 5.
  Voor deze hypothese werd in de beschrijvende analyse niet voldoende steun gevonden in het eerstejaar doch wel in het zesdejaar. De conclusie moet derhalve luiden dat er geen substantiële verschillen waren in betrouwbaarheid van beide toetsen.

- hypothese 6.
  Deze hypothese werd in het eerstejaar volledig ondersteund, doch ondervond geen enkele steun in het zesdejaar. De conclusie moet luiden dat er in beide jaargroepen een significant verschil in gemiddelde toetsscores van Eerstejaarstoetsen en Algemene Kennistoetsen was.

- hypothese 7.
  Werd volledig door de resultaten ondersteund. Derhalve luidt de conclusie dat er alleen in het eerstejaar verschillen waren in de varianties van de toetsscores. In dit studiejaar waren de varianties van de toetsscores in de Algemene Kennistoetsen groter dan in de Eerstejaarstoetsen.

- hypothese 8.
  De beschrijving van de data was in het eerstejaar gedeeltelijk in overeenstemming met de geformuleerde voorspelling: Alle Eerstejaarstoetsen konden worden getypeerd als "voortgangstoetsen", echter slechts twee van de Algemene Kennistoetsen bleken getypeerd te kunnen worden als "selectietoetsen".
  In het zesdejaar was de databeschrijving in overeenstemming met de voorspelling: alle 8 toetsen bleken getypeerd te kunnen worden als "voortgangstoetsen".

- hypothese 9.
  Werd volledig door de analyses ondersteund. De conclusie luidt derhalve dat er alleen in het eerstejaar verschil was de in gemiddelde p-waarde tussen een natoets en een voortoets.

Dit samenvattend kan de conclusie luiden dat, met betrekking tot de verwachte verschillen en overeenkomsten tussen Eerstejaarstoetsen en Algemene Kennistoetsen, er in de beschreven interne analyses voldoende steun voor de begripsvaliditeit werd gevonden (6 van de 9 hypothesen werden volledig en de overige 3 werden gedeeltelijk door de resultaten ondersteund).

4. VERGELIJKING EERSTEJAARSTOETSEN MET ALGEMENE KENNISTOETSEN (EXTERNE ANALYSES)

Inhoud

4.1	Inleiding
4.2	Het toetsen van hypothesen over correlaties (hypothesen 10a en 10b)
4.3	Statistische modellen
4.3.1	Simplex modellen
4.4	Resultaten
4.4.1	Correlaties tussen Eerstejaarstoetsen en Algemene Kennistoetsen
4.4.2	Twee toetsen, één simplex voor de latente variabelen
4.4.3	Conclusie
4.4.4	Parameterschattingen in de simplex modellen
4.4.5	Kennis- en toetsmethode variantie
4.4.6	Samenvatting en conclusies (hypothesen 10a en 10b)
4.5	Voorspellen van studieprestaties (hypothese 11)
4.5.1	Inleiding
4.5.2	Methode
4.5.3	Resultaten
4.5.4	Samenvatting en conclusies (hypothese 11)
4.6	Medische kennis in het eerstejaar (hypothese 12)
4.6.1	Inleiding
4.6.2	Resultaten
4.6.2.1	Correlaties tussen bloktoetsen, Eerstejaarstoetsen en Algemene Kennistoetsen
4.6.3	Samenvatting en conclusies (hypothese 12)
4.7	Samenvatting en conclusies (externe analyses)

4.1 Inleiding

In dit hoofdstuk worden de correlaties tussen de toetsscores (= logitscore van het % goed) binnen ieder tijdstip en tussen de vier tijdstippen beschreven en geanalyseerd.
In hoofdstuk 1 werd met betrekking tot het correlatiegedrag tussen de Eerstejaarstoets en de Algemene Kennistoets verondersteld dat er zich een specifieke structuur zou voordoen (hypothesen 10a en 10b). Deze specifieke structuur wordt verondersteld op grond van een aantal overwegingen.
Eerstejaarstoetsen en Algemene Kennistoetsen verschillen in theorie slechts van elkaar in moeilijkheid. Beide toetsen zijn bedoeld om hetzelfde begrip te meten nl., de medische kennis waarover iemand beschikt. Meetinstrumenten die hetzelfde meten zouden onderling hoog moeten correleren.
Bij twee onderling hoog correlerende variabelen, waarvan men in theorie aanneemt dat ze hetzelfde meten, worden ook wel indicatoren van eenzelfde latente variabele genoemd (zie o.a. Heise, 1975, Jöreskog en Sörbom, 1979, Kenny, 1979). In het onderhavige onderzoek is er op elk van de vier tijdstippen sprake van twee metingen van hetzelfde begrip. Eén met behulp van de Eerstejaars toets en één met behulp van de Algemene Kennistoets. Verder is het redelijk om aan te nemen dat het kennisbestand van een student, over een bepaalde tijdsperiode gezien, aan veranderingen onderhevig is. Hoe langer de periode, des te groter de veranderingen. Voor de structuur in de correlaties betekent dit hoge correlaties tussen opeenvolgende tijdstippen en lagere correlaties tussen tijdstippen die verder van elkaar verwijderd liggen. Deze overwegingen hebben geleid tot het formuleren van een hypothese over de structuur in de correlatiematrix. Centraal hierin staat de veronderstelling dat er slechts één latente variabele verantwoordelijk is voor de geobserveerde samenhang tussen de variabelen. Eventuele verschillen in structuur tussen eerstejaars- en zesdejaars- correlatiematrices moeten tot uiting komen in de niet door het gepostuleerde model verklaarde variantie. Het verdisconteren van deze niet verklaarde variantie leidt in dat geval voor eerstejaars tot een model met minder restricties. In het zesdejaar zou het meer restrictieve model de data goed moeten beschrijven. Deze theoretische overwegingen kunnen worden vertaald in een formeel statistisch model voor de structuur in de geobserveerde correlatiematrices. Dit statistisch model wordt in paragraaf 4.3 verder uitgewerkt.
De resultaten van het onderzoek van de modelfit komen aan de orde in paragraaf 4.4.2. De waarden van de geschatte parameters van de best passende modellen worden beschreven in paragraaf 4.4.4 en de implicaties daarvan voor het schatten van kennis-, toetsmethode en unieke varianties komen in paragraaf 4.4.5 aan de orde.
Verderop in het hoofdstuk wordt in paragraaf 4.5 de voorspellende waarde van de Eerstejaarstoets en Algemene Kennistoets bestudeerd (hypothese 11). De resultaten

van de data- analyse met betrekking tot de houdbaarheid van hypothese 11 worden beschreven in paragraaf 4.5.3. Tenslotte worden in paragraaf 4.6 de correlaties tussen Eerstejaarstoetsen, Algemene Kennistoetsen en andere toetsen voor medische kennis in het eerstejaar bestudeerd (hypothese 12).

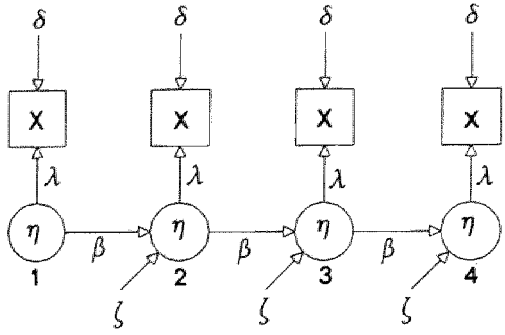
4.2 Het toetsen van hypothesen over correlaties (hypothesen 10a en 10b)

Hypothesen omtrent correlaties tussen al dan niet latente variabelen, kunnen op houdbaarheid worden onderzocht door middel van het specificeren van structurele modellen.
Structurele modellen bevatten twee basiselementen: variabelen en parameters. De variabelen vormen de ruwe materie. De parameters geven de theoretisch verwachte verbanden weer. Parameters worden geschat aan de hand van de geobserveerde variabelen. Met behulp van de geschatte parameters kan men het structurele model gebruiken om de data (bijvoorbeeld een correlatiematrix) te reconstrueren. Bij geringe afwijking tussen model en observatie concludeert men dat het model de data goed beschrijft.
Bekende boeken over deze materie zijn Heise (1975), Jöreskog en Sörbom (1979), Kenny (1979) en Saris en Stronkhorst (1984). Het toetsen van de in hoofdstuk 2 geformuleerde en hierboven nog eens toegelichte hypothese met betrekking tot de correlaties tussen de Eerstejaarstoetsen en de Algemene Kennistoetsen, zal geschieden met behulp van het statistisch model dat in paragraaf 4.3 wordt beschreven. De parameters van het model werden geschat met behulp van het LISREL V programma (Jöreskog en Sörbom, 1981).

4.3 Statistische modellen

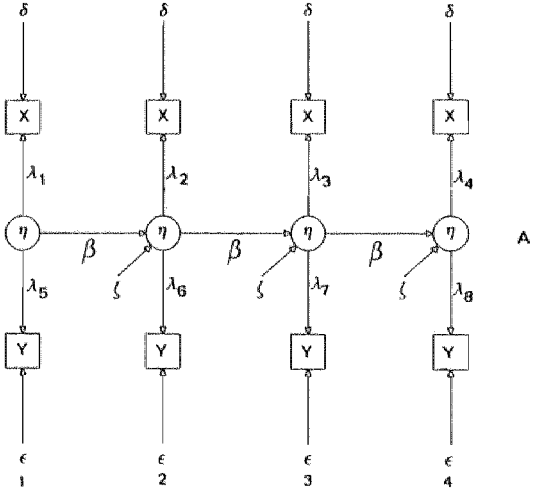
4.3.1 Simplex modellen

In de geformuleerde hypothese wordt verondersteld dat de correlaties tussen de Eerstejaarstoetsen en tussen de Algemene Kennistoetsen een simplex structuur vertonen. Een dergelijk model is adequaat in situaties waarbij eenzelfde variabele op verschillende tijdstippen wordt gemeten. Guttman (1954) noemde deze modellen simplex modellen vanwege het "typische" patroon van intercorrelaties in een matrix waarvoor deze structuur geldt. Het "typische" van een matrix met een simplex structuur is dat de correlaties gezien vanaf de diagonaal naar beneden afnemen. Correlatiematrices met een dergelijk patroon kunnen worden gerepresenteerd door een simplexmodel, waarvan een grafische weergave is gegeven in figuur 4.1.



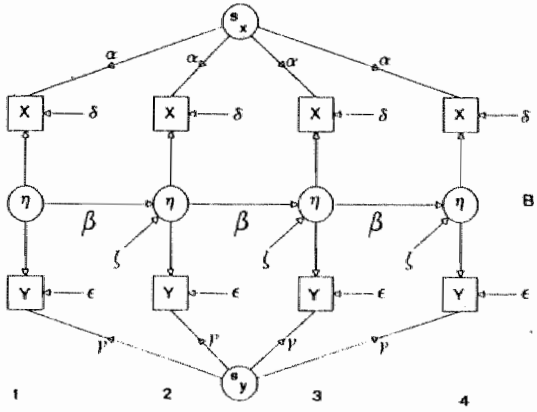
Figuur 4.1 Een quasi-simplex model voor vier tijdstippen.

Figuur 4.1 is een afbeelding van een simplex model voor een niet perfect gemeten variabele (= een variabele met meetfouten). Een dergelijk model wordt vanwege de aanwezigheid van een foutencomponent ook wel een "quasi simplex model" genoemd (Jöreskog en Sörbom, 1981).
Bovenstaand simplex model voor 4 tijdstippen werd gebruikt als uitgangspunt voor het analyseren van de correlatiematrix van Eerstejaarstoetsen en Algemene Kennistoetsen. In plaats van één geobserveerde variabele per tijdstip is er in het onderhavige geval sprake van twee geobserveerde variabelen op ieder tijdstip. Het model uit figuur 4.1 kan eenvoudig gegeneraliseerd worden naar modellen die grafisch zijn weergegeven in figuur 4.2.



Figuur 4.2

Modellen voor twee variabelen op vier tijdstippen zonder (A) en met (B) gecorreleerde meetfouten.



Zoals in figuur 4.2 te zien is, is de structuur van een simplex model ook in deze modellen herkenbaar. Het verschil met het model van figuur 4.1 is dat er voor ieder tijdstip één indicator variabele voor de latente variabele bij is gekomen. De essentie van model A en model B is dat Eerstejaarstoetsen en Algemene Kennistoetsen beide indicatoren zijn voor één en dezelfde onderliggende variabele. Het verschil tussen model A en model B is dat in het eerste model de meetfouten ongecorreleerd zijn, terwijl ze in het tweede model gecorreleerd zijn. Deze correlaties worden verklaard door twee toegevoegde latente variabelen die in model B geïnterpreteerd kunnen worden als toets-specifieke-factoren. Wanneer de gestandaardiseerde scores op de vier Eerstejaarstoetsen worden omschreven als x-variabelen, met $\mathbf{x}'$ = (x1,...,x4) en de Algemene Kennistoetsen als y-variabelen, met $\mathbf{y}'$ = (y1,...,y4), en wanneer de latente variabele η op dezelfde wijze wordt geschaald als de Eerstejaarstoetsen, dan is het model

$$\mathbf{x} = \eta + \boldsymbol{\delta} \qquad (4.1)$$

$$\mathbf{y} = D_{\lambda}\eta + \boldsymbol{\epsilon} \qquad (4.2)$$

waar η de latente variabele is die aan de twee geobserveerde variabelen ten grondslag ligt, $\boldsymbol{\delta}$ en $\boldsymbol{\epsilon}$ de meetfouten, of unieke factoren behorende bij de x-variabelen (= Eerstejaarstoetsen) respectievelijk de y-variabelen (= de Algemene Kennistoetsen) en D_{λ} = diag($\lambda_5, \lambda_6, \lambda_7, \lambda_8$), met de λ_i's als de factorladingen van de y-variabelen op de latente variabele. De correlatiematrix $(\mathbf{x}',\mathbf{y}')'$ is dan te

schrijven als

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \qquad (4.3)$$

met

$$\Sigma_{xx} = \Phi + \Theta_{\delta},\ \Sigma_{yy} = D_{\lambda}\Phi D_{\lambda} + \Theta_{\epsilon},$$

$$\Sigma_{yx} = D_{\lambda}\Phi D_{\lambda}$$

waar Φ, Θ_{δ} en Θ_{ϵ} de covariantie-matrices zijn van η, δ en ϵ. De matrix Φ van η wordt de restrictie opgelegd dat ze gegenereerd wordt door een simplex of eerste-orde autoregressief model, inhoudende dat,

$$\eta_i = \beta_i \eta_{i-1} + \zeta_i \quad i = 2,3,4,$$

waar β_i een regressieparameter is van η_{i-1} op η_i, een parameter die de stabiliteit van de latente variabele tussen de twee tijdstippen representeert. Tenslotte representeert ζ_i de fout in de structurele vergelijking zodat,

$$\Phi = \begin{bmatrix} \phi_1 & & & \\ \beta_2\phi_1 & \phi_2 & & \\ \beta_2\beta_3\phi_1 & \beta_3\phi_2 & \phi_3 & \\ \beta_2\beta_3\beta_4\phi_1 & \beta_3\beta_4\phi_2 & \beta_4\phi_3 & \phi_4 \end{bmatrix} \qquad (4.4)$$

waar, $\phi_i = \text{var}(\eta_i)$, i = 1,2,3,4.
In het programma LISREL V worden niet de varianties var(ϕ_i), geschat, doch var(ζ_i) = ψ_i. Schattingen van ϕ_i worden echter verkregen als bijproduct. De relaties tussen ϕ_i en ψ_i, i = 1,2,3,4 zijn

$$\phi_1 = \psi_1$$

$$\phi_i = \psi_i + \beta^2_i \phi_{i-1}, \qquad i = 2,3,4$$

Zoals al in figuur 4.2 werd weergegeven veronderstelt model B in tegenstelling tot model A dat de meetfouten van de Eerstejaarstoetsen onderling gecorreleerd zijn en dat deze verklaard kunnen worden door een toetsspecifieke factor S_x (in tegenstelling tot de situatiespecifieke factoren η_i (Jöreskog, 1970)). Van de meetfouten van de Algemene Kennistoetsen wordt eveneens aangenomen dat ze gecorreleerd zijn en verklaard kunnen worden door de toetsspecifieke factor S_y. S_x en S_y zijn in het model niet gecorreleerd met η, δ en ϵ. Zodat de vergelijkingen voor model B zijn

$$\mathbf{x} = \eta + \alpha S_x + \delta,$$
$$\mathbf{y} = D_\lambda + \gamma S_y + \epsilon \qquad (4.5)$$

waar α en factorladingen zijn die de geobserveerde variabelen **x** en **y** verbinden met de toetsspecifieke factoren. Model A is hierarchisch genest in model B, zodat de hypothese $\alpha = 0$ en $\gamma = 0$ getoetst kan worden met het aantal vrijheidsgraden dat het verschil is tussen de vrijheidsgraden van de twee modellen. De modellen A en B werden gebruikt om de totale correlatiematrices voor eerstejaars en zesdejaars te analyseren.

4.4 Resultaten

4.4.1 Correlaties tussen Eerstejaarstoetsen en Algemene Kennistoetsen

De basis van de analyses in deze paragraaf wordt gevormd door de twee correlatiematrices, één voor eerstejaars en één voor zesdejaars, die in tabel 4.1 worden weergegeven. De submatrices in de tabel zijn de matrices met de correlaties tussen respectievelijk Eerstejaarstoetsen en Algemene Kennistoetsen.

TABEL 4.1
Correlaties tussen Eerstejaarstoetsen (Ej) en Algemene kennistoetsen (Ak) op vier tijdstippen

jaar 1

	Ej1	Ej2	Ej3	Ej4	Ak1	Ak2	Ak3	Ak4
Ej1	1.00							
Ej2	0.37	1.00						
Ej3	0.24	0.48	1.00					
Ej4	0.08	0.22	0.42	1.00				
							N = 151	
Ak1	0.59	0.37	0.20	0.06	1.00			
Ak2	0.34	0.63	0.33	0.14	0.42	1.00		
Ak3	0.25	0.44	0.54	0.37	0.29	0.33	1.00	
Ak4	0.19	0.29	0.42	0.63	0.23	0.22	0.56	1.00

jaar 6

	Ej1	Ej2	Ej3	Ej4	Ak1	Ak2	Ak3	Ak4
Ej1	1.00							
Ej2	0.57	1.00						
Ej3	0.57	0.55	1.00					
Ej4	0.58	0.52	0.47	1.00				
							N = 56	
Ak1	0.68	0.53	0.52	0.58	1.00			
Ak2	0.50	0.53	0.44	0.58	0.70	1.00		
Ak3	0.45	0.51	0.65	0.51	0.65	0.61	1.00	
Ak4	0.54	0.59	0.53	0.66	0.52	0.53	0.58	1.00

x = correlatie tussen Ej en Ak op hetzelfde tijdstip

Bovenstaande matrices werden geanalyseerd met behulp van de modellen die hierboven werden besproken. Gedetailleerde beschrijvingen met betrekking tot de gebruikte modellen zijn te vinden in tabel 4.2. Het toetsen van de modellen uit tabel 4.2 levert informatie op omtrent de hypothesen met betrekking tot de structuur in de correlatie- matrices van eerstejaars en zesdejaars (hypothesen 10a en 10b). De informatie die uit deze tabel afgeleid kan worden, wordt in paragraaf 4.4.2 aan de orde gesteld.

TABEL 4.2
Samenvatting van de LISREL-modellen voor de correlaties tussen de toetsen

	Modelbeschrijving	maten voor de fit van het model					
		jaar 1			jaar 6		
		χ^2	df	p	χ^2	df	p
I	De totale correlatiematrix van Eerstejaarstoetsen en Algemene kennistoetsen gezamenlijk, heeft een simplexstructuur met voor ieder tijdstip 1 latente variabele (= situatie specifieke variabele). Alle meetfouten zijn gelijk aan elkaar; de 3 stabiliteits-coëfficienten zijn eveneens gelijk	33.41	26	0.15	29.72	26	0.28
II	De totale correlatiematrix van Eerstejaarstoetsen en Algemene kennistoetsen gezamenlijk, heeft een simplexstructuur, met een latente variabele voor ieder tijdstip, plus twee toetsspecifieke variabelen	17.16	18	0.51	17.49	18	0.49

4.4.2 Twee toetsen, één simplex voor de latente variabelen

De hypothese over het correlatiegedrag van de toetsen die het onderwerp zijn van deze studie, heeft betrekking op de gehele correlatiematrix van de acht toetsen gezamenlijk. Verondersteld werd dat de Eerstejaarstoetsen en de Algemene Kennistoetsen op hetzelfde tijdstip afgenomen, beschouwd mogen worden als indicatoren van één latente variabele, die een simplex structuur zou hebben. Deze verwachting werd hier gespecificeerd als model I. Model I is een restrictief model, omdat gelijke meetfouten en gelijke stabiliteitsparameters worden verondersteld. Verwacht werd verder dat model I een adequate beschrijving zou geven voor de zesdejaars data en wellicht ook voor de eerstejaars data (hypothese 10a).

Voor de eerstejaars data zouden indien model I de data niet goed beschrijft er nog twee latente variabelen gepostuleerd moeten worden; één voor de Eerstejaarstoetsen en één voor de Algemene Kennistoetsen (hypothese 10b). Dit model werd gespecificeerd als model II en heeft de structuur van model B in figuur 4.2. Model I heeft de structuur van model A in figuur 4.2. De opgelegde restricties houden in dat alle δ_j's gelijk zijn aan de ϵ_i's en dat $\beta_2 = \beta_3 = \beta_4$.
De resultaten van de toetsingen met behulp van beide modellen worden vermeld in tabel 4.2. Uit deze tabel blijkt dat model I reeds een adequate beschrijving geeft van de structuur in de correlatiematrix in het eerstejaar. De overschrijdingskans was .15.
Voor de correlatiematrix in het zesdejaar was de fit van dit meest restrictieve model eveneens adequaat. De overschrijdingskans was .28.
Versoepeling van model I tot model II, gaf voor de correlatiematrix in het eerstejaar een betere fit. De overschrijdingskans was .51. Ten opzichte van model I was de daling van de toetsingsgrootheid significant ($\chi^2 = 16.25$; df = 8; $p < .05$). Voor de zesdejaars data bleek de versoepeling naar model II geen echte verbetering. De χ^2 toetsingsgrootheid daalde wel, doch de daling was niet significant. ($\chi^2 = 12.23$; df = 8; $p > .10$).

4.4.3 Conclusie

Met betrekking tot de matrix uit tabel 4.1 kan de conclusie zijn dat Eerstejaarstoetsen en Algemene Kennistoetsen, die op hetzelfde tijdstip zijn afgenomen, inderdaad zo hoog correleren dat ze opgevat kunnen worden als indicatoren van één latente variabele.
Een latente variabele als die in model I en II wordt door Jöreskog (1970) omschreven als een situatie-specifieke-factor. Dit ter onderscheiding van een factor, die aan het gebruikte meetinstrument is gekoppeld en door Jöreskog wordt omschreven als test-specifieke-factor. Gezien de toepassing van de modellen I en II in dit onderzoek, kan de latente variabele worden geïnterpreteerd als tijdstip-afhankelijke-medische kennis (TMK). De TMK gedraagt zich, zoals verwacht, als een eerste-orde-autoregressief proces, of een Markov proces. Dit houdt in dat men, om het niveau van de kennis te voorspellen op tijdstip t, voldoende heeft aan de informatie op het direct voorafgaande moment, t-1. Alle eerdere tijdstippen verbeteren de voorspelling niet. Het postuleren van één tijdstip-afhankelijke latente variabele, zoals gespecificeerd in model I, bleek de correlaties tussen alle toetsen in het eerstejaar en in het zesdejaar adequaat te beschrijven. Voor het eerstejaar gaf de toevoeging van twee extra variabelen, zoals in model II, echter wel een duidelijke verbetering. Deze toegevoegde latente variabelen kunnen in deze studie geïnterpreteerd worden als toets-specifieke-factoren. Met andere woorden, een factor die niet bijdraagt aan de correlaties

tussen Eerstejaarstoets en Algemene Kennistoets op hetzelfde tijdstip, maar aan de correlaties over toetsen van hetzelfde type tussen de verschillende tijdstippen. Eén van deze variabelen is gekoppeld aan de Eerstejaarstoetsen. De andere aan de Algemene Kennistoetsen. In figuur 4.2 wordt de eerste variabele aangeduid met S_x en de tweede variabele met S_y. Uit figuur 4.2 blijkt eveneens dat deze twee latente variabelen ongecorreleerd zijn. Vandaar de omschrijving toetsspecifieke factoren.
Met betrekking tot de geformuleerde hypothesen 10a en 10b betekenen deze resultaten dat in beide jaargroepen hypothese 10a ondersteund wordt. Strikt genomen is een model zoals verwoord in hypothese 10b ook voor de correlatiematrix in het eerstejaar niet nodig. Het postuleren van één latente variabele per tijdstip bleek voldoende te zijn.

4.4.4 Parameterschattingen in de simplex modellen

In de voorgaande paragrafen werd alleen gesproken over de fit van de modellen en niet over de schattingen van de parameters in die modellen. Als het gaat om de interpretatie van de modellen zijn uiteraard met name deze schattingen van belang.
De parameterschattingen voor de gehele matrix met correlaties tussen alle acht toetsen, met behulp van de modellen I en II, worden vermeld in de tabellen 4.3 en 4.4. Tabel 4.3 bevat de parameterschattingen voor model I. Tabel 4.4 bevat de schattingen van de parameters voor model II.

TABEL 4.3
Parameterschattingen en standaardfouten ( ) van model I voor de correlatiematrices in het eerste- en zesdejaar

schatting	jaar 1 (N = 151)			jaar 6 (N = 56)		
parameter	ML		ST	ML		ST
$\hat{\lambda}_1$	1.00"		0.76	1.00"		0.77
$\hat{\lambda}_2$	1.00"		0.79	1.00"		0.76
$\hat{\lambda}_3$	1.00"		0.79	1.00"		0.75
$\hat{\lambda}_4$	1.00"		0.75	1.00"		0.78
$\hat{\lambda}_5$	1.01	(.11)	0.78	1.07	(.16)	0.82
$\hat{\lambda}_6$	0.97	(.09)	0.77	1.00	(.14)	0.76
$\hat{\lambda}_7$	0.98	(.09)	0.77	1.04	(.14)	0.79
$\hat{\lambda}_8$	1.06	(.11)	0.80	1.03	(.16)	0.80

TABEL 4.3 (vervolg)

$\hat{\beta}_2$	0.66	(.05)	0.63	0.95	(.06)	0.96
$\hat{\beta}_3$	0.66	(.05)	0.66	0.95	(.06)	0.95
$\hat{\beta}_4$	0.66	(.05)	0.69	0.95	(.06)	0.91
vâr(ζ_1)	0.58	(.11)	1.00"	0.59	(.16)	1.00"
var(ζ_2)	0.38	(.08)	0.60	0.05	(.06)	0.08
vâr(ζ_3)	0.35	(.08)	0.57	0.05	(.06)	0.09
vâr(ζ_4)	0.29	(.07)	0.52	0.11	(.08)	0.17
vâr(δ_1)	0.39	(.02)	0.43	0.39	(.04)	0.41
vâr(δ_2)	0.39	(.02)	0.38	0.39	(.04)	0.42
vâr(δ_3)	0.39	(.02)	0.38	0.39	(.04)	0.44
vâr(δ_4)	0.39	(.02)	0.44	0.39	(.04)	0.39
vâr(ϵ_1)	0.39	(.02)	0.39	0.39	(.04)	0.33
vâr(ϵ_2)	0.39	(.02)	0.41	0.39	(.04)	0.42
vâr(ϵ_3)	0.39	(.02)	0.41	0.39	(.04)	0.38
vâr(ϵ_4)	0.39	(.02)	0.36	0.39	(.04)	0.36

ML = Maximum Likelihoodoplossing
ST = Gestandaardiseerde oplossing
" = Gespecificeerd door het model

TABEL 4.4
Parameterschattingen en standaardfouten ( ) van model II voor de correlatiematrix in het eerstejaar

schatting parameters	(N = 156) ML		ST
$\hat{\lambda}_1$	1.00"		0.73
$\hat{\lambda}_2$	1.00"		0.80
$\hat{\lambda}_3$	1.00"		0.82
$\hat{\lambda}_4$	1.00"		0.81
$\hat{\lambda}_5$	1.08	(.13)	0.80
$\hat{\lambda}_6$	1.01	(.09)	0.81
$\hat{\lambda}_7$	0.82	(.10)	0.67
$\hat{\lambda}_8$	0.91	(.10)	0.74
$\hat{\beta}_2$	0.64	(.06)	0.58
$\hat{\beta}_3$	0.64	(.06)	0.62
$\hat{\beta}_4$	0.64	(.06)	0.64

TABEL 4.4 (vervolg)

vâr(ζ_1)	0.54	(.11)	1.00"
vâr(ζ_2)	0.42	(.08)	0.66
vâr(ζ_3)	0.41	(.09)	0.61
vâr(ζ_4)	0.38	(.09)	0.59
vâr(δ_1)	0.33	(.03)	0.35
vâr(δ_2)	0.33	(.03)	0.33
vâr(δ_3)	0.33	(.03)	0.30
vâr(δ_4)	0.33	(.03)	0.34
vâr(ϵ_1)	0.33	(.03)	0.34
vâr(ϵ_2)	0.33	(.03)	0.34
vâr(ϵ_3)	0.33	(.03)	0.34
vâr(ϵ_4)	0.33	(.03)	0.31

Toetsspecifieke parameters jaar 1

$\hat{\alpha}_1$	= 0.36	(.12)	$\hat{\gamma}_1$ = .15	(.11)
$\hat{\alpha}_2$	= 0.18	(.12)	$\hat{\gamma}_2$ = -.01	(.10)
$\hat{\alpha}_3$	= 0.16	(.14)	$\hat{\gamma}_3$ = .47	(.09)
$\hat{\alpha}_4$	= 0.01	(.14)	$\hat{\gamma}_4$ = .37	(.09)

ML = Maximum Likelihoodoplossing
ST = Gestandaardiseerde oplossing
" = Gespecificeerd door het model

Reeds eerder werd vermeld dat beide modellen zo gespecificeerd zijn dat de latente variabelen voor ieder tijdstip (zie figuur 4.2) op dezelfde schaal werden gezet als de Eerstejaarstoetsen. In de tabellen wordt dit aangegeven door voor de $\hat{\lambda}$'s 1 t/m 4 de waarde 1.00 te specificeren. De volgende vier parameters $\hat{\lambda}_5$ t/m λ_8 zijn de ladingen van de Algemene Kennistoetsen op de latente variabelen. De parameters $\hat{\beta}_i$ zijn de stabiliteitsparameters in het model. De termen vâr(ζ_1) t/m vâr(ζ_4) zijn de resttermen van de structurele vergelijkingen, met andere woorden dat deel van de variantie dat niet door de structurele vergelijking verklaard wordt. De laatste parameters vâr(δ_i) (i = 1 t/m 4) en vâr(ϵ_i) (i = 1 t/m 4) zijn schattingen van de variantie van de meetfouten van de geobserveerde variabelen in de modellen. In model II worden dan nog acht additionele parameters geschat. Deze worden in tabel 4.4 onderaan vermeld en zijn op te vatten als de ladingen van de geobserveerde variabelen op de toets-specifieke latente variabelen S_x en S_y. De $\hat{\alpha}_i$ zijn de ladingen van de Eerstejaarstoetsen en de $\hat{\gamma}_i$'s zijn de ladingen voor de Algemene Kennistoetsen.

Een eerste, opvallend, resultaat met betrekking tot de gestandaardiseerde oplossing (d.m.v. standaardisatie van de latente variabelen) uit tabel 4.3 is dat de schattingen van de parameters $\hat{\lambda}_i$ in de twee jaargroepen weinig van elkaar verschilden en bovendien redelijk hoge waarden hadden. Dit wijst erop dat in beide jaargroepen de geobserveerde variabelen opgevat mogen worden als goede indicatoren van de latente variabelen voor ieder tijdstip. Hoe hoger de waarden van de $\hat{\lambda}_i$'s, des te beter de geobserveerde variabelen de latente variabele in het model meten.
De schattingen van de parameters $\hat{\alpha}_i$ en $\hat{\gamma}_i$ voor model II (zie tabel 4.4) in het eerstejaar, waren beduidend kleiner dan de schattingen van de $\hat{\lambda}_i$'s in hetzelfde model. Dit betekent dat de toets-specifieke-variantie in vergelijking tot de tijdstip-afhankelijke-variantie in alle geobserveerde variabelen, kleiner was.
Een tweede, opvallend resultaat uit tabel 4.3 is dat de schattingen van de stabiliteitsparameters in beide jaargroepen sterk verschilden. In verband met de vraagstelling van het onderhavige onderzoek zijn vooral de schattingen van deze parameters interessant. Immers, ze geven informatie omtrent de fluctuatie van de latente variabele in het model. Met andere woorden, de parameters $\hat{\beta}_i$ geven aan hoe groot de verandering in medische kennis van tijdstip tot tijdstip was. Hoe lager de $\hat{\beta}_i$'s, des te groter de verandering in kennis en omgekeerd. De verschillen waren in alle drie gevallen significant. Het 95% betrouwbaarheidsinterval van $\hat{\beta}_i$ werd benaderd met behulp van:

$$\hat{\beta}_i \pm 2*\sqrt{\text{geschatte gepoolde variantie van } \hat{\beta}_i}$$

Inhoudelijk betekent dit dat de kennis, gemeten met behulp van Eerstejaarstoetsen en Algemene Kennistoetsen in het eerstejaar veel minder stabiel is, meer fluctueert, dan in het zesdejaar. Aan het eind van de medische studie verandert de kennis, zoals gemeten met deze twee toetsen, niet zoveel meer.
Deze constatering wordt nog eens onderstreept door de correlaties tussen de latente variabelen die in tabel 4.5 staan vermeld.

Uit de tabel blijkt dat de correlaties in het eerstejaar beduidend lager waren dan die in het zesdejaar. De schatting van de correlatie tussen de kennis op het eerste tijdstip en het vierde tijdstip was in het eerstejaar .23. In het zesdejaar was de schatting van de correlatie tussen deze twee tijdstippen .83. Uit deze voorbeelden blijkt inderdaad dat de kennis in het zesdejaar slechts weinig meer fluctueert, terwijl de veranderingen in kennis in het eerstejaar, zoals men van een adequate toetsprocedure zou mogen verwachten, aan sterke veranderingen onderhevig is.

TABEL 4.5
Schattingen van de correlaties tussen de tijdstip-afhankelijke latente variabelen voor eerstejaars (model II) en zesdejaars (model I)

	JAAR 1 (N = 151)					JAAR 6 (N = 56)			
	1	2	3	4		1	2	3	4
1	1.00				1	1.00			
2	0.58	1.00			2	0.96	1.00		
3	0.36	0.62	1.00		3	0.91	0.95	1.00	
4	0.23	0.40	0.64	1.00	4	0.83	0.87	0.91	1.00

4.4.5 Kennis- en toetsmethode-variantie

Op grond van de resultaten die in de voorgaande paragraaf gepresenteerd werden, kan men voor elk van de geobserveerde variabelen, in dit geval de acht toetsen afgenomen op vier tijdstippen, een schatting maken van het deel van de totale variantie dat toe te schrijven is aan de kennis op het betreffende tijdstip en het deel dat aan de soort toets (Eerstejaarstoets of Algemene Kennistoets) is toe te schrijven.
Door in de modellen I en II de parameters $\tilde{\lambda}_i$, en voor zover aanwezig de parameters $\tilde{\alpha}_i$ en $\tilde{\gamma}_i$ te kwadrateren krijgt men een schatting van het percentage variantie dat door de twee latente variabelen wordt verklaard. De schattingen worden vermeld in tabel 4.6.
Uit die tabel blijkt dat in alle gevallen de variantie, die toegeschreven kan worden aan de latente variabele kennis, groter was dan de toets-specifieke-variantie en de variantie van de meetfouten. Dit betekent dat de gemeenschappelijke variantie van de Eerstejaarstoetsen en de Algemene Kennistoetsen op ieder tijdstip het grootst was. Het percentage gemeenschappelijke variantie variëerde van 44% tot 67%.
De unieke varianties van de Eerstejaarstoetsen en de Algemene Kennistoetsen waren ongeveer even groot. In het eerstejaar was de meetfout van zowel de Eerstejaarstoetsen als van de Algemene Kennistoetsen gemiddeld .33.
In het zesdejaar was de gemiddelde meetfout of de unieke variantie van Eerstejaarstoetsen .42 en die van de Algemene Kennistoetsen .37. Dit betekent dat zowel in het eerstejaar als in het zesdejaar de Eerstejaarstoetsen, naast een sterke gemeenschappelijke component met de Algemene Kennistoetsen, eveneens een

redelijk deel unieke of niet verklaarde variantie hadden.

TABEL 4.6
Schattingen van de variantiecomponenten voor de modellen I en II

		Variantie componenten					
		JAAR 1			JAAR 6		
	tijd-stip	kennis	toets	meet-fout	kennis	toets	meet-fout
Ej	1	.53	.12	.35	.59	---	.41
	2	.64	.03	.33	.58	---	.42
	3	.67	.03	.30	.56	---	.44
	4	.66	.00	.34	.61	---	.39
Ak	1	.64	.02	.34	.67	---	.33
	2	.66	.00	.34	.58	---	.42
	3	.44	.22	.34	.62	---	.38
	4	.55	.14	.31	.64	---	.36

Uit tabel 4.6 blijkt verder dat het aandeel van de toets-specifieke-variantie in de data van het eerstejaar in vergelijking tot de twee andere bronnen van variantie, uitermate gering was. Het gemiddelde van Eerstejaarstoetsen was .05, met als uitschieter de toets op het eerste tijdstip. Voor de Algemene Kennistoetsen was het gemiddelde van de toets-specifieke-variantie .10, met als uitschieter de toets op het derde tijdstip.

4.4.6 Samenvatting en conclusies (hypothesen 10a en 10b)

De structuur in de correlaties tussen Eerstejaarstoetsen en Algemene Kennistoetsen worden door de gepostuleerde modellen goed beschreven. De houdbaarheid van de beschreven modellen vormt dan ook een ondersteuning van de in hoofdstuk 2 veronderstelde correlatiestructuur tussen de onderzochte toetsen. Inhoudelijk betekent dit dat beide toetsen gezien kunnen worden als goede indicatoren van de latente variabele in model I.
Deze latente variabele kan gezien de inhoud van de gebruikte meetinstrumenten

geïnterpreteerd worden als "tijdstip-afhankelijke-medische kennis". De stabiliteit van deze variabele bleek in het zesdejaar groter te zijn dan in het eerstejaar. Dit volgt rechtstreeks uit de hoogte van de parameters β_i en kwam tot uiting in het verschil tussen de op grond van het model geschatte correlaties tussen de latente variabele (tabel 4.5).
De toets-afhankelijke-latente variabelen in model II wijzen erop dat vooral in het eerstejaar andere factoren dan kennis een rol kunnen spelen. De inhoudelijke interpretatie van deze variabelen is niet eenvoudig. Het feit dat deze factor vooral van invloed is bij de Algemene Kennistoetsen op de laatste twee tijdstippen, zou kunnen duiden op een neiging van studenten om vragen te beantwoorden op grond van partiële kennnis. Deze tentatieve conclusie wordt ondersteund door analyses van tweede tot en met vijfdejaars correlatiematrices. Tot en met het vijfdejaar gaf model II een substantiële verbetering van de modelfit te zien. In hoofdstuk 7 wordt op deze conclusie verder ingegaan.
Tenslotte is gebleken dat beide toetsen op alle vier onderzochte tijdstippen substantiële unieke varianties bevatten. Gezien het feit dat de voortgangstoetsen nieuwe nog nooit eerder gebruikte items bevatten, behoeft dit resultaat geen verwondering te wekken. Verwacht mag worden dat het samenstellen van toetsen met behulp van reeds gebruikte en gecalibreerde items, de invloed van de unieke variantie in elke toets zal terugdringen. Hoewel er uiteraard altijd enige unieke variantie, ten gevolge van toevalsfluctuaties in de toetsen, aanwezig zal blijven.

4.5 Voorspellen van studieprestaties (hypothese 11)

4.5.1 Inleiding

Als onderdeel van de externe analyses werd de voorspellende waarde van de Eerstejaarstoetsen en de Algemene Kennistoetsen bestudeerd en vergeleken met de voorspellende waarde van de totale voortgangstoets en de residuele voortgangstoets (dit is de totale voortgangstoets minus de Eerstejaarstoets).

4.5.2 Methode

De voorspellende waarde van de genoemde toetsen werd onderzocht met behulp van het multiple lineaire regressie model. Dit model is algemeen bekend zodat de specificatie ervan hier verder buiten beschouwing wordt gelaten.
Als voorspellers werden gebruikt de vier Eerstejaarstoetsen, de vier Algemene Kennistoetsen en de combinatie van vier Eerstejaarstoetsen en vier Algemene Kennistoetsen (zie tabel 4.7). Deze drie voorspellingsmethoden werden vergeleken met de totale voortgangstoets en met een combinatie van Eerstejaarstoetsen en residuele voortgangstoetsen. De laatste combinaties werden toegevoegd om de vraag

te kunnen beantwoorden in hoeverre de voorspelling verbetert als gevolg van een grotere toetslengte. Deze in totaal 5 voorspellingsmethoden worden vermeld in tabel 4.7.
Criteria voor de voorspellingen waren de studieprestaties in het cognitieve domein uit het tweede- en derde studiejaar van het hier onderzochte cohort 1983-1984. Als indicatoren voor de studieprestaties werden de voortgangstoeten en de bloktoetsen uit het tweede en derde studiejaar gebruikt. Voor het tweede studiejaar waren er in totaal 10 criteria (vier voortgangstoetsen en 6 bloktoetsen) beschikbaar. Voor het derde studiejaar waren er in totaal 8 criteria (vier voortgangstoetsen en vier bloktoetsen) beschikbaar.
Om het succes van de voorspelling te kunnen bepalen werd gebruik gemaakt van de multiple correlatiecoëfficiënt R en de proportionele reductie in de standaardfout, die werd bepaald als

$$1 - \hat{S}_p/\hat{S}_o$$

waarbij $\hat{S}_p$ = standaard predictiefout en $\hat{S}_o$ = standaardfout in de criteriumvariabele. Deze index is in vergelijking met R, minder gevoelig voor de verschillen in spreiding tussen de criteriumvariabelen (Muijtjens, Imbos, Theunissen en Roos, 1988).

4.5.3 Resultaten

De resultaten van de voorspellingen van het studiesucces in het tweede studiejaar worden vermeld in tabel 4.7.

TABEL 4.7
Predictie van studieresultaten in het tweede studiejaar (N = 136)

	VOORTGANGSTOETSEN							
	1		2		3		4	
Predictoren*	$\hat{R}$	$\hat{C}_1$	$\hat{R}$	$\hat{C}_1$	$\hat{R}$	$\hat{C}_1$	$\hat{R}$	$\hat{C}_1$
1 Ej	.61	.21	.60	.19	.56	.16	.50	.12
2 Ak	.60	.21	.66	.24	.60	.19	.52	.15
3 Ej/Ak	.65	.24	.69	.24	.63	.22	.55	.15
4 Vgt	.64	.24	.66	.24	.61	.22	.52	.15
5 Ej/Res	.65	.24	.67	.24	.62	.19	.54	.12

TABEL 4.7 (vervolg)

BLOKTOETSEN

Predictoren*	2.1		2.2		2.4		2.5		2.6		2.7	
	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$
1 Ej	.43	.10	.40	.07	.42	.08	.40	.08	.34	.05	.33	.02
2 Ak	.38	.07	.41	.07	.34	.05	.38	.08	.30	.03	.23	.00
3 Ej/Ak	.45	.10	.46	.07	.44	.08	.44	.08	.38	.04	.36	.02
4 Vgt	.38	.07	.35	.05	.36	.05	.36	.06	.29	.03	.27	.02
5 Ej/Res	.44	.07	.41	.05	.43	.08	.43	.08	.37	.03	.34	.02

$\tilde{C}_1 = 1 - \tilde{S}_p/\tilde{S}_o$;
$\tilde{S}_p$ = standaard predictie fout; S_o = standaardfout van criterium variabele;
$\tilde{R}$ = multiple correlatiecoefficient;
Vgt = standaard voortgangstoets;
Res = standaard voortgangstoets minus Eerstejaarsitems;
Ej = Eerstejaarstoets; Ak = Algemene Kennistoets.
* = predictoren op tijdstippen 1,2,3 en 4

In tabel 4.7 kunnen een aantal opvallende patronen worden opgemerkt.
In de eerste plaats valt het op dat de voorspelling slechter was naarmate de criteriumvariabele op een later tijdstip werd afgenomen. De studieprestaties op de eerste twee tijdstippen in het tweede studiejaar werden in de meeste gevallen het best voorspeld. Bij de voortgangstoetsen daalde $\tilde{R}$, ongeacht de predictoren, van de eerste afname tot en met de laatste afname ± .10. Een zelfde daling van $\tilde{R}$ trad op bij de bloktoetsen.
Een tweede opvallend resultaat van de voorspellingen in het tweede studiejaar was dat voortgangstoetsen beter voorspeld werden dan bloktoetsen. Het bereik van $\tilde{R}$ voor de voortgangs-toetsen was van .69 tot .50. Voor de bloktoetsen was het bereik van .46 tot .23.
Een derde resultaat uit tabel 4.7 was dat er tussen de verschillende predictoren over het geheel genomen geen verschillen in kwaliteit van de voorspellingen waren. Alle multiple correlatiecoëfficiënten waarvoor de voortgangstoetsen als criteria dienden, lagen in het bereik van bruikbare coëfficiënten (Visser e.a., 1982). Genoemde auteurs noemen validiteitscoëfficiënten in het bereik van .40 tot .70 bruikbaar als het gaat om voorspellingsdoeleinden (Visser, e.a. blz. 15).
Gemeten naar deze norm waren een aantal coëfficiënten voor de bloktoetsen in het tweede studiejaar te laag.
Of men nu de ene groep voorspellers neemt of de andere maakt in principe weinig uit. Tegen de achtergrond van het verschil in toetslengte tussen de

Eerstejaarstoetsen en de Algemene Kennis-toetsen enerzijds, en de totale voortgangstoetsen anderzijds is dit resultaat opmerkelijk. Immers, Eerstejaarstoetsen en Algemene Kennistoetsen hebben een toetslengte van slechts 1/5 van de lengte van de totale voortgangstoets.
De resultaten van de voorspellingen van de studieresultaten in het derde studiejaar staan vermeld in tabel 4.8.
Uit de resultaten in de bovenstaande tabel komt hetzelfde beeld naar voren als bij de voorspellingen van studieresulaten uit het tweede studiejaar. Ook in het derde studiejaar waren de waarden van $\tilde{R}$ het hoogst voor de voortgangstoetsen, die qua afname tijdstip het dichtst bij het eerstejaar lagen. De waarden van $\tilde{R}$ waren voor deze criteriumvariabelen voor het merendeel in het bereik dat door Visser, e.a. acceptabel wordt genoemd.

TABEL 4.8
Predictie van studieresultaten in het derde studiejaar (N = 136)

	VOORTGANGSTOETSEN							
	1		2		3		4	
Predictoren*	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$
1 Ej	.50	.12	.52	.13	.41	.07	.37	.06
2 Ak	.53	.15	.57	.16	.51	.14	.46	.10
3 Ej/Ak	.55	.15	.60	.16	.52	.10	.48	.10
4 Vgt	.55	.15	.56	.16	.45	.10	.37	.06
5 Ej/Res	.56	.15	.57	.13	.46	.07	.40	.06

	BLOKTOETSEN							
	3.1		3.2		3.3		3.4	
Predictoren*	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{R}$	$\tilde{C}_1$
1 Ej	.27	.02	.27	.03	.27	.02	.31	.04
2 Ak	.28	.02	.15	.00	.31	.02	.35	.06
3 Ej/Ak	.35	.05	.31	.01	.37	.04	.38	.04
4 Vgt	.25	.02	.17	.00	.23	.00	.32	.04
5 Ej/Res	.32	.02	.30	.02	.34	.03	.34	.04

Legenda: zie tabel 4.7

De $\tilde{R}$'s voor de bloktoetsen in het derde studiejaar bereikten in geen van de gevallen, acceptabele waarden. Ook bij deze criteriumvariabelen is sprake van een in de tijd systematisch dalende trend in de $\tilde{R}$'s. Gemeten naar de maatstaven van Visser, e.a. waren de hier gebruikte predictoren niet bruikbaar als voorspellers van de resultaten op de bloktoetsen in het derde studiejaar.
Ook in dit geval bleek dat de kwaliteit van de predictoren elkaar niet veel ontliep. De Eerstejaarstoetsen en de Algemene Kennistoetsen deden het gezamenlijk niet slechter dan de totale voortgangstoets of de Eerstejaarstoetsen tezamen met de residuele toetsen. De Eerstejaarstoetsen en de Algemene Kennistoetsen waren ook voor wat het voorspellen van de studieresultaten betreft in het derde studiejaar gelijkwaardig. Ook bij de totale voortgangstoetsen deed zich in de tijd dalende tendens van $\tilde{R}$ voor. Gaande de studie worden de behaalde resultaten uit het eerstejaar steeds minder belangrijk.

4.5.4 Samenvatting en conclusies (hypothese 11)

In de resultaten met betrekking tot het voorspellend vermogen van de onderzochte toetsen waren een aantal lijnen te ontdekken.
In de eerste plaats bleek dat de Eerstejaarstoetsen en de Algemene Kennistoetsen qua voorspellend vermogen gelijkwaardig waren. Met deze resultaten werd de in hypothese 11 geformuleerde voorspelling ondersteund. Voor beide types toetsen was de kwaliteit van de voorspelling van de studieresultaten in het tweede en derde studiejaar acceptabel. De resultaten op de bloktoetsen bleken in het tweede studiejaar nog wel op acceptabele wijze voorspeld te kunnen worden, doch in het derde studiejaar niet meer. In het derde studiejaar bleken de waarden van $\tilde{R}$ gedaald tot beneden het niveau, dat hier als acceptabel werd beschouwd ($.40 \leq \tilde{R} \leq .70$). Bij dit alles dient wel bedacht te worden dat de proportionele reductie van de standaardfout in de criteriumvariabelen op z'n hoogst 24% was. In de tweede plaats bleek dat de Eerstejaarstoetsen en de Algemene Kennistoetsen geen slechtere voorspellers waren dan de veel langere totale voortgangstoetsen of de residuele voortgangstoetsen.
In de derde plaats bleek dat de voorspellingen slechter werden naarmate het tijdsverschil tussen afname van de voorspellers en de afname van de criteriumvariabele groter was.
In de vierde plaats bleken ongeacht het studiejaar, de resultaten op de voortgangstoetsen beter voorspelbaar dan de resultaten op de bloktoetsen. Elders (Imbos en Stalenhoef, 1988) was al gebleken dat bloktoetsen uit het eerste studiejaar goede voorspellers zijn van bloktoetsen in het tweede studiejaar ($\tilde{R} = .60$). Het voorspellen van de bloktoetsresultaten in het derde studiejaar door de bloktoetsen uit het eerste studiejaar bleek weer heel matig te zijn ($\tilde{R} = .38$). Met

andere woorden ook voor bloktoetsen geldt kennelijk hoe verder weg in de tijd hoe slechter de voorspellingen. Deze observatie is geheel in overeenstemming met de resultaten van het longitudinale gebruik van toetsen (Humphreys, 1968; Werts, Linn en Jöreskog, 1977). Bovendien is deze observatie in lijn met de resultaten die hiervoor met betrekking tot de simplex structuur van de correlaties binnen het eerstejaar werden vermeld. Als de correlaties binnen de relatief korte periode van één jaar al sterk dalen, dan mogen van voorspellingen op nog langere termijn evenmin hoge verwachtingen gekoesterd worden. De resultaten die met betrekking tot het voorspellend vermogen van de bloktoetsen uit het eerste studiejaar werden gevonden, wijzen eveneens in de richting van een simplex structuur bij bloktoetsen. Elders en bij een andere data set werd dit ook al eens geconstateerd (Imbos en Stalenhoef, 1988). In de tijd dalende correlaties komen niet alleen voor bij toetsen die, zoals voortgangstoetsen, een min of meer losse relatie met onderdelen van het onderwijsprogramma hebben. Ook toetsen die wel een duidelijke relatie met onderdelen van dat programma hebben vertonen een dergelijk correlatiepatroon (zie ook Humphreys, 1968).

4.6 Medische kennis in het eerstejaar (hypothese 12)

4.6.1 Inleiding

De kennis van eerstejaars studenten wordt behalve met voortgangstoetsen, gemeten met bloktoetsen. Eerder, in hoofdstuk 1, werd betoogd dat beide instrumenten opgevat kunnen worden als indicatoren van eenzelfde latente variabele, welke geïnterpreteerd kan worden als medische kennis. In deze laatste paragraaf van het hoofdstuk, over de externe analyses van de toetsen, wordt de houdbaarheid van deze opvatting nader onderzocht aan de hand van de analyses van de correlaties tussen Eerstejaarstoetsen, Algemene Kennistoetsen en de bloktoetsen uit het eerste studiejaar. In hypothese 12 werd naar analogie van Guttman's eerste wet voor intelligentie-items, voorspeld dat alle correlaties > 0 zouden moeten zijn. Met andere woorden: voorspeld wordt dat er geen negatieve correlaties zijn.

4.6.2 Resultaten

4.6.2.1 Correlaties tussen bloktoetsen, Eerstejaarstoetsen en Algemene Kennistoetsen

De correlaties tussen de 5 bloktoetsen, de 4 Eerstejaars- en de 4 Algemene Kennistoetsen worden vermeld in tabel 4.9.
In correlatiematrix in tabel 4.9 zijn de drie blokken, die op dezelfde toetsmethode betrekking hebben, door middel van lijnen aangegeven. Zo heeft het eerste blok

alleen betrekking op de correlaties tussen de bloktoetsen, het tweede op de correlaties tussen Eerstejaarstoetsen en het derde op de correlaties tussen de Algemene Kennistoetsen. Binnen elk van de drie blokken was de, al eerder besproken, in de tijd dalende trend in de correlaties het meest opvallend. Verder blijkt dat de correlaties binnen de blokken in vrijwel alle gevallen substantieel (vanaf .15 is de correlatie significant, $\alpha = .05$). In de gehele matrix kwamen slechts twee negatieve correlaties voor, namelijk de correlaties tussen bloktoets 1.6 en de Algemene Kennistoetsen 1 en 2.

TABEL 4.9
Correlaties tussen de bloktoetsen, Eerstejaarstoetsen en Algemene Kennistoetsen.

	toets												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.00												
2	0.55	1.00											
3	0.48	0.52	1.00										
4	0.48	0.56	0.51	1.00									
5	0.33	0.23	0.13	0.28	1.00								
6	0.17	0.03	0.27	0.05	0.05	1.00						N = 141	
7	0.32	0.25	0.27	0.15	0.05	0.38	1.00						
8	0.45	0.46	0.48	0.29	0.07	0.22	0.48	1.00					
9	0.36	0.46	0.36	0.27	0.08	0.10	0.30	0.47	1.00				
10	0.26	0.05	0.18	0.05	-0.01	0.61	0.40	0.2	0.08	1.00			
11	0.30	0.15	0.26	0.06	-0.04	0.38	0.66	0.3	0.21	0.46	1.00		
12	0.38	0.25	0.31	0.21	0.07	0.25	0.44	0.5	0.41	0.31	0.34	1.00	
13	0.45	0.35	0.34	0.23	0.14	0.20	0.36	0.4	0.64	0.28	0.28	0.58	1.00

1 = Bloktoets 1.2
2 = Bloktoets 1.3
3 = Bloktoets 1.4
4 = Bloktoets 1.5
5 = Bloktoets 1.6
6 = Eerstejaarstoets 1
7 = Eerstejaarstoets 2
8 = Eerstejaarstoets 3
9 = Eerstejaarstoets 4
10 = Algemene Kennistoets 1
11 = Algemene Kennistoets 2
12 = Algemene Kennistoets 3
13 = Algemene Kennistoets 4

Dat was ook het geval met de meeste correlaties tussen de blokken. De gemiddelde correlaties tussen Eerstejaarstoetsen en bloktoetsen was .24; die tussen Algemene Kennistoetsen en bloktoetsen .20. Het verschil in gemiddelde correlatie tussen Eerstejaarstoetsen versus bloktoetsen en Algemene Kennistoetsen versus bloktoetsen was dus niet erg groot. De laagste correlaties in de tussentoets blokken kwamen voor rekening van bloktoets 1.6.

4.6.3. Samenvatting en conclusies (hypothese 12)

Uit de analyses van de correlaties tussen bloktoetsen, Eerstejaarstoetsen en Algemene Kennistoetsen kon geen hard "bewijs" worden geleverd met betrekking tot de opvatting dat alle meetinstrumenten eigenlijk dezelfde variabele meten.
Wel werd aangetoond dat de correlaties tussen de toetsen uit hetzelfde domein niet negatief correleren. Met dit resultaat wordt de "eerste wet" voor het domein van kennisitems, zoals werd voorspeld in hypothese 12 ondersteund. Hieruit blijkt dat er aan de eerste voorwaarde voor toetsen uit eenzelfde empirisch domein is voldaan: ze correleren in ieder geval niet negatief.

4.7 Samenvatting en conclusies (externe analyses)

In dit hoofdstuk werden de hypothesen met betrekking tot de externe structuur van Eerstejaarstoetsen en Algemene Kennistoetsen op houdbaarheid onderzocht. Hypothesen 10a en 10b werden geformaliseerd met twee statistische modellen (paragraaf 4.3). Model I, met één latente variabele per tijdstip bleek beide onderzochte correlatiematrices goed te beschrijven. Model II met nog twee extra latente variabelen, voor elk type toets één, verbeterde de fit in het eerstejaar, echter niet in het zesdejaar. De conclusie luidde dat ook in het eerstejaar model II strikt genomen niet nodig is om de structuur in de correlatiematrix te beschrijven (paragraaf 4.4.3). In beide jaargroepen werd hypothese 10a door de fit van model I ondersteund. Deze conclusie werd onderstreept door het geringe percentage van de totale variantie dat door de (in model II) toegevoegde variabelen werd verklaard (paragraaf 4.4.5). Met betrekking tot de tweede toets-specifieke factoren in model II werd gesuggereerd, dat deze variabelen interessant zouden kunnen zijn als meting van de "neiging" bij sommige studenten om te moeilijke vragen te beantwoorden. Uit de fit van de beschreven modellen kon worden geconcludeerd dat Eerstejaarstoetsen en Algemene Kennistoetsen één latente variabele meten (paragraaf 4.4.6).
Vervolgens werd het voorspellend vermogen van beide soorten toetsen ten aanzien van later studiesucces onderzocht. Zoals voorspeld in hypothese 11 was er wat dit betreft tussen de beide toetsen geen verschil te constateren. Naarmate het te

voorspellen studiesucces verder van het eerstejaar was verwijderd werden de voorspellingen slechter. Verder bleek dat de Eerstejaarstoetsen en de Algemene Kennistoetsen de resultaten op de voortgangstoetsen beter voorspellen dan de resultaten op de bloktoetsen (paragraaf 4.5.3 en 4.5.4).
Tenslotte werden in dit hoofdstuk de correlaties tussen de verschillende toetsen die in het eerste studiejaar werden gebruikt onderzocht. In hypothese 12 was voorspeld dat er tussen deze toetsen geen negatieve correlaties zouden zijn. Dit bleek inderdaad het geval (paragraaf 4.6.2.1 en 4.6.3).

5. MODELLEN ALS HULPTHEORIE VOOR MEDISCHE KENNIS

Inhoud

5.1 Inleiding
5.2 De interactie tussen personen en items
5.3 Modellen voor het beantwoorden van toetsitems
5.3.1 De Guttmanschaal als basismodel
5.3.2 Het VGT-model (hypothese 13)
5.3.3 Het Raschmodel (hypothesen 14a, 14b en 14c)
5.3.4 Het Mokkenmodel (hypothesen 15a en 15b)
5.4 Procedure
5.4.1 De personen
5.4.2 De items
5.4.3 De statistische procedures
5.4.3.1 Procedures voor het VGT-model
5.4.3.2 Procedures voor het Raschmodel
5.4.3.3 Procedures voor het Mokkenmodel
5.5 Samenvatting

Hoofdstuk 5: Modellen

5.1 Inleiding

In de voorgaande hoofdstukken werd een antwoord gezocht op de centrale probleemstelling van dit onderzoek door middel van het toetsen van verschillende hypothesen, die in hoofdstuk 1 werden ontleend aan de probleemstelling, de analyse van het begrip-zoals-bedoeld en het instrument om dit begrip te meten.
Hoe nuttig het toetsen van hypothesen op zich ook is, het leidt slechts tot globale kennis van het probleemgebied. Tegenover dit verlies aan specificiteit in de beantwoording van de vraagstelling, staat de relatieve winst van het behoud van alle onderzoekseenheden, de toetsitems en de aan de toetsen deelnemende personen. In dit hoofdstuk wordt dit laatste opgeofferd met de expliciete bedoeling om de vraagstelling specifieker te kunnen beantwoorden en dus meer te weten te komen over de geselecteerde itemgroepen, eerder omschreven als Eerstejaarstoetsen en Algemene Kennistoetsen. Tegenover het waarschijnlijke verlies aan onderzoekseenheden (omdat ze niet aan de modelassumpties voldoen) staat dat met het gebruik van niet-triviale modellen meer voorspellingen over het antwoordgedrag van de respondenten gedaan kunnen worden. "We kopen kennis met assumpties" (Molenaar, 1983a).
Drie modellen worden hier gebruikt. Het aan de R.U. Limburg ontwikkelde Voortgangstoetsmodel, (VGT-model), het Raschmodel en het Mokkenmodel. De in hoofdstuk 1 geformuleerde hypothesen hebben betrekking op deze modellen. Hypothese 13 heeft betrekking op het VGT-model, de hypothesen 14a, 14b en 14c op het Raschmodel en de hypothesen 15a en 15b op het Mokkenmodel. De drie modellen hebben gemeenschappelijk dat antwoorden op toetsitems worden opgevat als een functie van een latente variabele. Ze kunnen daarom alle drie worden gerekend tot de grote groep van item response modellen, die het laatste decenium steeds meer in de belangstelling van onderzoekers is komen te staan. Aan de "wieg" van deze modellen staat het model van Guttman (1950 en 1954). De modellen vervullen in dit hoofdstuk de rol van een minitheorie over de variabele medische kennis. Deze theorie voorspelt hoe de variabele zich zou behoren te gedragen. Onderzoek naar de fit tussen het model en de gegevens levert een toetsing op van de (mini)theorie.
Hieronder worden de modellen slechts in essentie beschreven. Voor meer details zij verwezen naar de literatuur. Voor het VGT-model kan worden verwezen naar Albers, Does, Imbos en Janssen (1987). Over het Raschmodel is zeer veel literatuur beschikbaar. Goede inleidingen op dit model worden gegeven door Rasch (1960), Wright en Stone (1979), Hullin, Drasgow en Parsons (1983), Molenaar (1983a), Hambleton en Swaminathan (1985). Specifieke toepassingen van meer algemene versies van het Raschmodel zijn te vinden in Van der Kamp en De Gruijter (1980), Lord (1980) en Weiss (1983). Een handboek op het gebied van het Raschmodel is Fisher (1974), terwijl Lord en Novick (1968) gezien kan worden als het handboek

voor algemenere item response modellen waarvan het Raschmodel als een verbijzondering gezien kan worden.
Literatuur met betrekking tot het derde, hier gebruikte Mokkenmodel, is te vinden in Mokken (1971) en Molenaar (1982a, 1982b, 1983a). Een boeiende discussie over de (on)bruikbaarheid van dit model werd ten onzent gevoerd door Jansen (1982), Molenaar (1982c), Jansen, Roskam en Van den Wollenberg (1982) en Molenaar (1982d).
Voordat in paragraaf 5.3 de drie modellen aan de orde worden gesteld, zal in paragraaf 5.2 eerst stil worden gestaan bij het gedrag van studenten, die toetsvragen trachten te beantwoorden, met de bedoeling daarop een zo hoog mogelijk resultaat te behalen. Immers het is voor deze situatie dat de modellen in de betekenis van geformaliseerde theorie te hulp worden geroepen.
In paragraaf 5.3.1 worden de eigenschappen van de Guttmanschaal beschreven. In paragraaf 5.3.2 wordt het VGT-model beschreven, daarna het Raschmodel (paragraaf 5.3.3) en tenslotte het Mokkenmodel (paragraaf 5.3.4). De procedure volgens welke er met deze modellen werd gewerkt komt in paragraaf 5.4 aan de orde. Paragraaf 5.5 geeft tenslotte een overzicht van dit hoofdstuk.

5.2 De interactie tussen personen en items

Toetsen worden gebruikt om uitspraken over personen of - bij een averechts gebruik (Hofstee, 1969) - om uitspraken over instituten, bijvoorbeeld een school, te doen. De eigenschappen of constructen die met behulp van een toets gemeten kunnen worden, zijn doorgaans niet in één uitdrukking onder te brengen. Om de verschillende aspekten van een construct gerepresenteerd te krijgen in het meetinstrument, worden in een toets veel items gebruikt. Door alle relevante aspecten op deze wijze op te nemen mag verwacht worden dat de toets in zijn geheel het te meten construct adequaat weergeeft. Van elk item wordt nu aangenomen dat het een deelaspect van het construct vertegenwoordigt. Dit kan een eenvoudig aspect, of een moeilijk aspect van het construct zijn. Items kunnen met betrekking tot deze moeilijkheid worden gerangordend van gemakkelijk naar moeilijk. Vandaar dat wordt aangenomen dat ieder item gekarakteriseerd kan worden door de plaats die het in de rangorde van items inneemt. De positie van alle items gezamenlijk definiëren het construct.
Ook de personen brengen iets mee in de toetssituatie: hun cognitieve vaardigheid, kennisbestand of cognitieve structuren. Relatief ten opzichte van de toetsitems, kan men ook van de sterkte van een kennisbestand spreken. Afgezien van raadgedrag, is het het eenvoudigste om aan te nemen dat de uitkomst van de interactie tussen persoon en item door deze twee krachten wordt bepaald, de moeilijkheid van het item en de sterkte van de cognitieve structuur van de persoon. Een persoon met sterke cognitieve structuren zal doorgaans veel items beantwoorden en bovendien

veel daarvan goed beantwoorden. Daarentegen zal een persoon met minder sterke cognitieve structuren minder items beantwoorden. Verder kan worden verwacht dat de items die door een persoon beantwoord worden precies die items zijn die het best passen bij de kennisstructuren van de persoon in kwestie. Zo'n item wordt herkend als beantwoordbaar.
Wanneer personen en items op eenzelfde schaal gemeten kunnen worden, dan leidt bovenstaande redenering tot de voorspelling dat de respondenten doorgaans items beantwoorden met schaalwaarden (itemmoeilijkheden) die kleiner of net iets groter zijn dan hun persoonsschaalwaarde. Items met schaalwaarden die veel groter zijn dan de eigen persoonswaarde zullen als regel niet beantwoord worden.
Voor de doorsnee respondent betekent dit dat het beantwoorden van items opgevat kan worden als de uitkomst van een subjectieve vergelijking van de persoonlijke cognitieve structuren met de moeilijkheid van het item. De persoon beschouwt zichzelf sterker en besluit om het item te beantwoorden.
In deze lijn doorredenerend kan het niet beantwoorden van een item worden opgevat als een negatieve uitkomst van het vergelijkingsproces tussen de persoon en het item. De persoon beschouwt zichzelf als zwakker dan het item en besluit om het item niet te beantwoorden. De fouten die in het antwoordproces op kunnen treden worden in deze redenering verwacht bij de items die sterkere cognitieve structuren veronderstellen dan die waarover de persoon in kwestie de beschikking heeft. Hoe groter het aantal fouten, des te groter de kans dat de persoon zichzelf heeft overschat.
Onderwijs is bedoeld om de cognitieve structuren van studenten en leerlingen te ontwikkelen, sterker, rijker en algemener te maken. Door de bank genomen zullen de sterkste cognitieve structuren in de hogere studiejaren worden aangetroffen en de zwakste in het eerste studiejaar. Echter, verwacht mag worden dat de ontwikkeling van de cognitieve structuren zich niet strikt aan de afbakening van studiejaren zal houden. Dat betekent dat een strikte indeling van de studenten naar studiejaren bij het bestuderen van de ontwikkeling van de cognitieve structuren los gelaten moet worden. Bij het meten van medische kennis met behulp van voortgangstoetsen, mogen op grond van deze redenering gedeeltelijk overlappende scoreverdelingen van de jaargroepen worden verwacht. Voorwaarde daarbij is wel dat er een duidelijk stijgende tendens in de gemiddelde waarden van de scoreverdelingen aanwezig moet zijn. De empirische bevestiging hiervan in de scoreverdelingen van een toets voor medische kennis is in overeenstemming met de aanname dat onderwijs in het medische domein leidt tot de ontwikkeling van kennis. Hoe langer men bezig is des te meer problemen kunnen worden opgelost.
De hier gehanteerde redenering kan ook worden toegepast op een willekeurig probleem uit het medische domein. Ook voor één bepaald probleem geldt dat het bedenken van de juiste oplossing afhangt van de mate van ontwikkeling van de kennis. Per probleem uit het medische gebied geldt de verwachting dat de

succeskans voor dit probleem afhangt van het niveau van de kennis. Deze redenering leidt tot de empirische voorspelling dat voor elk probleem de gemiddelde proportie goede antwoorden in de jaargroepen een stijgende tendens vertoont. Of nog specifieker dat de propertie goede antwoorden over alle kennisniveaus (dus niet alleen gemiddeld) heen een stijgende tendens vertoont. Onder de additionele voorwaarde van voldoende specificiteit (de hogere kans op goed geldt alleen voor medisch geschoolde personen) moet voor ieder meetinstrument voor medische kennis worden aangetoond dat de beschreven relaties tussen de problemen (de items) en de gemeten kennis bestaan.
Voor het bestuderen van het antwoordgedrag van studenten op toetsitems en voor het meetbaar maken van de variabelen die hierop van invloed zijn, moet gebruik worden gemaakt van een statistisch model. De modellen waarnaar hier gezocht wordt, moeten in staat zijn om de uitkomsten van een interactie tussen personen en items te voorspellen.
Een bekend model, dat voor situaties zoals hierboven beschreven gebruikt kan worden en ook door verschillende onderzoekers (Wright en Stone, 1979 en Hambleton en Swaminathan, 1985) met succes is toegepast is het Raschmodel. Dit model gaat er van uit dat de kans op een goed antwoord op een toetsitem afhankelijk is van de positie van de persoon en de positie van het item op een of andere schaal. Beide posities kunnen in het model grafisch worden weergegeven door een punt op de gebruikte schaal. Hoe verder naar rechts het persoonspunt nu van het itempunt verwijderd is hoe groter de kans om dat item goed te beantwoorden. Het model kent dus twee grootheden, iets dat het item kenmerkt en iets dat de persoon kenmerkt. Beide fungeren in het model als parameters, die met behulp van de data geschat worden. Deze parameters hebben binnen het model de rol van enerzijds de sterkte van het item en anderzijds de sterkte van de persoon. Beide krachten worden dus door middel van het model meetbaar gemaakt, oftewel geschat. De relatie tussen de positie van een persoon op de schaal, de moeilijkheid van het item en de kans op een goed antwoord, wordt in het model beschreven met behulp van de logistische verdelingsfunctie.
Verwant aan het Raschmodel is het model van Mokken (1971). De functie die de kans op een goed antwoord weergeeft is binnen dit model niet anders omschreven dan dat het een monotoon niet-dalende functie van de latente variabele moet zijn. Om deze reden noemt men het Mokkenmodel wel de niet-parametrische variant van het Raschmodel (Molenaar, 1982d en 1983a).
Het aan de R.U. Limburg ontwikkelde VGT-model gaat eveneens uit van een situatie waarin bij het beantwoorden van toetsvragen twee krachten een rol spelen, die van de persoon en die van het item. Binnen dit model worden beide echter niet behandeld als parameters, maar als stochastische variabelen, waarvan gemiddelde en spreiding geschat kunnen worden. Ook in dit model wordt een expliciete wiskundige relatie verondersteld tussen de kans op een goed antwoord,

de sterkte van de cognitieve structuren waarover de persoon de beschikking heeft en de moeilijkheid van het item. Deze relatie wordt in het model weergegeven met behulp van de normale verdelingsfunctie.
Deze drie modellen bieden een goede statistische theorie voor het beschrijven van het antwoordgedrag van studenten op de Eerstejaarstoetsen en de Algemene Kennistoetsen. Met behulp van deze modellen kunnen de relevante variabelen of parameters van personen en items meetbaar gemaakt worden. Met het meetbaar maken van de sterkte van items en personen kunnen er voor alle gemeten niveaus van de latente variabele, uitspraken worden gedaan met betrekking tot de efficiëntie van de Eerstejaarstoets en de Algemene Kennistoets. De modellen worden in dit onderzoek naast elkaar gebruikt. Dit maakt het mogelijk om met behulp van het VGT-model allereerst de relatie tussen alle items gezamenlijk en het gemeten construct te onderzoeken. Vervolgens kan dan met behulp van het Raschmodel en het Mokkenmodel heel specifiek deze relatie tussen elk item afzonderlijk en het construct bestudeerd worden. Tenslotte heeft deze strategie als voordeel dat parametrische modellen en een niet-parametrisch model naast elkaar gebruikt kunnen worden, waardoor waarschijnlijk zal blijken dat meer items gunstige meeteigenschappen hebben dan bij het gebruik van enkel een parametrisch model het geval zou zijn.
In de volgende paragraaf worden de drie modellen achtereenvolgens aan de orde gesteld. De kenmerken van deze modellen kunnen het best worden begrepen door uit te gaan van de ideeën van Guttman met betrekking tot het meten van constructen met behulp van items.

5.3 Modellen voor het beantwoorden van toetsitems

5.3.1 De Guttmanschaal als basismodel

Het basisidee van de hieronder te bespreken modellen is dat personen en items schaalbaar zijn. Onder een schaal kan volgens Jansen e.a. (1982) worden verstaan:

> "een schaal is de (numerieke) representatie van een eigenschap of hoedanigheid van voorwerpen (i.c. personen en/of items)" blz. 39.

Een schaalmodel kan worden beschouwd als een formeel voorschrift voor empirische relaties tussen items en personen, gebaseerd op een meettheoretisch ideaal. De vraag naar de schaalbaarheid van items en personen kan derhalve alleen worden beantwoord met behulp van een schaalmodel. Een schaalmodel in deze visie is een instrument om een empirische hypothese omtrent de structuur die een onderzoeker in de data verwacht aan te treffen, te toetsen. Schaalbaarheid "is not to be desired or constructed" (Guttman, 1977). Schaalbaarheid is met behulp van een schaalmodel wél toetsbaar.
Ieder schaalmodel is gebaseerd op een meettheoretisch ideaal. Het meetideaal van

Guttman (1950 en 1954) is sterk verwant aan de ideeën die ten gronslag liggen aan wat men wel aanduidt met de term fundamenteel meten (Krantz, Luce, Suppes en Tversky, 1971). Het basiselement in deze opvatting over meten, is het vergelijken van de gemeten objecten (i.c. items en personen) en de constructie van meetinstrumenten die dergelijke kwantitatieve vergelijkingen mogelijk maken. Het resultaat van een fundamentele meetprocedure is dat de relaties tussen de gemeten variabelen multiplicatief of additief zijn. Verder zijn de vergelijkingen tussen twee objecten niet afhankelijk van de andere objecten die in dezelfde procedure gemeten zijn. Tenslotte biedt een fundamentele meetprocedure de mogelijkheid dat de vergelijkingen in termen van een constante eenheid gemaakt kunnen worden en dat die eenheid constant is over het gehele bereik van de variable.
Dit ideaal voor een meetprocedure is herkenbaar in het onderstaande citaat van Guttman:

> "A person's score tells what his responses were to each question [...] A person with a higher score is "favourable" on all questions that a person with a lower score answers "favourably", and on at least one more in addition. This rank order, furthermore, exists not only for the given series of questions, but is the same as the order that would be obtained with any other series of questions in the same area" (Guttman, 1950, blz. 20).

Naast de genoemde eisen stelt Guttman nog als eis dat de items die het concept meten ééndimensionaal zijn. In de praktijk komt het Guttman ideaal neer op de eis, dat een datamatrix met reakties van personen op items de structuur moeten vertonen, die is weergeven in tabel 5.1.

TABEL 5.1
Patroon van een perfecte Guttmanschaal

	Items					
Personen	1	2	3	4	5	totaal
1	1	1	1	1	1	5
2	1	1	1	1	0	4
3	1	1	1	0	0	3
4	1	1	0	0	0	2
5	1	0	0	0	0	1
Totaal	5	4	3	2	1	

Het patroon in de bovenstaande matrix komt inderdaad neer op het kernidee uit

het aangehaalde citaat. Het model, omschreven als het deterministische Guttmanmodel (Mokken, 1971) voorspelt dat een persoon met een score van 4 precies de vier gemakkelijkste items goed zal hebben, een persoon met een score van 3 precies de drie gemakkelijkste items, enz. Elk itempaar waarbij een gemakkelijk antwoord wordt gemist terwijl een moeilijker item wel goed wordt gemaakt telt in het deterministische Guttmanmodel als een fout. Zo bevat het antwoordpatroon 1 0 1 1 0 twee fouten want er zijn twee itemparen, niet noodzakelijkerwijs naast elkaar, met het patroon 0 1.
Deze Guttman schaalmethode noemt men ééndimensionaal omdat de personen i met de parameters s_i en de items j met de parameters t_j worden geordend op één latente variabele. Men zegt dat persoon i item j domineert wanneer geldt $s_i > t_j$. In plaats van met de parameters s_i en t_j werkt men in de Guttmanschaal met p_j, de proportie personen, die een bepaald item j domineren (= een 1 i.p.v een 0 gescoord op dat item). Deze proportie noemt men de itempopulariteit (Mokken, 1971).
De definitie van het Guttman scalogram of de perfecte Guttmanschaal impliceert dat iedereen die een moeilijk item i goed heeft, eveneens het minder moeilijke item j goed zal hebben, dus dat geldt $p_{ij} = p_i$, waarbij p_{ij} de proportie personen met beide items goed is. In dit geval noemt men de itemparen Guttmanhomogeen ($p_i < p_j$).
Omdat de perfecte Guttmanschaal uit tabel 5.1 een erg restrictief model is heeft men gezocht naar maten die aangeven in hoeverre een scorepatroon bij toeval van het perfecte Guttmanpatroon afwijkt. Daarvoor zijn verschillende maten voorgesteld en weer verworpen (Mokken, 1971). Een coëfficiënt die het heeft overleefd is de H-coëfficiënt van Loevinger (1947, 1948). In de coëfficiënt van Loevinger wordt de geobserveerde proportie p_{ij} (= proportie personen met i goed en j fout; i moeilijker dan j) vergeleken met de bij volledige onafhankelijheid van de item i en j verwachte proportie.
De hieronder te bespreken modellen kunnen worden gezien als probabilistische varianten van het deterministische Guttmanmodel. Gezien het feit dat perfecte Guttman schalen bijna niet en het bestaan van imperfecte schalen veelvuldig is aangetoond (Mokken, 1971, blz. 70-71), is men het deterministische criterium voor de schaalbaarheid gaan vervangen door een probabilistisch criterium. "The quasi scale (bedoeld wordt de imperfecte Guttmanschaal) is in fact nothing but a stochastic response model in embryonic form" (Mokken, 1971, blz. 71).
In plaats van de eerder genoemde Guttmanhomogeniteit stelt Mokken als meetideaal de "holomorfie" en het toetsen daarvan in de praktijk van de schaalconstructie. Wanneer de zogenaamde itemkarakteristieke curves van een item die de kans op een goed antwoord, gegeven de positie op latente variabele weergeeft, wordt gedefinieëerd als $\pi_j(s_i,t_j)$, dan noemt Mokken (blz. 118) de verzameling items holomorf als ze de volgende eigenschappen hebben:

- de itemkarakteristieke functies $\pi_j(s_i,t_j)$ hebben alle dezelfde functionele vorm;
- $\pi_j(s_i,t_j)$ is monotoon stijgend in s_i;
- $\pi_j(s_i,t_j)$ is monotoon dalend in t_j.

De laatste twee eigenschappen van item j vormen de zogenaamde dubbele monotonie (Mokken, 1971, blz. 118). Wanneer een item-karakteristieke curve alleen stijgend is in s_i, spreekt Mokken van "monotone homogeniteit".
De hieronder te bespreken modellen gaan alle drie uit van het meettheoretisch ideaal dat de verzameling items holomorf moeten zijn. Ze verschillen van elkaar in de specificatie van de functionele vorm van de itemkarakteristieke functies. In het VGT-model wordt er van uitgegaan dat $\pi_j(s_i,t_j) = H$, waarbij $H = \Phi$, de standaard normale verdelingsfunctie en bovendien s en t niet als parameters worden opgevat maar als stochastische variabelen.
Het Raschmodel gaat er vanuit dat de itemkarakteristieke curves parallelle logistische functies zijn. Naast de genoemde eisen van holomorfie, stelt het Raschmodel dat de totaalscore van het item en van de persoon voldoende stochastische grootheden zijn. Door deze aanvullende eis kan Raschhomogeniteit als een speciaal geval van holomorfie worden gezien.
Het Mokkenmodel tenslotte laat de functionele vorm van de itemkarakteristieke functie vrij en eist alleen dat de curves monotoon niet-dalend zijn en elkaar bovendien niet snijden.

5.3.2 Het VGT-model

Dit model werd door Albers e.a. (1987) speciaal voor de voortgangstoetsen ontwikkeld. Voornaamste doelstelling bij het ontwikkelen van het model was, het schatten van de individuele groei in kennis mogelijk te maken en statistisch te modeleren. De beoogde toepassing was een situatie waarin een cohort studenten op maximaal 24 verschillende tijdstippen een groot aantal vragen ($\pm$ 250) beantwoordt. Het beoogde model zou voor elk tijdstip een schatting van de kennis van de studenten op eenzelfde schaal moeten opleveren. Omdat er in de praktijk van de voortgangstoets steeds met een nieuwe steekproef van toetsvragen wordt gewerkt, moest er in het model rekening worden gehouden met de mogelijkheid dat er schommelingen in de moeilijkheid van de toetsvragen zouden kunnen optreden. Om tot een juiste schatting van de kennis te kunnen komen zou het model voor de verschillen in moeilijkheid moeten corrigeren.
Het model is het eenvoudigst uit te leggen door uit te gaan van de situatie waarin m personen n toetsvragen beantwoorden. De variabele $X_{ij} = 1$ als de i-de persoon de j-de toetsvraag goed beantwoordt. De kans op een goed antwoord in het Raschmodel wordt geschreven als

$$P(X_{ij}=1 \mid S_i=s_i, T_j=t_j) = H(s_i-t_j), \qquad (5.1)$$

waarbij H de standaard logistische verdelingsfunctie is, s_i de persoonsparameter en t_j de moeilijkheid van het item. In plaats van de standaard logistische verdelingsfunctie in 5.1, wordt in het VGT-model voor de functie $H = \Phi$ genomen, waar Φ de standaard normale verdelingsfunctie is. Verder, en dat is een belangrijker verschil met het Raschmodel worden $s_1,...,s_m$ en $t_1,...,t_n$ beschouwd als uitkomsten van twee onafhankelijke steekproeven $S_1,...,S_m$ en $T_1,...,T_n$ corresponderend met de stochastische variabelen S en T. Gezien de toepassing van het model leek dit een redelijk uitgangspunt. Immers, bij voortgangstoetsen worden de items door middel van een gestratificeerde steekproef uit een voorraad items getrokken. De deelnemende personen kunnen worden gezien als een denkbeeldige steekproef uit de personen, die toegang zoeken tot de medische studie.
In het model wordt er van uitgegaan dat S verdeeld is als $N(\Theta,\sigma^2)$ en T is verdeeld als $N(0,\tau^2)$, verder is de hulpvariabele V_{ij}, nodig om de kans $X_{ij} = 1$ uit te rekenen, verdeeld als N(0,1). De verschillende kansen in het model kunnen dan worden geschreven als,

$$
\begin{aligned}
&P(X_{ij}=1 \mid S_i=s_i, T_j=t_j) = \quad \Phi(s_i - t_j),\\
&P(X_{ij}=1) = \Phi\left(\frac{\Theta}{(1+\sigma^2+\tau^2)^{1/2}}\right)\\
&P(X_{ij}=1 \mid S_i=s_i) = \Phi\left(\frac{s_i}{(1+\tau^2)^{1/2}}\right) \qquad (5.2)\\
&P(X_{ij}=1 \mid T_j=t_j) = \Phi\left(\frac{\Theta - t_j}{(1+\sigma^2)^{1/2}}\right)
\end{aligned}
$$

Dit alles geldt in een situatie waarin een persoon i slechts één toets bestaande uit n toetsvragen aflegt.
Voor de situatie waarin de m personen meer toetsen maken, wordt er aan de verschillende stochastische variabelen een index k toegevoegd, die de p verschillende tijdstippen aangeeft. In deze situatie geldt dat de V_{ijk} evenals in de situatie van één toets onafhankelijk zijn. Dat geldt evenzeer voor de variabele T_{ijk}. De variabelen S_{ik} zijn niet onafhankelijk, omdat $S_{i1},...S_{ip}$ alle betrekking hebben op dezelfde persoon i.
Omdat het zelfs in een situatie van één toets niet redelijk lijkt om $S_{ik} = S_i$ te nemen voor alle k, wordt er in het model aangenomen dat bij verschillende afnames van één en dezelfde toets de toetsscores van persoon i

fluctureren rondom zijn intrinsieke of ware score. Zodat geldt,

$$S_{ik} = A_i + Z_{ik}, \qquad (5.3)$$

waarbij A_i verdeeld is als $N(\alpha,\sigma^2_A)$ en Z_{ik} als $N(0,\sigma^2_Z)$. In 5.3 vertegenwoordigt de eerste term A_i de inter- en de tweede term Z_{ik} de intrapersoonsvariabiliteit. Ook in deze situatie geldt dat alle variabelen V_{ijk}, T_{jk}, Z_{ik} en A_i, i = 1,...,m, j = 1,...,n, k = 1,...p, worden beschouwd als onafhankelijke variabelen.
Als laatste stap wordt de individuele groei in het model geïncorporeerd. Hierdoor wordt het model,

$$S_{ik} = A_i + B_i/p + Z_{ik}, \qquad (5.4)$$

waar B_i verdeeld is als $N(\beta,\sigma^2_B)$ en bovendien onafhankelijk is van de andere toevalsvariabelen in het model. A_i vertegenwoordigt nu het kennisniveau bij het begin van de studie en B_i de individuele groei in kennis gedurende de studie. In het model is de kans op een goed antwoord op de j-de vraag door persoon i op tijdstip k,

$$P(X_{ijk} = 1) = P(V_{ijk} < A_i + B_{ik}/p + Z_{ik} - T_{jk}), \qquad (5.5)$$

i = 1,...,m, j = 1,...,n, k = 1,...,p en alle variabelen rechts in (5.5) zijn onafhankelijk en verdeeld zoals eerder beschreven.
Met het bovenstaande is het VGT-model in essentie beschreven. Alle relevante parameters blijken te kunnen worden geschat op de wijze zoals beschreven in Albers e.a. (1987). De schattingsprocedures zijn opgenomen in een speciaal voor dit model ontwikkeld computerprogramma ABEST (Ability ESTimation) (Janssen, Imbos, Does en Albers, 1988). De wijze waarop er in het kader van het onderhavige onderzoek met dit model en de schattingen die het opleverde is gewerkt, wordt nader beschreven in paragraaf 5.4.3 van dit hoofdstuk.

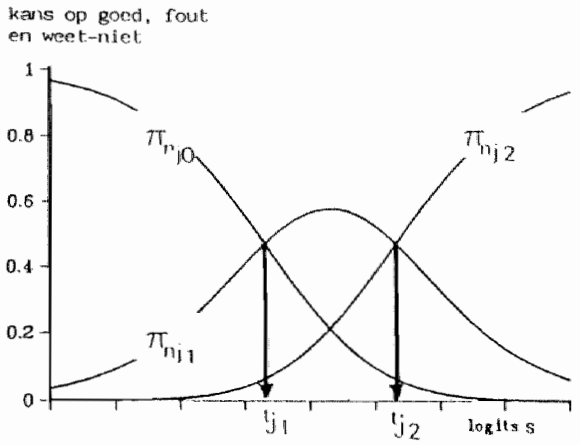
5.3.3 Het Raschmodel (hypothesen 14a, 14b en 14c)

Het Raschmodel is eveneens een model dat gebruikt kan worden om de relevante parameters die de interactie tussen personen en toetsitems verklaren, te schatten. In dit model spelen twee parameters s_i en t_j een rol. Eén voor de persoon en één voor het item. Zodat de kans op een goed antwoord voor persoon i op item j geschreven kan worden als,

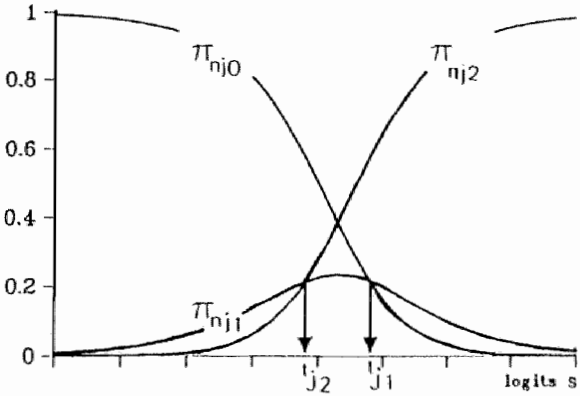
$$P(X_{ij} = 1 \mid S_i = s_i, T_j = t_j) = \frac{\exp(s_i - t_j)}{1 + \exp(s_i - t_j)} \qquad (5.6)$$

waar s_i en t_j, net als in het VGT-model opgevat kan worden als de kennis van persoon i en de moeilijkheid van item j. Formule 5.6 geeft de kans op een goed antwoord in het geval het item j dichotoom (1 = goed, 0 = niet goed) is gescoord.
Het dichotome Raschmodel kan gegeneraliseerd worden naar het Partial Creditmodel (PC-model) (Wright en Masters, 1982). Daardoor wordt het mogelijk om de 3 antwoord mogelijkheden, die bij de voortgangstoetsen worden gebruikt, apart te onderzoeken. De wiskundige formulering van het model zou te veel ruimte en uitleg vragen. Daarom wordt hier volstaan met een verbale beschrijving van het model.
Het model definieert voor ieder item j ($m_j + 1$) curves (m = het aantal stappen in het item), die net als in het dichotome Rasch-model logistische curves zijn. Ze verschillen slechts van elkaar in termen van de locatie t_{jk} op de variabele die mede door item j wordt gedefinieerd. Iedere curve geeft de kans aan op een antwoord in categorie k, in plaats van een antwoord in categorie k-1.
Wanneer de antwoord mogelijkheden op alle toetsvragen in plaats van dichotoom (1 = goed, 0 = vraagteken of fout), trichotoom gescoord worden, zodanig dat een vraagteken gescoord wordt als 0, een fout antwoord als 1 en een goed antwoord als 2, dan kunnen de antwoordcurves in het PC-model er uit zien zoals geïllustreerd wordt door figuur 5.1.
In het geval de scoring is zoals omschreven, dan is $m_j = 2$ en definieert het model 3 antwoordcurves. De parameters t_{jk} , k = 1,2 ook wel itemstappen genoemd (Molenaar, 1983b), vormen de intersecties van de curves 0 en 1 en van de curves 1 en 2. Uit figuur 5.1 blijkt welke de rol is van de parameters t_{j1} en t_{j2} in het PC-model. Het verschil tussen beide parameters bepaalt nl. hoe groot de kans is op een antwoord in categorie 1, met andere woorden in het onderhavige geval, hoe groter de kans op een fout antwoord voor de verschillende waarden van s_i. Bij een groot verschil tussen t_{j2} - t_{j1} met ($t_{j2} > t_{j1}$) is voor veel waarden van s_i de kans op een fout antwoord relatief groot.
In het PC-model is het geenszins noodzakelijk dat de tweede itemstap moeilijker ($t_{j2} > t_{j1}$) is dan de eerste. Door in het model het verschil t_{j2} - t_{j1} kleiner te laten worden daalt de kans op een antwoord in de midden categorie, i.c. de kans op een fout antwoord.
Het is zelfs mogelijk dat de tweede itemstap eenvoudiger te nemen is dan de eerste. In dat geval is $t_{j2} < t_{j1}$, i.c. t_{j2} ligt op de schaal van figuur 5.1 links ten opzichte van t_{j1}. In die situatie neemt voor alle personen i de kans om een fout antwoord te geven nog meer af. In het geval dat itemstap t_{j2} extreem links en t_{j1} een extreem rechts ligt op de schaal van de variabele, zal de antwoordcurve voor een fout antwoord een zeer vlak verloop hebben. Dit betekent bij een scoring zoals omschreven dat de kans op een fout antwoord over het gehele gebied van de variabele heel klein is en het responseproces gedomineerd wordt door hetzij een vraagteken hetzij een goed antwoord op de vraag. Deze rol van de parameters t_{jk} in het PC-model maakt dat ze moeten worden opgevat als parameters van item j,

die niet los van elkaar geïnterpreteerd kunnen worden (Molenaar, 1983b).



itemstap 1 gemakkelijker dan itemstap 2



itemstap 1 moeilijker dan itemstap 2

Figuur 5.1 Antwoordcurves voor goed, fout en weet-niet antwoorden in het PC-model (π_{nj0} = kans op weet-niet, π_{nj1} = kans op fout en π_{nj2} = kans op goed, t_{j1} = itemstap 1 en t_{j2} = itemstap 2).

5.3.4 Het Mokkenmodel (hypothesen 15a en 15b)

Ook het Mokkenmodel behoort tot de groep van latente trek modellen, die zich direct baseren op het discrete karakter van de items (Molenaar, 1983a). Evenals Rasch gaat Mokken (1971) ervan uit dat de kans op een goed antwoord afhangt van

enerzijds de positie van de persoon op de latente eigenschap en anderzijds de moeilijkheid van het item. Via vergelijkbare eisen aan een schaal, gaat ook Mokken ervan uit dat de kans op een goed antwoord een monotoon niet dalende functie is van S, de latente variabele. Zodat ook in het Mokkenmodel geldt

$$P(X_{ij} = 1 \mid S_i = s_i, T_j = t_j) = \quad H(s_i - t_j), \tag{5.7}$$

waar H(x) een niet nader omschreven functie is van s_i en t_j als in 5.1.
Uitgaande van de assumpties dubbele monotonie en locale onafhankelijkheid (d.w.z. onafhankelijke item responsen binnen een groep personen met dezelfde waarde van s_i) en bovendien van een populatieafhankelijke eis van voldoende samenhang tussen de items resulteert in het Mokkenmodel een schaaltype, waarin maar een beperkt aantal schendingen van de perfecte Guttmanschaal optreden (Mokken, 1971, blz 173 e.v.v., en Molenaar, 1983a).
Omdat het Mokkenmodel slechts de ordening van de items en de ordening van de personen als uitgangspunt voor de schaalconstructie neemt en de onderzoeker door middel van monotonieonderzoek (Molenaar, 1983a) van de scoregroepen en de items al dan niet besluit tot een Mokkenschaal, spreekt Molenaar van een niet-parametrisch latente trekmodel. Over de vraag of dit wel of niet gerechtvaardigd is werd in het Tijdschrift voor Onderwijsresearch een levendige discussie gevoerd (Jansen, 1982, Molenaar, 1982c, Jansen e.a., 1982 en Molenaar, 1982d).
Een duidelijke conclusie heeft deze discussie niet opgeleverd. De discussianten lijken het er over eens te zijn dat niet alle "Raschitems" in een Mokkenschaal worden geaccepteerd. Op zich is dat een onbevredigende situatie. Immers het Mokkenmodel zou een algemener model moeten zijn, met als gevolg dat items die in het Raschmodel passen, zeker in het Mokkenmodel zouden moeten passen. Molenaar (1982d) wijst er echter op dat alleen in het geval de spreiding van de totaalscores van de personenschaal klein is ten opzichte van de spreiding in itempopulariteiten, het Mokkenmodel de "Raschitems" niet accepteert.
Het is niet zo zeer interessant of Mokken nu wel of niet als "voorwas" (Jansen e.a., 1982) voor Rasch een zeeffunctie kan hebben. Veel nuttiger is het te constateren dat de discussie lijkt te suggeren dat beide modellen, wanneer ze passen op een bepaalde dataset, schalen opleveren met aantrekkelijke meeteigenschappen.
De beste onderzoekstrategie lijkt daarom te zijn een verzameling items allereerst met het sterkste model te analyseren om vervolgens met het zwakkere model na te gaan of er onder de resterende items nog een deelverzameling schuil gaat met eveneens gunstige meeteigenschappen (dubbele monotonie). Deze strategie werd bij het analyseren van de gegevens die in dit hoofdstuk worden gerapporteerd gevolgd.
Ook voor het Mokkenmodel is er een polytome versie beschikbaar (Molenaar,

1982b, 1986). Evenals in het Raschmodel voor meerdere antwoordcategorieën staat in het polytome Mokkenmodel het begrip itemstap centraal.

5.4 Procedure

In deze paragraaf wordt beschreven op welke wijze de in de vorige paragrafen besproken modellen feitelijk in het onderzoek werden gebruikt. Allereerst wordt beschreven op welke steekproef personen dit deel van het onderzoek betrekking had. Vervolgens wordt aangegeven welke de items waren die in dit deel van het onderzoek werden betrokken. Tenslotte wordt in deze paragraaf beschreven op welke wijze de analyses met behulp van de verschillende modellen werd uitgevoerd en op welke wijze de fit van de modellen met betrekking tot de data werd bestudeerd.

5.4.1 De personen

In de onderzoeken die in dit hoofdstuk worden beschreven, werden die studenten betrokken die aan alle vier voortgangstoetsen in het academisch jaar 1983-1984 hadden deelgenomen. Dit werd gedaan vanuit de overweging dat de besproken modellen het best tot hun recht komen, wanneer over verschillende niveaus van cognitief functioneren in het medische domein informatie beschikbaar is. Alleen in dat geval kunnen de te schatten itemkarakteristieke curves over het gehele bereik van de latente variabele worden geschat. In totaal hadden in het onderzochte academiejaar 612 studenten aan alle vier voortgangstoetsen deelgenomen.
Bij het presenteren van de resultaten zal blijken dat in sommige gevallen dit aantal kleiner is. Dit is steeds een gevolg van het feit dat bij het gebruik van het Raschmodel de waarde op de latente variabele van personen met een perfecte score (alle items goed beantwoord) en van personen met een nul score (geen enkel item goed beantwoord) niet kan worden geschat.

5.4.2 De items

In de voorafgaande hoofdstukken werd gewerkt met het aantal toetsvragen dat door de commissie van deskundigen met betrekking tot het eerstejaars programma werd aangewezen. In dit hoofdstuk wordt gewerkt met een toetslengte die voor alle 4 tijdstippen gelijk is. De reden hiervoor is een puur pragmatische: het voor het VGT-model ontwikkelde computerprogramma ABEST (Janssen, e.a., 1988) beschikte nog niet over de mogelijkheid om toetsen van een ongelijke lengte te verwerken. Als standaard toetslengte werd daarom gekozen voor de lengte van de kortste toets, die 51 items bevatte. De overtollige items uit de langere toetsen werden aselect uit het databestand verwijderd.

5.4.3 De statistische procedures

De twee itemgroepen Eerstejaarstoetsen en Algemene Kennistoetsen werden op ieder van de vier tijdstippen onderzocht met behulp van de hiervoor beschreven modellen.

5.4.3.1 Procedures voor het VGT-model

Voor het VGT-model werden de analyses uitgevoerd met behulp van het voor dit model ontwikkelde computerprogramma ABEST (Janssen, e.a., 1988). Dit programma schat voor ieder tijdstip de positie van de respondenten op de latente variabele. Daarnaast wordt per individu het kennisniveau bij aanvang van de studie A_i en de groei in kennis B_i (zie formule 5.4) geschat, alsmede de niet door het individuele regressiemodel verklaarde variantie. Details worden hier niet gegeven, daarvoor zij verwezen naar (Albers, e.a., 1987).
Beide itemgroepen werden over de vier tijdstippen apart geanalyseerd. Hierdoor werden de waarden van de genoemde variabelen zowel voor de Eerstejaarstoetsen als voor de Algemene Kennistoetsen verkregen. De fit van het model werd op twee manieren bestudeerd. In de eerste plaats werd de fit van het individuele regressiemodel bepaald aan de hand van de proportie verklaarde variantie gebaseerd op de latente variabele S_{ik} zelf en op basis van de schatting Y_{ik} van de S_{ik} (Albers, e.a.). Beide proporties geven aan in hoeverre het simpele lineaire regressiemodel (5.4) een adequate representatie geeft van de data. Ze geven met andere woorden per individu de fit van het model over de vier tijdstippen heen.
Ook per tijdstip werd nagegaan of het VGT-model een goede beschrijving geeft van de data. Daarbij werd een procedure gevolgd, analoog aan die welke wordt gebruikt om de fit van een individueel item in het Raschmodel te bepalen (Molenaar, 1983c). In deze procedure wordt de onder het model verwachte proportie goede antwoorden vergeleken met de geobserveerde proportie.
Met behulp van het VGT-model kan op basis van de schatting van de positie van de persoon op de latente variabele en de variantie van de itemmoeilijkheden τ^2, de kans op een goed antwoord op een willekeurige vraag worden uitgerekend met behulp van formule 5.2. Als het model klopt dan geldt voor iedere scoregroep in een nieuwe steekproef vragen precies deze kans op een goed antwoord. De interne modelfit kan dan bestaan uit het vergelijken van de onder het model verwachte kans met de geobserveerde proportie goede antwoorden.
De afwijking tussen verwachting en observatie kan dan vervolgens worden getoetst met behulp van

$$B^* = (B\text{-}NP_0)/(P_0(1\text{-}P_0))^{1/2} \qquad (5.8)$$

waar B het aantal successen (goede antwoorden) in de steekproef van omvang N is en P_0, de onder het VGT-model verwachte proportie successen (goede antwoorden).
Deze procedure werd hier toegepast. De fit van het VGT-model werd bepaald door het vergelijken van de onder het model verwachte proportie goede antwoorden, met de geobserveerde proportie goede antwoorden in een nieuwe aselect getrokken steekproef uit de residuële voortgangstoetsitems. Vervolgens werd de geobserveerde proportie goede antwoorden van de Eerstejaarstoetsen vergeleken met de onder het model verwachte proportie goede antwoorden. Deze vergelijking geeft informatie over de afwijking van de Eerstejaarsitems ten opzichte van een aselecte steekproef van items.

5.4.3.2 Procedures voor het Raschmodel

Voor de analyses met behulp van het Raschmodel werd de procedure gehanteerd die hieronder wordt beschreven.
De analyses werden in het dichotome geval voor elke itemgroep afzonderlijk uitgevoerd met behulp van het computerprogramma PML (Gustafsson, 1979). Begonnen werd met een Raschanalyse van alle 51 items. De modelcontrole bestond uit het berekenen van de voorwaardelijke likelihood ratio test voor twee scoregroepen, een hoog- en een laag scorende groep. De toets levert een globale controle op een van de centrale assumpties van het Raschmodel, nl. invariantie van de itemkarakteristieke curves (de engelse afkorting ICC, zal hier worden overgenomen). Deze toets kan voor verschillende doeleinden worden gebruikt(Gustafsson, 1980 en Molenaar, 1983c). Afhankelijk van de wijze waarop de scoregroepen worden gevormd is de toets gevoelig voor schendigen van de assumptie van eendimensionaliteit dan wel voor schendingen van de assumptie van parallelle ICC's. (Gustafsson, 1980). Worden de groepen gevormd op basis van de toetsprestatie, i.c. de ruwe score, dan is de toets gevoelig voor variaties in de hellingen van de ICC's. De laatste variant van de Andersontoets, wordt door Gustafsson aangeduid met A-ICCSL. Deze variant werd hier gebruikt.
Om bij een significante afwijking van het Raschmodel een beslissing te kunnen nemen over de vraag welke items nu voor verwijdering uit de dataset in aanmerking kwamen, werd naast de A-ICCSL toets, de U-toets van Molenaar (1983c) gebruikt. Deze toets is een maat voor de systematische afwijking van de succeskans van item j als een functie van de totaalscore r van de succeskans onder het Raschmodel. Deze toetsingsgrootheid, welke is ingebouwd in het programma PML, biedt de mogelijkheid om voor iedere scoregroep r, de afwijking van de model ICC te bestuderen en op basis daarvan te besluiten of een item eventueel voor verwijdering in aanmerking komt. Grote positieve waarden van U_j wijzen er op dat de helling van de ICC niet stijl genoeg is of dat het item te weinig of zelfs

negatief correleert met de resterende items in de schaal. Grote negatieve waarden wijzen er op dat de helling van de ICC te stijl is, dat het item "te goed" past (Molenaar, 1983c).
Het op de beschreven wijze gecombineerde gebruik van de twee model toetsen biedt de mogelijkheid om die items te selecteren die voldoen aan de assumpties van het Raschmodel.
Gebaseerd op de combinatie van de resultaten van de A- ICCSL toets en de U_j-toets, werden de twee itemgroepen per tijdstip in een aantal stappen geanalyseerd, een procedure die sterk lijkt op die welke door Sytsma (1983) wordt beschreven. Het uiteindelijke resultaat van de verschillende analysestappen, was voor ieder tijdstip twee Raschhomogene itemgroepen: één afkomstig uit de Eerstejaarsitems en één afkomstig uit de Algemene Kennisitems.
Deze twee itemgroepen werden vervolgens onderzocht door middel van de Martin-Löf toets, door Gustafsson (1980) aangeduid als ML-PCCSL-toets, waarbij de afkorting PCC staat voor Personal Characteristic Curve en het suffix SL (van slope) betrekking heeft op de helling van de PCC's. Deze toets kan ook weer op verschillende manieren gebuikt worden. Het basisprincipe van de toets is dat verschillende itemgroepen met elkaar vergeleken worden. Is de indeling van items in groepen gebaseerd op een inhoudelijk criterium dan is het een toets voor eendimensionaliteit van de twee itemgroepen. Is, zoals hier het geval was, de indeling gebaseerd op de itemscore, i.c. moeilijkheid van de items, dan is het een toets op de homogeniteit van de hellingen van de PCC's (Gustafsson, ibid). In dit geval is het een toets van de hypothese dat een model zoals voorgesteld door Lumsden (1977) nodig is om de observaties te verklaren. In dit model van Lumsden zijn de PCC's, waarin voor iedere persoon de kans op een goed antwoord wordt weergeven als een functie van de itemmoeilijheid, niet parallel. Het verschil in helling wordt in het model weergegeven door een zogenaamde gevoeligheidsparameter (sensitivity parameter, Lumsden, 1977). Wanneer zo'n gevoeligheidsparameter nodig is om de observaties te verklaren, dan is dat een aanwijzing voor meerdimensionaliteit van de itemgroepen.
Tenslotte werden op ieder tijdstip de twee Raschhomogene itemgroepen vergeleken met betrekking tot de informatie die de itemgroepen voor de verschillende niveaus van de latente variabele bevatten. Dit werd gedaan door de testinformatie uit beide itemgroepen te vergelijken, i.c. door de relatieve efficiëntie van beide te bepalen. Deze vergelijking werd gedaan op basis van de door Oosterloo (1984) ontwikkelde methode, gebruik makend van het computerprogramma's Rasch3 en Raspl3 (Steen, 1986).
Centraal in deze methode staat het begrip iteminformatie of testinformatie, die uit de iteminformatie geschat kan worden. Het begrip informatie is in de item response theorie het analogon van het begrip betrouwbaarheid in de klassieke testtheorie. Informatie kan worden gedefinieerd in termen van de onzekerheid met betrekking

tot het optreden van een bepaalde gebeurtenis. Hoe minder onzeker men over het optreden van een gebeurtenis is des te groter de informatie en omgekeerd. In de statistiek wordt de informatie gedefinieerd als de reciproke van de gekwadrateerde standaardfout van de schatter;

$$I = 1/\sigma^2.$$

Het informatie begrip wordt in de item response theorie gebruikt om de nauwkeurigheid waarmee de positie van de parameters s_i en t_i op latente variabele is gemeten te evalueren. De beschikbare informatie met betrekking tot een bepaald niveau s_i wordt gegeven door $I(s_i) = 1/\sigma^2$, of wel de reciproke van de gekwadrateerde standaardfout van s_i, gegeven s_i, welke kan worden geïnterpreteerd als de standaardeviatie van de verdeling van s_i voor alle personen met dezelfde s_i. Lord (1980, hoofdstuk 5) bewijst dat de test informatie, $I(s_i)$,

$$I(s_i) = \Sigma \frac{(p_j(s_i)')^2}{p_j(s_i)q_j(s_i)} \qquad (5.9)$$

waar $p_j(s_i)'$ de afgeleide of de helling van de ICC van item j is in het punt s_i, $p_j(s_i)$ de kans op een goed antwoord en $q_j(s_i)$ de kans op een fout antwoord op het kennisniveau s_i. Uit 5.9 blijkt dat de testinformatie functie ontstaat door de iteminformatiefuncties te sommeren.
Item- en testinformatie zijn schaal afhankelijke begrippen (Lord, 1980). Voor het vergelijken van twee toetsen kan daarom beter gebruik worden gemaakt van een schaalonafhankelijke maat. Zo'n maat is de relatieve efficiëntie (RE). Voor twee toetsen is indien de parameters s en t bekend zijn, de relatieve efficiëntie van toets y ten opzichte van toets x de ratio van de informatiefuncties van de twee toetsen;

$$RE(Y,X) = I(s,y)/I(s,x) \qquad (5.10)$$

De door Oosterloo (1984) ontwikkelde methode voor het schatten van toetsinformatiefuncties en relatieve efficiëntie functies werd hier gebruikt om de twee itemgroepen te kunnen vergelijken. Deze methode is geprogrammeerd in de computerprogramma's Rasch3 en Raspl3 (Steen, 1986). Deze programma's werden gebruikt om de schattingen van de informatiefuncties te verkrijgen. Om rekentechnische redenen bleek het onmogelijk om alle raschhomogene items simultaan te analyseren. Daarom werd er gekozen om voor alle vier tijdstippen steeds drie gebieden van de latente variabele te onderzoeken; het lage gebied, het midden gebied en het hoge gebied. De itemgroepen van ieder vijf items die deze

gebieden het best representeerden werden in deze procedure met elkaar vergeleken. Bijvoorbeeld, voor alle tijdstippen werden de vijf eenvoudigste items uit de Raschomogene Eerstejaarsitems vergeleken met de vijf eenvoudigste items uit de eveneeens Raschhomogene Algemene Kennisitems.
Na de dichotome Raschanalyse, waarbij de scoring was 1 = goed, 0 = niet goed, dus fout of vraagteken, werden de Raschhomogene items vervolgens als trichotome items gescoord. Overwegingen hierbij werden in paragraaf 5.3.3 reeds besproken. De trichotome scoring was alsvolgt: een goed antwoord werd gescoord als 2, een fout antwoord als 1 en een vraagteken als 0. Op deze wijze werden de niet goede antwoorden bij de dichotome scoring op gesplitst in twee categorieën.
Deze trichotome items werden geanalyseerd met behulp van het computerprogramma MICROSCALE (Wright en Linacre, 1985). Globale goodness-of-fit maten, zoals de hiervoor besproken toets van Anderson, worden daarbij niet berekend. Wel worden voor de personen en de items fitmaten geschat. De trichotome scoring werd als aanvulling op de dichotome scoring gebruikt, met de bedoeling om een antwoord te kunnen geven op de vraag in welk gebied van de latente variabele, de meeste fouten worden gemaakt en bovendien om vragen op te sporen, met een ongewenst foutenpatroon (De Gruijter, 1987). In het algemeen zijn dit items die voor een groot gebied van de latente variabele een relatief hoge kans op een fout antwoord te zien geven of items die voor een bepaald gebied van de latente variabele een hogere kans op een fout antwoord dan op een goed antwoord hebben.

5.4.3.3 Procedures voor het Mokkenmodel

Het resultaat van de in de vorige paragraaf besproken procedure was dat op ieder tijdstip de itemgroepen Eerstejaarsitems en Algemene Kennisitems, gesplitst kon worden in twee itemgroepen: een Raschhomogene groep items en een groep items, die niet in het Raschmodel pasten.
Van deze residuële items werden de meeteigenschappen verder onderzocht met behulp van het Mokkenmodel. Daarbij werd gebruik gemaakt van het computerprogramma MSP (Debets en Brouwer, 1986).
De overgebleven Eerstejaars- en Algemene Kennisitems werden op ieder tijdstip separaat geanalyseerd. De door het programma gevonden Mokkenschalen werden geëvalueerd met behulp van de door Molenaar (1983a) voorgestelde methode. Deze methode komt neer op het onderzoeken van de monotonie-eigenschappen van de items door middel van inspectie van de proporties goed per scoregroep binnen een item en over items. Het monotonieonderzoek werd vervolgens voortgezet met een inspectie van de P_{++} (= matrix met de fracties personen die beide items goed hebben) en P_{--} (= de matrix met de fracties personen die beide items fout hebben) matrices. Geconstateerde afwijkingen in de monotonie werden getoetst met behulp

van de eveneens door Molenaar voorgestelde toetsen: de Fisher exact toets voor onhankelijke groepen en de McNemar toets voor dezelfde - en dus afhankelijke - groepen.
Hierna werden de residuële items, op dezelfde wijze als de Raschitems, trichotoom gescoord. Deze trichotoom gescoorde items werden vervolgens op dezelfde manier met behulp van MSP geanalyseerd. Ook in dit geval werden de trichotome Mokkenschalen op monotonie onderzocht.
Aldus ontstonden er op alle vier tijdstippen binnen de Eerstejaarsitems en de Algemene Kennisitems drie itemgroepen: Raschitems, Mokkenitems en niet-schaalbare items. Als laatste stap in de analyses werden de correlaties tussen deze drie itemgroepen berekend.
De resultaten van de hierboven beschreven analyse procedures worden in het volgende hoofdstuk beschreven.

5.5 Samenvatting

In dit hoofdstuk werden drie modellen beschreven die beschouwd kunnen worden als geformaliseerde hulptheorie voor de variabele die centraal staat in dit onderzoek: de medische kennis.
De drie modellen zijn: het VGT-model, het Raschmodel en het Mokkenmodel. De drie modellen zijn te beschouwen als minitheorieën, die de interactie tussen personen en items in het beantwoordingsproces trachten te verklaren (paragraaf 5.2). Om die verklaring te kunnen geven gaan de modellen uit van een sterk vereenvoudigde situatie. Het antwoordgedrag van respondenten op kennisitems wordt bepaald door twee factoren: de kennisstructuur van de persoon en de (cognitieve) moeilijkheid van het item. De relatieve sterkte van de één ten opzichte van de ander bepaalt de uitkomst van het antwoordproces. Deze assumpties in de modellen leiden tot voorspellingen die al dan niet in overeenstemming zijn met de feitelijke gegeven antwoorden. Dit aspect van het model is toetsbaar. Niet toetsbaar is de andere kant van het model, de correspondentie met het te meten begrip of construct; de medische kennis. De correspondentie tussen het begrip-zoals-bedoeld en de gebruikte modellen werd beredeneerd en aannemelijk gemaakt (paragraaf 5.2).
Om de kenmerken van de drie psychometrische modellen duidelijk te maken werd als vertrekpunt de Guttmanschaal gekozen. Dit deterministische model voorspelt welke items goed beantwoord worden, gegeven de totaalscore op een itemgroep. De drie modellen, die in dit hoofdstuk besproken werden zijn te beschouwen als probabilistische varianten van het deterministische Guttmanmodel (paragraaf 5.3.1).
In het VGT-model is de kans op een goed antwoord net als in het Raschmodel afhankelijk van de persoonsparameter en de moeilijkheid van het item. In het VGT-model werden beide echter opgevat als stochastische variabelen en niet als parameters. Het VGT-model kan daardoor worden gebruikt in situaties waar met

behulp van grote aantallen items op verschillende tijdstippen wordt gemeten. Over de tijdstippen heen wordt de groei in de kennisvariabele gemeten met behulp van een eenvoudig lineair regressiemodel. Deze groei in kennis wordt in het model voor een ieder individueel geschat (paragraaf 5.3.2).
In het Raschmodel kan de kans op een goed antwoord gegeven de totaalscore op ieder item, worden geschat. Bovendien kan dit model gegeneraliseerd worden naar items met meer dan twee (goed, niet goed) categorieën. Hierdoor bleek het mogelijk om voor elk item de kans op een goed, een fout en een weet-niet antwoord te schatten (paragraaf 5.3.3).
Het Mokkenmodel tenslotte definieert niet expliciet de kansrelatie tussen het item en de gemeten variabele. In dit model zijn alleen de rangorde van de itemmoeilijkheden (itempopulariteiten) en de rangorde van de personen van belang. Items die niet aan de in, het model gestelde monotonie-eisen voldoen, kunnen worden opgespoord (paragraaf 5.3.4).
Tenslotte werd in paragraaf 5.4 beschreven op welke wijze er met deze drie modellen werd gewerkt. Het aantal onderzochte personen werd beperkt tot de studenten die aan alle vier toetsen hadden deelgenomen. Bovendien werd de toetslengte voor ieder tijdstip gelijk gesteld, namelijk 51 items.

6. RESULTATEN VAN HET TOETSEN VAN DE MODELLEN

Inhoud

6.1 Inleiding
6.2 Toetsing van het VGT-model (hypothese 13)
6.3 Samenvatting
6.4 Toetsing van het Raschmodel (hypothesen 14a, 14b en 14c)
6.5 Samenvatting
6.6 Toetsing van het Mokkenmodel (hypothesen 15a en 15b)
6.7 Samenvatting
6.8 Correlaties tussen Raschitems, Mokkenitems en niet-schaalbare items
6.9 Samenvatting
6.10 Constructrepresentatie van Medische Kennis
6.10.1 Inleiding
6.10.2 Grafische representatie van de variabele
6.10.3 De relatie tussen inhoud en moeilijkheid van de toetsitems
6.11 Samenvatting constructrepresentatie
6.12 Samenvatting en conclusies

6.1 Inleiding

In dit hoofdstuk worden de resultaten van de uitgevoerde analyses besproken. Begonnen wordt met de resultaten van het VGT-model, vervolgens komen de resultaten met behulp van het Raschmodel aan de orde en ten slotte de resultaten die met behulp van het Mokkenmodel werden verkregen.

6.2 Toetsing van het VGT-model (hypothese 13)

Eerder werd beschreven dat het VGT-model een schatting oplevert van de positie van de studenten op de latente variabele. Deze positie wordt net als in het Raschmodel geschat aan de hand van de totaalscore van de respondent op de toets. Een overzicht van de totaalscores en hun schaalwaarde geschat op basis van de gegevens uit de Eerstejaarstoetsen en de Algemene Kennistoetsen wordt gegeven in tabel 6.2.

TABEL 6.2
Overzicht van scoregroepen, schaalwaarden (si) en studentenaantal per scoregroep (totaal 612)

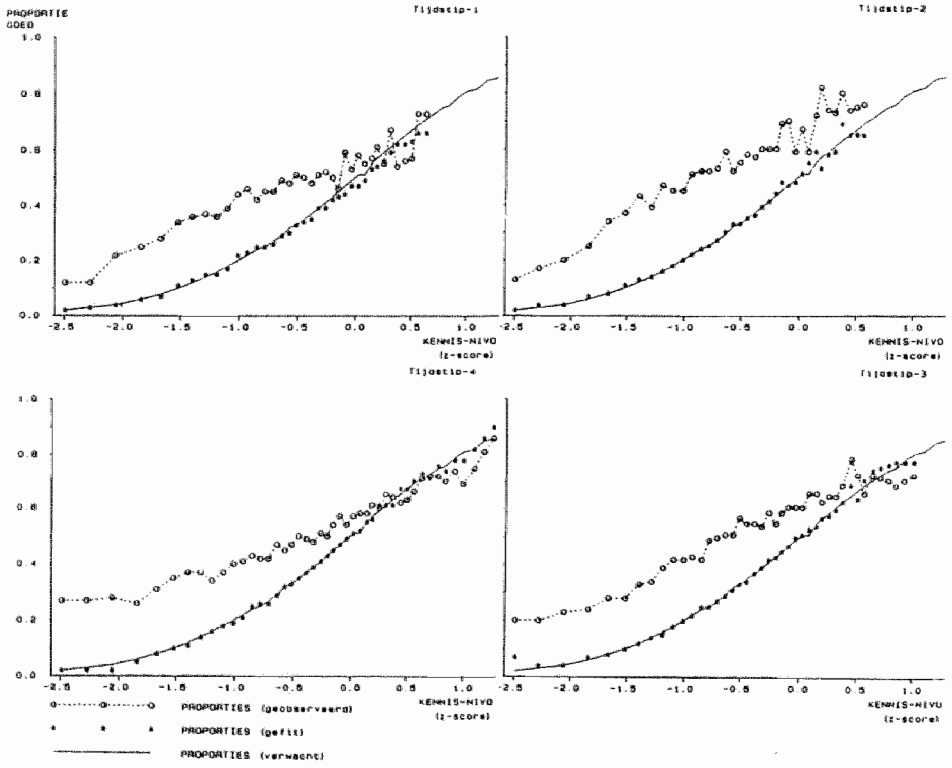
				aantal studenten per tijdstip							
				1		2		3		4	
score	schaal waarde Ej	schaal waarde Ak	succes kans	Ej	Ak	Ej	Ak	Ej	Ak	Ej	Ak
0		-2.49	.02	0	49	0	11	0	3	0	0
1	-2.34	-2.28	.03	4	37	0	34	0	8	0	1
2	-2.12	-2.06	.04	8	36	0	32	0	18	0	9
3	1.90	-1.84	.06	8	34	3	43	0	13	1	14
4	-1.72	-1.67	.08	8	33	2	34	0	26	0	25
5	-1.57	-1.52	.10	19	30	4	24	1	27	0	28
6	-1.44	-1.40	.12	16	36	13	28	0	30	0	18
7	-1.33	-1.29	.14	11	29	15	32	7	17	0	14
8	-1.22	-1.19	.16	23	30	11	31	9	18	2	15
9	-1.13	-1.10	.18	18	29	6	32	6	22	5	14
10	-1.04	-1.01	.20	7	18	18	26	10	17	3	17
11	-.96	-.93	.22	14	22	14	29	7	43	7	23
12	-.88	-.85	.24	7	15	21	30	17	18	13	22

TABEL 6.2 (vervolg)

13	-.80	-.78	.25	17	20	14	23	14	29	8	19
14	-.73	-.71	.27	14	19	14	20	18	30	11	20
15	-.66	-.64	.29	23	14	17	15	18	18	12	22
16	-.59	-.57	.32	20	11	14	19	18	22	19	20
17	-.52	-.51	.33	22	14	10	25	18	29	17	19
18	-.46	-.45	.35	37	14	14	22	16	15	21	17
19	-.39	-.38	.37	29	21	12	13	10	19	36	20
20	-.33	-.32	.39	23	13	14	14	20	14	23	26
21	-.27	-.26	.41	30	11	20	18	17	21	26	18
22	-.21	-.20	.43	38	10	22	12	27	14	28	22
23	-.15	-.15	.45	24	7	24	7	26	22	43	14
24	-.09	-.09	.47	18	7	30	8	25	20	39	20
25	-.03	-.03	.49	25	15	33	8	21	14	35	28
26	.03	.03	.51	27	8	27	5	29	13	38	13
27	.09	.09	.53	28	2	39	6	36	16	25	12
28	.15	.15	.55	26	4	36	1	28	13	28	23
29	.21	.20	.57	19	8	29	1	33	7	26	15
30	.27	.26	.59	7	3	22	3	32	7	26	8
31	.33	.32	.61	14	5	24	1	28	12	19	20
32	.39	.38	.63	11	2	19	1	24	7	16	7
33	.46	.45	.65	4	0	19	3	18	2	14	10
34	.52	.51	.67	5	2	11	0	21	2	19	10
35	.59	.57	.69	2	2	9	1	15	2	10	6
36	.66	.64	.71	2	2	10		15	1	9	7
37	.73	.71	.73	1		6		12	0	14	5
38	.80	.78	.75	2		5		7	0	7	3
39	.88	.85	.76	1		2		5	2	6	4
40	.96	.93	.79			2		4	0	4	2
41	1.04	1.01	.81			4		1	1	1	1
42	1.12	1.10	.82			2				1	0
43	1.22	1.20	.85			1					0
44		1.29	.86								1

Uit bovenstaande tabel blijkt dat het voor de schatting van de schaalwaarde s_i niet veel uitmaakt of deze nu gebaseerd is op de resultaten van de Eerstejaarstoetsen of de Algemene Kennistoetsen. In beide gevallen leidden de schattingen tot nagenoeg dezelfde schaalwaarden. Het VGT-model biedt de mogelijkheid om met behulp van formule 6.2, voor iedere positie op de latente

variabele de kans op een goed antwoord op een willekeurige vraag te schatten. Naast de schaalwaarde is daarvoor nodig een schatting van de variantie in itemmoeilijkheden τ^2. Voor de Eerstejaarstoetsen was deze schatting van τ^2 .48 met een standaardfout van .07. Voor de Algemene Kennistoetsen was deze schatting .39 met een standaardfout van .06. In de derde kolom van de bovenstaande tabel wordt voor ieder niveau van de latente variabele S, de modelkans gegeven.
Deze modelkans werd gebruikt om de fit van het model te bepalen, op de wijze zoals werd beschreven in paragraaf 5.4.3. De resultaten van de fit en de afwijking tussen de modelkans en de geobserveerde proporties goede antwoorden, is grafisch weergegeven in figuur 6.2.



Figuur 6.2 Modelfit voor nieuwe steekproef Algemene Kennisitems en afwijking van verwachte proportie goede antwoorden op Eerstejaarsitems.

Twee conclusies kunnen op grond van deze figuur worden getrokken: In de eerste plaats, dat het model op alle tijdstippen in staat bleek om de geobserveerde proportie goede antwoorden in een nieuwe steekproef van vergelijkbare items goed te voorspellen. In de tweede plaats bleek dat de geobserveerde proportie goede antwoorden op de Eerstejaarsitems op nagenoeg alle niveaus van de latente variabele afweek van de door het model voorspelde kans. Dit betekent dat de kans op een goed antwoord op een willekeurige vraag uit de eerstejaars itemgroep, voor de meeste niveaus van de variabele significant, hoger was dan de door het model voorspelde kans.
De afwijking tussen geobserveerde en verwachte proporties werd getoetst met behulp van formule 5.8. Voor slechts een paar kennisniveaus leverde dit een significante afwijking op. Met andere woorden de resultaten op de Algemene Kennistoets bleken voor de meeste niveaus representatief voor de prestaties op de rest van de voortgangstoetsen. Een geheel andere conclusie moet worden getrokken bij het vergelijken van de modelkans en de geobserveerde proportie goede antwoorden op de Eerstejaarstoetsen. Voor 3/4 van de scoreniveaus bleek het aantal geobserveerde antwoorden op de Eerstejaarstoetsen significant groter dan door het model werd voorspeld. Voor de hoogste score niveaus was de afwijking niet meer significant of ging het verschil de andere kant uit, met andere woorden was het voorspelde aantal goede antwoorden groter dan het geobserveerde aantal.
Eerder werd in paragraaf 5.2 de redenering gevolgd dat de ontwikkeling van de cognitieve structuren van studenten slechts gemiddeld genomen gecorreleerd is aan het aantal jaren dat men studeert. Individueel, zo werd verwacht, zijn er ook binnen jaargroepen met eenzelfde "curriculumleeftijd" grote verschillen in cognitieve ontwikkeling. Informatie met betrekking tot de geldigheid van deze redenering wordt gegeven in tabel 6.3.
Deze tabel bevestigt in de eerste plaats, dat de Eerstejaarstoetsen voor nagenoeg alle scoreniveaus gemakkelijker waren. Verder bleek er in alle jaargroepen een groot verschil te bestaan tussen het laagste- en het hoogste scoreniveau. Deze conclusie was zowel van toepassing op de schattingen die waren gebaseerd op de Eerstejaarsitems als op die van de Algemene Kennisitems. In het eerstejaar bijvoorbeeld varieerden de schaalwaarden, geschat met behulp van de Eerstejaarsitems, van - 2.34 tot .15. Dit komt neer op een variatie in gemiddelde antwoordkans van .03 tot .55. Dit betekent dat er in deze steekproef eerstejaars studenten waren die een cognitief niveau hadden dat vergelijkbaar was met dat van een gemiddelde zesdejaars student. Ook het omgekeerde bleek voor te komen; een zesdejaars student met een vrij laag cognitief niveau van -.96, hetgeen neerkomt op een gemiddelde kans op een goed antwoord van .22.

TABEL 6.3
Verdeling van de cognitieve niveau's over de zes jaargroepen

tijd-stip	jaar-groep	m	sd	min	max	med	Q1	Q3	Trm
(Eerstejaars toets)									
	1	-1.29	0.47	-2.34	0.15	-1.22	-1.57	-1.04	-1.29
	2	-0.42	0.31	-1.13	0.66	-0.46	-0.66	-0.21	-0.43
	3	-0.29	0.32	-1.04	0.59	-0.33	-0.52	-0.03	-0.30
1	4	-0.11	0.30	-1.22	0.73	-0.12	-0.27	0.09	-0.10
	5	0.02	0.27	-0.52	0.80	0.03	-0.21	0.21	0.02
	6	0.14	0.32	-0.96	0.88	0.15	0.00	0.36	0.15
	1	-0.98	0.36	-1.90	0.21	-0.96	-1.22	-0.73	-0.98
	2	-0.21	0.36	-1.44	0.73	-0.21	-0.46	0.03	-0.20
	3	0.04	0.27	-0.80	0.73	0.03	-0.15	0.15	0.03
2	4	0.13	0.30	-0.59	1.22	0.09	-0.09	0.33	0.12
	5	0.24	0.25	-0.46	0.80	0.21	0.09	0.39	0.25
	6	0.45	0.35	-0.27	1.13	0.46	0.18	0.69	0.44
	1	-0.71	0.36	-1.57	0.27	-0.73	-0.96	-0.52	-0.73
	2	-0.08	0.36	-0.88	0.96	-0.09	-0.33	0.21	-0.09
	3	0.07	0.31	-0.73	0.86	0.09	-0.15	0.21	0.07
3	4	0.21	0.32	-0.88	0.80	0.21	0.00	0.46	0.21
	5	0.35	0.29	-0.33	0.96	0.33	0.21	0.59	0.36
	6	0.39	0.32	-0.46	1.04	0.43	0.18	0.59	0.41
	1	-0.46	0.33	-1.22	0.27	-0.46	-0.66	-0.27	-0.44
	2	-0.19	0.39	-1.90	0.73	-0.15	-0.39	0.03	-0.17
	3	-0.06	0.25	-0.52	0.59	-0.09	-0.27	0.15	-0.07
4	4	0.09	0.32	-0.73	0.88	0.09	-0.15	0.33	0.09
	5	0.39	0.27	-0.09	0.96	0.39	0.15	0.52	0.38
	6	0.47	0.36	-0.39	1.33	0.46	0.21	0.73	0.48

$\tau^2 = 0.48\ (.07)$; $\tau^2_1 = 0.68$; $\tau^2_2 = 0.37$; $\tau^2_3 = 0.41$; $\tau^2_4 = 0.42$

Aantal respondenten:
jaar 1: 152, jaar 2: 139, jaar 3: 103, jaar 4:100, jaar 5: 62, jaar 6: 56.

Vervolg TABEL 6.3

tijd-stip	jaar-groep	m	sd	min	max	med	Q1	Q3	Trm
(Algemene kennistoets)									
	1	-2.07	0.47	-2.49	-0.03	-2.27	-2.49	-1.84	-2.16
	2	-1.41	0.42	-2.49	-0.32	-1.40	-1.67	-1.09	-1.42
	3	-1.09	0.37	-2.06	-0.09	-1.09	-1.29	-0.85	-1.11
1	4	-0.71	0.44	-2.49	0.32	-0.71	-0.93	-0.45	-0.69
	5	-0.34	0.36	-1.40	0.32	-0.32	-0.57	-0.03	-0.31
	6	-0.04	0.38	-1.29	0.63	-0.06	-0.32	0.20	-0.04
	1	-1.88	0.43	-2.49	0.09	-1.84	-2.28	-1.67	-1.93
	2	-1.23	0.43	-2.28	0.15	-1.29	-1.52	-0.93	-1.23
	3	-1.00	0.32	-1.67	-0.26	-1.00	-1.19	-0.85	-1.02
2	4	-0.66	0.33	-1.40	0.26	-0.64	-0.89	-0.41	-0.65
	5	-0.46	0.31	-1.10	0.45	-0.45	-0.71	-0.20	-0.47
	6	-0.17	0.31	-0.93	0.57	-0.20	-0.35	0.03	-0.18
	1	-1.56	0.39	-2.49	-0.71	-1.52	-1.84	-1.29	-1.56
	2	-0.85	0.36	-2.06	0.03	-0.93	-1.10	-0.64	-0.86
	3	-0.61	0.32	-1.29	0.15	-0.64	-0.85	-0.38	-0.62
3	4	-0.33	0.32	-0.93	0.38	-0.32	-0.57	-0.09	-0.33
	5	-0.12	0.30	-0.93	0.64	-0.12	-0.26	0.09	-0.11
	6	0.19	0.29	-0.38	1.01	0.20	0.03	0.32	0.18
	1	-1.40	0.37	-2.28	-0.32	-1.40	-1.67	-1.19	-1.42
	2	-0.67	0.34	-1.52	0.57	-0.71	-0.93	-0.45	-0.69
	3	-0.33	0.33	-1.52	0.57	-0.32	-0.57	-0.09	-0.33
4	4	-0.07	0.35	-0.85	1.01	-0.03	-0.32	0.14	-0.07
	5	0.15	0.29	-0.51	0.78	0.15	-0.03	0.32	0.15
	6	0.39	0.34	-0.32	1.29	0.38	0.17	0.64	0.40

$\tau^2 = 0.39$ (.06); $\tau^2_1 = 0.37$; $\tau^2_2 = 0.23$; $\tau^2_3 = 0.60$; $\tau^2_4 = 0.35$

Aantal respondenten:
jaar 1: 152, jaar 2: 139, jaar 3: 103, jaar 4:100, jaar 5: 62, jaar 6: 56.

Naast een schatting van de positie van de studenten op de latente variabele, die door de gebruikte itemgroepen wordt gedefinieerd, kan met behulp van het VGT-model een schatting worden verkregen van de groei in individuele kennis (zie daarvoor formule 5.4). De door het individuele regressiemodel verklaarde variantie in termen van $_i$ wordt vermeld in tabel 6.4.

TABEL 6.4
Overzicht van de door de lineaire regressie verklaarde variantie

	jaar-groep	N	Med	Q1	Q3	Min	Max
	1	152	.78	.55	.92	.00	.99
	2	139	.36	.09	.67	.00	.99
	3	103	.32	.06	.54	.00	.98
Ej	4	100	.34	.08	.54	.00	.97
	5	62	.48	.15	.81	.00	1.00
	6	56	.36	.09	.60	.00	.98
	1	152	.67	.53	.86	.00	.99
	2	139	.79	.58	.88	.00	.99
	3	103	.80	.56	.93	.00	1.00
Ak	4	100	.73	.50	.90	.00	1.00
	5	62	.59	.30	.85	.00	1.00
	6	56	.56	.32	.78	.00	1.00

Ej = Eerstejaarstoets; Ak = Algemene Kennistoets.

De conclusie die aan de gegevens uit de bovenstaande tabel kan worden verbonden is dat de eerstejaars gemiddeld genomen lineair groeien, ongeacht het soort items dat wordt gebruikt. Studenten uit de vijf overige jaargroepen bleken vooral op de Algemene Kennisitems lineair te groeien. Verder vallen ook hier de grote individuele verschillen op. Binnen alle jaargroepen varieerde de proportie verklaarde variantie van .00 tot bijna 1.00. Dit duidt op grote verschillen in het individuele groeipatroon.
Interessant met het oog op de vraagstelling van dit onderzoek is de groeisnelheid zelf, die in formule 5.4 wordt gerepresenteerd door de variabele B_i. De voornaamste kenmerken van de verdeling van deze parameter en de verschillen in groei in kennis op de twee itemgroepen worden gegeven in tabel 6.5.

TABEL 6.5
Gemiddelde groei in kennis op Eerstejaars toetsen en Algemene Kennistoetsen

Eerstejaarstoets							
Jaargroep	N	M	sd	Min	Max	Q1	Q3
1	152	1.12	.72	-1.01	3.14	.69	1.52
2	139	.32	.47	-0.90	1.40	.00	.68
3	103	.28	.39	-0.84	1.42	.05	.50
4	100	.27	.43	-1.00	2.49	.00	.50
5	62	.49	.34	-0.26	1.38	.19	.73
6	56	.37	.44	-0.60	1.42	.10	.59

Algemene kennistoets									
Jaargroep	N	M	sd	Min	Max	Q1	Q3	t	p*
1	152	.92	.71	2.22	2.72	.60	1.37	2.50	.009
2	139	1.04	.51	-0.52	2.28	.74	1.44	-13.33	<.0005
3	103	1.07	.51	-0.11	2.36	.72	1.40	-13.16	<.0005
4	100	.90	.51	-0.44	3.31	.58	1.23	- 9.00	<.0005
5	62	.75	.48	-0.39	1.87	.40	1.03	- 3.25	.0009
6	56	.67	.46	-0.30	2.03	.40	.95	- 3.75	<.0005

* overschrijdingskans berekend d.m.v. interpolatie

In alle jaargroepen was de groeisnelheid, gemeten met Eerstejaarsitems en Algemene Kennisitems, verschillend. Echter, alléén de eerstejaars groeiden het sterkst op de Eerstejaarsitems. Alle andere jaargroepen vertoonden de sterkste groei op de Algemene Kennisitems. Bovendien groeiden deze jaargroepen bijna niet op de Eerstejaarsitems. Verder valt op dat er grote individuele verschillen waren in de mate van groei in kennis. Dit was voor beide itemgroepen het geval.

6.3 Samenvatting

Het VGT-model bleek de gemiddelde proportie goede antwoorden op een verse steekproef van vragen uit de resterende Algemene Kennisitems goed te kunnen voorspellen. De geobserveerde proportie goede antwoorden op de Eerstejaarsitems bleek voor de meeste scoreniveaus significant te verschillen van de door het

verschil of ze getoetst werden met behulp van de ene of met de andere itemgroep. Met betrekking tot hypothese 13 betekenen deze resultaten dat de geformuleerde voorspelling voor de Algemene Kennistoetsen volledig ondersteund wordt. Gezien het nagenoeg identieke verloop van de kansen op een goed antwoord op de Eerstejaarsitems, valt te verwachten dat het model ook voor deze items de kansen voor de verschillende kennisniveaus in de populatie van Eerstejaarsitems goed voorspelt. Dit kon echter niet getoetst worden omdat er geen mogelijkheid bestond om nog eens een steekproef van Eerstejaarsitems te trekken.
Verder bleken er binnen de jaargroepen grote individuele verschillen te bestaan met betrekking tot het cognitieve niveau. Een enkele eerstejaars bleek te functioneren op een niveau dat vergelijkbaar was met een gemiddelde zesdejaars student. De groei in kennis was voor eerstejaars lineair op beide itemgroepen. Voor ouderejaars was de groei in kennis alleen lineair op de Algemene Kennisitems.
De gemiddelde groei bleek voor de eerstejaars het sterkst op de Eerstejaarsitems en voor ouderejaars het sterkst op de Algemene Kennisitems te zijn.

6.4 Toetsing van het Raschmodel (hypothesen 14a, 14b en 14c)

Met behulp van de procedure die werd beschreven in paragraaf 5.4.3.2 bleken er op ieder tijdstip binnen beide itemgroepen twee Raschhomogene itemgroepen te bestaan. De modelfit en een aantal andere eigenschappen van de itemgroepen worden vermeld in tabel 6.6.

TABEL 6.6
Raschhomogene subschalen in Eerstejaarstoetsen en AlgemeneKennistoetsen

	Eerstejaarstoetsen				Algemene Kennistoetsen			
tijdstip	1	2	3	4	1	2	3	4
Aantal items	31	24	23	22	20	21	23	27
KR20	.80	.76	.71	.72	.85	.76	.83	.84
ISS	.82	.78	.72	.65	.72	.61	.78	.81
Andersen toets =	39.31	21.43	28.95	28.47	19.04	32.45	27.38	18.81
df =	30	23	22	21	19	20	22	26
p =	.12	.55	.15	.13	.45	.04	.20	.84

KR 20 = betrouwbaarheids index(Kuder-Richardson 20)
ISS = index of subject separation

Uit tabel 6.6 blijkt dat alle geconstrueerde itemgroepen, gemeten aan het criterium van de A-ICCSL toets (zie paragraaf 5.4.3), bleken te voldoen aan de assumpties van het Raschmodel. Hoewel het in alle gevallen relatief kleine itemgroepen betrof, bleken de betrouwbaarheden en de Index of Subject Separation (ISS), de item response pendant van de interne consistentie van de Klassieke Test Theorie (KTT), (Gustafsson, 1977) zeer acceptabele waarden te hebben.
Door middel van de eveneens in paragraaf 5.4.3 beschreven procedure werden de Raschhomogene itemgroepen per tijdstip vergeleken met als criterium de Martin-Löf toets voor parallelle PCC's. De resultaten zijn te vinden in tabel 6.7.

TABEL 6.7
Martin-Löf toets (ML-PCCSL) voor parallelle PCC's op Eerstejaars items (Ej) en Algemene Kennisitems (Ak)

tijdstip	1	2	3	4
aantal items (Ej)#	30	24	23	22
aantal items (Ak)	20	21	23	27
correlatie	.70	.71	.68	.70
ML X^2 *	551	278	542	485
df	599	503	528	593
p	.92	1.0	.33	.99
$M(t_j)$ Ej	-.49	-.53	-.59	-.40
$M(t_j)$ Ak	.72	.49	.59	.33
$M(t_j)$ totaal	0	0	0	0
betrouwbaarheid (KR 20)	.89	.86	.86	.87
effectief aantal personen	611	610	612	612

* Martin-Löf toets (ML-PCCSL) voor itemgroepen.
# In verband met de uitvoerbaarheid van de toets moest 1 item worden verwijderd: het maximale aantal items in het programma PML bedraagt 30.
$M(t_j)$ = gemiddelde itemmoeilijkheid

Uit de tabel blijkt dat de Raschitems in beide itemgroepen op ieder tijdstip homogeen, in de betekenis van parallelle PCC's, genoemd mogen worden. Er was met andere woorden weinig reden om te zoeken naar alternatieve en ingewikkelder modellen zoals het model dat werd voorgesteld door Lumsden (1977). De conclusie dat per tijdstip de Eerstejaarsitems en de Algemene Kennisitems homogeen waren,

wordt nog eens onderstreept door de hoge correlaties tussen de scores op beide itemgroepen. Uit de hoge betrouwbaarheden van de samengevoegde itemgroepen blijkt dat ook de gemiddelde itemcorrelaties hoog waren. Verder blijkt uit de tabel dat de gemiddelde itemmoeilijkheid $M(t_j)$ op de samengevoegde schaal in alle vier gevallen voor de Eerstejaarsitems lager was dan voor de Algemene Kennisitems. Dat betekent dat de Eerstejaarsitems op de samengevoegde schaal gemiddeld genomen links van het centrum van de schaal verwacht mogen worden. Het is goed om dit in gedachten te houden wanneer verderop de Raschschalen van elke itemgroep apart worden gepresenteerd. Deze schalen zijn met andere woorden separaat gecalibreerd en laten om die reden geen rechtstreekse onderlinge vergelijking toe. De geschatte itemmoeilijkheden uitgedrukt in eenheden van de logitschaal, worden vermeld in tabel 6.8.

TABEL 6.8
Schattingen van de itemmoeilijkheden en itemstappen van de Raschitems

		Tijdstip 1			
item nr	item-moeilijk-heid	code* voor t_{j2}-t_{j1}	item nr	item-moeilijk-heid	code* voor t_{j2}-t_{j1}
Eerstejaarsitems			Algemene kennisitems		
101	-3.38	1	45	-2.17	1
67	-2.32	1	51	-1.89	1
102	-2.11	1	60	-1.74	1
108	-2.05	1	22	-1.53	1
114	-1.94	1	30	-1.47	1
76	-1.56	1	47	-0.63	1
78	-1.48	1	33	-0.35	1
70	-1.33	1	14	-0.31	1
82	-0.76	1	25	-0.21	1
71	-0.47	-	58	-0.21	1
68	-0.20	1	36	0.07	1
111	-0.11	1	23	0.13	1
89	-0.03	1	19	0.21	1
110	-0.01	1	53	0.23	1
91	0.14	1	49	0.24	1
72	0.18	1	18	0.98	1
109	0.30	1	13	1.37	1

TABEL 6.8 (vervolg)

73	0.42	1	12	2.06	0
88	0.46	1	34	2.06	1
99	0.48	1	32	3.09	0
95	0.54	1			
87	0.66	0			
112	0.71	1			
104	0.75	-			
103	1.10	-			
66	1.27	1			
65	1.28	1			
81	1.40	1			
85	2.52	-			
98	2.68	1			
97	2.86	-			

Tijdstip 2

item nr	item-moeilijkheid	code* voor t_{j2}-t_{j1}	item nr	item-moeilijkheid	code* voor t_{j2}-t_{j1}
Eerstejaarsitems			Algemene kennisitems		
90	-3.17	1	31	-1.82	1
89	-2.73	1	46	-1.44	1
87	-1.09	1	14	-1.04	1
92	-0.97	1	44	-0.89	1
77	-0.76	1	30	-0.64	1
83	-0.48	1	12	-0.52	1
69	-0.47	1	25	-0.34	-
111	-0.42	1	29	-0.05	1
108	-0.23	1	18	0.00	1
91	-0.13	1	38	0.08	1
96	0.06	0	35	0.29	1
73	0.16	1	47	0.30	-
67	0.18	0	39	0.39	1
84	0.19	1	48	0.52	1
70	0.25	1	60	0.54	1

TABEL 6.8 (vervolg)

107	0.38	1	59	0.59	1
88	0.58	1	34	0.59	0
81	0.77	1	61	0.64	1
76	1.05	0	49	0.65	-
103	1.13	1	56	0.69	1
82	1.14	1	50	1.47	1
86	1.32	1			
109	1.48	1			
106	1.79	1			

		Tijdstip 3			
item nr	item-moeilijk-heid	code* voor t_{j2}-t_{j1}	item nr	item-moeilijk-heid	code* voor t_{j2}-t_{j1}
Eerstejaarsitems			Algemene kennisitems		
81	-2.40	1	34	-3.68	1
78	-1.40	1	20	-2.61	1
66	-1.11	1	59	-1.96	1
101	-0.79	1	45	-1.87	1
87	-0.74	1	43	-1.60	1
80	-0.72	1	36	-1.15	1
109	-0.66	1	32	-1.02	1
100	-0.66	1	24	-0.62	1
85	-0.59	1	56	-0.07	1
92	-0.22	1	12	0.05	1
67	-0.07	0	37	0.32	1
102	-0.01	1	54	0.44	1
79	0.19	1	57	0.66	1
90	0.19	1	40	0.72	1
73	0.23	1	16	0.77	1
107	0.29	1	18	0.85	1
104	0.47	1	42	1.28	1
68	0.96	-	53	1.33	-
99	1.12	1	25	1.44	0

TABEL 6.8 (vervolg)

76	1.36	1	51	1.46	1
74	1.43	1	52	1.70	0
113	1.54	0	50	1.85	1
86	1.54	1	21	2.09	0

Tijdstip 4

item nr	item-moeilijkheid	code* voor t_{j2}-t_{j1}	item nr	item-moeilijkheid	code* voor t_{j2}-t_{j1}
Eerstejaarsitems			Algemene kennisitems		
79	-2.48	1	41	-1.65	1
112	-1.51	1	50	-1.58	1
98	-1.13	1	38	-1.45	1
114	-0.86	1	62	1.40	1
97	-0.77	1	39	-1.33	1
113	-0.58	1	16	-1.08	1
68	-0.48	1	24	-0.80	1
84	-0.47	1	58	-0.75	1
70	-0.40	1	20	-0.74	1
71	-0.35	1	61	-0.56	1
92	-0.24	1	36	-0.46	1
78	-0.01	1	33	-0.36	0
82	-0.01	1	26	-0.31	1
65	0.27	1	13	-0.06	1
96	0.36	1	37	0.12	1
74	0.70	1	46	0.13	1
75	0.82	0	35	0.18	1
67	1.14	1	49	0.25	1
115	1.42	0	34	0.26	1
76	2.96	1	32	0.37	1
			12	0.49	1
			14	0.72	0
			23	0.95	1
			26	1.84	-
			48	2.08	1

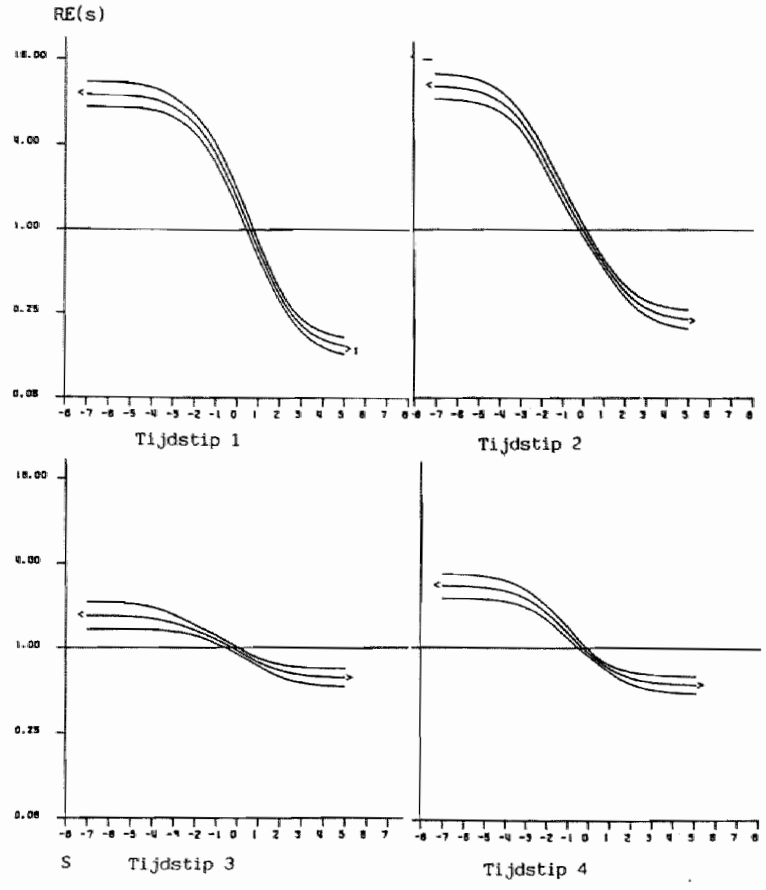
TABEL 6.8 (vervolg)

	56	2.38	0
	52	2.77	1

* 1 = $t2 < t_1$; 0 = $t_2 > t_1$; - = item vervallen.

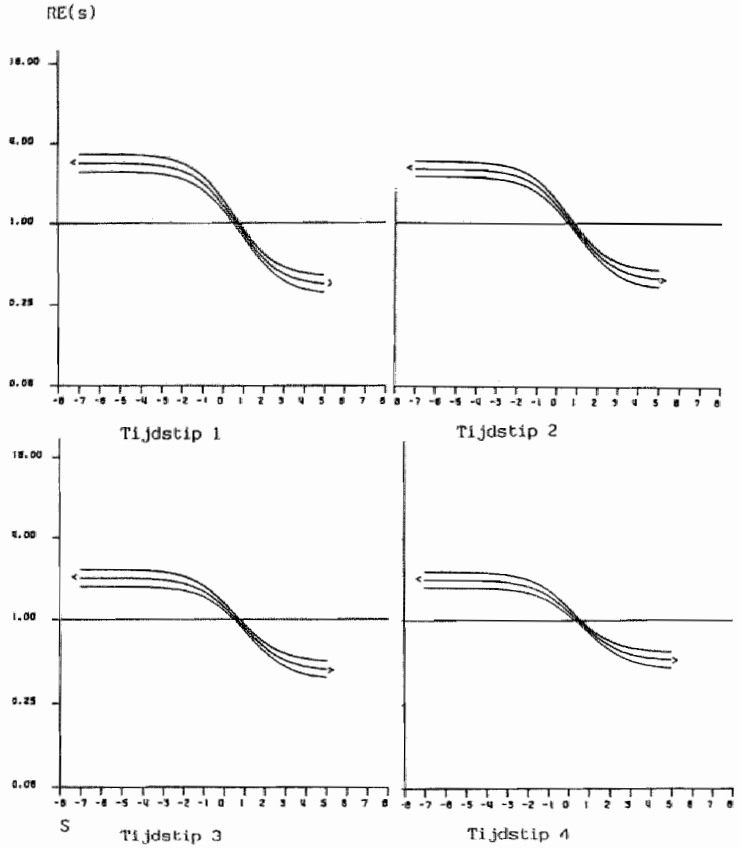
In de bovenstaande tabel zijn de items geordend van gemakkelijk naar moeilijk. De geschatte parameters in de tweede kolom van de tabel zijn gebaseerd op het dichotome Raschmodel. In de derde kolom zijn de verschillen tussen de itemstappen in het Partial Credit model weergegeven. (zie paragraaf 5.4.3.2).
Met behulp van de eerder beschreven procedure (zie eveneens paragraaf 5.4.3.2) werden van beide itemgroepen steeds vijf items vergeleken met betrekking tot de informatie die de items verstrekken over de geschatte persoonswaarden. Deze relatieve efficiëntie van de gemakkelijkste vijf items uit de Eerstejaarsitems en uit de Algemene Kennisitems op alle vier tijdstippen is grafisch weergegeven in figuur 6.3. De relatieve efficiëntie werd in dit geval gedefinieerd als de toetsinformatie (over vijf items) van de Eerstejaarsitems, gedeeld door de toetsinformatie van de Algemene Kennisitems. In figuur 6.3 en de volgende twee figuren wordt behalve een schatting van de relatieve efficiëntie ook een schatting van het betrouwbaarheidsinterval gegeven.

De relatieve efficiëntie functies hadden op alle vier tijdstippen dezelfde vorm. Voor het linker deel van de persoonsschaal had de functie waarden groter dan 1 en voor het linker deel van de persoonsschaal waren de waarden van de functie kleiner dan 1. Voor dat deel van de persoonsschaal waar de waarde van de functie gelijk is aan 1, hebben beide itemgroepen een gelijke efficiëntie, bevatten ze met andere woorden evenveel informatie over de waarden van s, de latente variabele. De conclusie uit figuur 6.3. is dat in alle vier gevallen de gemakkelijkste vijf items uit de Eerstejaarstoetsen meer informatie gaven over de personen met relatief weinig kennis, terwijl de vijf gemakkelijkste vragen uit de Algemene Kennistoetsen meer informatie gaven over de personen met relatief meer kennis. Aan de vorm van de functies op de verschillende tijdstippen is te zien dat de verschillen in informatie het grootst waren op de eerste twee tijdstippen.
De hier gegeven schattingen van de relatieve efficiëntie bestreken een groot gebied van de personenschaal, met waarden van -8 tot +8. In de praktijk van het toetsen van medische kennis komen dergelijke extreme verschillen niet voor. Om een voorbeeld te geven; op het eerste tijdstip variëerden de schattingen van de persoonswaarden van -4.37 tot 2.56. De mediaan van de eerstejaars was -2.61 en die van bijvoorbeeld de zesdejaars .31. Voor de mediane waarde van de eerstejaars was de Eerstejaarstoets ruim vier keer zo efficiënt als de Algemene Kennistoets.



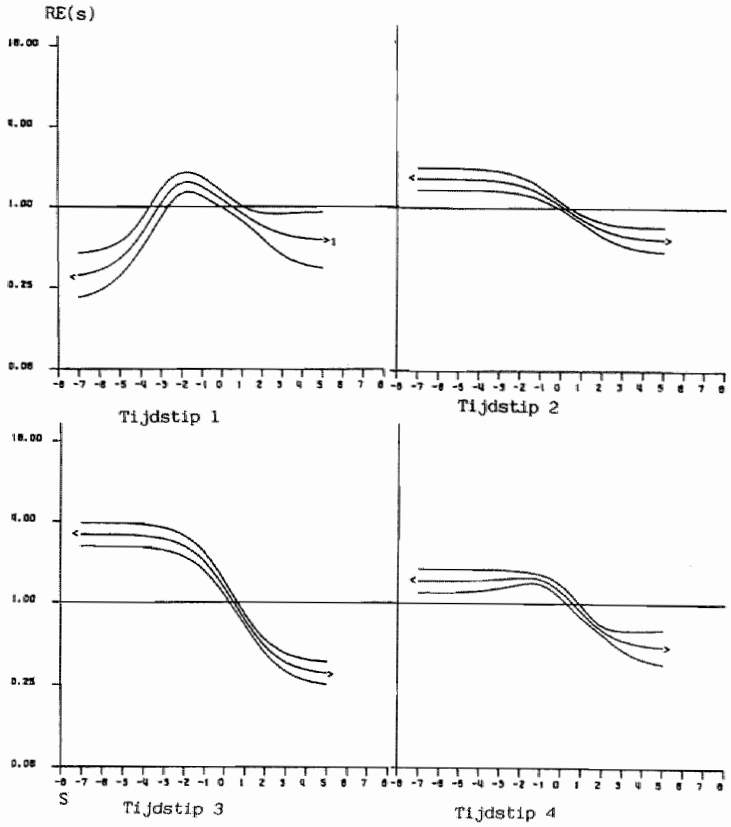
Figuur 6.3 Relatieve efficiëntie van de vijf gemakkelijkste items (s = kennisniveau).

Voor de mediane waarden van de zesdejaars, daarentegen waren beide versies even efficiënt. Slechts voor studenten met een waarde ver boven de mediaan van het zesdejaar was de Algemene Kennistoets efficiënter.
De relatieve efficiëntie van de vijf items met een gemiddelde moeilijkheid uit beide itemgroepen wordt weergegeven in figuur 6.4.
Uit deze figuur komt hetzelfde patroon naar voren als in de vorige figuur. Dit betekent dat ook de midden items uit de Eerstejaarstoetsen voor de lagere kennisgroepen efficiënter waren dan de items uit de Algemene Kennisitems.



Figuur 6.4 Relatieve efficiëntie van vijf items met een gemiddelde moeilijkheidsgraad (s = kennisniveau).

De verschillen waren echter beduidend kleiner dan bij de gemakkelijke itemgroepen. Dit bleek uit het veel vlakkere verloop van de relatieve efficiëntie functie.
De vijf moeilijkste items uit beide itemgroepen, tenslotte worden vergeleken in figuur 6.5.



Figuur 6.5 Relatieve efficiëntie van de vijf moeilijkste items (s = kennisniveau).

Deze figuur bevestigt de eerdere constatering: de Eerstejaarsitems waren efficiënter voor de lagere kennisniveaus en de Algemene Kennisitems voor de hogere kennisniveaus. De functies verliepen echter nog weer vlakker dan bij de midden items. Een uitzondering op het algemene beeld werd gevormd door de moeilijkste itemgroep op het eerste tijdstip. Hier bleken de Eerstejaarsitems efficiënter voor een relatief klein gebied van de persoonsschaal lopend van ongeveer -3 tot 0. Buiten dit interval bleken de Algemene Kennisitems efficiënter te zijn. Opvallend was het relatief grote betrouwbaarheidsinterval van de functie op het eerste

tijdstip. Dit duidt er op dat de data op dit tijdstip een nauwkeurige schatting van de relatieve efficiëntie niet toe lieten.
Dezelfde Raschhomogene items werden trichotoom gescoord en geanalyseerd met behulp van het Raschmodel voor Partial Credit. De gemiddelde schaalwaarden van de personen die met behulp van deze twee modellen werden verkregen worden vermeld in tabel 6.9.

TABEL 6.9
Samenvatting van de Raschanalyses: dichotoom en trichotoom

DICHOTOOM								
Tijdstip	1		2		3		4	
	Ej	Ak	Ej	Ak	Ej	Ak	Ej	Ak
M pers.	-0.90	-1.73	-0.57	-1.48	-0.27	-1.51	-0.15	-0.95
sd pers.	1.20	1.35	1.14	1.05	0.95	1.43	0.85	1.24
ISS	0.82	0.72	0.78	0.61	0.72	0.78	0.65	0.81
N	611	493	609	531	612	577	611	601
Modelfit	in alle gevallen acceptabel							

TRICHOTOOM								
Tijdstip	1		2		3		4	
	Ej	Ak	Ej	Ak	Ej	Ak	Ej	Ak
M pers.	-0.19	-0.84	-0.05	-0.69	0.09	-0.67	0.13	-0.39
sd pers.	0.69	0.76	0.69	0.62	0.52	0.85	0.47	0.76
ISS	0.86	0.81	0.85	0.73	0.79	0.85	0.75	0.87
N	612	510	611	546	612	587	612	601
Modelfit *	4	2	0	3	6	3	2	4

* = Aantal items met significante afwijking
ISS = Index of Subject Separation

Het effect van de scoringsregel (012) waarbij een vraagteken wordt gescoord als 0, een fout antwoord als 1 en een goed antwoord als 2, is direct te zien aan het

gemiddelde van de schaalwaarden van de personen. In alle gevallen was, zoals van deze scoring verwacht mag worden het gemiddelde bij de trichotome scoring hoger dan bij de dichotome scoring. Omdat de items in beide gevallen gecentreerd werden rondom 0, betekent dit dat het effect van de (012) scoring was dat de items gemiddeld genomen gemakkelijker werden. Op het eerste tijdstip werden bijvoorbeeld de Eerstejaarsitems .71 logits gemakkelijker. Het blijkt dat de standaarddeviatie bij de (012) scoring kleiner werd, en de Index of Subject Separation iets groter werd. Dit betekent dat de nieuwe scoringsregel de foutenvariantie van de persoonswaarden kleiner maakte. Dit houdt dan weer in dat door de (012) scoring de informatie met betrekking tot de schaalwaarden van de personen groter werd.
Het verschil tussen de itemstappen (zie paragraaf 5.4.3.2) die met behulp van het Partial Credit model werden verkregen zijn reeds vermeld in tabel 6.8. Uit deze tabel blijkt dat voor vrijwel alle items de tweede itemstap, die van een goed antwoord, gemakkelijker was dan de eerste itemstap, die van een fout antwoord. In paragraaf 5.4.3.2 werd al uiteengezet dat dit duidt op een gunstige eigenschap van de items. Items waarvoor dit geldt hebben n.l. voor de persoonsparameters, die liggen tussen de waarden van deze parameters, een relatief kleine kans op een fout antwoord. Hoe verder de itemstappen (met tweede stap kleiner dan de eerste stap) op de logitschaal uit elkaar liggen, des te kleiner de kans op een fout antwoord in het tussenliggende gebied.
De items waarvoor de eerste stap gemakkelijker is dan de tweede stap, hebben een relatief hoge kans op een fout antwoord. Hoe dichter de itemstappen bij elkaar liggen hoe hoger de kans op een fout antwoord in een relatief klein gebied van de personenschaal. Zoals blijkt hadden slechts weinig items deze eigenschap. Waar dit verschijnsel wel optrad, betrof het doorgaans de relatief moeilijker items.

6.5 Samenvatting

De analyses met behulp van het Raschmodel leverde op alle vier tijdstippen twee itemgroepen op die bleken te voldoen aan de assumpties van het Raschmodel. Deze itemgroepen bevatten gemiddeld de helft van het totaal aantal geanalyseerde items. Bij het toetsen van de twee itemgroepen met behulp van de Martin-Löf toets voor itemgroepen, bleek dat de nulhypothese van parallelle PCC's voor beide groepen niet verworpen kon worden. Er was met andere woorden geen reden om aan te nemen dat de personen die op de Eerstejaarstoetsen alleen de gemakkelijkste items goed beantwoordden, op de Algemene Kennistoetsen goede antwoorden zouden hebben op de moeilijkste items. Dezelfde conclusie kan uiteraard gelden voor personen die items over een breed bereik van de variabele gemeten met de Eerstejaarstoetsen, goed beantwoorden. Van deze personen mag worden verwacht dat ze dat ook zullen doen op de Algemene Kennistoetsen. Door deze resultaten

werden alle voorspellingen zoals geformuleerd in de hypothesen 14a, 14b en 14c bevestigd.
Uit het relatieve efficiëntie onderzoek bleek dat de Eerstejaarsitems voor de lagere kennisniveaus in alle opzichten het meest efficiënt waren. Voor de hogere kennisniveaus waren de Algemene Kennisitems efficiënter.
De Raschhomogene itemgroepen bleken ook bij een trichotome scoring redelijk aan de assumpties van het Raschmodel te voldoen. Uit inspectie van de geschatte itemstappen bleek dat bij de meeste items de tweede itemstap gemakkelijker was dan de eerste. Dit betekent dat de kans op een fout antwoord bij de meeste items vrij klein was.

6.6 Toetsing van het Mokkenmodel (hypothesen 15a en 15b)

Zoals beschreven werd in de procedure paragraaf 5.4.3, werden de overblijvende niet-Raschhomogene items geanalyseerd met behulp van het Mokkenmodel. Ook in dit geval werden de items dichotoom en trichotoom gescoord. Begonnen wordt met de resultaten van de dichotoom gescoorde items. In tabel 6.10.

Tabel 6.10
Samenvatting van de Mokkenschaalanalyse van de niet-Raschitems

Tijdstip 1	Eerstejaarsitems
items*	schaaleigenschappen
schaal 1 106,80,84,69,96,105 86,83,110	H:.45, H_i tussen .33 en .68 monotonie: goed betrouwbaarheid: .78 goed onderscheidende schaal, score-verdeling positief scheef.
schaal 2 107,74	H:.41 = H_i, monotonie: goed betrouwbaarheid: .33 slecht onderscheidende schaal, scoreverdeling symmetrisch rond 1.
schaal 3 93,90	H:.37 = H_i monotonie: goed betrouwbaarheid: .38, slecht onderscheidende schaal, score-verdeling uniform.

Tijdstip 1	Algemene Kennisitems
niet schaalbaar 113,115,79,75,92,94,77	H_i tussen .10 en .22, itempopulariteiten tussen .11 en .66.
	H: .41 H_i tussen .33 en .51
schaal 1 56,41,24,50,43,48,35,55,38 39,20,17,52,37,21,29,54	monotonie: goed betrouwbaarheid: .86 goed spreidende schaal, scoreverdeling positief scheef.
	H:.39 = H_i
schaal 2 26,15	betrouwbaarheid: .32 slecht onderscheidende schaal scoreverdeling positief scheef.
niet schaalbaar 46,16,28,42,27,40,31,62 44	H_i tussen .02 en .25 itempopulariteiten tussen .08 en .31.

Tijdstip 2	Eerstejaarsitems
items	schaaleigenschappen
	H:.43, H_i tussen .33 en .55
schaal 1 65,99,95,71,66,74,72,98,93 75,79,94,100,85	monotonie: goed betrouwbaarheid: .87 goed onderscheidende schaal, scoreverdeling meertoppig.
	H:.57 = H_i
schaal 2 101,102	monotonie: goed betrouwbaarheid: .46 slecht onderscheidende schaal, scoreverdeling redelijk symmetrisch.

schaal 3 114,105	H:.37 = H_i monotonie: goed betrouwbaarheid: .36 slecht onderscheidende schaal scoreverdeling symmetrisch.
niet schaalbaar 68,97,110,104,113,112,78, 115,80	H_i tussen .00 en .14, item populariteiten tussen .19 en .68.

Tijdstip 2	Algemene Kennisitems
schaal 1 57,51,41,55,40,16,13,45,17 54,22,52,19,43,26	H:.37, H_i tussen .30 en .45 monotonie: goed betrouwbaarheid: .83 goed onderscheidende schaal scoreverdeling positief scheef itempopulariteiten tussen .03 en .45.
schaal 2 37,15	H:.33 = H_i, monotonie: goed betrouwbaarheid: .32, slecht onderscheidende schaal, score- verdeling positief scheef.
niet schaalbaar 23,24,21,32,36,62,42 27,28,33,58,53,20	H_i tussen .07 en .19 itempopulariteiten tussen .12 en .37.

Tijdstip 3	Eerstejaarsitems
items	schaaleigenschappen
schaal 1 88,65,96,71,97,114,72,111, 94,95,89,91,93,70	H:.42, H_i tussen .35 en .55 monotonie: goed betrouwbaarheid: .84 goed spreidende schaal, score- verdeling symmetrisch.

	H:.56 = H_i
schaal 2 82,83	monotonie: goed betrouwbaarheid: .62 slecht spreidende schaal scoreverdeling positief scheef.
	H:.41 = H_i
schaal 3 110,77	monotonie: goed betrouwbaarheid: .32 slecht spreidende schaal scoreverdeling negatief scheef. H:.32 = H_i
schaal 4 106,105	monotonie: goed betrouwbaarheid: .33 slecht spreidende schaal scoreverdeling negatief scheef.
niet schaalbaar 103,69,75,115,112,84,108,98	H_i tussen .02 en .19 itempopulariteiten tussen .18 en .71.

Tijdstip 3	Algemene Kennisitems
	H:.39, H_i tussen .30 en .50
schaal 1 62,31,19,23,61,44,17,22,55, 27,41,36,15,58,33,26	monotonie: goed betrouwbaarheid: .85 goed spreidende schaal scoreverdeling positief scheef.
	H:.38, H_i tussen .31 en .48
schaal 2 14,60,28	monotonie: goed betrouwbaarheid: .39 slecht onderscheidende schaal scoreverdeling positief scheef.
	H:.44 = H_i
schaal 3 48,47	monotonie: goed betrouwbaarheid: .38 slecht spreidende schaal scoreverdeling positief scheef.

	H:.31 = H_i
schaal 4 39,49	monotonie: goed betrouwbaarheid: .28 slecht onderscheidende schaal scoreverdeling symmetrisch rond 1.
niet schaalbaar 46,29,13,30,38	H_i tussen .06 en .24 itempopulariteiten tussen .21 en .51.

Tijdstip 4	Eerstejaarsitems
items	schaaleigenschappen
	H:.42, H_i tussen .31 en .57
schaal 1 106,95,94,105,73,103,93 110,83,81	monotonie: goed betrouwbaarheid: .78 goed spreidende schaal scoreverdeling symmetrisch.
	H:.59, H_i tussen .39 en .67
schaal 2 89,90,91	monotonie: goed betrouwbaarheid: .68 slecht spreidende schaal scoreverling positief scheef.
niet schaalbaar 85,72,102,87,66,107,101, 100,104,69,109,108,86, 111,77,80	H_i tussen .05 en .12 itempopulariteiten tussen .12 en .79.

Tijdstip 4	Algemene Kennisitems
	H:.43, H_i tussen .33 en .57
schaal 1 43,42,55,19,21,60,15,47, 57,22,30,17	monotonie: goed betrouwbaarheid: .85 goed spreidende schaal scoreverdeling positief scheef.

	H:.33 = H_i monotonie: goed
schaal 2 54,59	betrouwbaarheid: .42 slechtspreidend schaal scoreverdeling positief scheef.
	H:.32 = H_i
schaal 2 44,51	monotonie: goed betrouwbaarheid: .34 slecht spreidende schaal.
	H_i tussen .13 en .25
niet schaalbaar 27,18,45,53,25,40,28,29	itempopulariteiten tussen .18 en .48.

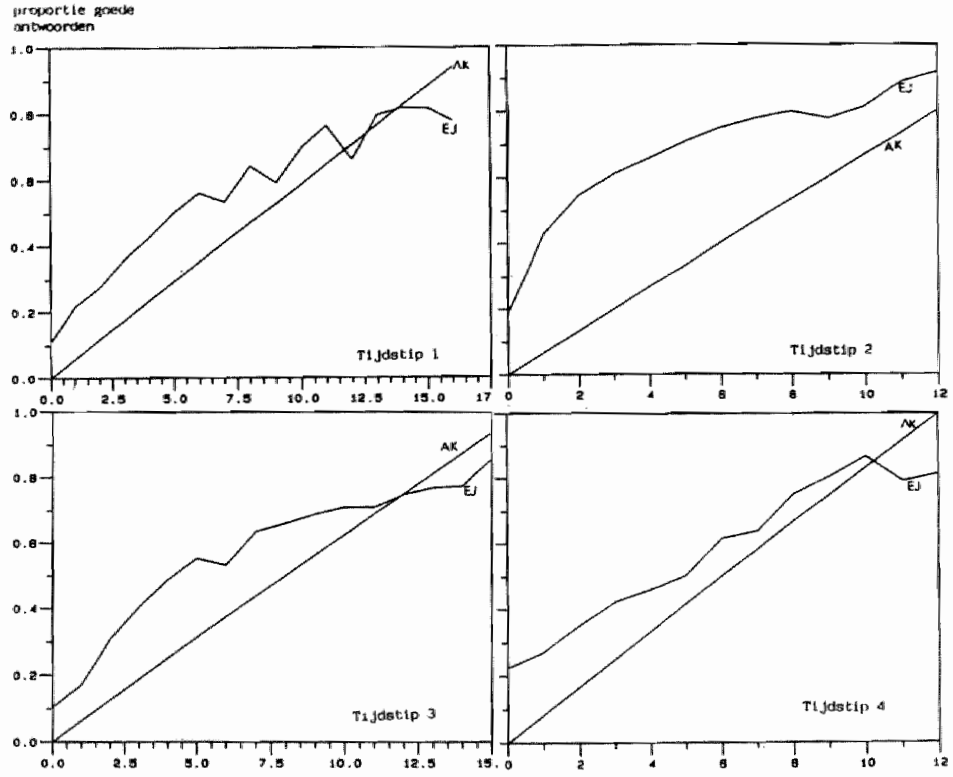
* geordend naar itempopulariteit (moeilijke items voorop).

Uit de samenvatting die wordt gegeven in de bovenstaande tabel blijkt dat er op ieder tijdstip uit de resterende niet-Raschitems een aantal schalen geconstrueerd kon worden, dat voldeed aan de eisen van het Mokkenmodel. Op ieder tijdstip bleek dat er zowel uit de Eerstejaarsitems als uit de Algemene Kennisitems een schaal van enige omvang, aangevuld met een paar kortere schalen, samengesteld konden worden. De korte schalen zijn praktisch gezien onbruikbaar en worden derhalve hier verder buiten beschouwing gelaten.
De Mokkenschalen die in de tabel vermeld worden onder het label schaal 1, hadden overwegend goede schaaleigenschappen. Deze conclusie werd vooral gebaseerd op het verrichte monotonieonderzoek (zie paragraaf 5.4.3). Ook bleek de betrouwbaarheid, hier bepaald door middel van de methode van aangrenzende proporties goede antwoorden (Molenaar en Sijtsma, 1987) zonder uitzondering zeer acceptabel te zijn. De gemiddelde H-coëfficiënten voldeden uiteraard aan de in het computerprogramma MSP vastgelegde default waarde (H = .30). In termen van Mokken (1971) kan men de mate van schaalbaarheid beoordelen aan de hoogte van de H-coëfficiënt. Mokken geeft daarvoor zelf de criteria:

$.30 < H \leq .40$: "zwakke schaal",
$.40 < H \leq .50$: "matige schaal",
$H > .50$: "sterke schaal".

Beoordeeld naar deze maatstaven bleken de meeste schalen gekwalificeerd te kunnen worden als "matige schalen". Een enkele bleek een zwakke schaal te zijn, en geen enkele schaal voldeed aan het hoogste door Mokken gestelde criterium.
De niet-schaalbare items zijn in de Mokkenschaalprocedure de items met lage H-coëfficiënten. Vat men de H-coëfficiënt op als een soort gestandaardiseerde

interitemcorrelatie (Jansen, 1982), dan kan uit tabel 6.10 worden geconcludeerd dat de niet-schaalbare items de items waren met de laagste interitemcorrelaties. De itempopulariteiten van de niet-schaalbare items bleken vooral bij de Eerstejaarsitems over een groot gebied verspreid te liggen. Op het vierde tijdstip bijvoorbeeld varieerde de itempopulariteit tussen .12 en .79. Met andere woorden de niet-schaalbaarheid van de items bleek niet samen te hangen met de moeilijkheid van de items. De Eerstejaars Mokkenschalen en de Algemene Kennis Mokkenschalen worden in figuur 6.6 met elkaar vergeleken.



Figuur 6.6 Geobserveerde proportie goede antwoorden als functie van de totaalscore op de Algemene Kennisitems (Ej = Eerstejaarstoets en Ak = Algemene Kennistoets).

Uit de figuur blijkt dat ook de Mokkenitems zich op eenzelfde wijze gedroegen als de Raschitems. De gemiddelde kans op een goed antwoord op de Eerstejaarsitems was voor de lagere scoregroepen in alle gevallen hoger dan die

op de Algemene Kennisitems. Voor de hogere scoregroepen was deze kans ongeveer gelijk.
De resterende niet-Raschitems werden eveneens trichotoom gescoord en op dezelfde wijze geanalyseerd. In tabel 6.11 wordt de schaalbaarheid van de items bij trichotome scoring vergeleken met de schaalbaarheid van dezelfde items bij een dichotoom scoringssysteem. Op welke wijze de trichotome scoring plaats vond, werd hierboven in paragraaf 5.4.3 reeds besproken.

TABEL 6.11
Percentage* items dat schaalbaar is bij zowel trichotome als dichotome scoring (cq. niet schaalbaar is bij trichotome en dichotome scoring)

tijd-stip	schaal	eerste jaars-items	algemene kennis-items
1	1	100	100
	2	0	50
	3	0	
	ns	57	35
2	1	93	73
	2	99	0
	3	100	
	ns	88	77
3	1	100	81
	2	100	0
	3	0	100
	4		100
4	1	100	100
	2	100	0
	3		0
	ns	100	88

ns = niet schaalbaar
* het percentage dichotome items dat in eenzelfde categorie valt bij trichotome scoring.

De voornaamste conclusie die uit deze tabel getrokken kan worden is dat de items die met de dichotome scoring Mokkenhomogeen bleken te zijn, dit voor het merendeel ook waren met de trichotome scoring. De grootste verschillen traden op bij de "korte" schalen en bij de items die bij het dichotome scoringssysteem niet-schaalbaar bleken. De gehanteerde wijze van scoring bleek weinig van invloed te zijn op de schaalbaarheid van de items.Gezien het feit dat de schaalconstructie bij Mokken gebaseerd is op de hoogte van de H-coëfficiënt, betekent dit resultaat in concreto, dat de gehanteerde trichotome scoring de intercorrelaties tussen de items niet verstoort: items die hoog met elkaar correleren onder het dichotome systeem blijven dat voor het merendeel ook onder het trichotome systeem doen. Als voorbeeld kan dienen schaal 1 op het eerste tijdstip van de Eerstejaarsitems. Tabel 6.11 laat zien dat deze schaal onder het trichotome scoringssysteem dezelfde items bevatte. De H-coëfficiënt bij dichotoom gescoorde items was .45 en die bij de trichotoom gescoorde items .41. De betrouwbaarheid geschat volgens de methode van aangrenzende proporties (Molenaar en Sijtsma, 1988) van de schaal was bij dichotome scoring .78 en bij trichotome scoring .82.
Op de betekenis van deze resultaten en die van de trichotome Raschitems wordt in het slothoofdstuk nader ingegaan.

6.7 Samenvatting

De Mokkenanalyses van de niet-Raschitems leverden op ieder tijdstip twee schalen op van een redelijke omvang. De Eerstejaars Mokkenschalen bleken voor de lagere scoregroepen eenvoudiger te zijn dan de Algemene Kennis Mokkenschalen. De meeteigenschappen van de geconstrueerde schalen bleken op grond van het uitgevoerde monotonieonderzoek goed te zijn.
Het toegepaste trichotome scoringssysteem leverde bij benadering dezelfde Mokkenschalen op. Deze resultaten betekenen dat de hypothesen 15a en 15b ondersteund worden.

6.8 Correlaties tussen Raschitems, Mokkenitems en niet-schaalbare items

De in de vorige paragrafen besproken analyses leidden tot een driedeling van de onderzochte items. In de eerste plaats bleken er items te zijn die voldeden aan de strenge assumpties van het Raschmodel. Vervolgens waren er onder de overgebleven items weer een aantal die aan de assumpties van het zwakkere Mokkenmodel voldeden. Tenslotte bleef er een groep items over die in geen enkel van de hier onderzochte modellen paste. Deze laatste bleek een groep op zich zelf staande items te zijn die met geen enkele groep items op een zinvolle manier te combineren viel. Het berekenen van een somscore over dergelijke items is op zich niet aan te bevelen. Dat het hier toch werd gedaan was om te illustreren dat de,

door veel onderzoekers voorgestane werkwijze om somscores te berekenen zonder zich te bekommeren om de psychometrische eigenschappen van de items, in veel gevallen geen goed meetinstrument oplevert.
Voor ieder tijdstip wordt in tabel 6.12 een overzicht gegeven van het aantal items in ieder van de zojuist aangeduide drie itemgroepen.

TABEL 6.12
Overzicht van aantal en percentage () Raschitems, Mokkenitems en niet-schaalbareitems

	Eerstejaarsitems				Algemene kennisitems			
	1	2	3	4	1	2	3	4
Rasch-items	31 (61)	24 (47)	23 (45)	22 (43)	20 (39)	21 (41)	23 (45)	27 (53)
Mokken-items	13 (25)	18 (35)	20 (39)	13 (25)	19 (37)	17 (33)	23 (45)	16 (31)
niet schaal-baar	7 (14)	9 (18)	5 (16)	16 (32)	12 (24)	13 (26)	5 (10)	8 (16)
Totaal	51	51	51	51	51	51	51	51

Uit de gegevens in tabel 6.12 kan worden geconcludeerd dat de meeste items schaalbaar bleken te zijn, hetzij volgens het Raschmodel hetzij volgens het Mokkenmodel. Gemiddeld over vier tijdstippen bleek slechts 19% van de items uit de Algemene Kennistoetsen niet-schaalbaar. Dit percentage kan worden gezien als een schatting van het percentage niet-schaalbare Algemene Kennisitems in de voortgangstoetsen. Ook bij de Eerstejaarsitems bleek, gemiddeld over de vier tijstippen slechts 20% van de items niet-schaalbaar te zijn. Dit percentage kan worden gezien als een schatting van het percentage niet-schaalbare Eerstejaarsitems in de voortgangstoetsen.
De vraag kan nu gesteld worden of het hier gehanteerde onderscheid in schaalbare items en niet-schaalbare items, zinvol is. Om deze vraag enigszins te kunnen beantwoorden werden de correlaties tussen de onderscheiden subschalen berekend. De correlaties tussen Raschschalen, Mokkenschalen en niet-schaalbare items worden vermeld in tabel 6.13.

TABEL 6.13
Correlaties tussen de verschillende subschalen (N = 612)

TIJDSTIP 1

	Eerstejaarstoetsen*				Algemene Kennistoetsen		
	1	2	3		1	2	3
1. Rasch	1.00			r_{11} = .71	1.00		
2. Mokken	0.76	1.00		r_{22} = .68	0.86	1.00	
3. NS	0.48	0.39	1.00	r_{33} = .71	0.48	0.44	1.00

TIJDSTIP 2

	Eerstejaarstoetsen				Algemene Kennistoetsen		
	1	2	3		1	2	3
1. Rasch	1.00			r_{11} = .70	1.00		
2. Mokken	0.77	1.00		r_{22} = .71	0.76	1.00	
3. NS	0.31	0.30	1.00	r_{33} = .26	0.51	0.47	1.00

TIJDSTIP 3

	Eerstejaarstoetsen				Algemene Kennistoetsen		
	1	2	3		1	2	3
1. Rasch	1.00			r_{11} = .67	1.00		
2. Mokken	0.74	1.00		r_{22} = .77	0.86	1.00	
3. NS	0.25	0.12	1.00	r_{33} = .12	0.44	0.44	1.00

TIJDSTIP 4

	Eerstejaarstoetsen				Algemene Kennistoetsen		
	1	2	3		1	2	3
1. Rasch	1.00			r_{11} = .70	1.00		
2. Mokken	0.62	1.00		r_{22} = .79	0.79	1.00	
3. NS	0.36	0.21	1.00	r_{33} = .23	0.52	0.49	1.00

* dichotomescoring: 1 = goed; 2 = niet goed
r_{11} = correlatie tussen Eertsejaarsitems en Algemene Kennistoetsen (Rasch)
r_{22} = idem (Mokken)
r_{33} = idem (NS: Niet-schaalbaar)

Uit bovenstaande tabel blijkt dat het onderscheid naar schaalbaarheid ook tot uitdrukking kwam in het correlatiegedrag van de schalen. De conclusie die uit de tabel getrokken kan worden, is dat Rasch en Mokken sterk met elkaar correleerden terwijl de niet-schaalbare items beduidend lager met de eerste twee schalen correleerden.

Deze conclusie betreft zowel de Eerstejaarstoetsen als de Algemene Kennistoetsen. Verder bleken de Raschschalen van Eerstejaarsitems en Algemene Kennisitems hoog te correleren. Dit was ook het geval met de Mokkenschalen van Eerstejaarsitems en Algemene Kennisitems. De correlaties tussen de niet-schaalbare items van Eerstejaarsitems en Algemene Kennisitems waren daarentegen laag.

6.9 Samenvatting

Uit de analyse van de correlaties tussen de drie types schalen bleek dat schalen met goede meeteigenschappen (= voldoen aan de criteria volgens een schaalmodel) eveneens een correlatiegedrag vertoonden dat meer in overeenstemming is met de verwachting dat toetsen die hetzelfde meten sterk moeten samenhangen. Niet alleen gemeten naar interne maatstaven, doch ook in termen van correlaties met externe variabelen bleek dat de constructie van goede schalen de moeite waard is.

6.10 Constructrepresentatie van Medische Kennis

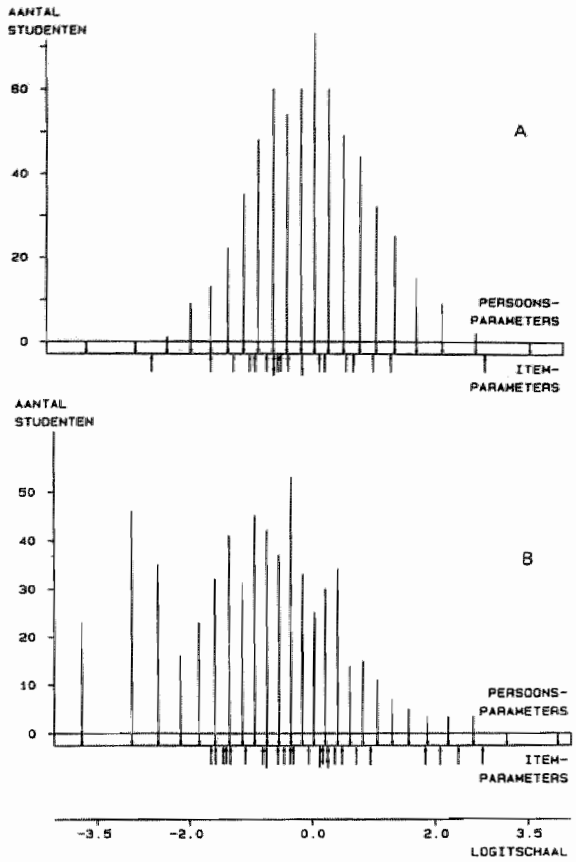
6.10.1 Inleiding

In de vorige paragrafen is gebleken dat het mogelijk is om het merendeel van de items die in het onderzoek werden betrokken, ééndimensionaal te ordenen naar moeilijkheid. Tegelijkertijd bleek het mogelijk om de in het onderzoek betrokken personen te ordenen naar de mate waarin deze in staat bleken om de geordende vragen te beantwoorden. In de - als hulptheorie - gebruikte modellen kan de relatie tussen personen en items worden opgevat als een dominantie relatie. Een persoon met een parameter s_i, die groter is dan een itemparameter t_j domineert dat item, anders gezegd, is sterker dan dat item. Omgekeerd domineert het item de persoon wanneer het itemparameter t_j groter is dan de persoonsparameter s_i.
De modellen die bij het onderzoek werden gebruikt zijn op zichzelf genomen niet meer dan formele theorie, met behulp waarvan aan de items en aan de personen getallen kunnen worden toegekend. In wezen is dit een inhoudloze bewerking van het onderzoekmateriaal.
In deze afsluitende paragraaf zal getracht worden om de stap naar de inhoudelijke variabele te zetten. Daarbij gaat het om de inhoudelijke interpretatie van de latente variable die, in de hier gebruikte modellen de centrale variabele was die het response proces bleek te kunnen verklaren.
Onderzoek naar de mechanismen die aan het beantwoordingsproces van items ten grondslag liggen wordt constructrepresentatie genoemd (Embretson, 1983). "Constructrepresentation is concerned with identifying the theoretical mechanisms that underly item responses, such as information processes, strategies, and knowledge stores" (blz. 179).

Het gaat in het onderzoek naar de constructrepresentatie, om het vinden van een verklaring voor het verschil in moeilijkheid van de toetsvragen, die de gemeten variabele definiëren. Embretson geeft een overzicht van de modellen die gebruikt kunnen worden om verschillen in moeilijkheden tussen items te verklaren. Tot de beste modellen behoort het lineaire logistische latente trek model (LLTM), waarvan het Raschmodel een speciaal geval is (zie ook Fisher, 1974).
Het gebruik van dit model dat zeer zeker mogelijkheden biedt voor verder onderzoek naar verklaringen voor itemmoeilijheden vergt echter additioneel onderzoek. Bijvoorbeeld naar de cognitieve complexiteit van de items als een mogelijke verklaring voor het verschil in moeilijkheden tussen de items. Dit aanvullend onderzoek gaat, hoe boeiend ook, het bestek van deze studie te buiten. Vandaar dat er gekozen werd voor een oppervlakkiger benadering van het beantwoorden van de vraag naar de constructrepresentatie.
De manier waarop deze vraag hier beantwoord zal worden is die van de beschrijving van de verschillen tussen moeilijke items en gemakkelijke items. Omdat de items zich op alle vier tijdstippen op een vergelijkbare manier laten beschrijven en een volledige beschrijving te veel ruimte in beslag zou nemen, werd hier gekozen voor het beschrijven van de constructrepresentatie op het vierde tijdstip.
De meeste aandacht gaat daarbij uit naar de Raschhomogene schalen. De reden daarvoor ligt voor de hand; een Raschschaal geeft de meeste informatie over items en personen, en over de afstand van het ene item ten opzichte van het andere item.
In paragraaf 6.10.2 wordt een grafische weergave van de gemeten variabelen op het vierde tijdstip gegeven. In paragraaf 6.10.3, wordt de inhoud van een aantal gemakkelijke en moeilijke items weergegeven.

6.10.2 Grafische representatie van de variabele

De twee geconstrueerde Raschschalen op het vierde tijdstip worden weergeven in figuur 6.7. Daaruit blijkt dat de items over een groot gebied van de variabele verspreid lagen. Voor de Eerstejaarsitems lag het bereik van de moeilijkheden tussen - 2.48 en 2.96 en voor de Algemene Kennisitems tussen -1.65 en 2.77. Ondanks dit grote verschil in itemmoeilijkheden, hetgeen er op duidt dat er zowel gemakkelijke als moeilijke vragen in de schalen vertegenwoordigd waren, bleken er in beide gevallen persoonspunten ter linker zijde van de itemspunten te liggen. Dit betekent dat er personen waren voor wie de makkelijkste items nog te moeilijk waren. Het gemakkelijkste Eerstejaarsitem bleek voor 2 personen te moeilijk. Het gemakkelijkste Algemene Kennisitem bleek voor 154 personen te moeilijk te zijn. Een overzicht van de verdeling van de persoonsparameters wordt gegeven in tabel 6.14.



Figuur 6.7 Grafische weergave van de variabele Medische Kennis gemeten met Eerstejaarsitems (A) en Algemene Kennisitems (B).

De afstanden tussen de itempunten en de persoonspunten in figuur 6.7 kunnen worden gebruikt om de informatie die het item over het kennisniveau van de persoon geeft te schatten. De items die de meeste informatie over een persoon geven zijn de items die niet verder dan 1 logit van het persoonspunt verwijderd zijn (Gustafsson, 1977; Wright en Stone, 1979). Deze items zijn voor de persoon in kwestie "op maat gesneden". Voor items in het gebied $\pm$ 1 logit rondom het persoonspunt is de relatieve efficiëntie ($=400p(1-p)$), waarbij p = kans op een goed antwoord onder het Raschmodel, Wright en Stone, 1979) van de observatie ten minste 75%. In vergelijking hiermee is de relatieve efficiëntie van een item dat 2 logits van het persoonspunt verwijderd is slechts 45%.

TABEL 6.14
Mediaan en kwartielspreiding van de persoonsparameters

	Eerstejaarsitems			Algemene Kennisitems		
jaar-groep	mediaan	Q1	Q3	mediaan	Q1	Q3
1	-0.96	-1.22	-0.26	-2.94	-2.94	-2.15
2	-0.26	-0.96	0.19	-1.13	-1.59	-0.73
3	-0.03	-0.48	0.42	-0.54	-0.93	-0.35
4	0.19	-0.26	0.66	-0.35	-0.73	0.03
5	0.42	-0.03	0.92	0.22	-0.16	0.42
6	0.66	0.19	1.19	0.42	0.03	1.06

Dat betekent dat de meest informatieve items voor de gemiddelde eerstejaars student, met een geschat kennisniveau van -.96, lagen in het gebied tussen -1.96 en .04. Uit figuur 6.7. blijkt dat dit een redelijk groot gebied van de totale schaal was, waar eigenlijk alleen de moeilijkste items buiten vielen.
Uit de gegevens die worden vermeld in figuur 6.7 en tabel 6.14 kan worden geconcludeerd dat de Raschschaal van Eerstejaarsitems voor de meeste scoreniveaus items "op maat" bevatte. Alleen voor het allerlaagste scoreniveau met een geschatte schaalwaarde van -2.75 bleek slechts het gemakkelijkste item met een schaalwaarde van -2.48 in het "informatierijke" gebied te liggen. Het op een na makkelijkste item bleek 1.24 logits van de persoonswaarde verwijderd te liggen. De relatieve efficiëntie van dit item voor het laagste scoreniveau was 65%. Het moeilijkste item bleek voor dit scoreniveau een relatieve efficiëntie te hebben van 1%.
Bij dit alles dient bedacht te worden dat het Raschmodel een hiërarchisch model is, hetgeen inhoudt dat personen met een relatief grote kans op een goed antwoord op een moeilijk item een nog hogere kans hebben op een goed antwoord op een gemakkelijk item. Een persoon met een niveau van 2.26 (het hoogste niveau dat werd geobserveerd) heeft een kans van .99 om het gemakkelijkste item, dat 4.75 logits van het persoonspunt verwijderd ligt, goed te beantwoorden.
De onderste grafiek in figuur 6.7 geeft een afbeelding van de door de Algemene Kennisitems gemeten variabele. Ook in dit geval lagen de itempunten en de persoonspunten over een relatief groot gebied van de variabele verspreid.
Uit tabel 6.14 en figuur 6.7 kan ook voor deze variabele een conclusie worden getrokken met betrekking tot de informatie van de items ten opzichte van bijvoorbeeld het gemiddelde niveau in de verschillende jaargroepen. Het eerstejaar

bleek, gemeten met de Algemene Kennisitems, een gemiddeld niveau te hebben van -2.94. Van dit punt was het gemakkelijkste item nr 41 2.1 logits verwijderd. Dit houdt in, een relatieve efficiëntie van 40% van een item dat precies "op maat" zou zijn. Alleen voor die eerstejaars met een score -2.65 lag het gemakkkelijkste item in het meest informatieve gebied van de schaal. Op het vierde tijdstip bleken 80 studenten een lager niveau dan het zojuist genoemde te halen. Hieruit kan worden geconcludeerd dat de Algemene Kennistoets voor deze 80 laag scorende studenten weinig informatie bevatte. Alle items bleken voor studenten van dit niveau te moeilijk te zijn. Genoemde 80 studenten waren afkomstig uit drie jaargroepen; eerstejaars, tweedejaars en derdejaars.
In dezelfde lijn door redenerend kan uit de bovenstaande figuur en de tabel worden geconcludeerd dat de informatiewaarde van de Algemene Kennisitems toenam met het aantal studiejaren; al vanaf het gemiddelde niveau van het tweede studiejaar waren er aanzienlijk meer items, die voldeden aan het gestelde criterium van 1 logit boven of onder de gemiddelde schaalwaarde van de jaargroep.
In het kader van het onderzoek naar de constructrepresentatie dient de vraag aan de orde te komen, waarin de items, die nog redelijk beantwoordbaar bleken te zijn, verschilden van die items die te ver van de lage niveaus verwijderd lagen. Een antwoord op die vraag geeft bovendien informatie over de gemeten variabele als zodanig. Naar het antwoord op deze vraag zal in de volgende paragraaf worden gezocht.

6.10.3 De relatie tussen inhoud en moeilijkheid van de toetsitems

Eerder bleek dat de beide Raschhomogene itemgroepen op alle vier tijdstippen één variabele representeerden. Er was met andere woorden geen enkel bezwaar tegen het tot een nieuwe toets samenvoegen van de Eerstejaarsitems en de Algemene Kennisitems. Dit samenvoegen van twee Raschschalen houdt in dat de items opnieuw moeten worden gecalibreerd. De nieuwe schattingen van de itemmoeilijkheden van de gecombineerde schaal kunnen vervolgens rechtstreeks met elkaar worden vergeleken. Uit de vergelijking tussen de oude schaalwaarden en de nieuwe schaalwaarden van de items bleek dat het samenvoegen in de meeste gevallen leidde tot enigszins extremere waarden. Met andere woorden, de gemakkelijke items uit de schaal van Eerstejaarsitems kregen op de samengevoegde schaal lagere waarden. Hetzelfde was het geval met de moeilijke items uit de schaal van Algemene Kennisitems. Die kregen hogere waarden. In tabel 6.15 wordt de tekst van de vragen weergegeven en de "nieuwe" schaalwaarden. De "oude" waarden kunnen uit figuur 6.7 worden afgelezen. De gemiddelde schaalwaarden van de zes jaargroepen op de gecombineerde schaal waren respectievelijk: -1.71, -.73, -.31, .01, .32 en .65. Door bij deze gemiddelde waarden 1 op te tellen

respectievelijk af te trekken kunnen, op dezelfde wijze als hierboven werd beschreven, de meest optimale items bij de gemiddelde jaarniveaus worden gezocht.

TABEL 6.15
Overzicht van de inhoud en moeilijkheid van de Raschitems op tijdstip 4

nr.	tekst	moeilijkheid op gezamenlijke schaal
	Vijf gemakkelijkste vragen	
	eerstejaarsitems	
79	Tot de effecten van de mediatoren die vrijkomen bij de ontstekings-reaktie behoort/behoren: dilitatie van capillaren.	-3.00
112	De commensale bacteriele flora in de nasopharijnx verlaagt de de weerstand tegen infecties met pathogene bacteriën.	-2.00
98	Tijdens de diastole stroom het bloed geleidelijk van het atrium in de ventrikel en van de aorta in het perifere vasculaire bed. Deze stroming van het bloed wordt met behulp van een stethoscoop als een geruis waargenomen.	-1.60
114	Bij poliomylitis worden de verlammingsverschijnselen veroorzaakt door een aantasting van de motorische voorhoorncellen	-1.31

97	Deze curve laat zien hoe het zuurstofgehalte van een bloedmonster varieert met de hoogte van de partiele zuurstofspanning in dit monster waarvan de koolzuurspanning op 40 mmHG (5.3 kPa) wordt gehouden. Wanneer men de koolzuurspanning in het bloedmonster verhoogt, zal, bij een gegeven zuurstofspanning, het zuurstofgehalte toenemen.	-1.21

Algemene Kennisitems

41	Autonurie bij een kind met diabetes melitus wijst op een slechte regulatie van de diabetes.	-1.21
50	Op de plaats waar de ureter de blaas binnentreedt is de ureter omgeven door prostaatweefsel.	-1.15
38	De afgifte van het neurohypophysiare hormoon oxytocine staat onder controle van de bijnierschors.	-1.03
62	Gordelroos wordt veroorzaakt door een endogene virale re-infectie. Deze infectie wordt veroorzaakt door het waterpokkenvirus.	-0.99

39	Een langdurig tekort aan jodium (langer dan één jaar) leidt in de meerderheid der gevallen tot volumevermindering van de schildklier.	-0.92

Vijf items van gemiddelde moeilijkheid

eerstejaarsitems

71	Bij de gezonde volwassen man bestaat gemiddeld 54-56% van het bloedvolume uit bloedplasma.	-0.77
78	In het kader van het rijksvaccinatieprogramma wordt op het ogenblik in Nederland het mazelenvaccin toegediend aan kinderen van 6-7 maanden.	-0.41
82	Bij luxatie van het schoudergewricht gaat het in de meerderheid van de gevallen om luxatie naar voren (ventraal).	-0.40
65	Bij percussie van de rechter thoraxhelft is er sprake van demping ter hoogte van de 4e rib, zowel aan de voorzijde als lateraal. Deze zone komt overeen met de grens tussen rechteronderkwab en rechtermiddenkwab.	-0.11

92	Lympheknoop-pakketten, juist craniaal van de inmonding van de vanae renales, hebben de bloedstroom in de vena curva inferior belemmerd. Tot de organen waarvan de veneuze afvloed verstard zal zijn behoort/behoren: het rectum.	-0.65

Algemene Kennisitems

33	Zowel een mitraalklepinsufficientie als een aortaklepstenose leidt tot linker ventrikel hypertrofie. Bij aortaklepstenose is de verhouding linker ventrikel-hole diameter/wanddikte in de meerderheid der gevallen groter dan bij mitraalklep-insufficientie.	-0.03
26	Gebrek aan assertiviteit kan zowel een gevolg van gedragsinhibitie, als van een gedragstekort zijn.	0.02
13	Bij patienten met (chronisch) longemfyseem is de inspiratoire 1-sekonde waarde meestal normaal.	0.25
37	Bij gebruik van het antihypertensivum alfa-methyl-dopa (Aldomet R) kan een autoimmuun hemolytische anemie optreden. Deze anemie is het gevolg van autoantilichamen tegen rode bloedcellen.	0.42
46	Een invaginatie van de dikke darm bij volwassenen gaat vrijwel altijd gepaard met koliekpijnen.	0.43

	De vijf moeilijkste vragen	
	eerstejaarsitems	
74	Klassieke compliment-aktivatie start indien de eerste complement component een interaktie aangaat met antigeen.	0.35
75	Bacteriele luchtweginfecties van de bovenste luchtwegen worden frequenter gezien bij een deficientie van de cellulaire immuniteit dan bij een deficientie van humorale immuniteit.	0.47
67	Van de longcarcinomen is het kleincellige longcarcinoom het meest chemotherapie gevoelig.	0.80
115	Om immuniteit tegen thetanus te onderhouden wordt een toxoid booster gegeven. Dit heeft na 3 jaar meer effect dan na 6 jaar.	1.09
76	Chronische lymfatische leukemie wordt in de meerderheid der gevallen gekarakteriseerd door een monoclonale woekering van T-lymphocyten.	2.68
	Algemene Kennisitems	
23	Het klinisch beeld van een patient met een rotatorcuff-ruptuur lijkt meer op een nervus axillaris letsel dan op een nervus thoracicus longus letsel.	1.21

31	Abortus provocatus van een foetus met erfelijke afwijkingen is een voorbeeld van negatieve eugenese.	2.08
48	Zowel bij colitis ulcerosa als bij Crohn'se collitis kunnen een aantal begeleidende verschijnselen optreden. Hiertoe behoort/behoren: pyodermie.	2.32
56	Een 68-jarige patient gebruikt sinds enkele maanden een thiazide-diureticum. Er wordt bij hem een lichte metabole acidose vastgesteld. Een causaal verband tussen het thiazide gebruik en de acidose is waarschijnlijk.	2.61
52	Unilaterale agenesie van de nier berust in de meerderheid der gevallen op een ontwikkelings-stoornis in de 8e week na de conceptie.	3.00

Uit tabel 6.15 blijkt dat de moeilijkheden van de items lagen tussen -3.00 en 3.00. Deze ordening van items naar moeilijkheid lijkt samen te gaan met een inhoudelijke van basale kennis naar kennis van ziekten en therapeutische ingrepen. De gemiddelde kennisniveaus van de onderzochte jaargroepen bevonden zich tussen deze itempunten. Het gemiddelde kennisniveau van eerstejaars op dit tijdstip werd geschat op -1.71. Optimale items voor deze studenten zijn items met schaalwaarden tussen -2.71 en .71. Items met schaalwaarden < -2.71 zijn te gemakkelijk, bijna iedereen zal zo'n item goed beantwoorden en items met schaalwaarden > .71 zijn juist te moeilijk. Inhoudelijk betekent dit dat items, zoals itemnummer 71, eigenlijk te gemakkelijk zijn. Onder het Raschmodel is de kans op een goed antwoord voor een student met de kennis van een gemiddelde eerstejaars .78. Een item met de inhoud van itemnummer 67 (Eerstejaarsitem met een geschatte moeilijkheid van .80) is eigenlijk te moeilijk. De kans op een goed antwoord onder het Raschmodel is .07. De kans op een goed antwoord voor een zesdejaars met een voor dit jaar gemiddeld kennisniveau op vraagnummer 71 is .97 en op vraagnummer 67 .46. Deze gegevens suggereren dat kennis van basale biologische zaken gemakkelijker

is dan kennis van therapeutische en diagnostische zaken. Deze suggestie werd bij inspectie van de Raschitems op de overige drie tijdstippen bevestigd.
Tenslotte is nog vermeldenswaard dat dit verschijnsel zich zowel bij de Eerstejaarsitems als bij de Algemene Kennisitems voordeed.
Hierboven werd gesproken van een suggestie, die uitgaat van de gepresenteerde variabele op tijdstip vier. Dit werd gedaan om te benadrukken dat deze constatering tentatief is, die nadere verificatie behoeft (zie ook hoofdstuk 7). Dit laatste substantiële hoofdstuk kan worden afgesloten met de conclusie dat de latente variabele, die door middel van het VGT-model, het Raschmodel en het Mokkenmodel meetbaar werd gemaakt, geïnterpreteerd kan worden als Medische Kennis of de cognitieve structuren van personen die worden opgeleid tot medicus. Alleen gevorderden in dit gebied hebben een redelijk kans op een goed antwoord op de moeilijkste vragen, over ziekten, diagnose en therapie.

6.11 Samenvatting constructrepresentatie

De antwoorden op Raschhomogene items kunnen worden verklaard door één onderliggende latente variabele. Deze latente variabele wordt door de items waarmee men het construct tracht te meten gedefinieerd. Het bleek dat de items die het construct definiëren over een relatief groot gebied verspreid lagen. Het dichtst bij de gemiddelde schaalwaarde van de Eerstejaars lagen de itempunten van de Eerstejaarsitems. Een groot aantal van deze items bleek informatief voor het gemiddelde kennisniveau. Wanneer de latente variabele gedefiniëerd werd door de Algemene Kennisitems alleen, dan bleek het eenvoudigste item nog te moeilijk voor het gemiddelde kennisniveau.
Bij het samenvoegen van Eerstejaarsitems en Algemene Kennisitems bleken een aantal items toch dichter bij het gemiddelde niveau te liggen dan op basis van de Algemene Kennistoets alleen verwacht kon worden.
De itemmoeilijkheden leken verklaard te kunnen worden door de inhoud van items. De gemakkelijkste items vragen naar basale medisch-biologische zaken, die te leren zijn ook zonder dat men veel ervaring in het "veld" heeft opgedaan. De moeilijkste items daarentegen zijn items die vragen naar het handelen van de medicus of naar de meest waarschijnlijke diagnose, gegeven de gepresenteerde verschijnselen.
Het construct of de latente variabele dat door deze items wordt gemeten kan op basis van de inhoud van de items en hun plaats op de schaal worden omschreven als Medische Kennis of als de Cognitieve structuren van studenten in een Faculteit der Geneeskunde. Het gemakkelijkste item is te moeilijk voor common sense kennis en het moeilijkste item is te moeilijk voor medische kennis, die kan worden opgedaan in een opleiding tot basisarts. Daar tussen ontwikkelen zich de kennisstructuren van de studenten.

6.12 Samenvatting en Conclusies

In dit laatste hoofdstuk werden drie psychometrische modellen op haalbaarheid getoetst. Het meest algemene model, het VGT-model bleek een adequate beschrijving te geven van de kansen op een goed antwoord in een nieuw getrokken steekproef van Algemene Kennisitems. Voor de Eerstejaarsitems kon er geen nieuwe steekproef getrokken worden. Derhalve was modeltoetsing voor deze itemgroep niet mogelijk. De resultaten gaven echter ook voor deze itemgroep geen aanleiding voor twijfel aan het model, als beschrijving van de kansen voor de verschillende kennisniveaus.
Gebleken is dat voor nagenoeg alle kennisniveaus de kansen op een goed antwoord voor de Eerstejaarsitems hoger was dan voor de Algemene Kennisitems. Alleen voor de hoogste niveaus waren deze kansen ongeveer gelijk. Verder bleek dat de groei in kennis op Eerstejaarsitems alleen in het eerste studiejaar het grootst was. Alle andere jaargroepen vertoonden de meeste groei op de Algemene Kennisitems (paragraaf 6.2). Uit de resultaten werd geconcludeerd dat hypothese 13, inhoudende dat de antwoorden op beide itemgroepen met behulp van een item response model te beschrijven zou zijn, ondersteund wordt (paragraaf 6.3).
Toetsing van de modelassumpties op itemniveau werd beschreven in paragraaf 6.4. In beide itemgroepen bleek ongeveer de helft Raschhomogeen te zijn. De hypothesen 14a en 14b werden door deze resultaten ondersteund.
Beide Raschhomogene itemgroepen bleken voor de verschillende kennisniveaus parallelle persoons karakteristieke curves op te leveren. Hiermee werd de voorspelling, geformuleerd in hypothese 14c, door de empirie ondersteund.
Trichotome scoring bleek de Raschhomogeniteit niet te verstoren. Bij de meeste items bleek de kans op een fout antwoord klein te zijn (paragraaf 6.5).
Toetsing van het niet-parametrische Mokkenmodel leverde voor beide itemgroepen op alle onderzochte tijdstippen een Mokkenschaal op met een substantiële lengte. De hypothesen 15a en 15b werden hiermee ondersteund. Ook hier leverde trichotome scoring bij benadering dezelfde Mokkenschalen op (paragraaf 6.6 en 6.7). Uit de analyse van de correlaties tussen de drie "schalen" bleek dat de schalen bestaande uit items met goede meeteigenschappen hoog correleren. De somscore van de niet-schaalbare items correleerde laag positief met de Rasch- en de Mokkenschalen.
In paragraaf 6.10 tenslotte werd de constructrepresentatie van de gemeten variabele op tijdstip vier onderzocht.
Volstaan werd met een grafische representatie (paragraaf 6.10.2). Het bleek dat de items die het construct definiëren over een groot gebied verspreid lagen. De Eerstejaarsitems bleken het dichtst te liggen bij het gemiddelde scoreniveau in het eerstejaar. Wanneer het construct wordt gemeten met de Algemene Kennisitems bleken de gemakkelijkste items te moeilijk te zijn voor het gemiddelde

scoreniveau in het eerstejaar. Er leek een verband te bestaan tussen de inhoud van de items en de moeilijkheid. Basale feitenkennis bleek gemakkelijker dan kennis met betrekking tot het medisch handelen (paragraaf 6.10.3 en 6.11).

7. ALGEMENE DISCUSSIE

Inhoud

7.1 Inleiding
7.2 Samenvatting van de onderzoeksresultaten
7.3 Het gebruik van einddoeltoetsen in het eerstejaar: bespreking van de onderzoeksresultaten
7.4 Bespreking onderzoeksresultaten vanuit het cognitief psychologisch perspectief
7.5 Bespreking onderzoeksresultaten vanuit het psychometrisch perspectief
7.6 Bespreking onderzoeksresultaten vanuit het data-analytisch perspectief
7.7 Slotbeschouwing

Hoofdstuk 7: Algemene discussie

7.1 Inleiding

In dit laatste hoofdstuk worden de bevindingen besproken en in een breder kader geplaatst. In paragraaf 7.2 worden de resultaten bediscussieerd met het oog op het gebruik van voortgangstoetsen in het eerste studiejaar. Daarna volgt een bespreking van de resultaten vanuit de drie theoretische niveaus die in deze studie voor het formuleren van de hypothesen werden gebruikt: de theorie over de ontwikkeling van medische kennis (paragraaf 7.3), de psychometrische theorie (paragraaf 7.4) en tenslotte de data-analytishe theorie (Paragraaf 7.5).

7.2 Samenvatting van de onderzoeksresultaten

Het in de voorgaande hoofdstukken beschreven onderzoek was gericht op het beantwoorden van twee vraagstellingen:

I. Bevatten voortgangstoetsen items die door beginnelingen in het medisch domein beantwoord kunnen worden?

Op grond van de resultaten die in hoofdstuk 2 werden beschreven luidde het antwoord op de eerste onderzoeksvraag bevestigend. Dit positieve resultaat maakte het mogelijk om het onderzoek verder uit te diepen en een antwoord te zoeken op de tweede vraagstelling, die als volgt luidde:

II. Zijn er voorspelbare verschillen en overeenkomsten in het "gedrag" van de in vraagstelling I bedoelde items en andere items uit de voortgangstoets?

Aan deze vraag werd een aantal deelvraagstellingen onderscheiden. Om de deelvragen te beantwoorden werd een empirisch domein voor het onderzoek gecreëerd. Dit werd omschreven als de antwoorden van respondenten op twee itemgroepen uit de vier voortgangstoetsen, die werden afgenomen in het academisch jaar 1983-1984. De itemgroepen werden aangeduid als Eerstejaarsitems en Algemene Kennisitems. De Eerstejaarsitems waren op grond van hun correspondentie met het onderwijsprogramma in het eerstejaar geselecteerd door blokcoördinatoren. De Algemene Kennisitems waren aselect getrokken uit de resterende items.
In totaal werden vijftien hypothesen opgesteld, ontleend aan drie theoretieën:

- theorie over kennis en kennisontwikkeling
- psychometrische theorie
- data-analytische theorie

De hypothesen worden beschouwd volledig in overeenstemming met de resultaten te zijn, als de voorspellingen in alle onderzochte op tenminste drie tijdstippen bevestigd werden.
Als de gegevens in sommige groepen wel en in andere niet met de hypothese overeenstemmen dan wordt dit beschouwd als gedeeltelijke overeenstemming. Als de gegevens in geen van de onderzochte groepen de hypothese bevestigen, dan wordt er gesproken van in het geheel geen overeenstemming. Bij de interne analyse betekent overeenstemming in het eerstejaar dat de nulhypothese wordt verworpen. In het zesdejaar en in de artsengroep betekent overeenstemming dat de nulhypothese juist niet wordt verworpen. Bij de externe analyses betekent overeenstemming.

- of, het onderzochte model kan niet worden verworpen,
- of, het voorspellend vermogen van Eerstejaarstoetsen en Algemene Kennistoetsen zijn gelijk
- of de correlaties zijn in overeenstemming met de voorspelling

Tenslotte betekent overeenstemming bij de psychometrische modellen dat het model niet verworpen kan worden, de data zijn in overeenstemming met het model. In Tabel 7.1 wordt een overzicht gegeven van de mate van overeenstemming tussen hypothesen en de gegevens.

TABEL 7.1
Overzicht van de overeenstemming tussen de gegevens en de 15 hypothesen.
+ betekent volledig in overeenstemming, ± is gedeeltelijke overeenstemming.

	Hypothese	
	Interne structuur	
1.	Eerstejaarsitems meer beantwoord dan Algemene Kennisitems	+
2.	Item p-waarden van Eerstejaarsitems hoger dan van Algemene Kennisitems	+
3.	Variantie van item p-waarden van Eerstejaarsitems groter dan van Algemene Kennisitems	+
4.	Item-restcorrelaties van Eerstejaarsitems kleiner dan van Algemene Kennisitems	±
5.	Betrouwbaarheid van Eerstejaarsitems lager dan van Algemene Kennisitems	±
6.	Toetsscores van personen op Eerstejaarsitems hoger dan op Algemene Kennisitems	+

TABEL 7.1 (vervolg)

7.	Variantie van toetsscores van personen op Eerstejaarsitems kleiner dan op Algemene Kennisitems	+
8.	Voor Eerstejaarsitems itemvariantie groter dan personen-variantie. Voor Algemene Kennisitems personenvariantie groter dan itemvariantie	+
9.	Natoets van Eerstejaarsitems hogere item p-waarde dan voortoets	+
	Externe analyses	
10.	a. een latente variabele per tijdstip, met simplex structuur	+
	b. als 10a met twee toegevoegde latente variabelen	± *
11.	Voorspellend vermogen van tweede- en derdejaar studiesucces van Eerstejaarstoetsen en Algemene Kennistoetsen is gelijk	+
12.	Alle correlaties met andere toetsen voor medische kennis zijn niet negatief	+
	Psychometrische modellen als hypothese	
13.	Beide itemgroepen voldoen aan assumpties van item response model, het VGT-model	+
14.	a. Deel Eerstejaarsitems voldoet aan Raschmodel	+
	b. Deel Algemene Kennisitems voldoet aan Raschmodel	+
	c. Rasch eerstejaarsitems en Rasch Algemene Kennisitems zijn homogeen	+
15.	a. Deel van niet-Rasch Eerstejaarsitems voldoet aan Mokkenmodel	+
	b. Deel van niet-Rasch Algemene Kennisitems voldoet aan Mokkenmodel	+

* significante verbetering in eerstejaar echter niet in zesdejaar

Deze samenvatting laat zien dat in slechts drie van de in totaal negentien voorspellingen sprake was van een gedeeltelijke afwijking tussen de voorspelling en de resultaten. Er kan geconcludeerd worden dat de onderzochte itemgroepen, ten aanzien van de meeste aspecten, voorspelbare verschillen en overeenkomsten laten zien. Het antwoord op vraagstelling II moet dan ook bevestigend luiden.

In drie gevallen was er sprake van een gedeeltelijke overeenstemming tussen hypothesen en de onderzoeksresultaten. In hypothese 4 werd voorspeld dat de itemrestcorrelaties van de Eerstejaarsitems gemiddeld kleiner zouden zijn dan die van de Algemene Kennisitems. In het eerstejaar zou de nulhypothese verworpen moeten worden en in het zesdejaar niet. Op alle vier tijdstippen bleek de voorspelling in het zesdejaar correct te zijn. In het eerstejaar was de voorspelling slechts op de eerste twee tijdstippen correct. Er zou sprake kunnen zijn van twee toevallige "treffers" op grond van een verkeerde redenering. Dit lijkt niet erg waarschijnlijk. De redenering is ontleend aan de klassieke testtheorie, die voorspelt dat alle indices die op de productmoment correlatie gebaseerd zijn in een homogene groep kleiner zijn dan in een heterogene groep. Bovendien leidt dezelfde redenering in het zesdejaar tot vier succesvolle voorspellingen. Waarschijnlijker lijkt het dan ook dat er hier sprake is van een tijdelijk verschijnsel, dat zich alleen op de eerste twee tijdstippen voordoet. Op de laatste twee tijdstippen zijn de verschillen in heterogeniteit kleiner dan in het begin van de studie. Dit blijkt ondermeer uit de varianties van de toetsscores (zie tabel 3.7). Het lijkt er dus op dat de voorspelling over de verschillen in itemrestcorrelaties moet worden herzien en alleen van toepassing moet worden geacht op de eerste maanden van de studie.
Ook hypothese 5 over de verschillen in betrouwbaarheid werd slechts gedeeltelijk bevestigd. In het zesdejaar waren de resultaten in overeenstemming met de onderzoeksresultaten. In het eerstejaar werd de voorspelling niet bevestigd. Ook in dit geval is het verschil in heterogeniteit van de toetsscores de basis waarop de redenering, die aan de hypothese ten grondslag ligt, steunt. Zoals hiervoor ook werd opgemerkt is er geen reden om aan deze redenering die ontleend is aan de klassieke testtheorie te twijfelen. Dit betekent dat de verschillen in heterogeniteit van de toetsscores tussen de beide toetsen niet groot genoeg zijn om tot de voorspelde verschillen in de betrouwbaarheid te leiden. Derhalve moet wel worden geconcludeerd dat de voorspelling over de verschillen in betrouwbaarheid onjuist was.
Ook hypothese 6 over verschillen in toetsscores van de respondenten bleek slechts gedeeltelijk in overeenstemming met de onderzoeksresultaten te zijn. In het eerstejaar waren de onderzoeksresultaten in overeenstemming met de hypothese, in het zesdejaar echter niet. Dit betekent dat er ook in dit jaar significante verschillen waren. Er zijn geen argumenten aanwezig om te twijfelen aan de gehanteerde redenering. De conclusie moet zijn dat de voorspelling onjuist was. Eerstejaarsitems hebben ook in het zesdejaar een hogere gemiddelde toetsscore.

7.3 Het gebruik van voortgangstoetsen in het eerstejaar: bespreking van de onderzoeksresultaten

Nagenoeg alle hypothesen werden door de empirische gegevens ondersteund. De

ondersteuning van de hypothesen is van belang voor de begripsvaliditeit van de voortgangstoets. Kort samengevat betekent begripsvaliditeit dat toetsen zich "gedragen" overeenkomstig de theorie over het gemeten construct. In deze studie is het gemeten construct de medische kennis, opgevat als het "gedragsrepertoire" (De Groot, 1986) of de cognitieve structuren waarover studenten en afgestudeerden beschikken. Het construct is hier de niet direct meetbare variabele die een verklaring biedt voor de observeerbare toetsresultaten. Ondersteuning van de hypothesen door de onderzoeksresultaten verleent -zij het indirect- steun aan de theorie over het construct. In deze studie waren dat de drie hiervoor reeds genoemde theorieën. (zie de paragrafen 7.4 tot en met 7.6).
Uit de in dit onderzoek gevonden ondersteuning van de hypothesen kunnen een aantal conclusies worden getrokken, die betekenis hebben voor het gebruik van einddoeltoetsen bij aanvang van de studie. In de eerste plaats is daarbij van belang de constatering dat voortgangstoetsen items bevatten, die volgens docenten corresponderen met het onderwijsprogramma in het eerstejaar. In de tweede plaats is gebleken dat deze items door meer studenten beantwoord worden en ook beter beantwoord worden. In de derde plaats bleek dat de kans op een goed antwoord op de meeste items monotoon stijgt met het niveau van de gemeten variabele. In de vierde plaats bleek de gemeten variabele door deze items sterk te correleren met de door de Algemene Kennisitems gemeten variabele. Vooral voor de lage scoregroepen bleken de Eerstjaarsitems efficiënter te zijn om de personen te meten. Bovendien is de groei in kennis van vooral eerstejaars het grootst op deze items. De geringe variantie die toe te schrijven is aan het gebruik van Eerstejaarsitems of Algemene Kennisitems als toetsmethode is nog een extra argument voor de conclusie dat beide itemgroepen hetzelfde construct meten. Deze conclusie krijgt nog meer ondersteuning door de observatie dat zowel Eerstejaarstoetsen als Algemene Kennistoetsen positief correleren met andere toetsen voor medische kennis i.c. de bloktoetsen. Deze hier beschreven conclusies hebben uiteraard betrekking op de bestudeerde itemgroepen uit het academisch jaar 1983-1984. Er zijn echter geen argumenten voorhanden die aanleiding geven om te denken dat dit academisch jaar een gerichte selectie zou zijn uit de ruim tien jaar dat er voortgangstoetsen worden afgenomen. Er zijn geen redenen om aan te nemen dat het jaar waaruit deze toetsen zijn gekozen afwijkend is van andere jaren. Bovendien worden alle voortgangstoetsen volgens dezelfde blauwdruk samengesteld en worden de items door middel van een gestratificeerde aselecte steekproef uit de beschikbare items getrokken. Op grond van deze argumenten lijkt de conclusie gerechtvaardigd dat de hier beschreven onderzoeksresultaten gegeneraliseerd kunnen worden naar alle voortgangstoetsen.
Op grond hiervan samenvattend kan als conclusie worden geformuleerd dat voortgangstoetsen "medische kennis" toetsen op een zodanige manier, dat beginnende studenten op een door docenten aanwijsbaar deel een duidelijk hogere

succeskans hebben dan op een willekeurige groep items. Deze verhoogde succeskans geldt voor de gehele groep van eerstejaars studenten, alsmede voor alle onderscheidbare kennisniveaus binnen dit eerstejaar.
In het onderzoek werd deze itemgroep omschreven als Eerstejaarsitems. Deze omschrijving suggereert dat alleen eerstejaars studenten een verhoogde kans op een goed antwoord hebben. Dit is onjuist. Gebleken is dat ook studenten met een hoger kennisniveau eveneens op deze items een verhoogde succeskans hebben. Slechts voor de hoogste in dit onderzoek geobserveerde kennisniveaus is de succeskans op Eerstejaarsitems gelijk of iets kleiner dan die op een willekeurige groep items uit de voortgangstoetsen (zie figuur). Dit wijst er op dat de kennis die met de Eerstejaarsitems wordt getoetst ook in de latere studiejaren nog bekrachtigd wordt. Kennelijk gaat het om vrij fundamentele kennis die door het gehele curriculum heen gestimuleerd en verdiept kan worden.
Een ander aspect dat aan de orde moet komen is de succeskans op de Eerstejaarsitems voor leken. Harde informatie daaromtrent bevat het onderhavige onderzoek niet. Wel kan uit de informatie van de laagst scorende studenten op het eerste tijdstip iets worden gezegd over de verwachte succeskans voor leken. Het laagste bij de Eerstejaarstoetsen geobserveerde kennisniveau is blijkens tabel 6.5 - 2.34. Met behulp van het VGT-model kan de succeskans van dit niveau worden geschat op .03. De succeskans op de Eerstejaarsitems voor een gemiddelde eerstejaars student in de geneeskunde wordt met hetzelfde model geschat op .14. Een beginnende student in de geneeskunde is natuurlijk doorgaans een leek. Wanneer deze "leken" op het eerste tijdstip gemiddeld een succeskans op de Eerstejaarsitems hebben van .14, dan betekent dit dat de Eerstejaarsitems voor een deel beantwoordbaar zijn op grond van kennis, die voor de aanvang van de medische studie wordt opgedaan op grond van middelbare schoolkennis. Dit werd ook door één van de beoordelaars opgemerkt (zie tabel 2.4).
In deze lijn doorredenerend kan bij wijze van hypothese voor vervolgonderzoek worden gesteld dat de succeskans op de Eerstejaarsitems voor personen zonder kennis van biologie en scheikunde .00 zal zijn.
De houdbaarheid van deze redenering betekent voor de hier onderzochte Eerstejaarsitems, dat ze een beroep doen op kennis die te moeilijk is voor de leek, gemakkelijker voor eerstejaars en relatief eenvoudig voor gevorderden op het gebied van de geneeskunde.
De aanwezigheid van een itemgroep van redelijke omvang met de eigenschappen zoals hierboven werd omschreven is wel een voorwaarde, maar niet een voldoende voorwaarde voor de acceptabiliteit van voortgangstoetsen in het eerstejaar. De in hoofdstuk 6 beschreven resultaten geven aanleiding om te denken dat de laagste scoreniveaus te weinig items beantwoorden. Voor deze scoreniveaus zijn de voortgangstoetsen weinig informatief (zie voor meer details de psychometrische bespreking van de resultaten). Vanaf het eerste kwartiel van de scoreverdeling

worden naar schatting meer dan 16 items van de in totaal 250 items beantwoord. Studenten met een hoger scoreniveau in het eerstejaar beantwoorden meer items. Op het hoogste niveau in het eerstejaar worden naar schatting ruim 80 items beantwoord. De vraag naar de bruikbaarheid van voortgangstoetsen in het eerstejaar kan om deze reden niet in zijn algemeenheid worden beantwoord. De toetsen zijn bruikbaar voor de meeste studenten, doch niet voor de laagstscorende studenten.
Deze conclusie roept de vraag op welke beleidsmaatregelen genomen kunnen worden om het gebruik van voortgangstoetsen ook voor de laagste scoreniveaus acceptabel te maken. Een rigoreuze maatregel zou zijn het afschaffen van de voortgangstoetsen. Deze maatregel zou de Faculteit der Geneeskunde weer in de situatie brengen, die zo'n jaar of twaalf geleden aanleiding gaf tot het doen van het voorstel om voortgangstoetsen af te nemen (zie bijlage II). Dit is voldoende reden om deze maatregel te verwerpen. De maatregelen zullen daarom moeten worden gezocht in het bijsturen van de huidige toetspraktijk. Daartoe kunnen een aantal suggesties worden gedaan:

- In ieder geval moeten de studenten en hun begeleiders op de aanwezigheid van een relatief eenvoudige subtoets binnen de voortgangstoetsen worden geattendeerd. De methodiek van dit onderzoek kan relatief eenvoudig in de gebruikelijke rapportage procedures worden ingebouwd. Voorwaarde daarvoor is wel de bereidheid van de docenten van het eerste studiejaar om jaarlijks de items te beoordelen.
- Alle studenten zouden ten minste een maal per jaar een grafische representatie van hun kennis temidden van de itemmoeilijkheden van deze eenvoudige subtoets moeten krijgen. en de resultaten hiervan met hun begeleider bespreken.
- Een wat verdergaande oplossing is laag scorende studenten te hertoetsen met een langere op maat gesneden toets, die uit de itembank kan worden samengesteld. Een andere oplossing zou zijn een onvoldoende prestatie op de voortgangstoets te laten compenseren door voldoende resultaten op de bloktoetsen. In de huidige praktijk van het toetsen van de kennis heeft men voor deze oplossing gekozen. Vanwege het eerder al in hoofdstuk 1 geschetste belangrijke verschil tussen bloktoetsen en voortgangstoetsen dient de voorkeur te worden gegeven aan hertoetsing met op maat gesneden items.
- Een andere mogelijkheid zou zijn om de toetsen in te delen naar moeilijkheidsgraad van de items. De toets zou dan moeten beginnen met de gemakkelijkste items en eindigen met de moeilijkste. Een dergelijke samenstelling van de toetsen maakt het voor studenten met relatief weinig kennis beter mogelijk om te beslissen welke items wel en welke niet beantwoord moeten worden. Een van de redenen waarom sommige studente

weinig items beantwoorden zou wel eens kunnen berusten op het verkeerde idee dat voortgangstoetsen alleen maar uit (te) moeilijke items bestaan. Deze laatste maatregel veronderstelt het hergebruik van items uit de itembank. Onnodig hierbij te vermelden dat de genoemde maatregelen ieder voor zich weer de nodige vragen oproepen die de moeite van nader onderzoek waard zijn.

7.4 Bespreking van onderzoeksresultaten vanuit het cognitief psychologisch perspectief

In de laatste paragrafen van hoofdstuk 6 werd de aandacht gevestigd op een tendens die in de inhoudelijke representatie van de items op de ééndimensionale Raschschaal aanwezig leek te zijn. Voor alle duidelijkheid ook hier geldt dat nader onderzoek is geboden. Niettemin zijn de resultaten te interessant om er niet een paar woorden aan te wijden.
Het interessante zit hier in dat de gesignaleerde tendens in inhoudelijke moeilijkheden van de items (van basale kennis naar medische toepassingen) voorspeld wordt door de moderne cognitieve psychologie (Anderson, 1982, Feltovich en Barrows, 1984, Messick, 1984 en Boshuizen en Schmidt, 1986). Deze theorie voorspelt dat het menselijk leren een aantal stadia doorloopt. De beginneling in een vakgebied beschikt over relatief weinig en bovendien vrij onsamenhangende kennis van basale zaken. De expert heeft niet alleen de beschikking over kennis van basale zaken maar kent ook de samenhangen en is op de hoogte van de toepassingen.
Belangrijk is dat deze cognitief psychologische opvatting een theorie van itemmoeilijkheden oplevert, waardoor inhoudelijke theorie en psychometrische theorie aan elkaar gekoppeld kunnen worden. De inhoudelijke theorie voorspelt dat eerst de basale feiten worden geleerd en pas later de toepassingen. Op grond van psychometrische theorie kan men moeilijkheden meten. De naar moeilijkheden geordende items zouden in overeenstemming moeten zijn met de inhoudelijke theorie.
Op het vierde tijdstip leek dit zowel bij Eerstejaarsitems als bij Algemene Kennisitems het geval te zijn (zie daarvoor paragraaf 6.10.3). Dat zou erop kunnen wijzen dat de genoemde tendens in itemmoeilijkheden een zeer algemeen verschijnsel is, dat ongeacht de opbouw van het curriculum optreedt. Ook een curriculum zoals dat van Maastricht, met door het gehele curriculum heen veel aandacht voor patiëntgerichte problemen, zal deze "rode draad" in de cognitieve ontwikkeling te zien geven. Het eerst ontwikkelt zich de basale kennis, de toegepaste kennis komt pas veel later. De recent voorgestelde nieuwe inrichting van het curriculum met meer aandacht voor het leren verwerven van basisbegrippen in de eerste twee studiejaren is op het niveau van het onderwijsprogramma een

ondersteuning van deze visie (Onderwijscommissie Faculteit der Geneeskunde, 1987). De rode draad die men nu op het niveau van het curriculum wil aanbrengen bestond, ondanks het curriculum, op het niveau van de studieresultaten al.
Het behoeft geen betoog dat de theorie over itemmoeilijkheden nog veel nader onderzoek behoeft. Inhoudelijk zal dergelijk onderzoek gericht moeten zijn op het meetbaar maken van cognitieve verschillen tussen de items. Op het psychometrische niveau kan dan vervolgens worden nagegaan of de cognitieve moeilijkheden een verklaring kunnen leveren voor de geconstateerde verschillen in de Raschparameters, welke op het psychometrische niveau de itemmoeilijkheden representeren. Voor een dergelijk onderzoek kan het Lineaire Logistische Test Model (LLTM) (Fisher, 1974), worden gebruikt. Een ander aspect van de cognitief psychologische opvatting, dat in verschillende van de getoetste hypothesen naar voren kwam is de opbouw van de kennis. Een van de consequenties van de lagentheorie (Feltovich en Barrows, 1984 en Boshuizen en Schmidt, 1986) is dat een cognitieve structuur op het hoogste niveau de twee onderliggende niveaus omvat. Met andere woorden dat de opbouw van de kennis hiërarchisch is. Wanneer men personen met verschillende cognitieve niveaus met problemen confronteert blijkt in een free-recall situatie dit niveauverschil uit de wijze waarop men de problemen analyseert en beschrijft. De novice geeft een gebrekkige beschrijving van het probleem, onderscheidt in het geheel een paar basale feiten die hem bekend voorkomen. De expert daarentegen is in staat om de verschillende aspecten van het probleem te overzien en te beschrijven. Het resultaat is een voor beide kennisniveaus zeer verschillend protocol van het free-recall process. Hieruit wordt dan vervolgens de cognitieve structuur afgeleid. Het behoeft geen betoog dat er wel gebieden kunnen zijn waar de novice de expert is en andersom.
Deze variabele komt natuurlijk niet alleen tot uiting in free-recall protocollen, maar speelt een centrale rol in alle cognitieve gedragingen van de persoon. De stap naar toetsresultaten is niet zo groot en ook hiervoor kan de lagentheorie van de kennis een kader bieden. Voor toetsresultaten voorspelt de theorie behalve dat experts meer weten en het anders weten, ook dat het verschil in kennis doorwerkt in elk probleem c.q. in elk item. Dit heeft op het psychometrische niveau geleid tot de keuze van de drie modellen, die deze cumulatieve structuur in de items en over alle items heen als voornaamste kenmerk hebben. De hier beschreven correspondentie tussen de cognitieve kennistheorie en de psychometrische theorie maakt het mogelijk om beide terreinen met elkaar te verbinden. De kennistheorie is een theorie voor toetsprestaties waarbij de psychometrische theorie als hulptheorie wordt gebruikt om de voorspelde relaties toetsbaar te maken. Eenzelfde intermediaire rol kan de psychometrische theorie vervullen in het cognitief psychologisch onderzoek zelf, waar ze de kansen op het optreden van bepaalde cognitieve relaties kan formaliseren en dus voorspellingen toetsbaar kan maken.

7.5 Bespreking onderzoeksresultaten vanuit het psychometrische perspectief

De drie gebruikte psychometrische (of statistische) modellen gaven voor de meeste items een adequate beschrijving. Dit betekent dat de antwoorden van de student psychometrisch gezien door één latente variabele worden verklaard. Voor alle modellen geldt, dat de kans op een goed antwoord alleen afhankelijk is van de moeilijkheid van het item en van de cognitieve vaardigheid van de persoon.
Het VGT-model gaat er van uit dat itemmoeilijkheid een stochastische normaal verdeelde variabele is, waarvan alleen de gemiddelde moeilijkheid en de variantie geschat behoeven te worden. In het Raschmodel is itemmoeilijkheid een parameter, waarvan de waarde voor elk item wordt geschat. Dit laatste is ook het geval in het Mokkenmodel. Echter hier wordt noch voor de persoon noch voor het item de positie op de latente variabele geschat. Dit vanwege het ontbreken van een specifieke functionele relatie tussen de persoon, het item en de kans op een goed antwoord.
Het opvatten van itemmoeilijkheid als een stochastische variabele in het VGT-model heeft als voordeel dat een groot aantal items op een eveneens groot aantal tijdstippen simultaan geanalyseerd kunnen worden en dat de positie van de personen op één latente variabele op verschillende tijdstippen kan worden geschat. Het nadeel van het VGT-model is uiteraard dat de itemparameters niet worden geschat. Juist deze schatting is voor het opbouwen van een grote gecalibreerde itembank essentieel.
Het beschikbaar komen van het VGT-model is psychometrisch gezien een vooruitgang. Met dit model wordt het immers mogelijk om over een groot aantal tijdstippen de individuele groei in één of andere latente variabele meetbaar te maken. Daarmee is een bijdrage geleverd aan een van de problemen op het gebied van de psychometrie; het meten van verandering. Met behulp van dit model is het verschil in groei op twee itemgroepen beantwoordbaar gebleken. Alleen eerstejaars groeiden sneller op de Eerstejaarsitems. In alle andere jaargroepen nam de kennis meer toe op de Algemene Kennisitems. Deze constatering op het psychometrische niveau is in overeenstemming met de inhoudelijke theorie.
Het op psychometrisch niveau beschikbaar komen van een meetprocedure voor de individuele groei maakt het beantwoorden van allerlei nieuwe inhoudelijke vragen mogelijk zoals: met welke inhoudelijke variabelen hangt de groei samen; is groei door het curriculum te beïnvloeden; zijn er in andere fases van het curriculum snel en langzaam groeiende kennisgebieden; kan er op basis van de groeiparameter een beslissingsmodel worden geconstrueerd?
Het is de vraag of de caesuur beslissingen met name in het eerstejaar kunnen worden geoptimaliseerd door onderscheid te maken tussen resultaten op de "eenvoudige" toets en die op willekeurige vragen. Het in dit onderzoek gerapporteerde voorspellend vermogen van Eerstejaarstoets en Algemene Kennistoets

biedt met betrekking tot optimale beslissingsprocedures geen oplossing. Om deze vraag te beantwoorden is verder inhoudelijk en psychometrisch onderzoek noodzakelijk. Docenten en studenten zou kunnen worden gevraagd om voor verschillende kennisniveaus gemeten met een Eerstejaarstoets en met een Algemene Kennistoets, hun subjectieve waardering voor het slagen of zakken voor de propedeuse aan te geven. Het vergelijken van de resulterende empirische utiliteitsfuncties zou dan uitsluitsel kunnen geven over een eventueel verschil in acceptabiliteit van de beslissingen. Op het psychometrische niveau zijn met betrekking tot het bepalen van de caesuur al verschillende procedures beschreven (zie o.a. Van der Linden en Mellenbergh, 1977, en Van der Gaag, 1987).
De precieze relatie tussen het VGT-model en het Rasch-model is nog niet duidelijk. Echter gezien het kleine verschil tussen de logistische verdelingsfunctie en de normale verdelingsfunctie, in het Raschmodel respectievelijk het VGT-model, zouden de empirische verschillen klein moeten zijn. Raschitems, geanalyseerd met het VGT-model, zouden voor alle waarden van de latente variabele een gemiddelde kans op een goed antwoord moeten geven welke ongeveer gelijk is aan de verwachte proportie goed onder het VGT-model.
Omgekeerd, als gebleken is dat het VGT-model de proportie goede antwoorden in een itempopulatie goed beschrijft, dan zouden de meeste van deze items in het Raschmodel moeten passen. Biedt het VGT-model geen goede beschrijving, dan zouden slechts weinig items uit de onderzochte populatie voldoen aan de assumpties van het Raschmodel.
Met andere woorden het zoeken naar Raschhomogene items zou met behulp van het VGT-model vergemakkelijkt kunnen worden. De onderzoeksstrategie zou kunnen worden VGT als "voorwas" voor Rasch!
Langs deze weg zou de relatie tussen beide modellen verder empirisch onderzocht kunnen worden.
Items die een acceptabele basis vormen om met behulp van het VGT-model de proportie goede antwoorden in een itempopulatie goed te voorspellen, zouden in ieder geval moeten voldoen aan de assumpties van het Mokkenmodel, dat immers als een algemener model kan worden opgevat. De relatie tussen het Mokkenmodel en het Raschmodel is nog steeds niet opgehelderd (zie ook hoofdstuk 6). Wel lijkt duidelijk dat de opvatting dat alleen "niet zinvolle" Mokkenschalen vanwege de geringe variantie in de personenschaal, niet in het Raschmodel zullen passen (Molenaar, 1982 c en d) in deze studie geen empirische ondersteuning vindt.
Dit werd bij toeval ontdekt door de Raschitems op het vierde tijdstip, met behulp van het Mokkenmodel te analyseren. De Raschschalen bleken geen Mokkenschalen, terwijl wel aan de eis van Molenaar (1982, c en d) n.l. voldoende spreiding in de persoonsparameter (zie hoofdstuk 6 figuur 6.7) was voldaan.
De relatie tussen de verschillende modellen is op het niveau van de psychometrische theorie nog geenszins duidelijk. Wellicht dat het VGT-model, dat

als parametrisch model tussen het Mokkenmodel en het Raschmodel in staat, de relaties tussen de modellen in empirische zin kan verduidelijken. Pas dan kan ook duidelijk worden wat het inhoudelijk gezien nu precies betekent of een item al dan niet in een model past.
Van items die hetzij in het Raschmodel passen, hetzij in het Mokkenmodel, is in ieder geval bekend dat de kans op een goed antwoord een monotone functie is van de totaalscore. Voor Raschitems geldt bovendien dat de ordening van personen op alle items dezelfde is (specifieke objectiviteit). Gezien het uitgevoerde monotonie-onderzoek (zie hoofdstuk 6) geldt dat ook min of meer voor de Mokkenitems.
Verder is bekend dat per tijdstip Raschschalen onderling, Mokkenschalen onderling en bovendien Raschschalen en Mokkenschalen zeer hoog correleren.
Naast de hoog gecorreleerde totaalscores, kon worden geconstateerd dat aan Eerstejaars-Raschitems en Algemene Kennis-Raschitems één latente variabele ten grondslag ligt. Wat dit betreft waren de psychometrische theorie en de inhoudelijke theorie met elkaar in overeenstemming: beide voorspelden immers hetzelfde.
Verder bleek dat de Eerstejaarsitems voor de lagere niveaus van de variabele efficiënter (= meer informatie bevatten) waren dan de Algemene Kennisitems, die bij de hoogste niveaus efficiënter waren. Dit deel van het onderzoek leidt tot de inhoudelijke vraag of het zinvol zou zijn om de voortgangstoetsen naar itemmoeilijkheid samen te stellen. Men zou bijvoorbeeld bij ieder item de moeilijkheidsgraad aan kunnen geven. De bedoeling is om de student te helpen bij zijn beslissing om het item te beantwoorden, dan wel het vraagteken alternatief te kiezen. Alleen het fout beantwoorden van te moeilijke items, zou dan moeten worden afgestraft.
De redelijke fit van het Partial Credit-model (PC-model, zie paragraaf 5.3.3) voor de trichotoom gescoorde Raschitems en van het Mokkenmodel voor items met meerdere categorieën, wijst er op dat de kansen op een goed, fout of vraagteken antwoord afhangen van één latente variabele. Hierbij moet worden aangetekend dat deze conclusie geldt onder het gehanteerde scoringssysteem, waarbij een fout antwoord wordt afgestraft. Toepassing van het in dit onderzoek gehanteerde scoringssysteem (zie hoofdstuk 6) zal vrij snel leiden tot een andere invulstrategie. Niettemin biedt het PC-model, ook al is het onrealistisch als practisch scoringsmodel, interessante mogelijkheden om goede, foute en weet-niet antwoorden te onderzoeken. Interessant is daarbij vooral het opsporen van items en respondenten, die zich niet conform het gespecificeerde model gedragen.
Voornaamste resultaat van de toepassing van het PC-model in dit onderzoek is de constatering dat de kans op een fout antwoord stijgt en vervolgens weer daalt in de latente variabele. Dit betekent dat vooral middelmatige studenten de meeste kans op een fout hebben. De echt goede studenten hebben de meeste kans op een goed antwoord en de beginnende student heeft de meeste kans op een vraagteken.

Tot slot van deze bespreking van de resultaten vanuit een psychometrisch perspectief nog de vraag naar de bruikbaarheid van de voortgangstoets als een instrument om eerstejaars mee te beoordelen. Of dit psychometrisch gezien verantwoord is, hangt af van het cognitieve niveau van de betreffende eerstejaars student.
Met behulp van de resultaten van het VGT-model kan er een schatting worden gemaakt van de proportie goede antwoorden op het Algemene Kennis deel en op het Eerstejaars deel van de voortgangstoets. Er vanuit gaande dat er op ieder kennisniveau ook wel wat fouten gemaakt worden levert de schatting van de proportie goede antwoorden een ondergrens voor de per kennisniveau te verwachten toetslengte.
Voor een aantal kennisniveaus die in dit onderzoek zijn geobserveerd is dit berekend met behulp van tabel 6.3 en figuur 6.2, uitgaande van een toetslengte van 50 items in een Eerstejaarstoets en van 200 items in een Algemene Kennistoets. Voor het hoogste kennisniveau levert dit een schatting van de individuele toetslengte gemiddeld over de vier tijdstippen op van grofweg 89 items; voor het niveau op het derde kwartiel 29 items; voor het gemiddelde niveau 21 items; voor het kennisniveau op het eerste kwartiel 16 items en tenslotte voor het laagste kennisniveau een geschatte toetslengte van globaal 14 items.
Aannemende dat niet alle items precies op het kennisniveau van de respondent zijn afgestemd, en dus niet de hoogst mogelijke informatiewaarde voor het betreffende kennisniveau hebben, moet de conclusie zijn dat deze "self-tailored" toetsen aan de korte kant zijn. Dit geldt met name voor de laagste kennisniveaus. Immers uit de grafische weergave van de gemeten variabele bleek dat voor deze lage kennisniveaus de meeste items te ver van het geschatte kennisniveau verwijderd liggen, waardoor zij zich niet echt goed met deze items kunnen meten.
Inhoudelijk gezien hoeft dit misschien niet zo'n probleem te zijn omdat de constatering dat iemand heel weinig items kan beantwoorden op zichzelf al betekenis heeft. Vanuit een psychometrisch perspectief is het gebruik van een korte, weinig informatieve toets wèl een probleem.

7.6 Bespreking onderzoeksresultaten vanuit een data-analytisch perspectief

De data-analytische theorie die in dit onderzoek werd gebruikt was de theorie over longitudinale data, geconcretiseerd in de simplex modellen en varianten daarvan (Guttman, 1954 en Jöreskog, 1970). Deze theorie voorspelt dat de correlaties tussen een variabele die op verschillende tijdstippen wordt gemeten, een in de tijd dalend patroon te zien geven, dat door een simplex model beschreven kan worden. In deze theorie is er voor ieder tijdstip één latente variabele, waarvan er één of meerdere indicatoren kunnen zijn. In het onderhavige onderzoek waren er twee indicatoren voor de latente variabele per tijdstip. Deze latente variabele verklaart in de

data-analytische theorie de gemeenschappelijke variantie tussen de indicatoren. De inhoudelijke theorie voorspelde sterke samenhang tussen de indicatoren. Vandaar dat de data-analytische theorie een, voor de inhoudelijke vraagstelling, bruikbare theorie leek. De inhoudelijke voorspellingen werden op het data-analytische niveau bevestigd. Dit betekent dat Eerstejaarstoetsen en Algemene Kennistoetsen ook op het data-analytische niveau, met aparte analyses voor eerstejaars en zesdejaars opgevat kunnen worden als metingen van één onderliggende latente variabele. De psychometrische theorie en de data-analytische theorie wijzen in dezelfde richting. Beide itemgroepen meten één construct, dat gezien de inhoud van de item totdat het tegendeel blijkt, inhoudelijk geïnterpreteerd kan worden als "medische kennis". De extra toets-methode-afhankelijke latente variabele, die de fit van het model voor de geobserveerde samenhang in het eerstejaar verbeterde, lijkt een interessante "nieuwe" variabele te zijn, die nader inhoudelijke en psychometrische beschrijving behoeft. Voorlopig werd deze variabele geïnterpreteerd als de neiging bij respondenten om moeilijke items te beantwoorden. Een andere formulering voor hetzelfde fenomeen zou zijn "de neiging tot risikogedrag".

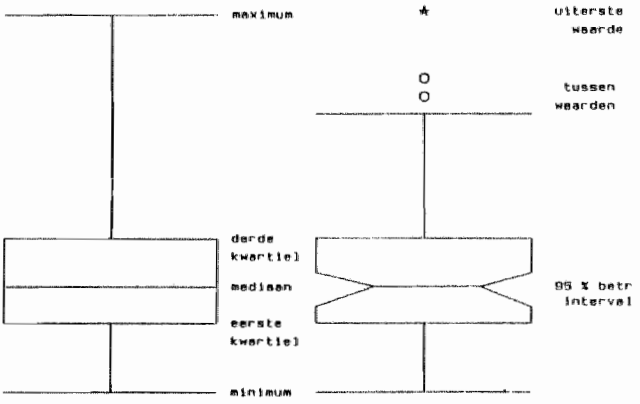
7.7 Slotbeschouwing

In het voorgaande is heel wat geschreven over observaties, onderzoek en resultaten. Het begin was een inhoudelijke vraagstelling. Daarbij werd een domein van observaties gecreëerd, waarvan verwacht werd, dat daarmede de inhoudelijke vraagstelling beantwoord zou kunnen worden.
Gaande het onderzoek is veel duidelijk geworden. Veel echter vraagt om nieuw onderzoek, zowel inhoudelijk, psychometrisch als data-analytisch. Het in dit onderzoek gecreëerde empirische domein is nog lang niet in voldoende mate beschreven. Het vormt een uitdaging voor nieuw onderzoek. Juist omdat de structuur ervan in verschillende opzichten duidelijk is. Dit kenmerk maakt verdere theorievorming mogelijk over het antwoord gedrag van studenten (en artsen) op voortgangstoetsitems. In dit opzicht is het uitgevoerde en gerapporteerde onderzoek vruchtbaar gebleken.

Hoofdstuk 7: Algemene discussie

Snorredozen (Box-and-wisker plots)

Om eendimensionale data te beschrijven worden gewoonlijk het gemiddelde, de standaarddeviatie en andere elementaire kenmerken van een verdeling berekend. In de jaren zeventig introduceerde Tukey (1977) een nieuwe manier om data weer te geven. In zijn boek worden eenvoudige en krachtige technieken besproken voor het exploratief analyseren van gegevens.
Een van die technieken is de Snorredoos (box-and-wisker plots).
Voor een eenvoudige snorredoos zijn vijf waarden van een databestand voldoende: de extreme waarden, het eerste en het derde kwartiel en de mediaan. In figuur 1.1 is een grafische voorstelling van een eenvoudige snorredoos weergegeven. Figuur 1.2 is een weergave van een wat ingewikkelder snorredoos. Hierin zijn twee verschillende "uitbijters" gelocaliseerd: uiterste waarden en tussenwaarden. Globaal kan worden gezegd dat in een normale verdeling uiterste waarden buiten het 99.999 percentiel liggen. Tussenwaarden liggen tussen het 99.3 en het 99.999 percentiel. De inkeping in het midden van figuur 1.2 representeert het 95% betrouwbaarheids-interval voor de mediaan.



Figuur 1.1 en 1.2

Referentie: Tukey, J.W. (1977). **Exploratory data analysis.** Addison Wesley

Bijlage 1

DE VOORTGANGSTOETS

achtergrond, constructie, itembanking,

mogelijkheden, problemen

Bijlage 2

DE VOORTGANGSTOETS

INHOUD

1. Inleiding
2. Het idee
3. Het begrip voortgangstoets
4. De blauwdruk van voortgangstoetsen
5. De samenstelling van voortgangstoetsen
6. Rapportage
7. Het bewaren van gegevens en itembanking
8. Mogelijkheden met voortgangstoetsen
9. Toetsbaarheid van voortgangstoetsen
10. Acceptabiliteit voor studenten
11. De kwaliteit van de voortgangstoets als meetinstrument

1. Inleiding

In deze bijlage wordt beschreven wat het begrip voortgangstoets inhoudt. Allereerst wordt aangegeven hoe het idee van de voortgangstoets werd geboren.Daarna worden de kenmerken en tenslotte de mogelijkheden van, en de problemen met de voortgangstoetsen beschreven.
Voortgangstoetsen vormen slechts een onderdeel van het evaluatiesysteem van de Faculteit der Geneeskunde. De kenmerken van het evaluatiesysteem werden elders beschreven (Verwijnen en Imbos, 1982). De ontwikkeling van dit systeem, vanuit een historisch perspectief, is uitvoerig beschreven door Imbos (1988).

2. Het idee

Gedurende de eerste twee jaar van het bestaan van de Faculteit der Geneeskunde der Rijksuniversiteit Limburg bestond de voortgangstoets niet. Om de prestaties van de studenten te beoordelen werden er alleen bloktoetsen gebruikt. Dit zijn toetsen die na afloop van een periode van thematisch onderwijs worden afgenomen. Ze toetsen de doelstellingen van de betreffende onderwijsperiode. Aan het eind van een studiejaar werd er een zogenaamde jaartoets afgenomen. Een jaartoets bestond uit toetsvragen afkomstig uit eerder in het jaar afgenomen bloktoetsen. De resultaten op de jaartoets en op de bloktoets, bepaalden gezamenlijk de studievoortgang. Met andere woorden, beide toetsen hadden een beslissend karakter. (Zie voor meer details Imbos, 1988).

Het laatst genoemde aspect van de bloktoetsen, leidde tot kritiek van de kant van studenten (en docenten). De uitgangspunten van het onderwijsprogramma: de zelfwerkzaamheid en de probleemoriëntatie, bleken zich moeilijk met een beslissende blokevaluatie te verdragen (Wijnen, 1978 en Imbos en Verwijnen, 1978). In januari 1976 reageerde Wijnen op de naar voren gebrachte kritiek met een nota over de evaluatieproblematiek (Wijnen, 1976). Dit rapport beoogde de discussie uit de emotionele sfeer te halen, door het aanreiken van een wetenschappelijk begrippenkader. Tegen de achtergrond van algemene wetenschappelijke eisen met betrekking tot de kwaliteit van de informatie uit studietoetsen (Ebel, 1966), werd de eigenlijke discussie gevoerd.
In de discussie over het evaluatiesysteem zouden volgens Wijnen drie problemen moeten worden onderscheiden en separaat opgelost. In de eerste plaats het meetprobleem, met als nadere onderscheidingen vragen met betrekking tot wat men wil meten en hoe men wil meten. In de tweede plaats het waarderingsprobleem, met als kernvraag welke waarde men aan de verkregen meetresultaten wenst toe te kennen. En tenslotte het beslissingsprobleem, met als kernvraag welke beslissingen omtrent individuele studenten men van plan is te nemen, nadat het meetprobleem en het waarderingsprobleem adequaat zijn opgelost.
Vervolgens somde Wijnen een aantal uitgangspunten voor een evaluatiesysteem op, waarvan hier alleen het laatste van belang is:

> "beschikbaarheid van eindtermen hoeft niet noodzakelijkerwijs gecompleteerd te worden door beschikbaarheid van eindtermen voor programmaonderdelen" (Wijnen, 1976).

Voor de evaluatie betekende dit, dat er gekozen zou moeten worden tussen twee mogelijkheden. Of men kiest voor een vorm van evaluatie die nauwkeurig is afgestemd op de eindtermen, die voor het betreffende programmaonderdeel geformuleerd zijn of men kiest voor een vorm van evaluatie, die vooral zichtbaar moet maken, dat er ten aanzien van het totale einddoel kwantitatieve voortgang is geboekt. In het laatste geval zou men kunnen volstaan met een verzameling problemen, die het einddoel zo goed mogelijk representeren.
De auteur vermeldde voorzichtig dat hij "enige affiniteit" voelde met het laatst geformuleerde standpunt; "het geregeld aanspreken van wenselijk geachte kennis en kunde, verdient de voorkeur boven het achtereenvolgens "aanbieden" en toetsen van wenselijk geachte kennis en kunde".
Met dit standpunt werd stelling genomen tegen het in bijna elke onderwijsinstelling vigerende systeem van vrijstellende tentamens. "Het afstreepeffect dat feitelijk uit kan gaan van vrijstellende tentamens, ontslaat iedereen van de verplichting om de gedemonstreerde kennis ook funktioneel te houden. Door een appèl, bij voortduring aan een breed scala van kennis en kunde, lijkt het mogelijk de wenselijkheid van

funktionerende kennis meer naar voren te halen, eventueel ten koste van het "niveau" van deze kennis. Het is immers niet onwaarschijnlijk, dat het "niveau" van voortdurend aanwezige kennis niet vergelijkbaar is met het "niveau" van kennis dat bij gelegenheid van een tentamen of toets extra is opgepoetst" (Wijnen, 1976).
Hoewel er in het bedoelde rapport uit januari 1976 met geen woord werd gerept over de voortgangstoets, is het duidelijk dat met de aangehaalde citaten, het idee achter de voortgangstoets in zijn huidige verschijningsvorm, was geboren.
Terwille van het oplossen van de problemen met de op dat moment vigerende blokevaluatie werd gezocht naar een systeem van toetsingen dat minder sturend zou zijn. Dit laatste werd vereist door de gekozen uitgangspunten van het curriculum.

3. Het begrip voortgangstoets

De voortgangstoets, zoals die sinds 1977 vier keer per studiejaar wordt afgenomen kan als volgt worden omschreven: [1]

> Een voortgangtoets is een schriftelijke toets, bestaande uit ongeveer 250 juist-onjuist vragen, die een gestratificeerde aselecte steekproef vormen uit het gehele gebied van de medische wetenschap, dat door een basisarts beheerst dient te worden. De afname vindt vier keer per studiejaar plaats en deelname aan de toets is verplicht voor studenten uit alle studiejaren (eerstejaars tot en met zesdejaars).

Een student die het onderwijs in Maastricht volledig volgt, neemt aldus tenminste vierentwintig keer deel aan een voortgangstoets, zodat kan worden vastgesteld hoe de groei de groei in kennis verloopt
De studenten kunnen bij het beantwoorden van de toetsvragen gebruik maken van het vraagteken-alternatief. De scoring van de toetsen geschiedt door middel van de goed minus fout scoring. Met andere woorden: een goed antwoord levert één punt op, een fout antwoord min één, en een vraagteken levert geen punten op.
De voortgangstoets wijkt op een drietal punten af van een gewone studietoets. Het eerste verschil met een gewone studietoets zit hierin dat voortgangstoetsen, het einddoel trachten te meten. Zoals bekend, worden studietoetsen doorgaans gebruikt om tussentijdse doelstellingen van het onderwijsprogramma te meten. Dit, op zich subtiele, verschil tussen voortgangstoetsen en gewone studietoetsen, heeft een aantal verstrekkende gevolgen. De beslissing om alleen het einddoel van een

---

[1] Hier is gekozen voor domein specifieke definitie. Men kan het begrip voortgangstoets uiteraard ook zonder verwijzing naar een bepaald domein definiëren

zesjarige opleiding te meten, leidt nagenoeg vanzelfsprekend ook tot de beslissing om herhaald te meten.
Het einddoel van een opleiding hoeft uiteraard pas aan het eind van het onderwijsprogramma gerealiseerd te zijn. In principe zou men dus de vaststelling van het al dan niet bereikt hebben van het einddoel kunnen beperken tot de eindfase van de opleiding. Overwegingen van acceptabiliteit (De Groot, 1982, 1986) maken dit echter ongewenst. Zonder dat men regelmatig (dus tussentijds) de leerling of de student op de hoogte heeft gesteld van de stand van zaken wat betreft zijn/haar kennis, is het niet verantwoord om een eindoordeel te geven. Voortgangstoetsen kunnen dus alleen zinvol worden toegepast indien men door middel van herhaalde metingen het einddoel van de opleiding toetst.
Een tweede verschil volgt uit het principe van herhaalde of longitudinale meting van het einddoel van een opleiding. Herhaalde meting van het einddoel in een opleiding die zich over meerdere jaren uitstrekt houdt in, dat men gelijktijdig de beschikking krijgt over de prestaties van groepen die qua gevorderdheid nogal wat van elkaar verschillen. Men kan met behulp van één meetinstrument oudere- en jongerejaars studenten vergelijken.
Voor zowel de studenten als de docenten levert dit nuttige informatie op. De docent krijgt als het ware in één meting een (zij het in dwarsdoorsnede) overzicht van de ontwikkeling van het vakgebied door het gehele curriculum heen. De student krijgt eveneens informatie ter beschikking die bij een gewone studietoets ontbreekt. Doordat er informatie van alle "niveaugroepen" beschikbaar is, kan iedere student zichzelf als het ware evalueren ten opzichte van het onderwijsprogramma. Een student in het tweede studiejaar met bijvoorbeeld een goedscore van 60% op het vakgebied Anatomie, weet dat het eigen kennisniveau vergelijkbaar is met dat van een gemiddelde zesdejaars student. Een zesdejaars met een goedscore van 20% op hetzelfde vakgebied weet eveneens dat het eigen niveau vergelijkbaar is met dat van een tweedejaars student.
Een derde verschil met de gewone studietoets is dat door middel van een voortgangstoets het gehele kennisdomein, steekproefsgewijs wordt getoetst. Dit kenmerk maakt het mogelijk om verschil in groei tussen vakgebieden te bestuderen en vast te stellen. Dit geeft aan een onderwijsorganisatie een belangrijke evaluatiemogelijkheid met betrekking tot het curriculum als geheel. Immers een eventueel achterblijven van de groei in een bepaald vakgebied, zou men via gerichte onderwijskundige verbeteringen kunnen veranderen. Het effect van dergelijke ingrepen is met behulp van voortgangstoetsen vervolgens weer te kwantificeren.
Deze algemene omschrijving van het begrip voortgangstoets zal in de volgende paragraaf meer inhoud krijgen.

Bijlage 2

4. De blauwdruk van voortgangstoetsen

Om de inhoudelijke parallelliteit van de voortgangstoetsen zo goed mogelijk te garanderen, is in het verleden een indelingssysteem ontwikkeld, dat ontleend werd aan de International Classification of Diseases. Dit systeem bevat 15 categorieën. Het totaal aantal vragen in een voortgangstoets wordt volgens een vaste sleutel over de categorieën verdeeld. Onderstaande tabel geeft een overzicht van de categorieën en de verdeling van de vragen over deze categorieën.

TABEL 1. Blauwdruk van de voortgangstoets. Verdeelsleutel voor het aantal vragen per categorie.

Cat -nr.	Omschrijving	Aantal vragen abs	%
1	Ademhalingsstelsel	28	11
2	Hematologisch en lymfatisch stelsel	15	6
3	Spier- en skeletstelsel	19	8
4	Geestelijke gezondheidszorg	18	7
5	Voortplantingsstelsel	15	6
6	Hart- en vaatstelsel	28	11
7	Hormoonstelsel en metabolisme	16	6
8	Huid en bindweefsel	10	4
9	Maatschappelijke factoren t.a.v. ziekte en gezondheid(szorg)	14	6
10	Spijsverteringsstelsel	22	9
11	Nieren en urinewegen	17	7
12	Zenuwen en zintuigen	21	8
13	Rest	14	6
14	Wetenschapsleer en methodologie	6	2
15	Persoonsgebonden factoren t.a.v. ziekte en gezondheid(szorg)	7	3
	TOTAAL	250	100

De verdeelsleutel die gehanteerd wordt, is indertijd vastgesteld door de gezamenlijke vakgroepen en was gebaseerd op een combinatie van een drietal gegevens:

- een enquête onder de vakgroepen
- morbiditeitsgegevens

- omvang van het gebied in algemeen-medische handboeken (Verwijnen en Fröberg, 1986).

5. De samenstelling van voortgangstoetsen

Uit de verzameling beschikbare vragen worden 250 toetsvragen getrokken. Voor elke toets worden bovendien nog eens 100 reserve vragen getrokken. Binnen iedere categorie wordt er aselect getrokken. De verschillende vakgroepen kunnen dus hun aandeel in de toets beïnvloeden, door er voor te zorgen dat ze in de verschillende categorieën goed vertegenwoordigd zijn. Dit "marktmechanisme" werd destijds ingevoerd om de vakgroepen tot productie te stimuleren.
De stand van zaken met betrekking tot de itembank wordt regelmatig aan de leveranciers van de toetsvragen bekend gemaakt door middel van overzichten, waarvan als voorbeeld tabel 2 is opgenomen.

TABEL 2. Overzicht samenstelling Itembank (datum: 03-02-88)

Catnr Vakg	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Totaal
AE	45	35	164	1	79	66	9	24	0	79	39	72	27	0	0	640
AG	56	6	9	2	4	13	7	7	59	7	1	16	26	37	22	272
AH	8	18	122	0	27	141	38	73	1	220	54	12	18	0	0	732
AN	29	2	0	0	0	23	0	1	0	0	2	6	0	0	0	63
BC	6	71	20	0	8	10	143	0	0	30	14	22	163	0	0	487
BF	15	4	23	0	0	91	3	0	0	11	14	29	19	1	0	210
CAR	0	0	0	0	1	293	0	0	0	0	0	0	3	0	0	297
DER	0	3	0	0	6	2	0	193	0	0	0	0	0	0	0	204
EGZ	0	0	0	0	0	0	0	0	237	0	0	0	0	11	0	248
EP	6	5	6	1	9	16	2	3	7	17	2	3	16	181	2	276
FAR	83	44	36	9	9	81	59	21	0	45	43	61	65	0	0	556
FYS	152	230	65	0	36	55	98	19	0	188	50	238	0	0	0	1131
GC	0	12	5	2	18	0	10	1	0	2	0	3	36	0	0	89
GO	2	7	0	0	282	1	49	0	0	0	7	3	24	0	0	375
GZR	0	2	0	21	10	1	1	0	112	0	1	1	22	0	26	197
HAG	65	18	34	9	54	43	67	38	5	72	41	78	12	10	6	572
IG	61	156	126	1	10	65	208	27	3	283	194	16	64	1	1	1216
IM	38	114	21	0	21	13	16	20	0	24	17	13	9	0	0	306
KG	99	41	22	6	24	43	60	23	8	63	51	80	43	2	2	567
KNO	71	0	17	0	0	0	0	2	2	12	0	204	0	0	0	308
KP	0	1	0	95	0	0	1	0	23	0	0	0	7	3	9	139
ME	0	0	0	14	8	0	0	0	70	0	0	2	1	186	16	297
MM	32	76	3	0	36	2	1	25	13	42	15	33	152	0	1	431
MP	6	0	1	184	5	27	2	1	7	0	1	16	71	21	50	392
MS	0	0	0	10	2	1	0	0	102	0	0	0	80	49	39	283
NEU	0	0	4	1	0	1	0	0	0	0	0	949	0	0	0	955
OHK	0	0	0	0	0	0	1	0	0	0	0	219	0	0	0	220
ORT	0	0	159	0	0	0	0	0	0	0	0	1	0	0	0	160

Bijlage 2

TABEL 2 (vervolg)

PA	20	24	9	0	41	47	37	37	3	50	31	29	45	0	0	373
PUL	201	0	0	0	0	0	0	0	3	0	0	0	0	0	1	205
REV	10	13	109	1	8	23	2	7	11	1	9	70	19	2	2	287
SP	0	0	0	215	2	2	1	0	17	1	0	15	4	0	32	289
TOT	1005	882	975	572	700	1060	815	522	683	1147	586	2191	926	504	209	12777

CAR = Cardiologie
DER = Dermatologie
EGZ = Economie van de Gezondheidszorg
EP = Epidemiologie
FAR = Farmacologie
FYS = Fysiologie
GC = Genetica/Celbiologie
GO = Gynaecologie/Obstetrie
GZR = Gezondheidsrecht
HAG = Huisartsgeneeskunde
ME = Gezondheidsethiek en Wijsbegeerte
MM = Medische Microbiologie
MP = Medische Psychologie
MS = Medische Sociologie
NEU = Neurologie
OHK = Oogheelkunde
ORT = Orthopedie
PA = Pathologie
PUL = Pulmonologie
REV = Revalidatie
SP = Sociale Psychiatrie

Om te voorkomen dat bepaalde vakgroepen als gevolg van deze selectieprocedure systematisch onder- of oververtegenwoordigd worden, is er een onder- en een bovengrens aan het aantal toetsvragen per vakgroep gesteld. Als ondergrens is gesteld 5 toetsvragen en als bovengrens 25 toetsvragen per vakgroep.
Alle toetsvragen die voor een bepaalde toets geselecteerd zijn, worden beoordeeld door de Voortgangstoets Beoordelings Commissie (VBC).
Dat dit noodzakelijk is blijkt wel uit het feit dat gemiddeld 65% van alle vragen alvorens in de toets te kunnen worden opgenomen, aanpassing behoeft, (Verwijnen en Fröberg, 1986). Het merendeel van de problemen met toetsvragen blijkt te bestaan uit vormtechnische problemen. Vaak gaat het daarbij om formuleringskwesties, waardoor een vraag niet eenduidig te beantwoorden is. Ook komt het voor dat de vraag zo is geformuleerd dat er ongewenste en onbedoelde suggesties voor het correcte antwoord in schuil gaan. Uit het overzicht van Verwijnen en Fröberg blijkt verder dat vormtechnische problemen in de toetsvragen nog al eens inhoudelijke problemen met zich mee brengen. Als voorbeeld hiervan noemen de auteurs dat de toevoeging van het woord "altijd" in een vraag inhoudelijk meestal aanvechtbaar is. Van alle toetsvragen die in het studiejaar 1984-1985 door de VBC werden beoordeeld bleek 23% inhoudelijke problemen te geven. Tenslotte bleek 11% van alle beoordeelde vragen tot relevantieproblemen aanleiding te geven. Alleen toetsvragen, die de toets der VBC-kritiek hebben doorstaan dan wel na overleg met de auteurs zijn aangepast, worden uiteindelijk in de toets opgenomen.
Als laatste stap, voordat de resultaten aan de studenten en de faculteit

gerapporteerd worden, houdt de VBC een nabespreking. De ervaring heeft geleerd dat bij de voorbereiding van de toetsen onvermijdelijk zaken over het hoofd worden gezien. Deze nabespreking vindt plaats aan de hand van itemanalyses en commentaren van de studenten. Nogmaals passeert in deze nabespreking elke vraag de revue. Daarbij wordt ieder commentaar beoordeeld en gecontroleerd. Indien men in deze fase alsnog op een probleem stuit, dan vindt er telefonisch overleg plaats met een inhoudsdeskundige contactpersoon van de vakgroep waaruit de betreffende vraag afkomstig is. Zo nodig wordt de vraag alsnog uit de toets verwijderd. (voor meer details zij verwezen naar Verwijnen en Fröberg, 1986)

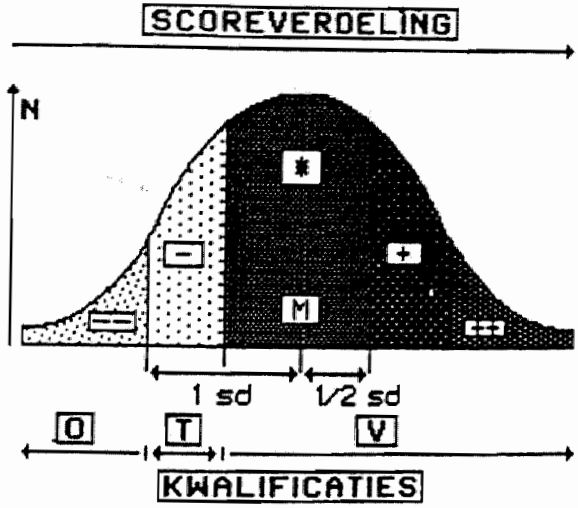
6. Rapportage

Nadat op grond van de genoemde nabespreking de toets definitief is samengesteld, worden de toetsscores van de studenten en de itemindices berekend.
Elke student kan door middel van het uitslagenformulier de eigen score op de verschillende onderdelen van de toets vergelijken met die van de jaargroep.
De beoordeling ten behoeve van de examenbeslissingen gebeurt per toets op basis van de totaalscore. Daarbij wordt een correctie-voor-raden toegepast. Het aantal correcte antwoorden wordt verminderd met het aantal foute antwoorden. Het vraagteken krijgt daarbij de waarde 0. Op het uitslagenformulier wordt de goed-, de fout- en de vraagtekenscore van de individuele student en van de jaargroep als geheel vermeld. Om de positie van iedere score tegen de achtergrond van de jaargroepscore te kunnen plaatsen, worden de scores voorzien van plus en min tekens. De betekenis daarvan is:

(+ +) = relatief zeer hoge score
(+) = relatief hoge score
(*) = gemiddelde score
(-) = relatief lage score
(--) = relatief zeer lage score

Een grafische samenvatting van de waardering van toetsscores is weergegeven in figuur 1.

De vakgroepen ontvangen een overzicht van de door de zes jaargroepen behaalde resultaten. Daarnaast ontvangen ze over iedere toetsvraag, die van de betreffende vakgroep in de toets was opgenomen, de resultaten.



Figuur 1 Gebieden in de verdeling van toetsscores en hun waardering.

7. Het bewaren van gegevens en itembanking

De resultaten van de studenten worden gearchiveerd en op microfiche bewaard. Ten behoeve van de opslag van de gegevens van de items wordt er een geautomatiseerde itembank ontwikkeld, die voor een belangrijk deel reeds operationeel is. De verschillende fases in de ontwikkeling van de itembank werden beschreven door Imbos en Verwijnen (1978), Janssen en Pollemans (1979, 1980, 1981), Pollemans (1982, 1983) en Cobben (1982, 1983).
Een itembank kan globaal gekarakteriseerd worden als een informatie- en documentatiesysteem voor toetsvragen en toetsen met de daarbij behorende moderne electronische en psychometrische hulpmiddelen om de opgeslagen informatie snel en efficiënt ter beschikking te kunnen stellen.
Voornaamste kenmerk van een itembank is dat alle informatie die omtrent individuele items in de loop van de tijd wordt verzameld, wordt opgeslagen en gebruikt ter verbetering van de kwaliteit van het systeem. De wijze waarop dat gebeurt, valt buiten het bestek van deze studie. De moderne electronische hulpmiddelen worden beschreven door Cobben (ibid), terwijl Van der Linden (1983) beschrijft hoe een electronische kaartenbak, met behulp van moderne psychometrische technieken in een itembank kan worden getransformeerd.
Aan de ontwikkeling van de itembank van de Faculteit der Geneeskunde Maastricht

hebben een aantal ontwikkelingen ten grondslag gelegen, die door Pollemans (1983) werden beschreven en die hier kort zullen worden samengevat. De onderwijskundige principes probleem-georiënteerdheid en zelfwaarkzaamheid leidden reeds in een vroege fase van de ontwikkeling van het evaluatiesysteem tot de conclusie dat een ideaal evaluatieprogramma voor dit onderwijs twee tegenstrijdige kenmerken zou moeten bezitten: enerzijds individualiseerbaarheid, om tegemoet te komen aan de keuzevrijheid van de studenten en anderzijds uniformiteit om een minimaal wenselijk geacht eindniveau te garanderen (Imbos en Verwijnen, 1978).
Voor het realiseren van een systeem, dat aan de zojuist omschreven eisen zou voldoen, was automatisering een absolute voorwaarde (Imbos en Verwijnen, 1978). Ontwikkeling van een geautomatiseerd systeem voor toetsen, dat flexibel genoeg is om aan de gestelde eisen te voldoen, is op zich niet meer dan een geavanceerde electronische "itembak". Voor een itembank is nodig dat moderne psychometrische technieken, gebaseerd op item response theorie een onderdeel vormen van het systeem (Van der Linden, 1983). Inderdaad kan men vanuit een psychometrisch gezichtspunt stellen dat het einddoel van alle toetsautomatisering een itembank zou moeten zijn. En daaronder verstaat men meer dan een verzameling toetsvragen: "But a bank of carefully calibrated test items is more than a collection. It is a composition of coordinated questions that develop, define, and quantify a common theme and thus provide an operational definition of a variable" (Wright en Bell, 1984).
Een itembank in de hierboven bedoelde betekenis kan men zich voor het medische domein als volgt voorstellen: Het totale domein wordt inhoudelijk in een aantal deelgebieden ingedeeld. Voor het samenstellen van voortgangstoetsen zouden daarvoor de 15 categorieën gebruikt kunnen worden. Uit elk deelgebied wordt voor iedere toets een steekproef ter grootte van het aantal vragen dat door de blauwdruk wordt voorgeschreven genomen. De toets wordt vervolgens afgenomen en met behulp van een geschikt gekozen item response model wordt onderzocht of de items ééndimensionaal zijn. Blijkt dit het geval, dan wordt de betreffende selectie van items in de bank opgenomen met hun, door middel van het item response model, geschatte schaalwaarde. Dergelijke gecalibreerde deelverzamelingen kunnen op een efficiënte wijze hergebruikt worden voor nieuwe toetsen, waarvan de moeilijkheidsgraad constant gehouden kan worden door gebruik te maken van de itemcalibraties. De nieuwe informatie die bij hergebruik beschikbaar komt, wordt gebruikt om de eerdere itemcalibraties te verbeteren.
Een itembank met deze mogelijkheden, is het uiteindelijke streven van de Faculteit der Geneeskunde. Op dit moment (februari 1988), is het systeem zover ontwikkeld dat items met de daarbij behorende inhoudelijke, administratieve en statistische kenmerken, kunnen worden opgeslagen en hergedrukt. Om dat mogelijk te maken werd er een thesaurus ontwikkeld, met behulp waarvan ieder item medisch inhoudelijk kan worden gekarakteriseerd en beschreven (Pollemans, 1983).

Deze thesaurus werd verder ontwikkeld en vervolmaakt door Van Hessen, Cobben en Verwijnen (1987). Het computersysteem dat de kern vormt van de itembank is ontworpen, in overleg met diverse gebruikers van het systeem, door Cobben (1982). Dit systeem wordt gekenmerkt door een grote mate van flexibiliteit. Men kan zich met behulp van de thesaurus op diverse wijzen toegang verschaffen tot de items. De enige beperking is gelegen in de kenmerken die aan de items zijn toegekend. Als dit goed is gebeurd kunnen door middel van het systeem alle items geselecteerd worden die men in een toets wil opnemen. Het systeem is gebruikersvriendelijk doordat er gebruik is gemaakt van een menu-schema, met behulp waarvan de gebruiker de specifieke wensen kenbaar kan maken.
Met de itembank in zijn huidige staat heeft de Faculteit der Geneeskunde als enige faculteit in Nederland de beschikking over een krachtig instrument om voor diverse doeleinden toetsen samen te stellen uit het gehele gebied van de medische wetenschap. De vervolmaking van het systeem zal bestaan uit het koppelen van de moderne psychometrische technologie aan het nu reeds bestaande systeem. Daarbij zal dankbaar gebruik worden gemaakt van het werk dat op dit terrein door anderen wordt verricht (Keizer, Van Thiel en Zwarts, 1985, Sluiter, 1986, Kelderman, 1986, Van der Linden en Zwarts, 1986, Knol, 1986, Buning, Eggen, Kelderman en Van der Linden, 1986). Hoewel het systeem werd ingegeven door de specifieke eigenschappen van het onderwijssysteem is het ook te gebruiken door andere Geneeskunde Faculteiten. Een samenwerking van verschillende faculteiten gericht op het ontwikkelen van een gezamenlijke itembank zou zeer aan te bevelen zijn. Itembanken voorkomen dat de informatie die bij iedere toets ter beschikking komt verloren gaat. Samenwerking tussen faculteiten met betrekking tot het verder ontwikkelen van de itembank van Maastricht tot een landelijk systeem, biedt tevens de mogelijkheid om de kwaliteit van het medisch onderwijs in ons land te bepalen. Een dergelijke samenwerking tussen de medische faculteiten in Nederland zou door de overheid, die de kwaliteit van het onderwijs zo zeer ter harte gaat, moeten worden gestimuleerd.
De itembank bevatte in de maand januari 1988 12777 toetsitems, alle afkomstig uit het medische domein.

8. Mogelijkheden met voortgangstoetsen

Het herhaald gebruik van een toets, die de einddoelen in het kennisdomein van een opleiding representeert, biedt een aantal mogelijkheden die het normale gebruik van studietoetsen niet heeft. Deze mogelijkheden zullen in deze paragraaf aan de orde worden gesteld.

Geen structurele studievertraging

Gesteld dat voortgangstoetsen tenminste voldoen aan de psychometrische kwaliteitseisen voor studietoetsen en tests, zoals onder meer beschreven door Visser (1982), bieden een alternatief voor de in veel onderwijsinstellingen gebruikelijke reeks van vrijstellende tentamens. Een systeem dat door De Groot (1972) en Wijnen en Van der Vleuten (1985) wordt getypeerd als een "hordenloop". "Een reeks van toetsen en tentamina wordt door de studenten van eenzelfde studierichting in nagenoeg dezelfde volgorde afgelegd. Heeft men voldaan aan de eisen van een tentamen, dan heeft dat veelal tot gevolg, dat de betreffende leerstof in het toetsprogramma niet meer als zodanig voorkomt".
Centraal in een dergelijk systeem staat het opdelen van de wenselijk geachte prestaties in verschillende deelprestaties, die door iedereen in een gelijke volgorde moeten worden afgelegd. Het niet halen van een tentamen heeft doorgaans voor de betreffende student, in tegenstelling tot een echte hordenloop, verstrekkende gevolgen. Studievertraging en, misschien zelfs wel studiestaking, zijn noodzakelijke consequenties van een dergelijk systeem, leidend tot slechte rendementcijfers van de betreffende onderwijsinstelling.
Een systeem met voortgangstoetsen biedt hiervoor een alternatief. Door in de toetsen niet meer de afzonderlijke vakken centraal te stellen, wordt het mogelijk om bij een zelfde eindniveau, een hoger studierendement te halen. Vertragingen, die het gevolg zijn van een tekortkoming in een bepaald vakgebied, kunnen in een systeem met voortgangstoetsen worden vermeden. Herhalingstentamens die noodzakelijk leiden tot studievertraging zijn bij een stelsel van voortgangstoetsen niet langer noodzakelijk. Wanneer een student bij een volgende voortgangstoets kan laten zien dat de beoogde prestatie geleverd wordt, dan is er geen enkele reden om overbodige vertragingen structureel in het studieprogramma op te nemen. Studenten komen de eigen resultaten als het ware altijd weer tegen. Komt er dan een moment waarop de resultaten van een vierdejaars student niet meer onderscheiden kunnen worden van die van een derdejaars student, dan heeft men met een dergelijke constatering overtuigende argumenten in handen om gericht tot actie over te gaan. Kortom met voortgangstoetsen heeft een faculteit "een krachtig middel voor kwaliteitscontrole" (Van der Linden, 1987), zonder structureel ingebouwde vertragingsmechanismen.

Feedback en studiebegeleiding

Voortgangstoetsen bieden verder de mogelijkheid om studenten inzage te geven in hun sterke en zwakke gebieden in het totale domein van de medische kennis. Door de resultaten die met voortgangstoetsen worden behaald zorgvuldig te analyseren, doen zich een aantal mogelijkheden voor die het leerproces van de studenten

richting kunnen geven. De volgende mogelijkheden worden feitelijk voor dit doel aangewend:

- Voortgangstoetsen kunnen door de studenten zelf als leermiddel worden gebruikt. Het toetsboekje met de vragen wordt door de studenten meegenomen. Bovendien wordt na afloop van iedere toets de antwoordsleutel beschikbaar gesteld. Iedere student kan daardoor zelf vast stellen welk resultaat feitelijk werd bereikt. Tenslotte kan een voortgangstoets als leermiddel fungeren, omdat bij iedere toetsvraag een literatuurverwijzing is afgedrukt, zodat iedereen de bij de vraag behorende redenering of gedachtengang nog eens kan nalezen of studeren.
- Zoals al eerder werd opgemerkt wordt het resultaat niet alleen gegeven in de vorm van een totaal waardering. De studenten krijgen kort na de afname hun eigen resultaten op gedetailleerde wijze verstrekt. Naast het totaal resultaat, worden er gegevens verstrekt over de resultaten, die op de verschillende categorieën van de blauwdruk en de verschillende disciplines, werden bereikt. De resultaten worden daarbij gerapporteerd in termen van procentuele goed-, fout- en vraagtekenscores. Om een kader te geven voor het beoordelen van de verschillende deelresultaten, worden het gemiddelde en de standaarddeviatie van de toetsscores die door de jaargroep waartoe de student behoort, eveneens gerapporteerd.
  Op die manier kan de student inzicht krijgen in de sterke en zwakke punten in vergelijking met degenen, die zich in dezelfde fase van de studie bevinden. De verder te volgen studiestrategie kan dan op deze bevindingen worden afgestemd.
- Verder worden de resultaten van iedere student op de voortgangstoetsen periodiek beoordeeld door een studiebegeleider. Iedere student heeft een studiebegeleider en tenminste eenmaal per jaar vindt er een gesprek plaats. Dit geschiedt niet alleen op basis van de voortgangstoetsresultaten. Ook de resultaten, die op de andere informatiebronnen werden behaald, worden in verband gebracht met de voortgangstoetsresultaten. Een grondige analyse vooraf door de begeleider, kan daarbij interessante zaken aan het licht brengen. Verwijnen (1987) laat zien dat studieresultaten "onvermoede diagnostische mogelijkheden in zich bergen" (blz. 109). In een vijftal casus toont Verwijnen aan dat een zorgvuldige analyse van de studieresultaten indicaties kunnen geven over de studieaanpak, inhoudelijk sterke en zwakke punten van de studenten en in een enkel geval zelfs over persoonlijke en situationele achtergronden. Vooral het naast elkaar bestaan van bloktoetsen en voortgangstoetsen acht Verwijnen van belang voor een gedetailleerd inzicht in de studieaanpak van studenten. Door beide soorten toetsen naast elkaar te hanteren, is er behalve informatie over de kennisontwikkeling op de korte termijn, ook informatie beschikbaar over kennisontwikkelingen op

de langere termijn. Dat beide ontwikkelingen lang niet altijd parallel verlopen wordt door Verwijnen beschreven.

Evaluatie van het onderwijsprogramma

Bij een averechts gebruik (Hofstee, 1969) bieden voortgangstoetsen mogelijkheden voor de evaluatie van het onderwijsprogramma die gewone studietoetsen niet, of in veel mindere mate, hebben. Omdat binnen een toets de scores op de verschillende categorieën uit de toetsblauwdruk, op een zelfde schaal gezet kunnen worden, kan men de effectiviteit van het onderwijsprogramma voor de verschillende deelgebieden bestuderen.
De groei in de verschillende gebieden van de blauwdruk kan worden gekwantificeerd. Ook kan worden nagegaan of het beoogde niveau voor alle gebieden ook daadwerkelijk wordt bereikt. Vastgesteld kan worden, of de vooruitgang zoals die door de studenten wordt bereikt een afspiegeling is van de intenties, die aan het onderwijsprogramma ten grondslag liggen. Zo blijkt bij iedere voortgangstoets opnieuw dat de kennis van de studenten van de klinische vakgebieden in de eerste vier jaren van het onderwijsprogramma achter blijft bij de kennis van de basisvakken.
Naast het vergelijken van verschillende jaargroepen binnen eenzelfde onderwijsprogramma bieden voortgangstoetsen ook de mogelijkheid om jaargroepen van vergelijkbare programma's te bestuderen. Dit is mogelijk omdat een voortgangstoets het einddoel van een medische opleiding tracht te meten, ongeacht het specifieke programma dat aan de realisering van dat einddoel ten grondslag heeft gelegen. Door een dergelijke vergelijking kan er zicht komen op de sterke en zwakke punten van de programma's. Bosschert (1987) doet verslag van een dergelijke exercitie. Zijn analyses waren gebaseerd op een voortgangstoets die zowel in Leiden (bij tweede- en derdejaars studenten) als in Maastricht werd afgenomen. Naast het vergelijken van de twee jaargroepen uit Leiden, in de zin zoals hierboven werd beschreven, vergelijkt hij ook de resultaten van de Leidse studenten met die van de studenten uit Maastricht. In zijn studie laat hij zien in welk opzicht de resultaten van Leiden en Maastricht van elkaar verschillen.
Deze verschillen zijn met name interessant omdat eerder onderzoek (Bender, Cohen-Schotanus, Imbos, Versfelt en Verwijnen, 1984) al aantoonde dat de studieprestaties van verschillende medische faculteiten in het laatste jaar van de opleiding weer vergelijkbaar zijn. Voortgangstoetsen blijken ook als het gaat om de evaluatie van onderwijsprogramma's bruikbare instrumenten te zijn.

Bijlage 2

Longitudinale evalutatie van studenten en onderwijsprogramma

Voortgangstoetsen bieden uiteraard ook de mogelijkheid om vakgebieden, curricula of studenten longitudinaal te volgen. Omdat voortgangstoetsen steeds volgens eenzelfde blauwdruk worden samengesteld en vier keer per studiejaar worden afgenomen, heeft men aan het einde van de studie in de geneeskunde van de meeste studenten de beschikking over 24 scores op de verschillende deelgebieden. Verschillende mogelijkheden doen zich daarbij weer voor. Men kan de prestaties van de studenten op zich bestuderen met het oog op het identificeren van snelle groeiers en langzame groeiers. Men kan ook verschillende deelgebieden nemen met het oog op het identificeren van deelgebieden met een snelle ontwikkeling in kennis en gebieden met een langzame ontwikkeling in kennis.
Voorbeelden van studies van dit type zijn te vinden bij Imbos (1982), Imbos, Drukker, Van Mameren en Verwijnen (1984, 1987), Muijtjens en Imbos (1987), Sprooten en Imbos (1988).

9. Toepasbaarheid van voortgangstoetsen

Voortgangstoetsen werden ontwikkeld aan een universiteit, waarvan het onderwijs op verschillende punten afwijkt van hetgeen gebruikelijk is. Bovendien werd hiervoor al aangegeven dat het invoeren van voortgangstoetsen in het programma van de Faculteit der Geneeskunde is gebeurd naar aanleiding van problemen die waren ontstaan door het gebruik van bloktoetsen met een beslissend karakter. De vraag of voortgangstoetsen in andere onderwijssituaties bruikbaar zijn is dan ook gerechtvaardigd. Er zijn inmiddels onderzoeksgegevens beschikbaar die aantonen dat voortgangstoetsen ook in meer gangbare onderwijsprogramma's bruikbaar zijn (Bender, Cohen-Schotanus, Imbos, Versfelt en Verwijnen, 1984). Verder worden er aan de Medical School van de University of Missouri in Kansas City al jaren (men begon daar in 1972) Quarterly Profile Examinations (QPE) afgenomen. Deze QPE's toetsen, net als de voortgangstoetsen het gehele domein van de medische kennis en bestaan uit een steekproef van 400 toetsvragen uit een itembank, die in totaal circa 11.000 toetsvragen omvat. De toets wordt ieder kwartaal afgenomen aan alle studenten van de Medical School, die net als in Nederland zes studiejaren omvat. Hoewel de QPE's tot in detail lijken op de voortgangstoets, is er een belangrijk verschil. De QPE wordt alleen gebruikt ten behoeve van formatieve doeleinden. Men gebruikt de toets niet om beslissingen over studenten te nemen met betrekking tot hun voortgang, doch alleen om het leerproces van de studenten te begeleiden. Ook het reeds eerder genoemde onderzoek van Bosschert (1987), toont aan dat voortgangstoetsen eveneens te gebruiken zijn in medische curricula, die qua opzet en organisatie totaal verschillend zijn van het onderwijsprogramma van de Faculteit der Geneeskunde in Maastricht.

Voortgangstoetsen in de verst doorgevoerde vorm onstaan, wanneer men de toetsen ontleend aan de einddoelstellingen van het programma als geheel. Wijnen (1977) sprak om die reden van einddoeltoetsen. Deze einddoel-gerichtheid van de voortgangstoetsprocedure zal in de ogen van critici de toepasbaarheid beperken. Onder handhaving van het voornaamste kenmerk van de toetsprocedure, nl. herhaalde meting van hetzelfde construct, kan aan deze kritiek op een aantal manieren tegemoet worden gekomen.
In de eerste plaats kan men ook voortgangstoetsen ontlenen aan de doelstellingen van een bepaald studiejaar. In dit geval trekt men een aantal keren per jaar een steekproef van vragen uit een vragenvoorraad, die betrekking heeft op de programmaonderdelen van dat betreffende studiejaar. Op deze wijze beperkt men de periode waarin men de toetsen afneemt. In plaats van een periode van verschillende jaren observeert men de studenten slechts gedurende één studiejaar. Het spreekt voor zich dat er in dit geval een sterkere sturing van de toets uitgaat dan wanneer men de voortgangstoets als einddoeltoets gebruikt. Een student die in de eerste jaren vooral geïnteresseerd is in de klinische vakken of onderdelen daarvan heeft in een systeem met voortgangstoetsen per studiejaar beduidend minder mogelijkheden. Een zo groot mogelijke aansluiting bij de belangstelling en motivatie van de studenten kan het best worden bereikt, wanneer men de voortgangstoets als einddoeltoets gebruikt (Wijnen en Van der Vleuten, 1985).
In de tweede plaats kan men voortgangstoetsen ontlenen aan de einddoelstellingen van een bepaald vakgebied. In dat geval beperkt men het domein van de toets. Dat binnen de reguliere voortgangstoetsen zonder veel moeite een voortgangstoets voor een bepaald vakgebied, in dit geval het vakgebied van de anatomie en embryologie, een subtoets geconstrueerd kan worden, wordt geïllustreerd door Imbos e.a. (1984).
In de derde plaats kan men voortgangstoetsen construeren door zowel de periode van toetsing als het domein te beperken. Men kan bijvoorbeeld voortgangstoetsen voor alleen de klinische vakken in de laatste twee jaren van de medische studie construeren.
Tenslotte dient in deze paragraaf over de toepasbaarheid van de voortgangstoetsprocedure de vraag aan de orde te komen of de procedure ook in andere studies dan de Faculteit der Geneeskunde gebruikt kan worden. Men kan zich namelijk voorstellen dat voortgangstoetsen als einddoeltoetsen voor het gehele curriculum domeinspecifiek zijn en bijvoorbeeld alleen toepasbaar in opleidingen zonder vergaande specialisaties. Er zijn inmiddels een aantal toepassingen van voortgangstoetsen in andere domeinen. Te noemen zijn in dit verband: de fysiotherapie (Besseling en Van der Vleuten, 1986), de gezondheidswetenschappen (Sprooten en Van der Vleuten, 1987), de rechtsgeleerdheid (Nuy, 1983, Cohen en Nuy, 1985) en de economie (Mullink, 1987). In slechts één van de hier aangehaalde toepassingen, die binnen het gebied van de fysiotherapie, kan gesproken worden van een succesvolle toepassing. In alle andere gevallen is het gebruik van

voortgangstoetsen problematisch gebleken. Niettemin ziet men kennelijk nog mogelijkheden om een goede voortgangstoets voor de verschillende studierichtingen te ontwikkelen (Van der Vleuten en Spooten, 1987 a en b; Mullink, 1987). Of deze pogingen slagen zal moeten worden afgewacht.

10. De acceptabiliteit voor studenten

De beschouwing over de toepasbaarheid van voortgangstoetsen is niet volledig zonder de acceptabiliteit van de procedure voor studenten onder de loep te nemen. Als uitgangspunt voor een beschouwing over de acceptabiliteit (De Groot, o.a. in 1968 en 1986) van de toetsen kunnen twee enquêtes worden genomen die in 1980 en in 1986 aan studenten van de Faculteit der Geneeskunde werden afgenomen. Uit de antwoorden van de studenten op een aantal vragen uit die enquêtes kan de acceptabiliteit van de voortgangstoets voor de studenten worden afgeleid. Uit de resultaten van deze twee enquêtes komt het volgende naar voren: De studenten beoordeelden het systeem van voortgangstoetsen positief. Men vond het beantwoorden van vragen op eindniveau niet frusterend. De vragen werden als relevant beoordeeld. Ze meten echter naar het oordeel van de studenten geïsoleerde feitenkennis, die vooral in de eerste studiejaren, wanneer men een beperkt aantal onderwerpen heeft bestudeerd, te beantwoorden zijn. De studenten gebruiken de toets niet voor het plannen van hun studieactiviteiten. De toetsen geven geen goed inzicht in de relatieve positie ten opzichte van het einddoel van de opleiding.
Een en ander betekent dat de perceptie van de studenten over de voortgangstoets gedeeltelijk overeenkomt met die van de Faculteit der Geneeskunde.
Voortgangstoetsen zijn instrumenten om kennis te meten, niet meer en niet minder. Op het punt van de geïsoleerde feitenkennis hebben de studenten de facto ongelijk. Statistische analyses van de toetsen tonen steeds opnieuw aan dat de toetsitems wel degelijk onderling samenhangen. Dat geldt eveneens voor de grotere onderdelen, zoals de verschillende subtoetsen, basisvakkennis, klinische kennis en gedragswetenschappelijke kennis.
De kritiek die in de antwoorden van de studenten naar voren komt, is de kritiek tegen de gehanteerde methode om met een grote hoeveelheid toetsitems de kennis van personen, die zelf hun kennis als één geheel zien en beleven, te meten. Het is met andere woorden de kritiek die altijd tegen studietoetsen naar voren wordt gebracht. De kern daarvan wordt weergegeven door het gevoel "ik kan niet laten zien wat ik allemaal weet". Een serieus punt van kritiek, dat echter moeilijk te ondervangen is. Docenten delen overigens dit punt van kritiek vaak met studenten. Het is de vraag of bijvoorbeeld het geven van goede voorlichting over de bedoeling en mogelijkheden van voortgangstoetsen deze kritiek weg kan nemen. Een betere voorlichting aan studenten over het gebruik van studieresultaten bij het inrichten en plannen van de verdere studie, zou wellicht een deel van de

kritiek die in de antwoorden op de enquêtevragen op dit punt doorklinkt, kunnen wegnemen. Het wordt uit de gegeven antwoorden duidelijk dat studenten de mogelijkheden die de voortgangstoetsen in zich hebben niet onderkennen. De voorlichting die Verwijnen (1987) aan studiebegeleiders geeft zou ook voor studenten van betekenis kunnen zijn.
Uitgaande van de naar voren gebrachte positieve houding van zeer veel studenten tegenover het systeem van de voortgangstoetsen, waarbij men erkent dat de gevraagde kennis relevant is en dat het niet frustrerend is om vragen op eindniveau te beantwoorden, zou de Faculteit op de geringe acceptabiliteit van de toetsen voor de studenten kunnen reageren met het geven van een meer adequate voorlichting over de bedoeling en mogelijkheden van de voortgangstoetsen.
Een andere manier om de acceptabiliteit te verbeteren, zou kunnen zijn de beoogde kennis op een meer geïntegreerde manier te toetsen. Bijvoorbeeld door gebruik te maken van zogenaamde casusvragen. Dit zijn vragen met een inleidend stukje casuïstiek als stam, waarover één of meer vragen worden gesteld. Uit de antwoorden van de studenten op de enquêtevraag, of dit type vraag beter geschikt zou zijn om de kennis te demonstreren kan worden afgeleid dat opname van meer casusvragen inderdaad de acceptabiliteit zou verbeteren. Vanaf het tweede studiejaar wordt deze vraag door de meerderheid positief beantwoord. Overigens worden vragen van dit type meer en meer in de toetsen opgenomen.

11. De kwaliteit van de voortgangstoets als meetinstrument

Een laatste aspect, dat in deze bijlage over de voortgangstoets aan de orde dient te komen is de kwaliteit van de voortgangstoets als meetinstrument. De kwaliteit van een psychologisch meetintrument moet blijken. Vanaf de introductie van de voortgangstoetsen is er systematisch onderzoek gedaan naar aspecten van de betrouwbaarheid en de validiteit van de voortgangstoetsen. Een cumulatie van empirische gegevens over de toetsen is daarvan het resultaat. De voornaamste empirische kenmerken van de voortgangstoetsen zullen hier worden samengevat.
Het eenvoudigst is het bestuderen van de betrouwbaarheid van een meetinstrument. Deze bleek vanaf de introductie van de voortgangstoetsen acceptabel tot zeer goed te zijn. Een illustratie daarvan is te vinden in tabel 1.4 waarin de gemiddelde betrouwbaarheden (Cronbachs Alfa) van acht voortgangstoetsen met een gestandaardiseerde toetslengte worden vermeld.
De resultaten in deze tabel geven een goede indruk van de betrouwbaarheden die doorgaans voor de voortgangstoets realiseerbaar blijken te zijn. De betrouwbaarheid van de goedscores is altijd beduidend hoger dan die van de goed-foutscores. Deze laatste blijken doorgaans in de lagere studiejaren het laagst te zijn.

Bijlage 2

Tabel 1.4

Gemiddelde betrouwbaarheden van acht voortgangstoetsen met een standaard toetslengte (N items = 200)

1	.91	.01	.94	.89	.51	.13	.67	.30
2	.89	.02	.91	.87	.58	.07	.66	.45
3	.85	.02	.88	.84	.65	.11	.73	.59
4	.85	.02	.88	.83	.68	.04	.73	.60
5	.86	.03	.90	.81	.67	.02	.70	.65
6	.84	.04	.94	.81	.69	.05	.78	.62
totaal	.97	.01	.98	.96	.87	.02	.91	.85
Basisvakken	.97	.01	.98	.94	.89	.05	.96	.80
Klinische vakken	.98	.004	.98	.97	.90	.02	.93	.86
Gedragsweten-schappelijke vakken	.95	.01	.96	.94	.80	.04	.85	.73

De betrouwbaarheden van de goedscores kunnen worden gekwalificeerd als zeer hoog. De betrouwbaarheden van de goed-foutscores zijn gemiddeld nog wel acceptabel maar sommige uitbijters, in met name het eerste- en tweede studiejaar, zijn onacceptabel laag.
Het is hier niet de plaats om het probleem van de betrouwbaarheid van de goed-foutscores op te lossen. Volstaan wordt met het signaleren van het probleem, waarbij kan worden opgemerkt dat er momenteel gewerkt wordt aan de oplossing van het probleem (Mullink, 1987, Mullink en Imbos, 1988, Wijnen, 1988).
Van een goed meetinstrument wordt meer dan een betrouwbare meting alléén geëist. Het voornaamste kenmerk van een goed meetinstrument is de validiteit. Daarbij gaat het vooral om de vraag of het instrument doet wat men ervan verwacht (o.a. De Groot en Van Naerssen, 1969 en Hofstee, 1971). Het definitieve antwoord op de validiteitvraag is, omdat het gerelateerd is aan het gebruiksdoel van het instrument, vrijwel nooit te geven. Wel is het mogelijk om een aantal empirische kenmerken van de voortgangstoetsen te beschrijven, waarmee aspecten van de validiteit van de instrumenten worden belicht.
Allereerst blijken de voortgangstoetsen in staat om systematisch ouderejaars en jongerejaars van elkaar te onderscheiden. Bovendien blijken artsen, die aan de toetsen deelnemen gemiddeld gezien altijd hoger te scoren dan de eerste vijf studiejaren.

Verder blijkt dat deze verschillen zich niet alleen voordoen in het curriculum, waar de toetsen werden ontwikkeld (en waardoor er een zekere onzuiverheid in de resultaten zou kunnen optreden), ook in andere medische faculteiten laten de voortgangstoetsen verschillen tussen de jaargroepen zien (Bender, e.a., 1984 en Bosschert, 1987). Hoewel de resultaten globaal gezien gelijk waren, bleken de toetsen specifiek (Ebel, 1965) genoeg te zijn om karakteristieke verschillen tussen de verschillende opleidingen te registreren (Imbos, Van Hessen, Muijtjens, Verwijnen en Wijnen, 1987).
Dat resultaten van dit type niet alleen verklaard kunnen worden door het leeftijdsverschil, in samenhang met het volgen van een opleiding waarin iedereen wel enige globale medische kennis opdoet, bleek uit de resultaten die Fokkema (1986) rapporteerde. Fokkema nam voortgangstoetsen af bij studenten van verschillende jaargroepen uit de studierichtingen Biologie, Psychologie en Rechten. Zijn conclusie was dat de voortgangstoetsen onderwijsspecifiek waren en alleen in het medische domein verschillen tussen de jaargroepen lieten zien. Binnen het medische domein toonde Imbos (1982 a) aan dat een geschikt gekozen subtoets uit de voortgangstoets gevoelig is voor zeer specifieke onderwijseffecten. Een onderwijsblok over hart- en vaatproblemen werd als de behandeling beschouwd. Als afhankelijke variabele werd de score op de subtoets hart- en vaatstelsel genomen. Er waren twee voor- en twee nametingen. Na de behandeling vertoonde de mediane score op deze subtoets, in tegenstelling tot een willekeurig gekozen subtoets met gelijke toetslengte, een duidelijke afwijking van de gebruikelijke groei in kennis. Voor andere blokken en andere subtoetsen konden soortgelijke effecten worden aangetoond door Sprooten en Imbos (1988).
Binnen hetzelfde medische domein correleren de voortgangstoetsen significant met andere maten voor medische kennis, i.c. bloktoetsen (Snellen, Stalenhoef en Imbos, 1986). De bruikbaarheid van voortgangstoetsen om voortgangsbeslissingen over studenten in het eerstejaar van hun studie te nemen, werd onderzocht door Van Boven, Imbos, Snellen, Stalenhoef en Van der Vleuten, (1987). Als criterium werd gebruikt het studiesucces in het tweedejaar, geoperationaliseerd als de hoogte van de gemiddelde score op vier voortgangstoetsen in het tweede studiejaar. Deze gemiddelde scores werden gedichotomiseerd bij de mediaan van de scoreverdeling. Boven het gemiddelde werd als "goed", onder het gemiddelde werd als "niet goed" beschouwd. Als voorspellers werden de als voldoende en onvoldoende gedichotomiseerde scores van de vier voortgangstoetsen in het eerste studiejaar gebruikt. Uit de resulterende vier, twee bij twee, tabellen konden de verschillende conditionele kansen werden berekend.
Uit de resultaten bleek de conditionele kans op een goed resultaat in het tweedejaar, gegeven een goed resultaat in het eerstejaar gemiddeld .84 te zijn. De conditionele kans op een slecht resultaat in het tweedejaar, gegeven een slecht resultaat in het eerstejaar, was gemiddeld .33. De kans op een voldoende

toetsresultaat in het eerstejaar, gegeven een voldoende resultaat in het tweedejaar was .98. De kans op een onvoldoende resultaat in het eerstejaar, gegeven een onvoldoende resultaat in het tweedejaar was .06.
Het patroon in de correlaties van de toetsen over verschillende jaren werd onderzocht door Sprooten (1984). Het merendeel van de onderzochte longitudinale reeksen bleek een zogenaamde simplex structuur te hebben. Dit betekent dat grote verschuivingen in de prestaties van de studenten, gemeten over een relatief korte periode, niet te verwachten zijn. Met andere woorden de studieprestaties, zoals gemeten met de voortgangstoets, zijn stabiel over een relatief korte periode van plus minus een jaar. Gemeten over een langere periode daalt de stabiliteit aanzienlijk (tussen opeenvolgende tijdstippen werden correlaties van .80 en hoger gerapporteerd, terwijl de correlaties tussen bijvoorbeeld eind tweedejaar en eind vierdejaar in de orde van grootte van .30 waren).
De mogelijkheid om individuele groei met behulp van voortgangstoetsen te meten, werd onderzocht door Verwijnen en Imbos (1979). Zij onderzochten de individuele groei door middel van de modellen die voor dat doel door Lord (1963) waren ontwikkeld. Aangetoond kon worden dat de ware groei, (d.i. de groei in de niet observeerbare "ware" scores) in tegenstelling tot de geobserveerde groei tussen opeenvolgende tijdstippen aanzienlijk was.
Hoewel deze resultaten op zich zelf genomen heel bevredigend waren, bleef er toch behoefte bestaan aan een omvattend statistisch model, dat in staat was om simultane schattingen op eenzelfde schaal te geven, van de kennis van de studenten. Een dergelijk model werd ontwikkeld en onderzocht door Albers e.a. (1987). Niet alleen bleek het mogelijk om een dergelijk model te maken en goede schatters voor de parameters in het model te ontwikkelen, ook de individele groei (uitgedrukt in een individuele groeiparameter) in capaciteiten door de jaren (in totaal zes jaar) bleek lineair te modeleren te zijn, met een gemiddeld percentage verklaarde variatie van ruim 80%.
De beschikbaarheid van dit model, dat in de hoofdstukken 5 en 6 van deze studie gedetailleerd is beschreven en toegepast, maakt verder onderzoek naar de groei van kennis in verschillende deelgebieden van de medische wetenschap mogelijk. Hierdoor kan het systeem van de voortgangstoetsen nog meer tot zijn recht komen en worden uitgebuit. Ook lijkt het waarschijnlijk dat het model gebruikt kan worden om snel groeiende- en langzaam groeiende kennisgebieden in de andere faculteiten van de Rijksuniversiteit Limburg op het spoor te komen.
Uit de bovenstaande samenvatting blijkt dat voor zover de kennis omtrent het meetinstrument nu reikt, geconcludeerd kan worden dat de voortgangstoetsen, in termen van basisbegrippen als betrouwbaarheid en validiteit, de toets der kritiek, kan doorstaan. Met het bestaan van de voortgangstoets heeft de Faculteit der Geneeskunde, als een van de weinige instellingen een vorm van kwaliteitsbewaking, die het niveau van de vakgroep overstijgt. Daarmee heeft de faculteit een "krachtig middel tot kwaliteitscontrole" (Van der Linden, 1987).

Referenties

Albers, W., Does, R.J.M.M., Imbos, Tj. & Janssen, M.P.E. (1987). **A stochastic model for knowledge growth in progress tests.** (Medical Informatics and Statistics Report no. 11). Maastricht: Rijksuniversiteit Limburg, vakgroep Medische Informatica en Statistiek. Verschijnt tevens in **Psychometrika, 53**.

Allen, M.J. en Yen, W.M. (1979). **Introduction to measurement theory**. Belmont: Wadsworth.

Anderson, J.R. (1982). Acquisition of cognitive skill. **Psychological Review, 89**, 369-406.

Ausubel, D.P. (1968). **Educational Psychology**. New York: Holt.

Bender, W., Cohen-Schotanus, J., Imbos, Tj., Versfelt, W.A. & Verwijnen, G.M., (1984). Medische kennis bij studenten uit verschillende medische faculteiten: van hetzelfde laken een pak? **Nederlands Tijdschrift voor de Geneeskunde, 128,** 917-921.

Berkel, H.J.M. van, (1984). **De diagnose van toetsvragen**. (Academisch Proefschrift). Universiteit van Amsterdam, Centrum voor Onderzoek van het Wetenschappelijk Onderwijs.

Besseling, C.W.M., Tromp, J.H.M., Vleuten, C.P.M. van der & Verbraeck, A., (1986). Voortgangstoetsen voor fysiotherapie-opleidingen: Ervaringen en gegevens. **Tijdschrift voor Hoger Onderwijs, 4,** 107-115.

Blalock, H.M. Jr. (1968). The measurement problem: a gap between languages of theory and research. In H.M. Blalock Jr. and A.B. Blalock (eds.), **Methodology in social research**. New York: McGraw-Hill.

Blalock, H.M. Jr. (1979). Measurement and conceptualization problems: the major obstacle to integrating theory and research. **American Sociological Review, 44**, 881-894.

Blalock, H.M. (1982). **Conceptualization and measurement in the social sciences.** London: Sages.

Borgatta, E.F. & Bohrnstedt, G.W. (1972). How one normally constructs good measures. **Sociological Methods and Research, 1,** 3-12.

Boshuizen, H.P.A. & Schmidt, H.G. (1986). **The developing structure of medical knowledge.** (Onderzoek van Onderwijs nr. 32). Maastricht: Rijksuniversiteit Limburg, Vakgroep Onderwijsontwikkeling en Onderwijsresearch. Verschijnt tevens in: Z. Nooman, H.G. Schmidt & E. Ezzat, (eds.), **Innovation in medical education: An evaluation of its present status.** New York: Springer.

Boshuizen, H.P.A. (1987). **Medical expertise and the structure of knowledge: a cognitive psychologist's view on its development.** (Ongepubliceerde voordracht). Maastricht, Rijksuniversiteit Limburg, vakgroep Onderwijsontwikkeling en Onderwijsresearch.

Bosschert, D. (1987). **Resultaten afname voortgangstoets.** Leiden: Rijksuniversiteit Leiden, Faculteit der Geneeskunde, Sectie Onderwijsontwikkeling.

Boven, C.P.A. van, Stalenhoef, B., Snellen, H., Imbos, Tj. & Vleuten, C.P.M. van der (1987). Taking decisions about students: Lottery or science? **Abstracts of the International Symposium on Evaluation in Medical Education,** blz. 10. Beer Sheeva, Israel.

Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). **Statistics for experimenters: an introduction to design, data analyses and model building.** New York: Wiley and Sons.

Buning, G.R., Eggen, T.J.H.M., Kelderman, H. & Linden, W.J. van der (1986). **Het gebruik van het Raschmodel voor een decentraal toetsservice systeem.** (Rapport 86-3). Enschede: Universiteit Twente, Vakgroep OMD.

Cobben, N., (1982). **Gebruikershandleiding Itembanksysteem.** (PES-bulletin nr. 4) Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Cobben, N., (1983). VAX 11/780; DBMS-32; **Rapportage en aanbevelingen.** (Dienst Informatie Verwerking; Rapport). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Cohen, M.J. & Nuy, H.J.P. (1985). **De voortgangstoets: eerste vergelijkende gegevens.** (Ongepubliceerd manuscript). Rijksuniversiteit Limburg, Faculteit der Rechtsgeleerdheid.

Cohen, J. (1988). De Juristen slingerend op het Pad van de Onderwijsvernieuwing. **De Observant,** 19, 3 e.v. Maastricht: Rijksuniversiteit Limburg.

Debets, P. & Brouwer, E. (1986). **MSP; A program for Mokken scale analysis for polychotomous items.** Amsterdam: Technisch Centrum.

Dessens, J.A.G. & Jansen, W. (1987). **Operationaliseren: Traditie en kritiek.** (Academisch proefschrift). Utrecht: Rijksuniversiteit Utrecht.

Does, R.J.M.M., Koning, A.J., Strijbosch, L.W.G. & Volovics, A. (1983). **Syllabus Data-analyse.** Maastricht: Rijksuniversiteit Limburg, vakgroep Medische Informatica en Statistiek.

Drenth, P.J.D. (1975). **Inleiding in de testtheorie.** Deventer: Van Loghum & Slaterus.

Ebel, R.L. (1965). **Measuring educational achievement.** Englewood Cliffs N.J.: Prentice-Hall.

Ebel, R.L. (1972). **Essentials of educational measurement.** Englewood Cliffs N.J.: Prentice-Hall.

Ebel, R.L. & Frisbie, D.A. (1986). **Essentials of educational Measurement.** Englewood Cliffs N.J.: Prentice-Hall.

Elshout, J.J. (1976). **Karakteristieke moeilijkheden in het denken.** (Academisch proefschrift). Amsterdam: Universiteit van Amsterdam.

Embretson, S. (1983). Construct Validity: construct representation versus nomothetic span. **Psychological Bulletin, 93,** 179-197.

Feldt, L.S., Woodruff, D.J. & Fathi, A.S. (1987). Statistical inference for coefficient alpha. **Applied Psychological Measurement, 11,** 93-103.

Feltovich, P.J. & Barrows, H.S. (1984). Issues of generality in medical problem solving. In H.G. Schmidt and M.L. de Volder (eds.), **Tutorials in problem-based learning.** Assen: Van Gorcum, 128-142.

Fisher, G.H. (1974). **Einführung in die Theorie Psychologischer Tests:** Grundlagen und Anwendungen. Bern: Hans Huber.

Fokkema, F. (1986). **Het gebruik van voortgangstoetsen bij het vergelijken van medische curricula: Een begripsvalidatiestudie.** (Heijmans Bulletin HB-86-812-SW). Groningen: Rijksuniversiteit Groningen, Psychologische Instituten Heijmans.

Gaag, N. van der (1987). **Empirische utiliteitsfuncties; 3 meetprocedures.** Amsterdam: Universiteit van Amsterdam, Vakgroep Psychologische Methoden

Gronlund, N.E. (1968). **Constructing achievement tests**. Englewood Cliffs: Prentice-Hall.

Groot, A.D. de (1966). **Vijven en zessen**. Groningen: Wolters-Noordhof.

Groot, A.D. de (1968). **Methodologie**. Den Haag: Mouton.

Groot, A.D. de & Naerssen, R.F. van (1969). **Studietoetsen, construeren, afnemen en analyseren**. Den Haag: Mouton.

Groot, A.D. de (ed.), (1972). **Selectie voor en in het hoger onderwijs**. Den Haag: Staatsuitgeverij.

Groot, A.D. de (1982). **Academie en forum**. Meppel: Boom.

Groot, A.D. de (1986). **Begrip van evalueren**. Den Haag: Vuga.

Grosse, M.E. & Wright, B.D. (1985). Validity and reliability of true-false tests. **Educational and Psychological Measurement, 45**, 1-14

Gruijter, D.N.M. de (1987). **Een model voor waar-onwaar vragen met goed-min-fout scoring**. (Memorandum 914-87). Leiden: Rijksuniversiteit Leiden, Bureau Onderzoek van Onderwijs.

Gustafsson, J.E. (1977). **The Raschmodel for dichotomous items: Theory applications and a computerprogram**. (Report No. 63). Göteborg: The institute of education.

Guttman, L.A. (1950). The basis for scalogram analysis. In: S.A. Stouffer (ed.), **Measurement and Prediction**, blz. 60-90. New York: John Wiley.

Guttman, L.A. (1954). A new Approach to Factor Analysis: The Radex. In P.F. Lazersfeld (ed). **Mathematical thinking in the social sciences**, blz. 258-348. New York: Colombia University Press.

Guttman, L.A. (1977). What is not what in statistics. **The Statistician, 26**, 81-107.

Guttman, L.A. (1981). What is and what not in theory construction. In J. Borg (ed.), **Multidimensional data representations: When and why**. Ann Arbor: Mathesis Press 47-64.

Hambleton, R.K. & Swaminathan, H. (1985). **Item response theory; Principles and applications**. Boston: Kluwer Publishing.

Heise, D.R. (1975). **Causal analysis**. New York: Wiley.

Hessen, P. van, Cobben, N. & Verwijnen, G.M. (1987). Development of a medical content classification system for itembanking. **Abstracts of the International Symposium on Evaluation in Medical Education**, blz. 30. Beer Sheeva, Israel.

Hofstee, W.K.B. (1969). Individuele verschillen en averechtse toepassing. **Nederlands Tijdschrift voor de Psychologie**, 24, 482-493.

Hofstee, W.K.B. (1971). Begripsvalidatie van studietoetsen: Een aanbeveling. **Nederlands Tijdschrift voor de Psychologie, 26**, 491-500.

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

Hox, J.J.C.M. (1986). **Het gebruik van hulptheorieën bij operationalisering: Een studie rond het begrip subjectief welbevinden**. (Academisch proefschrift). Amsterdam: Universiteit van Amsterdam.

Hullin, C.L., Drasgow, F. & Parsons, C.K. (1983). **Item response theory: Applications to psychological measurement**. Homewood, Illinois: Dow Jones-Irwin.

Humphreys, L.G. (1960). Investigations of the simplex. **Psychometrika, 4**, 313-323.

Humphreys, L.G. (1968). The fleeting nature of the prediction of college academic success. **Journal of Educational Psychology, 5**, 375-380.

Imbos, Tj. & Verwijnen, G.M. (1978). Evaluatie aan de Maastrichtse Medische Faculteit. **Metamedica, 57**, 21-32.

Imbos, Tj. & Verwijnen G.M. (1978). **De ontwikkeling van een geautomatiseerde itembank**. (Ongepubliceerd manuscript). Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Imbos, Tj., & Verwijnen, G.M. (1979) Evaluation in a new medical school. **Proceedings of the EARDHE Conference**. Klagenfurt.

Imbos, Tj., (1982a). **Effecten van computer simulaties in blok 1.5.** (PES-Bulletin nr. 22). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Imbos, Tj. (1982b). **De betekenis van voortgangstoetsen en bloktoetsen in het kader van een selektieve propedeuse**. (PES-Bulletin nr. 2). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Imbos, Tj., Drukker, J., Mameren, H. van & Verwijnen, G.M. (1984). The growth in Knowledge of Anatomy in a problem-based curriculum. In: H.G. Schmidt & M.L. de Volder (eds.), **Tutorials in problem-based learning a new direction in teaching the health professions**. blz. 106-115. Assen: van Gorcum.

Imbos, Tj., Drukker, J., Mameren, H. van & Verwijnen, G.M. (1987, februari). Anatomische kennis getoetst door clinici of anatomen: van hetzelfde laken een pak? Amsterdam: **Nederlandse Vereniging voor Medisch Onderwijs, Vrije Voordrachten Dag**, blz. 2.

Imbos, Tj., Hessen, P.A.W. van, Muytjens, A., Snellen, H., Verwijnen, G.M., & Wijnen, W.H.F.W. (1987). Some examples of the possiblities of program independent achievement testing with progress tests. **Abstracts of the International Symposium on Evaluation in Medical Education**, blz. 33. Beer Sheeva, Israel.

Imbos, Tj. (1988). **Het evaluatiesysteem van de Medische Faculteit Maastricht; Een historisch overzicht**. (PES-Bulletin nr. 88-2). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Imbos, Tj., Snellen, H.A.M. & Verwijnen, G.M. (1988, in voorbereiding). **Construct validity of progress tests in freshman years.**

Imbos, Tj. & Stalenhoef, B. (1988, in voorbereiding). **Veranderende correlatiepatronen tussen einddoeltoetsen en onderwijs-specifieke toetsen.**

Janssen, G.M.E. & Pollemans, M.C. (1979). **GIM-NOTA 1: geautomatiseerde itembank Maastricht**. (DPG-SE Bulletin nr.10). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Janssen, G.M.E. & Pollemans, M.C. (1980). **GIM-NOTA-2: geautomatiseerde itembank Maastricht: Ontwikkeling van een thesaurus**. (DPG-SE Bulletin nr.36). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Janssen, G.M.E. & Pollemans, M.C. (1981). **GIM-NOTA-3: geautomatiseerde itembank Maastricht**. (DPG-SE Bulletin nr.44). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Jansen, P.G.W. (1982). De onbruikbaarheid van mokkenschaalanalyse. **Tijdschrift voor Onderwijsresearch, 7**, 11-25.

Jansen, P.G.W., Roskam, E.E.Ch.I. & Wollenberg, A.L. van der (1982). De mokkenschaal gewogen. **Tijdschrift voor Onderwijsresearch, 7**, 31-43.

Jansen, M.P.E., Imbos, Tj., Does, R.J.M.M. & Albers, W. (1988, in voorbereiding). **ABEST; a computerprogram for ability estimation in longitudinal designs.**

Jansweijer, W.N.H., Elshout, J.J. & Wielinga, B.J. **Simulation of novice problem solving in physics**. (Memorandum 50). Amsterdam: Universiteit van Amsterdam, Instituut voor cognitie onderzoek.

Jong, J. de & Hermkens, P. (1983). Artefacten van de geest of de naakte feiten: De Groot en Roskam. **Mens en Maatschappij, 58**, 407-413.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. **The British Journal of Mathematical and Statistical Psychology, 23,** 121-145.

Jöreskog, K.G. & Sörbom, D. (1979). **Advances in factor analysis and structural equation models**. Cambridge, M.A.: Abt Books.

Jöreskog, K.G. & Sörbom, D. (1981). **LISREL: Analysis of structural relationships by the method of maximum likelihood: Usersguide**. Chicago: International Educational Resources.

Kamp, L.J.Th. van der, Langerak, W.F. & Gruijter, D.N.M. de (eds.), (1980). **Psychometrics for educational debates**. New York: John Wiley & Sons.

Kaplan, A. (1964). **The conduct of inquiry**. San Francisco: Chandler Publ. Co.

Keizer, H., Thiel, R. van & Zwarts, M. (1985). **Verslag vooronderzoek toetsservicesysteem**. (Specialistisch Bulletin nr. 32) Arnhem: Cito.

Kelderman, H. (1986). **Common item equating using the loglinear Rasch model.** (Research Report 86-9). Enschede: Universiteit Twente, Vakgroep OMD.

Kenny, D.A. (1979). **Correlation and causality**. New York: Wiley.

Knol, D.L. (1986). **Inventarisatie van automatische itemselectieprocedures voor Rasch-schalen**. (Rapport 86-2). Enschede: Universiteit Twente, Vakgroep OMD.

Krantz, D.H., Luce, P. Suppes, P. & Tversky, A. (1971). **Foundations of Measurement**, vol.1. New York: Academic Press.

Levy, P. & Goldstein, H. (1984). **Tests in education: a book of critical reviews.** London: Academic Press.

Linden, W.J. van der & Mellenbergh, G.J. (1977). Optimal cutting scores using a linear loss function. **Applied Psychological Measurement, 1,** 593-599.

Linden, W.J. van der. (1983). **Van standaardtest naar itembank.** Oratie. Universiteit Twente.

Linden, W.J. van der & Zwarts, M.A. (1986). **Some procedures for computerized ability testing.** (Research Report 86-10). Enschede: Universiteit Twente, Vakgroep OMD.

Linden, W.J. van der. (1987, november). **Het zwalkende niveau van het onderwijs.** Diësrede. Universiteit Twente.

Loevinger, J. (1947). A systematic approach to the construction and evaluation for tests of ability. **Psychological Monographs, 4,** 61.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. **Psychological Bulletin, 45,** 507-530.

Loftus, G.R. & Loftus, E.F. (1976). **Human memory.** Hillsdale, N.J.: Erlbaum.

Lord, F.M. (1963). Elementary models for measuring change. In: C.W. Harris. **Problems in measuring change.** Madison: The University of Wisconsin Press.

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

Lord, F.M. (1980). **Applications of item response theory to practical testing problems.** Hillsdale, New Jersey: Lawrence Erlbaum.

Lumsden, J. (1977). Person reliability. **Applied Psychological Measurement, 1,** 477-482.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. **Journal of the National Cancer Institute, 22,** 719-748.

Margenau, H. (1950). **The nature of physical reality.** New York: McGraw-Hill.

Mellenbergh, G.J. (1980). Theorie op verschillende niveau's. **Nederlands Tijdschrift voor de Psychologie, 35**, 275-288.

Messick, S. (1983). Constructs and their vicissitudes in educational and psychological measurement. **Psychological Bulletin, 89**, 575-588.

Messick, S. (1984). The psychology of educational measurement. **Journal of Educational Measurement, 21**, 215-237.

Miller, G.R. (1981). **Simultaneous statistical inference**. New York: Springer Verlag.

Molenaar, I.W. (1982a). Mensen die het beter meten: Een inleiding tot enige latente trekmodellen. **Kwantitatieve Methoden, 13**, 3-29.

Molenaar, I.W. (1982b). Mokken scaling revisited. **Kwantitatieve Methoden, 3**, 145-164.

Molenaar, I.W. (1982c). De beperkte bruikbaarheid van Jansen's kritiek. **Tijdschrift voor Onderwijsresearch, 7**, 25-31.

Molenaar, I.W. (1982d). Een tweede weging van de mokkenschaal. **Tijdschrift voor Onderwijsresearch, 7**, 172-182.

Molenaar, I.W. (1983a). Rasch, Mokken en schoolbeleving. In: S. Lindeman en F.N. Stokman (eds.). **Modellen in de Sociologie**. Deventer: Van Loghum Slaterus, 195-213.

Molenaar, I.W. (1983b). **Item steps**. Heyman Bulletin No. HB-83-630 Ex. Groningen: Rijksuniversiteit, Heymans Psychologische Instituten.

Molenaar, I.W. (1983c). Some improved diagnostics for failure of the Rasch model. **Psychometrika, 48**, 49-73.

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

Molenaar I.W. & Sijtsma, K. (1988). **Kwantitatieve Methoden, 28**, 115-126.

Mokken, R.J. (1971). **A theory and procedure of scale analysis: with applications in political research**. Den Haag: Mouton.

Mosteller, F. & Tukey, J.W. (1977). **Data analysis and regression: a second course in statistics**. Massachusetts: Addison-Wesley.

Muijtjens, A.A.M. & Imbos, Tj. (1987). **Exploratory analysis of progress data.** (PES-Bulletin nr. 176) Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Muijtjens, A.M.M., Imbos, Tj., Theunissen, M.J.A. & Roos, S.M.A. (1988). Exploratory analysis of progress test data. **Onderzoek van onderwijs, 38**. Maastricht: Rijksuniversiteit Limburg.

Mullink, J. (1987). **Achtergronden van de voortgangstoets**. (Ongepubliceerd manuscript). Rijksuniversiteit Limburg, vakgroep Onderwijsontwikkeling en Onderwijsresearch.

Northrop, F.S.C. (1947). **The logic of the sciences and the humanities**. New York: World Publishing Company.

Nunnally, J.C. (1967). **Psychometric theory**. New York: McGraw-Hill Book Company.

Nuy, H.J.P. (1983). **Vergelijkend onderzoek voortgangstoets.** (Ongepubliceerd manuscript). Rijksuniversiteit Limburg, Faculteit der Rechtsgeleerdheid.

Onderwijscommissie Faculteit der Geneeskunde (1987). **Nota: de rode draad III.** Maastricht: Rijksuniversiteit Limburg.

Oosterloo, S.J. (1984). Confidence intervals for test information and relative efficiency. **Statistica Neerlandica, 2**, 37-53.

Pollemans, M.C. (1982). **GIM-NOTA-4: Rapportage itembanking**. (PES-Bulletin nr. 15) Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Pollemans, M.C. (1983). **GIM-NOTA-5: Rapportage itembanking**. (PES-Bulletin nr. 41) Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Popper, K.R. (1978). **De groei van kennis**. Meppel: Boom.

Rasch, G. (1960). **Probabilistic models for some intelligence and attainment tests**. Copenhagen: The Danish Institute for Educational Research.

Roskam, E.E.Ch.I. (1982). Hypotheses non fingo: Een methodologische gevalsstudie over onderzoek van intelligentietests. **Nederlands Tijdschrift voor de Psychologie, 37**, 331-359.

Rovinelli, R.J. & Hambleton, R.K. (1977). On the use of context specialists in the assessment of criterion referenced test item validity. **Tijdschrift voor Onderwijsresearch, 2**, 49-60.

Saris, W. & Stronkhorst, H. (1984). **Causal modelling in nonexperimental research: An introduction to the LISREL approach.** Amsterdam: Sociometric Research Foundation.

Schouten, H.J.A. (1985). **Statistical measurement of inter-observer agreement.** Rotterdam, academisch proefschrift.

Sijtsma, K. (1983). Rasch-homogeniteit empirisch onderzocht. **Tijdschrift voor Onderwijsresearch, 8**, 104-121.

Sluiter, C. (1986). **Het bijwerken van een itembank: bijstellen van itemmoeilijkheid in het Rasch-model.** (Rapport 86-1). Enschede: Universiteit Twente, Vakgroep OMD.

Snellen, H.A.M. (1982). **Curriculumoverzicht Faculteit der Geneeskunde.** (Onderzoek van Onderwijs 19b). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde, vakgroep Onderwijsontwikkeling en Onderwijsresearch.

Sprooten, J. (1984). **Lisrel simplex analyses.** (Pes-Bulletin nr. 73). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Sprooten, J. & Imbos, Tj. (1988, in voorbereiding). **Onderwijseffecten gemeten met einddoeltoetsen.**

Stalenhoef, B., Snellen, H., & Imbos, Tj. (1985). **Jaarverslag Bloktoetsen 1983/84 en 1984/85.** (Pes-Bulletin nr.109)Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Steen, R. (1986). Rasch. Inf.; **Overzicht programma's Raschanalyse.** Enschede: Universiteit Twente, Vakgroep OMD.

Thorndike R.L. (1982). **Applied psychometrics.** Boston: Misslin Company.

Verwijnen, G.M., Imbos, Tj., Snellen, H., Stalenhoef, B., Sprooten, M., Leeuwen, Y.D. van & Vleuten. C.P.M. van der (1982). The evaluation system at the medical school of Maastricht, **Assessment and Evaluation in Higher Education, 3,** 225-244.

Verwijnen, G.M. & Fröberg-Gresnich, C. (1986). **Jaarverslag over het academisch jaar 1984/85 van de Voortgangstoets-beoordelingscommissie.** (Pes-Bulletin nr. 117). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Verwijnen, G.M., Imbos. Tj., Hessen, P.A.W. van & Wijnen, W.H.F.W. (1987) What's in the score! An exploratory Analysis of True-False Test Scores. **Abstracts of the International Symposium on Evaluation in Medical Education, blz. 85.** Beer Sheeva, Israel.

Verwijnen, G.M. (1987). Betekenis van studieresultaten bij studie-begeleiding: Een handvol ervaringen. In: W.S. de Grave en H.J.P. Nuy. **Leren studeren in het hoger onderwijs.** Almere: Versluys Uitgeverij.

Visser, R.S.H., Vliet-Mulder, J.C. van, Evers, A. & Haak J. Ter (1982). **Documentatie van tests en testresearch in Nederland.** Nijmegen: Van Mameren

Vleuten, M. van der & Sprooten, J. (1987a). **De voortgangstoets.** (Rapport van het Project Evaluatie Studieresultaten FdGW 1986/1987). Maastricht: Rijksuniversiteit Limburg, Faculteit der Gezondheidswetenschappen.

Vleuten, M. van der & Sprooten, J. (1987b). **Handleiding voor de constructie van voortgangstoetsitems.** (Rapport van het Project Evaluatie Studieresultaten FdGW). Maastricht: Rijksuniversiteit Limburg, Faculteit der Gezondheidswetenschappen.

Weiss, D.J. (1983). New horizons in testing: **Latent trait theory and computerized adaptive testing.** New York: Academic Press.

Werts, C.E., Linn, R.L. & Jöreskog, K.G. (1977). A simplex model for analyzing academic growth. **Educational and Psychological Measurement, 3,** 745-756.

Wijnen, W.H.F.W. (1971). **Onder of boven de maat.** Lisse: Swets en Zeitlinger.

Wijnen, W.H.F.W., (1976). **Voorstellen betreffende evaluatie.** (Onderzoeksrapport Nr. OC 76-011). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Wijnen, W.H.F.W., (1977). Einddoeltoetsen: Waarom en hoe? **Onderzoek van Onderwijs, 6,** 16-19.

Wijnen, W.H.F.W., (1978). Evaluatie van een Medische Opleiding. In: **Colloquium: Toekomst van de huisarts.** Brussel: Ministerie van Volksgezondheid en het gezin.

Wijnen, W.H.F.W., (1984). **Frequentieverdeling van de goede antwoorden bij eerstejaars studenten.** (PES-bulletin nr. 68). Maastricht: Rijksuniversiteit Limburg, Faculteit der Geneeskunde.

Wijnen, W.H.F.W. & Vleuten, C.P.M. van der (1985). Toetsing: Hordenloop of voortgangskontrole? **Universiteit en Hogeschool, 6,** 270-279.

Willoughby, T.L. & Hutcheson, S.J. (1978). Edumetric validity of the Quarterly Profile Examination. **Educational and Psychological Measurement, 38,** 1057-1061.

Willoughby, T.L., Gerrity, M. Durbian, Khalsa, S.J.K. & Buck, L. (1980, May). **Quarterly Profile Examination.** Kansas City: University of Missouri, School of Medicine.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. **Journal of Educational Measurement, 14,** 97-116.

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

Wright, B.D. & Masters, G.N. (1982). **Rating scale analysis: Rasch measurement.** Chicago: Mesa Press

Wright, B.D. & Bell, S.R. (1984). Itembanks: What, why and how. **Journal of Educational Measurement, 21,** 331-345.

Wright, B.D. & Linacre, J.M. (1985). **Microscale: manual.** Mediax Interactive Technologies.

SUMMARY

This thesis discusses certain aspects of the Maastricht Progress Tests. Progress tests test the end objectives in the cognitive domain of medicine. The use of these tests in freshman years could be questioned. The research questions adressed in this thesis are:

- is it reasonable to assume that the progress tests include test items which can be answered by freshmen?

And if so,

- are there predictable differences and/or similarities between these test items and randomly selected test items?

A group of test items that can be answered by freshmen is called a Freshmen Test (FT). A group of test items that is randomly chosen is called a General Knowledge Test (GKT). Both types of tests were studied in the academic year 1983-1984 on four occasions. The differences and similarities between the tests are hypothesised. Three different kinds of hypotheses can be distinquised:

- hypotheses concerning the internal structure of the tests, i.e. differences in terms of the p-values, test scores, reliability.

- hypotheses concerning the external structure of the tests, i.e. differences in correlation patterns.

- hypotheses concerning the fit of the test data to the assumptions of psychometric models **(CHAPTER 1)**.

To answer the first research question a panel of experts on the first year educational program were asked to judge 1029 test items. The inter rater reliability approved to be acceptable. Between 19% and 24% of the test items corresponded to material handled in the first year program. The item selection method using content experts is compared to an empirical method using the observed p-values as a criterion. It could be shown that both methods were related. Furthermore the items that were selected by the content experts did not show a bias concerning difficult or easy test items. On the contrary the experts selected test items that had a wide range in difficulties. A conclusion is that the first research question may be answered positively **(CHAPTER 2)**.

The differences concerning the internal structure between the FT and the GKT are studied. Hypotheses are tested in a sample of first year students, a sample of sixth

year students and a sample of physicians. Six out of nine hypotheses are fully confirmed by the results. Three hypotheses are only partly confirmed which means that the prediction was correct in some groups but incorrect in other groups. These results indicate that in the first year of medical school:

- FT test items are answered more than GKT test items.
- Ft test items are answered better than GKT test items.
- FT test items and GKT test items are equally reliable.
- FT test items indicate of educational effects more than GKT test items.
- The test scores on the FT test items differ significantly from the test scores on the GKT test items.

In the sample of sixth year students and the sample of physicians no differences between the FT and GKT test items regarding these aspects were found, except one: the test scores, which were significantly higher on the Freshmen Test, in both samples.
It was therefore concluded that as far as the internal differences between the tests are concerned the results indicate that the tests have construction validity. The Freshmen Test and the Genenal Knowledge Test behave in most cases as was predicted in the hypotheses. **(CHAPTER 3).**

The hypotheses about the structure in the correlations between the four FT's and the four GKT's are formalised in two statistical models. In model I, one latent variable per occasion is specified. The covariance matrix of the latent variables is assumed to be generated by a simplex or first-order autoregressive model. The errors of measurement are asssumed to be uncorrelated. In model II, the same structure as in model I is assumed, while, moreover, the errors of measurement of the same type of test are allowed to correlate. Two additional latent variables called test-specific factors (Jöreskog, 1970) are assumed to account for these correlations. It could be shown that model I fits the correlationmatrix in both the first year and the sixth year adequately. By using model II the fit of the model is significantly improved in the first year but not in the sixth year. Although, strictly speaking model, II is not needed to explain the correlations among the tests, the results indicate that some test-specific variance (variance due to the test being used) is present in the freshman years.
The predictive power of both the FT's and the GKT's appeared to be equal. The achievement of the cohort 1983-1984 could be predicted better for their second year than in the third year of medical school.
Finally the correlations between the FT's, the GKT's and other tests used to

measure medical knowledge during the first year of medical school are studied. As expected all the correlations were positive and most of them differed significantly from zero.
The hypothesised differences and similarities between the tests are strongly confirmed by the results. Indicating that, also from the external structure point of view, the tests demonstrate construct validity **(CHAPTER 4)**.

In the next chapter hypotheses about the structure in the test results are formulated as psychometric models. These models are assumed to be minitheories for the construct medical knowledge. They are used to predict and test the structure in the responses of the students on the test items. Three models are described: the recently developed Progress Test model (PTmodel), the well-known Raschmodel and the Mokkenscalemodel. When applied in this context all models assume that the outcome of the interaction between a respondent and a test item is determined by two factors: the knowledge of the respondent and the difficulty of the test item. In the Raschmodel and in the Mokkenscalemodel these factors are parameters, while in the PTmodel these factors are assumed to be normally distributed random variables. **(CHAPTER 5)**.

The testing results of the models showed that the fit of the PTmodel to all the test items is reasonable. Furthermore, at the level of the individual test items, about 50% of the items were Raschhomogeneous and about 34% of the items were congruent to the assumptions of the Mokkenscalemodel. These results indicate that about 80% of the individual test items have good measurment properties with monotoneously increasing probabilities of a correct answer in the latent variable. Furthermore the analysis with the PTmodel showed that the FT test items had higher probabilities of a correct answer at all levels, except the highest, of the latent variable. The growth in knowledge among freshmen were highest on FT items, while the other years showed a higher growth in knowledge on GKT items. The Raschhomogeneous test items of both tests appeared on one occasion to be unidimensional. The Mokkenscale items of both tests were highly correlated. These results indicate that also at the level of individual test items, the assumption that the two tests measure the same trait cannot be rejected.
The analysis of the construct representation (Embretson, 1983) indicate that easier test items seemed to be related to so-called basic knowledge, whereas more difficult test items are related to subjects dealing with the practice of medicine. The conclusion is that, also at the psychometric level, the tests behave as was hypothesised **(CHAPTER 6)**.

In the final chapter the results are discussed from four points of view:

- the meaning of the results for the use of progress tests in the freshman year.

- the results discussed from a cognitive point of view.

- the psychometric meaning of the results .

- the data-analytic findings in this study.

Although the results found in this study strongly indicate the construct validity of progress tests, the use of these tests for students with the low proficiency scores in the first year is not without problems. Suggestions to deal with the problem are discussed.
From the discussion of the results using the points of view mentioned, it is concluded that the results are congruent at the different theoretical levels. At each level new research questions emerged showing that the empirical domain of this subject is rich enough to justify further exploration.

Dankwoord

Nu het boek klaar is rest mij nog de plezierige taak om die mensen te bedanken zonder wie ik dit niet had kunnen volbrengen. Allereerst wil ik mijn gezinsleden noemen. Lieve Nelly voor jou was het allemaal niet nodig geweest. Maar toen ik na lang aarzelen het besluit had genomen dat het toch maar moest gebeuren heb jij je volledig achter deze onderneming gezet en het mij mogelijk gemaakt om dit project tot een goed einde te brengen. Ik ben jou hiervoor erg dankbaar. Met jou ben ik blij dat het voorbij is en we samen weer aan andere zaken kunnen gaan denken.
Nienke, Sanne en Jelle, jullie enthousiasme bij het groter zien worden van de hoeveelheid, in eerste instantie eigenhandig getypte, tekst was hartverwarmend. Evenzeer als jullie verschillende opmerkingen, waarvan ik de volgende nooit meer zal vergeten: "Pappa, jij bent aan het proeven voor doctorandus hè?".
Mijn promotor prof.dr. W. Wijnen en co-promotor dr. R. Does ben ik heel dankbaar voor de wijze waarop ze mij bij het schrijven van dit proefschrift hebben begeleid. Beste Wijnand en Ronald jullie waren voor mij een uitstekend duo, dat elkaar op een heel goede manier aanvulde. Zowel de onderwijskunde als de toegepaste statistiek op het terrein van onderzoek van onderwijs zouden er van kunnen profiteren als jullie ook in de toekomst gezamenlijk zouden gaan optreden!
De leden van de beoordelingscommissie ben ik zeer erkentelijk voor hun bereidheid om het manuscript te lezen en te beoordelen. Vooral prof.dr. M.J. Drop en prof.dr. G.M. Mellenbergh ben ik dankbaar voor het schriftelijk commentaar dat zij mij hebben gegeven. Dit commentaar heeft geleid tot een beter leesbaar boek.
Ook wil het JOC-I bedanken voor het meewerken aan dit onderzoek. De jaarcoördinator van destijds, dr. G.J. Tangelder, wil ik langs deze weg dank zeggen voor het coördineren van de itembeoordeling.
Michel Janssen was gedurende de laatste fase waarin de teksten en tabellen definitief bewerkt moesten worden mijn grote steun en toeverlaat. Beste Michel dankzij jou inzicht in en grote werklust op het terrein van tekstverwerking en lay-out kon mijn geschrijf uitgroeien tot het boek dat het nu is geworden. Wat ik hier ook zou schrijven, ik doe je altijd te kort! Toen jij in een kritieke fase even iets anders moest doen was Marliese van Wissen bereid om je gedurende twee dagen te vervangen. Ook haar wil ik bij deze bedanken voor de verrichte werkzaamheden.
Marcel Theunissen nam de diverse statistische berekeningen voor zijn rekening. Beste Marcel ik ben jou heel dankbaar voor het vele werk dat je voor me hebt verricht. Onze samenwerking kwam min of meer toevallig tot stand. Ik hoop nog vele jaren met jou te mogen samenwerken.
Hetty Snellen was de in het boek beschreven externe beoordelaar. Beste Hetty zonder jou had ik een aantal interessante zaken niet boven tafel gekregen. Toen ook jij met een selectie van toetsvragen kwam die in veel opzichten overeenkwam met de selectie van de blokcoördinatoren kreeg ik pas echt vertrouwen in het

onderzoek. Voor het door jou verrichte werk en je jarenlange collegialiteit ben ik je heel erg dankbaar.
Monique Laenen en Anneke Strouken speelden het klaar om uit mijn doorhalingen en toevoegingen toch steeds weer leesbare teksten te voorschijn te toveren. Waarlijk een staaltje van geduld en uithoudingsvermogen. Beste Monique en Anneke zonder jullie had ik me geen raad geweten. Heel erg bedankt voor jullie inzet.
Carla Fröberg zorde er voor dat ik over de tekst van de onderzochte toetsvragen kon beschikken. Beste Carla bedankt voor de snelheid en accuratesse waarmee je deze werkzaamheden hebt verricht.
Mijn collega's van het Project Evaluatie Studieresultaten ben ik dankbaar voor de ruimte die ze me hebben gegeven om dit onderzoek te voltooien. Beste collega's op deze wijze wil ik jullie bedanken voor jullie jarenlange collegialiteit. Ik hoop dat we ook in het nieuwe samenwerkingsverband dat we nu hebben afgeproken op een prettige wijze zullen kunnen samenwerken.
Betsy Kops en Nelly Kuijs waren bereid om de drukproeven te lezen en te corrigeren, waarvoor ik hun heel dankbaar ben. De fouten die er nu nog inzitten komen uiteraard geheel voor mijn rekening.
Micky Schim van der Loeff las de summary en corrigeerde de fouten. Beste Micky hartelijk dank voor jouw medewerking.
Tenslotte mijn paranimfen en vrienden Henk Meijer en Maarten Verwijnen. Henk mien jong, op dizze veur ons wat vrumde menaaier wil ik die bedanken veur de vrundschop dei wie aal zo'n dikke twinteg joar mit die en dien vrauw Ronnie hebben. Dast doe as peerdenimf optredst sprekt ja veur zuch. Beste Maarten al meer dan tien jaar werken jij en ik samen aan onze toets. De vele momenten van intensieve samenwerking bij jou of bij mij thuis en onze gezamenlijke reisjes, waarbij we altijd de hotelkamer deelden, hebben er toe bijgedragen dat we van collega's vrienden werden. Ik weet zeker dat ze ook in de toekomst nog van ons zullen horen. Maarten bedankt voor dit alles.

CURRICULUM VITAE

Ik werd in 1946 geboren te Sappemeer, waar ik de lagere school en de MULO bezocht. Na het behalen van het diploma van het laatst genoemde schooltype in 1962 ging ik naar de Rijkskweekschool voor Onderwijzers en Onderwijzeressen te Winschoten. In januari 1968 behaalde ik het diploma van Volledig Bevoegd Onderwijzer. Na het vervullen van de dienstplicht werd ik onderwijzer aan een LOM-school te Amsterdam (De Pieter Hoogland School).
In 1969 begon ik met de studie psychologie aan de Rijksuniveriteit Groningen. In april 1972 slaagde ik voor het kandidaatsexamen, waarna ik gedurende twee jaar op part-time basis als student-assistent heb gewerkt ten behoeve van het methodologie-onderwijs aan eerstejaars-studenten. Het doctoraalexamen psychologie behaalde ik in april 1976. Mijn afstudeerrichting was de persoonlijkheidsleer (Prof. dr W.K.B. Hofstee) met als specialisatie psychometrie en methoden en technieken (Prof. dr. I.W. Molenaar).
Nog voordat ik was afgestudeerd werd ik als wetenschappelijk-ambtenaar (later - medewerker) aangesteld bij de vakgroep Onderwijsontwikkeling en Onderwijsresearch van de Rijksuniversiteit Limburg (Prof. dr. W.H.F.W. Wijnen). In deze functie had ik samen met de artsen drs. H. Snellen, drs. B. Stalenhoef en drs. M. Verwijnen de verantwoordelijheid voor de implementatie en ontwikkeling van het evaluatiesysteem in de Faculteit der Geneeskunde. In 1980 werd ik leider van het Project Evaluatie Studieresultaten (PES). Dit projectleiderschap heb ik vrijwillig beëindigd toen ik in 1983 werd aangesteld als universitair docent bij de vakgroep Medische Informatica en Statistiek (dr. R.J.M.M. Does). De voornaamste accenten in de huidige functie zijn psychometrische ondersteuning van het Project Evaluatie Studieresultaten en psychometrisch onderzoek in het kader van het VF-programma Onderzoek van Onderwijs.