Toetsen en beslissen

Toetsing bij doorstroombeslissingen
in het voortgezet onderwijs

Cor Sluijter

Toetsen en beslissen

Toetsing bij doorstroombeslissingen

in het voortgezet onderwijs

ISBN 90-801795-4-x

Omslag: Daan Verwoert/Marije Hosper (Grafische Dienst Cito)

Toetsen en beslissen

Toetsing bij doorstroombeslissingen in het voortgezet onderwijs

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. J.J.M. Franse
ten overstaan van een door het college voor promoties ingestelde
commissie in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 27 november 1998 te 11.00 uur

door

Cornelis Sluijter

geboren te Amsterdam

promotores

Prof. dr. G.J. Mellenbergh
Prof. dr. P.F. Sanders

copromotor

Dr. P. Koele

overige leden promotiecommissie

Prof. dr. J.J.C.M. Hox
Prof. dr. W.J. van der Linden (Universiteit Twente)
Dr. M.A. Zwarts (Inspectie van het Onderwijs)

Voor Gerda, Soraya en Laryssa

Voorwoord

Een zin uit een gedicht van Judith Herzberg luidt: ’Het duurt altijd langer dan je denkt, ook als je denkt het zal wel langer duren dan ik denk dan duurt het toch nog langer dan je denkt’. Deze regel is volledig van toepassing op de totstandkoming van dit proefschrift. Evenzeer van toepassing is echter de volgende variant (van mijn hand): ’Je krijgt meer ondersteuning dan je denkt, ook als je denkt dat je wel meer ondersteuning zult krijgen dan je denkt dan krijg je toch nog meer ondersteuning dan je denkt’. Vele mensen hebben op een of andere wijze een bijdrage geleverd aan dit proefschrift.

Bijzonder erkentelijk ben ik mijn promotores Don Mellenbergh en Piet Sanders en mijn copromotor Pieter Koele voor hun vakkundige en inspirerende commentaar en hun geduld. Zij hebben mij bij het schrijven van dit proefschrift uitstekend begeleid. Piet Sanders was van dit drietal, als de behuizer van het aan mijn werkkamer grenzende vertrek, letterlijk het dichtst betrokken bij het tot stand komen van dit proefschrift. Omdat Piet vaak beschikbaar was, vonden er veelvuldig informele discussies plaats; soms over principiële zaken, maar vaak ook over formuleringen, woordkeus, of het al dan niet plaatsen van een komma. Mochten lezers dan ook in dit proefschrift stuiten op passages die in hun ogen op de laatstgenoemde aspecten te kort schieten, dan kunnen ze daarover met Piet in discussie treden.

De ontwikkeling van het itemmateriaal voor Nederlands, Engels en wiskunde en de toetsen waar dit proefschrift over gaat, was het werk van respectievelijk Willem van Roosmalen, Erna Gille en Harm Boertien en hun constructieteams. Voor het verzamelen, selecteren en aanpassen van het itemmateriaal voor de checklist studievaardigheid die in dit proefschrift zijdelings aan de orde komt, was Ton Heuvelmans verantwoordelijk. De ontwikkeling van de checklist zelf kwam voor rekening van Pien de Wit. Hedda van ’t Land en Jessica van Dijk voerden onder mijn supervisie een onderzoek uit waarvan een deel in hoofdstuk vijf van dit proefschrift ter sprake komt. Met Joke Harte werkte ik samen bij de opzet en uitvoering van een onderzoek dat eveneens in hoofdstuk vijf aan de orde komt. Marianne Sanders screende mijn Engelstalige samenvatting. Al deze mensen wil ik van harte danken voor hun inspanningen.

Naast Ton en Pien wil ik nog een aantal collega's van de afdeling OPD met name noemen. De formidabele Huub Verstralen ontwikkelde programmatuur waarvan ik bij de totstandkoming van dit proefschrift dankbaar gebruik heb gemaakt en voerde bovendien de OPLM-analyses uit. Frans Kamphuis heeft de ontbrekende gegevens in mijn databestanden geïmputeerd. Frans Kleintjes en mijn kamergenoot Alfred Verschoor hadden altijd een luisterend oor en soms goede suggesties, wanneer ik het nodig vond om hen deelgenoot te maken van mijn vorderingen en problemen bij het proefschriftschrijven. Alfred heeft bovendien de foto gemaakt die de voorkant van dit proefschrift siert. Mijn dank voor dit alles is groot.

De assistentie van Wil Kuysten-Speijers en Betty de Lavaletta was een noodzakelijke voorwaarde voor het welslagen van het project waarbinnen de plaatsingstoetsen en de checklist studievaardigheid ontwikkeld zijn. Zij hebben veel werk verzet en goed werk gedaan bij het ontwikkelen van het itemmateriaal, het opzetten en uitwerken van de proefafnames, het verwerken van de resultaten van de proefafnames en het opstellen van de handleidingen bij de ontwikkelde instrumenten. Ik ben Wil en Betty uitermate dankbaar. Zij waren voor het project en voor mij persoonlijk een onmisbare steun en toeverlaat.

Een andere noodzakelijke voorwaarde voor het welslagen van het project was de medewerking aan de proefafnames van vele docenten Nederlands, Engels en wiskunde en nog veel meer leerlingen. Ook naar hen gaat mijn dank uit.

Mijn werkgever, het Cito, ben ik erkentelijk voor de faciliteiten die het schrijven van dit proefschrift mede mogelijk hebben gemaakt.

Ten slotte wil ik drie mensen danken voor een bijdrage van geheel andere aard. Mijn vrouw Gerda dank ik voor vele dingen, maar met name voor het feit dat ze de laatste maanden streng toezicht heeft uitgeoefend op mijn werkzaamheden thuis, hetgeen de afronding van dit proefschrift aanzienlijk heeft bespoedigd. En mijn dochters Soraya en Laryssa wil ik ervoor danken dat ze de afgelopen maanden regelmatig gezorgd hebben voor wat (voor Gerda acceptabele) afleiding.

Cor Sluijter,
Amsterdam, september 1998

Inhoudsopgave

1 Inleiding

Bij het nemen van beslissingen over personen is vaak sprake van onzekerheid. Zo is het bijvoorbeeld nooit absoluut zeker dat een leerling die voor het vwo kiest deze onderwijsvorm ook met succes zal afronden. Evenmin is het ooit zeker dat een bepaalde persoon uit een groep sollicitanten ook werkelijk het meest geschikt is voor een of andere functie.

Psychologisch onderzoek van beslisprocessen op allerlei terreinen, waaronder het onderwijs (Cooksey, 1988), heeft duidelijk gemaakt dat de mens in de regel geen goede voorspeller en beslisser is en dat het mogelijk is de kans op een correcte beslissing te vergroten (Meehl, 1954; Kahneman, Slovic & Tversky, 1982; Arkes & Hammond, 1986; Brehmer & Joyce, 1988). Uiteraard geldt hierbij dat beslissingen onder onzekerheid per definitie een probabilistisch karakter hebben: de informatie die ter beschikking staat bij het nemen van dergelijke beslissingen levert nooit meer dan een feilbare indicatie van toekomstig gedrag. Het vergroten van de kans op een correcte beslissing is daarom slechts tot op zekere hoogte mogelijk.

Tests en toetsen

De kwaliteit van beslissingen over personen onder onzekerheid is te verbeteren door gebruik te maken van tests. Een test is een gestandaardiseerde en systematische meetprocedure die het mogelijk maakt kwantitatieve uitspraken te doen over eigenschappen van personen, met het doel om beslissingen te nemen. Deze definitie is uitgebreider dan de in de literatuur gangbare definities (zie bijvoorbeeld Cronbach, 1990, Drenth & Sijtsma, 1990, Nederlands Instituut van Psychologen, 1988). In de gangbare definities ligt het accent op de test als meetinstrument, terwijl de hier gehanteerde definitie expliciet aangeeft dat meten geen doel op zich is, maar een middel om beslissingen te nemen (vgl. Van der Linden & Mellenbergh, 1978).

Het begrip toets is een verbijzondering van het begrip test. De term toets is gereserveerd voor instrumenten die door onderwijs en studie verworven kennis, inzicht of vaardigheden meten. Wordt de term test in niet-generieke zin gebruikt, dan verwijst deze naar instrumenten die eigenschappen van personen meten die niet door intentioneel onderwijs en studie verworven zijn (vgl. De Groot & Van Naerssen, 1969).

Tests kunnen bij juist gebruik een belangrijke bijdrage leveren aan de kwaliteit van beslissingen over personen: '*... the proper use of well-constructed and validated tests provides a better basis for making some important decisions about individuals ... than would otherwise be available.*' (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985). De gedachte achter het gebruik van tests is dat het nemen van beslissingen over personen met de grootst mogelijke zorgvuldigheid dient te gebeuren, zowel in het belang van de persoon of instantie die de beslissing neemt, als in het belang van de persoon of personen op wie de beslissing betrekking heeft.
Al ver voor het begin van de jaartelling maakte de Chinese overheid voor het selecteren van ambtenaren gebruik van tests. Ook het gebruik van toetsen in het onderwijs kent al een lange traditie. De ontwikkeling van de eerste toetsen met bijbehorende vaste afnameregels wordt toegeschreven aan de orde der Jezuïeten en vond plaats in de zestiende eeuw (Du Bois, 1970; Wainer, 1987). Het gebruik van tests heeft echter pas deze eeuw een grote vlucht genomen. Binnen het Nederlandse taalgebied zijn bijvoorbeeld bijna vierhonderd tests geregistreerd (Evers, Van Vliet-Mulder & Ter Laak, 1992), die onder meer een rol spelen bij psychodiagnostiek en bij personeelsselectie in het bedrijfsleven en bij de overheid.

Het Nederlands onderwijs maakt op grote schaal gebruik van toetsen. Een bekend voorbeeld van een veelvuldig gebruikte toets is de jaarlijks door het Cito uitgebrachte Eindtoets Basisonderwijs, in de volksmond vaak de 'Cito-toets' geheten. Van deze toets, die informatie verschaft over leerlingen ten behoeve van de overgang naar het voortgezet onderwijs, verschijnt elk jaar een nieuwe versie. In de periode 1980-1991 lag het aantal deelnemers aan deze toets tussen de 75.000 en 100.000. Sinds 1992 ligt het aantal deelnemers zelfs boven de 100.000 (Uiterwijk, 1994). In 1997 name ruim 122.000 leerlingen deel (Cito Instituut voor Toetsontwikkeling, 1997). Twee andere

bekende en massaal gebruikte toetsen, die eveneens door het Cito worden uitgebracht, zijn de centrale eindexamens en de afsluitingstoetsen voor de basisvorming.

Tests kunnen de kwaliteit van beslissingen over personen in verschillende opzichten verbeteren. In de eerste plaats kan het gebruik van tests de kans op een correcte beslissing verhogen. Een tweede voordeel van het gebruik van tests is dat de objectiviteit van het beslisproces erdoor toeneemt. Willekeur bij het nemen van beslissingen kan voorkomen worden. Een derde voordeel is dat de transparantie van het beslisproces voor de betrokkenen groter is. Het valt aan de persoon of personen op wie de beslissing betrekking heeft duidelijk te maken volgens welke procedure een beslissing tot stand gekomen is of tot stand zal komen. Een vierde voordeel is dat het gebruik van tests efficiënte beslisprocessen mogelijk maakt. Tests bieden de gelegenheid om op snelle en relatief eenvoudige wijze beslissingen te nemen over grote groepen personen. Een vijfde voordeel is dat het gebruik van tests kwantitatieve gegevens oplevert. Met behulp van deze gegevens valt wetenschappelijk onderzoek uit te voeren om de kwaliteit en bruikbaarheid te verbeteren van tests en de beslisprocessen waar tests een rol bij spelen.

Doorstroombeslissingen in het onderwijs

Zowel aan het einde van het basisonderwijs, als in het voortgezet onderwijs doen zich voor leerlingen belangrijke keuzemomenten voor. Aan het einde van alle schooljaren die voorafgaan aan het eindexamenjaar vinden in het voortgezet onderwijs doorstroombeslissingen plaats. Leerlingen moeten op die momenten al dan niet bindende adviezen krijgen over de beste voortzetting van hun leerweg.

Welke aard een doorstroombeslissing heeft, is afhankelijk van het leerjaar waarin leerlingen zich bevinden en de categoriale onderwijsvormen die een school aanbiedt. Binnen categoriale vormen van voortgezet onderwijs dienen docenten aan het eind van een leerjaar te bepalen of leerlingen over kunnen gaan naar het volgende leerjaar van de betreffende onderwijsvorm. Indien leerlingen zich nog niet in een categoriale vorm van onderwijs bevinden, dan spreekt men bij doorstroombeslissingen van determinatie of van ‘streaming’.

Van determinatie is sprake, indien scholen aan het einde van een leerjaar moeten bepalen voor welke categoriale vorm van onderwijs hun leerlingen het meest geschikt zijn. Van streaming is sprake, indien scholen leerlingen indelen in klassen die nog niet volledig homogeen van samenstelling zijn, zoals bijvoorbeeld mavo/havo- en havo/vwo-klassen op een scholengemeenschap voor mavo, havo en vwo. Voor een recente gedetailleerde beschrijving van verschillende kenmerken van de doorstroming van leerlingen in de eerste vier leerjaren van het voortgezet onderwijs wordt verwezen naar Bos, Cremers-Van Wees en Lugthart (1996).

Het probabilistisch karakter van doorstroombeslissingen

Omdat het voorspellen van de toekomstige leerprestaties van leerlingen slechts tot op zekere hoogte mogelijk is, zullen zich altijd fouten voordoen bij het nemen van doorstroombeslissingen. Het is dan ook niet verrassend dat een zeker deel van de leerlingen vertraging oploopt gedurende de schoolloopbaan. De meest recente gegevens op dit terrein (Ministerie van Onderwijs, Cultuur en Wetenschappen, 1996) geven aan dat 23,9 procent van alle vbo-leerlingen en 31,2 procent van de mavo-leerlingen één jaar vertraging of meer oploopt. Voor havo en vwo bedraagt het percentage leerlingen dat na vijf jaar één jaar vertraging of meer heeft opgelopen respectievelijk 52,0 en 24,3 procent.

Zittenblijven is de belangrijkste oorzaak van het oplopen van vertraging. Het percentage zittenblijvers is in de leerjaren direct vóór het eindexamen het grootst. Zo bleef in het schooljaar 1993-1994 in het derde leerjaar van vbo en mavo respectievelijk 6,7 en 9,9 procent van de leerlingen zitten. In datzelfde schooljaar doubleerde in het vierde leerjaar van de havo 16,8 procent van de leerlingen en in het vijfde leerjaar van het vwo 9,5 procent (Ministerie van Onderwijs, Cultuur en Wetenschappen, 1996).

Verder lopen leerlingen in de regel ook vertraging op, wanneer zij ‘afstromen’ of ‘opstromen’. Van afstromen is sprake wanneer leerlingen naar een lagere categoriale vorm van voortgezet onderwijs gaan. Bij opstromen gaan leerlingen zonder diploma naar een hogere categoriale vorm van voortgezet onderwijs. De oorzaak van de vertraging die leerlingen

oplopen bij opstromen en afstromen is dat zij in het volgend schooljaar in de nieuwe onderwijsvorm veelal niet in een hoger leerjaar belanden. In het schooljaar 1993-1994 stroomde 2,7 procent van alle leerlingen die zich in de leerjaren twee tot en met vier van de mavo bevonden af naar het vbo, 3,0 procent van de leerlingen uit de leerjaren twee tot en met vijf van de havo stroomde af naar de mavo en 3,8 procent van de leerlingen uit de leerjaren twee tot en met zes van het vwo stroomde af naar de havo. In totaal stroomden ruim 15.000 leerlingen - ongeveer 1,7 procent van alle leerlingen in de leerjaren twee of hoger - af. Het opstromen van leerlingen is een verschijnsel dat minder vaak voorkomt. In het schooljaar 1993-1994 stroomde 0,5 procent van alle leerlingen die zich in de leerjaren twee tot en met vier van het vbo bevonden op naar de mavo. Een zelfde percentage leerlingen stroomde uit de leerjaren twee tot en met vier van de mavo op naar de havo en uit de leerjaren twee tot en met vijf van de havo naar het vwo. In totaal stroomden bijna 2000 leerlingen - ongeveer 0,2 procent - op (Ministerie van Onderwijs, Cultuur en Wetenschappen, 1996).

Het probabilistisch karakter van doorstroombeslissingen komt niet alleen tot uiting in het zittenblijven en op- en afstromen. Een deel van de leerlingen verlaat het onderwijs zelfs in het geheel zonder diploma. Zo valt op grond van de onderwijsmatrix 1994 (Kapel & Roessingh, 1996) en gegevens over de absolute uitval uit verschillende onderwijstypen (Ministerie van Onderwijs, Cultuur en Wetenschappen, 1996) te berekenen dat in het schooljaar 1992-1993 van alle leerlingen die het reguliere voortgezet onderwijs vanuit de leerjaren twee, drie en vier van de mavo verlieten circa 5,8 procent dit zonder diploma deed. Van alle leerlingen die het reguliere voortgezet onderwijs vanuit de leerjaren vier en vijf van de havo verlieten deed circa 10,0 procent dit ongediplomeerd. En van alle leerlingen die het reguliere voortgezet onderwijs vanuit de leerjaren vier, vijf en zes van het vwo verlieten deed circa 13,9 procent dit zonder diploma.

Het subjectieve karakter van doorstroombeslissingen

Het subjectieve karakter van doorstroombeslissingen komt tot uiting, wanneer docenten twijfelen over de juistheid van een beslissing, of het oneens zijn over de vraag welke leerweg het meest geschikt is voor een

leerling. Ook kunnen er verschillen van mening optreden tussen docenten enerzijds en de leerling of de ouders of verzorgers van de leerling anderzijds. Een enquête gehouden onder conrectoren onderbouw op scholengemeenschappen (Sluijter, 1988a) maakt duidelijk dat docenten regelmatig twijfelen over de juistheid van genomen doorstroombeslissingen. Op de vraag bij welk percentage van de leerlingen sprake was van twijfel binnen het docentenkorps omtrent de meest geschikte voortzetting van de leerweg, antwoordde slechts 1,2 procent van de respondenten dat twijfel zich nooit voordeed. Verder noemde 47,3 procent van de respondenten een percentage tussen de nul en tien; 26,4 procent vulde een percentage in tussen de tien en twintig en 10,8 procent gaf aan dat het percentage groter was dan twintig. Het percentage respondenten dat geen antwoord gaf op de betreffende vraag bedroeg 14,3.

Ook het optreden van verschillen van mening tussen docenten onderling over de te nemen beslissing is geen onbekend fenomeen. Op de vraag bij welk percentage van de leerlingen zich tussen docenten onderling verschillen van mening voordeden, antwoordde in het voornoemde onderzoek 3,5 procent van de respondenten dat daar nooit sprake van was. Daarnaast gaf 44,2 procent van de respondenten aan dat het percentage tussen de nul en tien lag; 18,0 procent noemde een percentage tussen de tien en twintig en 13,5 procent vulde een percentage van meer dan twintig in. Het percentage respondenten dat geen antwoord gaf op de betreffende vraag bedroeg 20,9.

Op de vraag bij welk percentage van de leerlingen zich verschillen van mening voordeden over de te nemen beslissing tussen docenten enerzijds en leerlingen of hun ouders/verzorgers anderzijds, antwoordde 3,7 procent van de respondenten dat dit probleem zich nooit voordeed. Verder noemde 67,1 procent van de respondenten een percentage tussen de nul en tien; 13,1 procent gaf een percentage tussen de tien en twintig op en 3,9 procent van de respondenten vulde een percentage van meer dan twintig in. Het percentage respondenten dat geen antwoord gaf op de betreffende vraag bedroeg 12,3.

Toetsen en doorstroombeslissingen

De overheid legt uit het oogpunt van kostenbeheersing een steeds grotere nadruk op het rendement van het voortgezet onderwijs. Het streven is de gemiddelde verblijfsduur van leerlingen in de verschillende categoriale onderwijsvormen terug te dringen en de uitstroom van ongediplomeerde leerlingen zoveel mogelijk te beperken. Door de steeds grotere nadruk op het rendement krijgen leerlingen en scholen steeds minder gelegenheid om keuzes die achteraf verkeerd blijken weer recht te zetten. Voor leerlingen wordt het steeds belangrijker om op ieder keuzemoment de juiste voortzetting van een leerweg te kiezen. Voor scholen neemt eveneens de noodzaak toe om de doelmatigheid van doorstroombeslissingen te verhogen. Naarmate scholen beter in staat zijn hun leerlingen in de juiste leerwegen onder te brengen, zullen zij immers minder moeite hebben de kwaliteit van hun onderwijs op het gewenste niveau te brengen of te houden.

Toetsen die specifiek ontwikkeld zijn om de toekomstige prestaties van leerlingen te voorspellen, kunnen een rol spelen bij het verhogen van het rendement van doorstroombeslissingen. Dergelijke toetsen bieden scholen de gelegenheid om op een efficiënte en transparante wijze tot een goed onderbouwd advies te komen over de meest geschikte voortzetting van een leerweg. Verder geeft de informatie die de toetsen leveren een concreet aanknopingspunt voor het voeren van discussies tussen docenten onderling, of tussen docenten enerzijds en leerlingen of hun ouders anderzijds. De toetsen bieden dus de gelegenheid om verschillen van mening te beslechten.

Het Cito houdt zich al een aantal jaren bezig met het ontwikkelen van toetsen die een rol kunnen spelen bij het nemen van doorstroombeslissingen. Om hun functie te onderstrepen hebben deze toetsen de naam plaatsingstoetsen gekregen. Zo zijn in 1993 twee reeksen plaatsingstoetsen op de markt gebracht die scholen konden gebruiken om onderscheid te maken tussen potentiële mavo- en havo-leerlingen en potentiële havo-en vwo-leerlingen aan het einde van het eerste leerjaar (Cito Instituut voor Toetsontwikkeling, 1993a, 1993b, 1993c, 1993d, 1993e, 1993f). En in 1996 zijn in opdracht van het ministerie van Onderwijs, Cultuur en Wetenschappen door het Cito twee reeksen plaatsingstoetsen uitgegeven ter ondersteuning van doorstroombeslissingen aan het einde van het derde leerjaar havo en vwo (Cito Instituut voor Toetsontwikkeling, 1996a; Cito Instituut voor Toetsontwikkeling, 1996b).

Inhoud van dit proefschrift

In dit proefschrift staan twee thema's centraal. Het eerste thema is de ontwikkeling van plaatsingstoetsen. Dit thema komt aan de orde in de hoofdstukken twee, drie en vier. Het onderwerp van hoofdstuk twee is de constructie van plaatsingstoetsen als hulpmiddel bij doorstroombeslissingen aan het einde van het derde leerjaar van havo en vwo. In dit hoofdstuk komt allereerst de vraag aan de orde binnen welke randvoorwaarden de plaatsingstoetsen moeten functioneren en aan welke eisen de toetsen, gegeven hun doelstelling, moeten voldoen. Daarna komen de opzet, uitvoering en resultaten aan bod van de proefafnames die gehouden zijn om de kwaliteit te bepalen van de opgaven die op grond van de toetsspecificaties ontwikkeld zijn. Vervolgens vindt een beschrijving plaats van de wijze waarop selectie van de opgaven voor opname in de toetsen heeft plaatsgevonden. Ten slotte wordt nader ingegaan op de wijze waarop de toetsscores genormeerd zijn. In de hoofdstukken drie en vier volgt een beschrijving van het onderzoek dat heeft plaatsgevonden om de kwaliteit van de toetsen te bepalen. Hoofdstuk drie heeft betrekking op de meetnauwkeurigheid van de toetsen en hoofdstuk vier heeft de validiteit van de toetsen als onderwerp

Het tweede thema dat in dit proefschrift aan de orde komt - in hoofdstuk vijf - is de besliskundige analyse van doorstroombeslissingen en de rol die plaatsingstoetsen bij het nemen van doorstroombeslissingen kunnen spelen. Binnen de besliskunde zijn drie benaderingen te onderscheiden: een normatieve, een descriptieve en een prescriptieve (Bell, Raiffa & Tversky, 1988). Hoofdstuk vijf behandelt de eerste twee van deze benaderingen.

De normatieve besliskunde heeft tot doel methoden te ontwikkelen met behulp waarvan personen, groepen en instanties hun vanuit rationeel oogpunt bezien suboptimale processen van besluitvorming kunnen optimaliseren. In het eerste deel van hoofdstuk vijf is sprake van een toepassing van de normatieve besliskunde op doorstroombeslissingen: aangegeven wordt hoe deze met behulp van de psychometrische besliskunde (Cronbach & Gleser, 1965) te optimaliseren zijn.

In het tweede deel van hoofdstuk vijf komt de descriptieve besliskunde aan bod. Deze heeft tot doel te onderzoeken op welke wijze feitelijke beslissin-

gen, oordelen en keuzes van personen, groepen en organisaties kunnen worden begrepen, verklaard en zo mogelijk voorspeld. In dit deel wordt verslag gedaan van onderzoek dat bedoeld is om de overeenkomsten en verschillen tussen docenten vast te stellen bij het bepalen van de meest geschikte categoriale onderwijsvorm voor hun leer-lingen. Het onderzoek beschrijft welke gegevens in het algemeen van belang zijn bij dit soort beslissingen en laat zien welke rol toetsresultaten daarbij spelen.

De prescriptieve besliskunde heeft tot doel methoden te ontwikkelen die de kwaliteit van beslissingen daadwerkelijk kunnen verbeteren, door gebruik te maken van de inzichten die binnen de twee andere benaderingen verworven zijn. De prescriptieve besliskundige benadering valt echter buiten het kader van dit proefschrift.

2 De constructie van de plaatsingstoetsen

Eind 1993 kreeg het Cito van het ministerie van Onderwijs, Cultuur en Wetenschappen de opdracht instrumentaria te ontwikkelen die scholen ondersteuning konden bieden bij het nemen van beslissingen over de doorstroming van leerlingen na het derde leerjaar van de havo en het derde leerjaar van het vwo.

Zoals al in hoofdstuk een aangeduid, is aan deze opdracht gevolg gegeven door zowel voor drie havo als voor drie vwo een reeks plaatsingstoetsen te ontwikkelen. Het betrof hier toetsen voor de vakken Nederlands, Engels en wiskunde. Daarnaast is een checklist studievaardigheid ontwikkeld die het docenten mogelijk maakt de studievaardigheid van een leerling systematisch te beoordelen en in een score uit te drukken. Dit hoofdstuk geeft een beschrijving van de wijze waarop de toetsen ontwikkeld zijn. De checklist komt in dit proefschrift alleen zijdelings aan de orde. Voor een verslag van de ontwikkeling, normering en validering van de checklist wordt verwezen naar De Wit, Heuvelmans & Sluijter (1995) en naar de handleidingen bij de instrumentaria die voor drie havo en drie vwo ontwikkeld zijn (Cito Instituut voor Toetsontwikkeling, 1996a, 1996b).

2.1 Uitgangspunten

Aan de keuze voor de ontwikkeling van toetsen voor de vakken Nederlands, Engels en wiskunde en een checklist studievaardigheid lagen verschillende overwegingen ten grondslag. Er is voor gekozen meer dan één instrument te ontwikkelen, omdat er vele vaardigheden van belang zijn voor het succesvol vervolgen van een havo- of vwo-opleiding. Het aantal te ontwikkelen instrumenten is beperkt tot vier, omdat de gezamenlijke voorspellende waarde van een reeks instrumenten steeds minder toeneemt naarmate het aantal instrumenten in de reeks groeit.

De voorspellende waarde van een reeks instrumenten is hoger, naarmate de vaardigheden waarop ze betrekking hebben sterker met het te voorspellen criterium in verband staan en conditioneel op deze samenhang onderling minder sterk met elkaar samenhangen. Er is gekozen voor de constructie van toetsen voor de vakken Nederlands, Engels en wiskunde, omdat de verwachting was dat deze instrumenten de gewenste eigenschappen zouden hebben.

Naast schoolse vaardigheden zijn ook andere eigenschappen van leerlingen bepalend voor toekomstig studiesucces. Om daaraan recht te doen, is naast de toetsen de voornoemde checklist studievaardigheid ontwikkeld. Zoals eerder aangegeven, besteedt dit proefschrift geen aandacht aan de ontwikkeling, normering en validering van deze checklist.

Toetsconstructie is een complex en tijdrovend proces dat uit een reeks opeenvolgende stappen bestaat (Millman & Greene, 1989; Sanders & Eggen, 1993). Belangrijke stappen zijn:
– het specificeren van het doel van de toets;
– het specificeren van de kenmerken van de toets;
– het construeren van items op basis van de toetsspecificaties;
– het onderzoeken van de eigenschappen van de geconstrueerde items door het houden van een proefafname;
– het samenstellen van de toets door het selecteren van de items die de toets het best aan de specificaties laten voldoen;
- het opstellen van normtabellen die hulp bieden bij het interpreteren van de scores op de toets;
– het onderzoeken van de kwaliteit (betrouwbaarheid en validiteit) van de toets.

In dit hoofdstuk komt een aantal van deze stappen aan bod, te beginnen met het specificeren van de functie en kenmerken van de toetsen die ontwikkeld zijn. Daarna volgt een bespreking van de constructie van de items en de opzet, uitvoering en resultaten van de proefafnames. De laatste twee onderwerpen van dit hoofdstuk betreffen het samenstellen en het normeren van de toetsen.

Het Cito houdt zich al een aantal jaren bezig met het ontwikkelen van toetsen die een indicatie geven van het te verwachten prestatieniveau van een leerling in verschillende leerwegen en die daarom een rol kunnen spelen bij het nemen van doorstroombeslissingen (Sluijter, 1988b; Sluijter, Boertien, De Klijn & van Roosmalen, 1991; Van Roosmalen & Sluijter, 1991; Sluijter, 1995; Sluijter, 1998).
Om hun functie te benadrukken hebben deze toetsen de naam plaatsingstoets gekregen. Er is voor gekozen zowel voor drie havo als voor drie vwo drie plaatsingstoetsen te ontwikkelen, omdat er dan sprake is van een goede balans tussen het streven naar een zo hoog mogelijke voorspellende waarde en het streven leerlingen niet overmatig te belasten. Verder is besloten in iedere toets een zodanig aantal items op te nemen dat de afnametijd per toets maximaal twee lesuren van 45 minuten zou bedragen, eveneens om leerlingen niet overmatig te belasten.

Een volgende eis was dat de toetsen curriculumonafhankelijk dienden te zijn. Dit betekent dat de items in de toetsen gebaseerd moeten zijn op leerstof die algemeen in het derde leerjaar van de havo of het vwo (of daarvoor) aan de orde geweest is. Alle leerlingen die in aanmerking komen om de toetsen te maken moeten in het voorafgaande onderwijs de gelegenheid gehad hebben om zich de kennis en vaardigheden eigen te maken waar de items in de toetsen betrekking op hebben. Plaatsingstoetsen mogen geen items bevatten die een aanspraak doen op kennis die slechts aan een deel van de leerlingen onderwezen is. Bevatten plaatsingstoetsen namelijk dergelijke items, dan is het mogelijk dat de voorspellende waarde van de toetsen vertekend zal worden voor leerlingen aan wie een deel van de relevante kennis en vaardigheden nooit onderwezen is. De toetsen hoeven het voorafgaande onderwijsaanbod echter niet te dekken. Het is voldoende als de toetsen zich richten op onderdelen van de leerstof waarvan de beheersing van belang is voor het succesvol vervolgen van een havo- of vwo-opleiding.

Ook is het gewenst dat de potentiële gebruikers van de toetsen de indruk hebben dat ze bruikbaar zijn voor het ondersteunen van doorstroombeslissingen aan het einde van het derde leerjaar havo en vwo. Scholen zullen immers slechts geneigd zijn toetsresultaten in hun oordeel te verdisconteren, als zij van mening zijn dat deze een positieve bijdrage kunnen leveren aan de kwaliteit van hun oordeel.

Ontwikkeling van de items voor het vak Nederlands

Bij het vak Nederlands is er voor gekozen toetsen voor leesvaardigheid te ontwikkelen, omdat de verwachting was dat dergelijke toetsen een voorspellende waarde zouden hebben ten aanzien van studiesucces in de bovenbouw van havo en vwo. Goed kunnen lezen is bij alle vakken in de bovenbouw van het voortgezet onderwijs belangrijk, omdat leerlingen de meeste leerstof op schrift krijgen aangeboden. Naarmate leerlingen leesvaardiger zijn, hebben zij volgens deze opvatting een grotere kans om in de bovenbouw naar behoren te presteren. Een tweede reden om voor het toetsen van leesvaardigheid te kiezen is dat deze vaardigheid een belangrijk onderdeel is van het vak Nederlands in de bovenbouw van de havo en het vwo en ook aan de orde komt in het centraal eindexamen.

Voor het meten van leesvaardigheid zijn drie soorten items ontwikkeld: meerkeuze-items bij teksten, kort-open-antwoord-items bij teksten en zogeheten cloze-items. Bij clozetoetsing zijn er woorden weggelaten uit een tekst en is het de taak van een leerling de ontbrekende woorden in te vullen. Clozetoetsing is gebleken een valide alternatieve procedure voor het meten van leesvaardigheid te zijn (zie Staphorsius, 1994). Het weglaten van woorden kan onder meer gebeuren volgens een vast stramien – bijvoorbeeld ieder vijfde woord – of volgens een procedure die gebaseerd is op tekstinhoudelijke overwegingen. Voor deze laatste vorm is hier gekozen.

De constructie van de kort-open-antwoord- en de cloze-items vond plaats in de eerste maanden van 1994. De items zijn ontwikkeld in samenwerking met een constructieteam bestaande uit docenten met ruime ervaring in het geven van onderwijs in de onder- en bovenbouw van havo en vwo. Het constructieproces leverde bij beide soorten items twee informatieve en twee beschouwende teksten op. Het totaal aantal items bij de vier cloze-teksten bedroeg 51. Het totaal aantal kort-open-antwoord-items bij de vier andere teksten bedroeg 45.

De meerkeuze-items zijn afkomstig uit een verzameling items die al in de jaren 1980-1985 was ontwikkeld. Ieder item uit deze verzameling bestaat uit een tekst van enkele regels met daarbij één meerkeuze-item dat betrekking

heeft op een belangrijk aspect van leesvaardigheid. Een deel van de items uit de verzameling zijn uitgegeven in de map Leestoetsen Nederlands Onderbouw Voortgezet Onderwijs (Cito Instituut voor Toetsontwikkeling, 1986). Uit de niet-gepubliceerde items zijn er 28 geselecteerd. Van deze items hebben er veertien betrekking op het kunnen herkennen van het hoofdonderwerp van een tekst en veertien op het herkennen van de hoofdgedachte in een tekst.

Uit elk van deze drie verzamelingen van items zijn zes toetsboekjes samengesteld volgens een hierna te beschrijven opzet. Het itemmateriaal voor de proefafname Nederlands bestond derhalve in totaal uit achttien toetsboekjes. Ieder toetsboekje had een geschatte afnametijd van één lesuur.

Ontwikkeling van de items voor het vak Engels

Om voor het vak Engels te bepalen welke vreemdtalige vaardigheden essentieel zijn voor studiesucces in de bovenbouw van havo of vwo, werd een voorstudie (Instituut voor Didactiek en Onderwijspraktijk, 1994) uitgevoerd. De conclusie van deze voorstudie was dat ook bij Engels de toetsing zich op verschillende aspecten van leesvaardigheid zou moeten richten. Omdat twee van deze aspecten - inzicht in tekstsamenhang en het hanteren van compenserende strategieën - ook aspecten van schrijfvaardigheid vormen, is ervoor gekozen eveneens een aantal items voor het meten van schrijfvaardigheid te construeren. De constructie van de items vond plaats in de eerste maanden van 1994. De items zijn ontwikkeld in samenwerking met een groep docenten met ruime ervaring in het geven van onderwijs in de onder- en bovenbouw van havo en vwo.

Voor het meten van leesvaardigheid Engels zijn naast traditionele meerkeuze-items ook cloze-items van het meerkeuzetype ontwikkeld. Bij de constructie van de cloze-items is er, net als bij het vak Nederlands, voor gekozen woorden uit teksten weg te laten op grond van tekstinhoudelijke overwegingen. Bij ieder ontbrekend woord konden leerlingen bij het invullen kiezen uit drie alternatieven. In totaal zijn er 49 cloze-items geconstrueerd bij vier Engelse teksten.

De verzameling geconstrueerde traditionele meerkeuze-items bestond uit:

– vijftien meerkeuze-items bij twee lange Engelse teksten;
– vijftien meerkeuze-items met Nederlandse alternatieven bij alinea's, waarbij de betekenis van een niet bestaand Engels woord op grond van de context bepaald moet worden;
– twintig meerkeuze-items met Nederlandse alternatieven, waarbij de betekenis van een woord uit een woordformatie met behulp van een gegeven verwant woord moet worden afgeleid.

Voor schrijfvaardigheid zijn in totaal 46 items ontwikkeld, te weten:
– tien items die zich richten op het gebruik van een woordenboek;
– tien items die zich richten op vaardigheid in het formuleren;
– dertien items die zich richten op vaardigheid in het parafraseren;
– dertien items die zich richten op de vaardigheid om zinnen of alinea's op de juiste wijze met elkaar te verbinden.

Uit elk van deze drie verzamelingen van items zijn zes toetsboekjes samengesteld volgens een hierna te beschrijven opzet. Het itemmateriaal voor de proefafname Engels bestond derhalve in totaal uit achttien toetsboekjes. Ieder toetsboekje had een geschatte afnametijd van één lesuur.

Ontwikkeling van de items voor het vak wiskunde

Bij het vak wiskunde is er voor gekozen toetsen te ontwikkelen die kennis en vaardigheden meten die een hoge gebruiksfrequentie hebben, nodig zijn in de toekomstige leerjaren en bovendien moeilijk aan te leren zijn. Deze keuze is gemaakt, omdat de verwachting bestond dat dergelijke toetsen een voorspellende waarde zouden hebben ten aanzien van studiesucces in de bovenbouw van havo en vwo. Immers, naarmate leerlingen de later frequent benodigde kennis en vaardigheden meer beheersen, zullen ze beter in staat zijn het toekomstig wiskunde-onderwijs te volgen. Dat geldt des te sterker voor kennis en vaardigheden die moeilijk zijn aan te leren, omdat het voor leerlingen zwaar zal zijn achterstanden in de beheersing van deze kennis en vaardigheden in te halen in de bovenbouw van havo en vwo. Een voorwaarde waar voornoemde kennis en vaardigheden bovendien aan dienden te voldoen, was dat ze binnen en buiten het vak toepassingen moesten hebben die los staan van de context waarin ze oorspronkelijk verworven zijn. Dit, omdat het kunnen gebruiken van verworven kennis en vaardigheden in nieuwe situaties een noodzakelijke voorwaarde is voor voortgang in de ontwikkeling van de leerlingen.

Bij de constructie van de items is geanticipeerd op de vakinhoudelijke vernieuwingen in de onder- en bovenbouw van havo en vwo. Op 1 augustus 1993 is voor het vak wiskunde namelijk een nieuw leerplan ingevoerd. Een probleem was dat de proefafname voor wiskunde plaats zou vinden bij

leerlingen die volgens het oude leerplan opgeleid zijn, terwijl de toetsen voorgelegd moeten kunnen worden aan leerlingen van wie een groot deel onderwijs gevolgd heeft volgens het nieuwe leerplan. Daarom was het noodzakelijk items te construeren die zowel binnen het oude leerplan als binnen het nieuwe leerplan pasten. In de items komt derhalve 'praktische' formele wiskunde aan de orde, zoals die op het moment van hun ontwikkeling naar verwachting in de toekomst in de onderbouw van de havo en het vwo aan de orde zou komen en in de bovenbouw van de havo en het vwo in de vakken wiskunde A en B door leerlingen beheerst zou moeten worden. Verder spelen contexten in de verzameling ontwikkelde items een prominente rol.

Bij de constructie van de items is uitgegaan van de twee algemene doelstellingen van de basisvorming die expliciet gericht zijn op de beheersing van het vak wiskunde. De eerste doelstelling is het verwerven van de wiskundige taal als communicatiemiddel. De tweede is het ontwikkelen van een wiskundige werkhouding, waarin plaats is voor systematisch methodisch werken, generaliseren, kritisch beoordelen van gegevens en uitkomsten en het creatief bedenken van oplossingen. Deze doelstellingen geven aan hoe leerlingen de formele wiskunde moeten leren gebruiken. De formele wiskunde waar het hier om gaat is vooral gericht op getalsmatige verhoudingen en verbanden, wiskundige formuleringen, formules en notaties, (soorten) functies en relaties, meetkundige (vlakke en ruimtelijke) figuren en statistiek en combinatoriek en kans.

Verder is bij de constructie van de items nog rekening gehouden met het feit dat in de bovenbouw van de havo en het vwo sprake is van een tweedeling in wiskunde A en wiskunde B. Wiskunde A richt zich op het kunnen werken in contexten waarbij niet veel abstracte wiskunde nodig is. Dit houdt in dat bij wiskunde A de nadruk ligt op het kunnen gebruiken van rekenvaardigheden en zaken als, schalen en verhoudingen, tabellen en schema's, het grafisch representeren van data en verbanden. Voor wiskunde B zijn meer technische vaardigheden nodig, zoals het kunnen manipuleren met algebraïsche uitdrukkingen en functies en het kunnen werken met goniometrische verhoudingen. Uiteraard moet deze wiskunde ook toepasbaar zijn in realistische problemen. In de verzameling van geconstrueerde items komen zowel aspecten van wiskunde A als van wiskunde B aan de orde.

Bij toepassingsgerichte items is het minder goed mogelijk meerkeuze-items te construeren. Daarom zijn de ontwikkelde items van het kort-open-antwoord-type. Constructie van de items vond plaats in de eerste drie maanden van 1994. De items zijn ontwikkeld in samenwerking met een groep docenten met ruime ervaring in het geven van onderwijs in de onder- en bovenbouw van havo en vwo. Er zijn in totaal 93 items ontwikkeld.

Uit deze itemverzameling zijn zes toetsboekjes samengesteld volgens een hierna te beschrijven opzet. Drie van deze boekjes concentreerden zich op aspecten van wiskunde A. De andere drie boekjes hadden betrekking op aspecten van wiskunde B. Ieder toetsboekje had een geschatte afnametijd van één lesuur.

2.2 De proefafnames

Om de eigenschappen te onderzoeken van de items die ontwikkeld waren voor de vakken Nederlands, Engels en wiskunde en voor het prototype van de checklist studievaardigheid zijn aan het einde van het schooljaar 1993-1994 proefafnames gehouden. Deze paragraaf bevat een beschrijving van de opzet, de uitvoering en de resultaten van de proefafnames van de voor de drie vakken ontwikkelde items.

2.2.1 Opzet en uitvoering

In januari 1994 is een aselecte steekproef getrokken van 400 scholen voor havo en/of vwo. Aan de vaksecties Nederlands, Engels en wiskunde van de geselecteerde scholen is via de directie gevraagd met klassen uit het derde leerjaar deel te nemen aan de proefafname voor hun vak. Deelname hield onder meer in dat een docent voor een van zijn of haar klassen twee lesuren van tenminste 45 minuten ter beschikking stelde, zodat de leerlingen uit deze klas twee toetsboekjes konden maken met voor het betreffende vak ontwikkelde items. De drie klastypen die voor deelname in aanmerking kwamen waren drie havo, drie havo/vwo en drie vwo. Om de belasting voor

docenten te beperken, gold de restrictie dat elke docent per klastype met niet meer dan één klas kon deelnemen. Waar nodig, is een op toeval gebaseerde selectie gemaakt uit de klassen die door vaksecties waren opgegeven.

Na afsluiting van de wervingsperiode bleek het mogelijk klassen die voor slechts één vak waren ingeschreven van deelname uit te sluiten. De respons op het verzoek om deelname was namelijk zo groot dat het aantal verwachte observaties per item al toereikend was bij het uitsluitend selecteren van klassen die voor twee of drie vakken waren ingeschreven. Door deze maatregel kon uiteindelijk de samenhang tussen de prestaties van leerlingen voor de verschillende vakken bepaald worden aan de hand van zoveel mogelijk waarnemingen.

De designs voor Nederlands en Engels

Zowel bij de proefafname voor Nederlands als bij de proefafname voor Engels was sprake van drie soorten items. Bij Nederlands betrof het kort-open-antwoord-cloze-items, kort-open-antwoord-items leesvaardigheid en meerkeuze-items leesvaardigheid. En bij Engels was sprake van cloze-items van het meerkeuzetype, meerkeuze-items leesvaardigheid en kort-open-antwoord-items schrijfvaardigheid. Bij elk van de zes itemverzamelingen zijn de items verdeeld over vier complementaire deelverzamelingen, zogeheten clusters. Ieder cluster had naar schatting een afnametijd van een half lesuur.

Bij de cloze-items en bij de kort-open-antwoord-items leesvaardigheid voor Nederlands vormden alle bij een en dezelfde tekst behorende items een cluster. Bij de meerkeuze-items leesvaardigheid voor Nederlands vormden zeven willekeurige items tezamen een cluster. Bij de cloze-items voor Engels vormden eveneens alle bij een en dezelfde tekst behorende items een cluster. Ook twee van de vier clusters bij de meerkeuze-items leesvaardigheid voor Engels bestonden uit items bij een tekst. De resterende twee clusters in dit toetsonderdeel bevatten de twee verschillende soorten items die al in paragraaf 2.1 zijn beschreven. Het betrof hier respectievelijk clusters van twintig en vijftien items. Bij de kort-open-antwoord-items schrijfvaardigheid voor Engels bevatten de vier clusters de vier verschillende soorten items die

eveneens al in paragraaf 2.1 zijn beschreven. Het betrof hier twee clusters met tien en twee met dertien items.

Door de clusters volgens een volledig gebalanceerd geblokt kettingdesign (block-interlaced anchoring design; Vale, 1986) te paren, ontstonden er per itemsoort zes toetsboekjes met een afnametijd van één lesuur. Omdat de proefafname voor een vak twee lesuren in beslag nam, was het mogelijk in de proefafnames voor Nederlands en Engels binnen iedere klas twee verschillende soorten toetsboekjes aan leerlingen voor te leggen.

Om afkijken te kunnen voorkomen, werden klassen in ieder lesuur in tweeën gesplitst. De twee resulterende groepen leerlingen kregen twee verschillende toetsboekjes voorgelegd die betrekking hadden op een en dezelfde itemsoort. Deze combinatie van toetsboekjes binnen een soort - de klasset - was zodanig dat alle items uit de betreffende itemsoort in een klas aan bod kwamen. Deze maatregel had als bijkomend voordeel dat de intraklassecorrelatie gereduceerd werd, hetgeen de nauwkeurigheid van de schattingen in de steekproef ten goede komt.

Tabel 2.1 geeft het bij de aanbieding van de zes itemsoorten gehanteerde geblokte kettingdesign in tabelvorm weer. De tabel laat voor elk van de zes itemverzamelingen met behulp van kruisjes zien welke clusters ieder toetsboekje bevat. Verder maakt de tabel duidelijk tot welke klasset ieder toetsboekje behoort. De indices bij de kruisjes geven de volgorde aan van de clusters in de toetsboekjes. De tabel maakt bijvoorbeeld duidelijk dat het toetsboekje met het nummer 1 de clusters 1 en 2 in die volgorde bevat en het toetsboekje met het nummer 6 de clusters 3 en 4, eveneens in die volgorde. Verder laat de tabel in de kolom met het opschrift ‘klasset’ bijvoorbeeld zien dat binnen een klas het toetsboekje 1 altijd in combinatie met het toetsboekje 6 werd afgenomen.

Tabel 2.1
Het proefafnamedesign voor Nederlands en Engels

	Cluster				
Toetsboekje	1	2	3	4	Klasset
1	X_1	X_2			1
2	X_1		X_2		2
3	X_2			X_1	3
4		X_1	X_2		3
5		X_1		X_2	2
6			X_1	X_2	1

Er is voor gezorgd dat zowel de drie klassets als de drie mogelijke combinaties van soorten items binnen de proefafnames voor Nederlands en Engels bij benadering even vaak voorkwamen. Hierdoor mocht voor ieder item een ongeveer gelijk aantal leerlingantwoorden verwacht worden.

Het design voor wiskunde

De zes toetsboekjes die voor wiskunde ontwikkeld waren, zijn volgens een onvolledig gebalanceerd geblokt kettingdesign met elkaar gecombineerd. Van de drie boekjes die zich concentreerden op aspecten van wiskunde A en de drie boekjes die betrekking hadden op aspecten van wiskunde B zijn alle mogelijke paren gevormd. In totaal konden derhalve negen verschillende boekjesparen aan klassen werden aangeboden. Er is voor gezorgd dat de negen verschillende boekjesparen binnen de proefafname wiskunde ongeveer even vaak voorkwamen. Hierdoor mocht voor ieder item een ongeveer gelijk aantal leerlingantwoorden verwacht worden.

Overzicht van het proefafnamemateriaal

Bij ieder vak gingen de voor de leerlingen bestemde toetsboekjes vergezeld van optisch leesbare formulieren, een instructieboekje voor docenten, een docentvragenlijst, een verslagformulier en tien exemplaren van het prototype van de checklist studievaardigheid. Indien sprake was van open-antwoord items werd eveneens een correctievoorschrift bijgesloten.

Bij de werving voor de proefafnames is het aantal leerlingen per klas geregistreerd. Elke klas kreeg per lesuur net zoveel toetsboekjes toegezonden als het aantal leerlingen dat de klas telde, aangevuld met extra exemplaren bestemd voor de docent. De optisch leesbare formulieren dienden voor het registreren van de antwoorden van leerlingen op de items.

Het aantal optisch leesbare formulieren correspondeerde bij Nederlands en Engels met het aantal toetsboekjes per lesuur, zodat per leerling twee optisch leesbare formulieren aangeboden werden. Bij wiskunde volstond één optisch leesbaar formulier per leerling, omdat het - in tegenstelling tot de twee andere vakken - mogelijk was de antwoorden van leerlingen op de items in de twee toetsboekjes op één enkel formulier weer te geven.

De leerlingen konden hun antwoorden op de meerkeuze-items rechtstreeks op de formulieren aanstrepen. De open-antwoord items moesten zij in de toetsboekjes beantwoorden. Docenten is gevraagd om de antwoorden van hun leerlingen na te kijken aan de hand van het bijgeleverde correctievoorschrift. Vervolgens moesten zij de met behulp van het correctievoorschrift bepaalde scores aanstrepen op de optisch leesbare formulieren. Verder is de docenten gevraagd op deze formulieren een reeks leerlinggegevens te verstrekken. Het betrof hier de sexe en thuistaal van de leerlingen, het cijfer op het eindrapport voor het betreffende vak en het oordeel van de docent over toekomstige prestaties van de betreffende leerling.

De docentvragenlijst bood de gelegenheid uitspraken te doen over de kwaliteit van de ontwikkelde items. Docenten konden onder meer aangeven dat ze een item niet geschikt vonden voor opname in een toets, omdat het item te moeilijk of te makkelijk was, of omdat de leerstof die in het item aan bod kwam niet behandeld was. Verder konden docenten uitspraken doen over het correctievoorschrift bij de kort-open-antwoord-items en onder meer aangeven dat het antwoordmodel bij een item inhoudelijk niet in orde was,

of niet duidelijk, of niet goed hanteerbaar. Bij de vakken Nederlands en Engels konden docenten vanwege het gehanteerde proef-afnamedesign alle items uit twee van de drie itemverzamelingen beoordelen.

Het verslagformulier was bestemd voor het verstrekken van informatie over het verloop van een proefafname. Docenten konden onder meer aangeven of de afnametijd van een toetsboekje toereikend was en of de proefafname ordelijk verlopen was. Ook is docenten van deelnemende klassen gevraagd voor de eerste tien leerlingen uit de klasselijst het prototype van de checklist studievaardigheid in te vullen.

Inschrijving en deelname

Tabel 2.2 laat voor ieder vak zien aan hoeveel klassen van verschillend type proefafnamemateriaal verzonden is en hoeveel van deze klassen daadwerkelijk hebben deelgenomen aan de proefafname. In totaal is er voor het vak Nederlands materiaal verzonden aan 200 klassen. Het betrof hier 75 havo, 21 havo/vwo- en 104 vwo-klassen. Niet meer dan 101 van deze klassen namen daadwerkelijk deel aan de proefafname; 31 havo-, veertien havo/vwo- en 56 vwo-klassen. Voor Engels is er proefafnamemateriaal verzonden aan 215 klassen; 84 havo-, 27 havo/vwo- en 104 vwo-klassen. Van deze klassen verleenden er 141 inderdaad medewerking aan de proefafname, te weten 55 havo-, achttien havo/vwo- en 68 vwo-klassen. Voor wiskunde is er aan 200 klassen materiaal gestuurd. Het betrof hier 83 havo-, 22 havo/vwo- en 95 vwo-klassen. Van deze klassen participeerden er 123 werkelijk; 49 havo-, veertien havo/vwo- en 60 vwo-klassen.

Tabel 2.2

Aantal ingeschreven en daadwerkelijk deelnemende klassen per vak en klastype

		Klastype			Totaal
		Havo	Havo/vwo	Vwo	
Nederlands	Ingeschreven	75	21	104	200
	Deelnemend	31	14	56	101
Engels	Ingeschreven	84	27	104	215
	Deelnemend	55	18	68	141
Wiskunde	Ingeschreven	83	22	95	200
	Deelnemend	49	14	60	123

De lage respons heeft een aantal oorzaken. Navraag leerde dat de belangrijkste reden voor docenten om af te zien van deelname was dat de uit te voeren werkzaamheden bij nader inzien als te belastend werden ervaren. In de tweede plaats bleek het voor een aantal klassen niet meer mogelijk te zijn nog twee lesuren in de rest van het schooljaar te reserveren voor de proefafname. Ook viel een klein aantal klassen uit, omdat docenten van deelname afzagen vanwege het feit dat de klas al medewerking verleend had aan een proefafname van afsluitingstoetsen basisvorming.

Gegevensverwerking

Na verzameling van het proefafnamemateriaal zijn de optisch leesbare formulieren, docentvragenlijsten en verslagformulieren verwerkt en geanalyseerd. Om het mogelijk te maken de gegevens op de optisch leesbare formulieren over en binnen de vakken te kunnen koppelen is gebruik gemaakt van een leerlingcode die uniek was per leerling. Deze code bestond uit een code voor de school, een volgnummer voor de klas binnen de school en een volgnummer voor de leerling binnen de klas. School- en klasnaam waren met bijbehorende schoolcode en volgnummer op de formulieren gedrukt. De voorbereidingstijd voor de proefafnames was echter te kort om de formulieren ook te voorzien van de leerlingnamen met bijbehorend volgnummer. Daarom is docenten gevraagd de namen van de leerlingen in te vullen op de formulieren en iedere leerling het volgnummer te geven dat correspondeerde met zijn of haar positie in de alfabetische ordening van de achternamen van leerlingen uit de betreffende klas.

De optisch leesbare formulieren voor de vakken Nederlands en Engels zijn bij binnenkomst gecontroleerd en de fouten die door docenten zijn gemaakt bij de eerder beschreven procedure zijn hersteld. Vervolgens zijn de optisch leesbare formulieren ingelezen. Bij het inlezen is gebruik gemaakt van inleesspecificaties waarin aangegeven werd op welke posities op het formulier informatie kon voorkomen. Aperte fouten bij het invullen van de formulieren konden derhalve gedetecteerd worden. Formulieren met fouten zijn gecontroleerd en de fouten zijn zo mogelijk hersteld. Het betrof hier bijvoorbeeld formulieren waar bij een meerkeuze-item een niet bestaand alternatief was aangestreept, formulieren waar de maximumscore bij een kort-open-antwoord-item was overschreden, of formulieren waar bij een niet in het betreffende toetsboekje voorkomend itemnummer iets was aangestreept.

Na het inlezen ontstonden drie bestanden met gegevens, voor elk vak één. Vervolgens zijn de records in deze drie bestanden met elkaar gekoppeld met behulp van de eerder beschreven unieke leerlingcode. Na de koppeling is voor records waarin zich gegevens bevonden voor twee of drie vakken gecontroleerd of deze gegevens ook werkelijk betrekking hadden op dezelfde leerling. Dit was mogelijk, omdat voor ieder vak de achtergrondgegevens sexe en thuistaal op de formulieren ingevuld waren. Deden zich binnen een

klas één of meer discrepanties voor wat betreft vermelde sexe en/of thuistaal, dan werden de optisch leesbare formulieren van alle leerlingen uit de betreffende klas aan een nader onderzoek onderworpen.

Door de volgnummers en de corresponderende leerlingnamen voor de verschillende vakken te vergelijken, was het mogelijk te bepalen waar de discrepantie uit voortkwam. In veruit de meeste gevallen bleken de volgnummers van dezelfde leerlingen over de vakken niet overeen te komen. Bij deze records zijn de leerlingcodes aangepast en zijn de juiste leerlingresponsen vervolgens met elkaar gekoppeld. In enkele gevallen was blijkens de leerlingnaam sprake van dezelfde leerling, maar was een vergissing gemaakt bij het aanstrepen van de sexe of de thuistaal. Bij deze records zijn de betreffende achtergrondgegevens verwijderd, indien niet te achterhalen viel bij welk vak het juiste gegeven verstrekt was. In enkele gevallen was niet te achterhalen wat de aard van de discrepantie was, omdat voor een of twee vakken leerlingnamen niet waren ingevuld of achtergrondgegevens ontbraken. De betreffende records zijn uit het gegevensbestand verwijderd.

Tabel 2.3
Aantal bruikbare leerlingrecords per vak en klastype

	Klastype			Totaal
	Havo	Havo/vwo	Vwo	
Nederlands	836	314	1367	2517
Engels	1489	464	1640	3593
Wiskunde	1217	405	1548	3170

In tabel 2.3 staat hoeveel bruikbare leerlingrecords er uiteindelijk na de gegevensverwerking overbleven voor elk vak en in welk klastype de betreffende leerlingen zich bevonden. Van de 2517 bruikbare records voor

Nederlands zijn er bijvoorbeeld 836 afkomstig van leerlingen uit drie-havo-klassen, 314 van leerlingen uit drie-havo/vwo- klassen en 1367 van leerlingen uit drie-vwo-klassen. De drie verschillende itemsoorten voor de vakken Nederlands en Engels worden in deze tabel buiten beschouwing gelaten. Omdat leerlingen in de proefafnames voor deze vakken items uit twee van de drie verzamelingen maakten, is het aantal leerlingrecords dat gegevens voor een specifieke itemsoort bevat ongeveer een factor 2/3 kleiner.

De samenstelling van de leerlingrecords verschilt. Een deel van de leerlingrecords bevat leerlingresponsen voor alle drie de vakken. In een ander deel van de records bevinden zich leerlingresponsen voor twee van de drie vakken. Daarnaast zijn er ook leerlingrecords die slechts voor een vak leerlinggegevens bevatten. In totaal zijn er wat samenstelling betreft zeven verschillende typen leerlingrecords mogelijk. Tabel 2.4 geeft door middel van kruisjes aan welke dat zijn en in welke hoeveelheden zij in het gegevensbestand voorkomen. In deze tabel wordt het klastype van de leerlingen buiten beschouwing gelaten. De tabel laat bijvoorbeeld zien dat er 800 leerlingrecords zijn die voor zowel Nederlands, Engels als wiskunde leerlingresponsen bevatten en dat het totale aantal bruikbare records uit de drie proefafnames 5544 bedraagt. Verder blijkt uit de tabel dat er relatief veel records zijn die slechts voor één vak een waarneming bevatten.

Tabel 2.4

Samenstelling van verschillende typen leerlingrecords en hun aantal

		Nederlands	Engels	wiskunde
	Aantal			
	800	X	X	X
	605	X	X	
	496	X		X
	1035		X	X
	616	X		
	1153		X	
	839			X
Totaal	5544	2517	3593	3170

De psychometrische analyses

Voor het bepalen van de psychometrische eigenschappen van de items is gebruik gemaakt van de itemresponstheorie (IRT). De IRT maakt het mogelijk de vaardigheid van leerlingen en de moeilijkheid van items op eenvoudige wijze te schalen. De antwoorden van leerlingen zijn geanalyseerd met behulp van het eenparameter logistisch model (OPLM: One-Parameter Logistic Model; Verhelst & Glas, 1995). De algemene vorm van het OPLM beschrijft de kans dat de score X_i op item i van een persoon met vaardigheid θ één van de waarden j (j = 0, 1, ..., m_i) aan zal nemen als:

$$f_{ij}(\theta) = \frac{\exp\left[a_i\left(j\theta - \sum_{g=1}^{j}\beta_{ig}\right)\right]}{1 + \sum_{h=1}^{m_i}\exp\left[a_i\left(h\theta - \sum_{g=1}^{h}\beta_{ig}\right)\right]}, \quad (j = 0, 1, \ldots, m_i), \qquad (2.1)$$

waarbij de β_{ig} (g = 1, ..., m_i), de locatieparameters van de categorieresponsfuncties van item i zijn en a_i een discriminatieparameter is met een gehele waarde. Deze wordt daarom de discriminatie-index genoemd. Indien j = 0, dan moeten de β_{ig} in de teller van uitdrukking (2.1) gesommeerd worden van g = 1 tot g = 0. De uitkomst van deze sommatie wordt verondersteld gelijk aan 0 te zijn. Hebben alle discriminatie-indices een identieke waarde, dan heeft uitdrukking (2.1) dezelfde vorm als het partial credit model (Masters, 1982). Geldt dat $m_i = 1$ voor alle items i, dan heeft uitdrukking (2.1) dezelfde vorm als het Raschmodel (Rasch, 1960).

Het model in uitdrukking (2.1) is alleen gebruikt bij het analyseren van de leerlingantwoorden op de wiskunde-items en de kort-open-antwoord-items Nederlands. Bij de overige twee itemsoorten voor het vak Nederlands en de drie itemsoorten voor het vak Engels was uitsluitend sprake van dichotome items en gold derhalve $m_i = 1$ voor alle items i. Het model dat in al deze gevallen gebruikt is bij het analyseren van de leerlingantwoorden beschrijft

de kans dat de score X_i van een persoon met vaardigheid θ op item *i* de waarde 1 aan zal nemen als:

$$f_i(\theta) = \frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]}. \qquad (2.2)$$

Het OPLM is een van de modellen die gebruikt kan worden voor het analyseren van ordinale polytome itemresponsen (Mellenbergh, 1995). Voor gebruik van het OPLM is gekozen, omdat het aantal items dat een gebrek aan modelfit vertoont bij gebruik van dit model in de regel klein is. Bovendien maakt gebruik van het OPLM het uitvoeren van CML (Conditional Maximum Likelihood)-schattingen mogelijk. Het voordeel van het gebruik van een CML-schattingsmethode is dat het mogelijk is de parameters van de modellen in (2.1) en (2.2) te schatten en te toetsen, zonder dat aannames over de vorm van de vaardigheidsverdeling nodig zijn.

Het specifieke kenmerk van het schatten van item- en persoonsparameters met behulp van het OPLM is dat de waarde van de discriminatie-index voor ieder item als een te toetsen hypothese beschouwd wordt. Er bestaat een heuristiek om op grond van de data te komen tot plausibele waarden van de discriminatie-indices (Verhelst, Verstralen & Eggen, 1991). Tevens bestaan er statistische toetsen die gevoelig zijn voor misspecificatie van de discriminatie-index (Verhelst & Glas, 1995). Calibratie van items volgens het OPLM is dan ook een iteratief proces waarin alternerend de modelfit van items wordt onderzocht met behulp van de voornoemde statistische toetsen en de waarden van discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden op basis van analyses op dezelfde data en er kan dus kanskapitalisatie optreden. Indien de steekproeven een voldoende omvang hebben, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen & Eggen, 1991).

Voor het uitvoeren van OPLM-analyses is gebruik gemaakt van een op het Cito ontwikkeld computerprogramma met de naam OPLM (Verhelst, Glas & Verstralen, 1995). In dit programma zijn de voornoemde toetsen en de heuristiek voor het bepalen van plausibele waarden voor de discriminatie-indices geïmplementeerd. Bij alle itemsoorten zijn CML-schattingen uitgevoerd. Bij het schatten van de itemparameters is het nulpunt van iedere

vaardigheidsschaal vastgelegd door de som van alle locatieparameters gelijk aan nul te stellen.

Om de grootte van de vaardigheidsverschillen tussen de klastypen drie havo, drie havo/vwo en drie vwo aan een nader onderzoek te onderwerpen is gebruik gemaakt van het eveneens op het Cito ontwikkelde programma SAUL (Structural Analysis of Unidimensional Latent variables; Verhelst & Verstralen, 1996). SAUL is een hulpprogramma bij het OPLM-programma, dat regressie-analyses uitvoert met de latente variabele θ als afhankelijke variabele en achtergrondvariabelen als onafhankelijke variabelen.

De eenheid van een met behulp van het OPLM gecalibreerde vaardigheidsschaal is afhankelijk van de waarden van de discriminatie-indices van de items bij deze schaal. Om de effecten van achtergrondvariabelen voor verschillende vaardigheidsschalen met elkaar te kunnen vergelijken moeten deze vaardigheidsschalen daarom gestandaardiseerd worden. Standaardisatie vindt binnen SAUL plaats met behulp van het geometrisch gemiddelde van de discriminatie-indices van de bij de schaal behorende items. Voor het geometrisch gemiddelde G geldt:

$$G = \left(\prod_{i=1}^{n} a_i \right)^{\frac{1}{n}}, \qquad (2.3)$$

waarbij *n* staat voor het aantal items dat bij de vaardigheidsschaal hoort. De standaardisatie van een vaardigheidsschaal wordt bereikt door de locatieparameters van de items bij deze schaal met G te vermenigvuldigen en iedere discriminatie-index door G te delen. De eenheid van aldus getransformeerde schalen is identiek en de effecten van achtergrondvariabelen voor dergelijke schalen zijn derhalve vergelijkbaar.

SAUL schat de gemiddelde vaardigheid in verschillende subpopulaties volgens een lineair structureel model. In het geval van twee achtergrondvariabelen, waarbij geen interactie verondersteld wordt, geldt voor de vaardigheid in iedere subpopulatie het structurele model:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}. \qquad (2.4)$$

Hierbij is μ het algemeen gemiddelde; α_i de effectparameter voor de achtergrondvariabele α met I categorieën; β_j de effectparameter voor de achtergrondvariabele β met J categorieën en ε_{ij} het $N(0, \sigma^2)$ verdeelde residu. Om effecten eenduidig schatbaar te maken stelt SAUL het effect van de eerste categorie van iedere achtergrondvariabele gelijk aan nul. Verder neemt SAUL aan dat binnen de subpopulaties die gevormd kunnen worden op basis van de achtergrondvariabelen de vaardigheid normaal verdeeld is. Bovendien neemt SAUL aan dat de binnengroepvariantie van de vaardigheid voor iedere subpopulatie gelijk is. Een en ander impliceert voor het model in (2.4) dat binnen alle $I \cdot J$ subpopulaties geldt dat $\theta_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$. De interpretatie van de effectparameters in SAUL is identiek aan de interpretatie van regressiecoëfficiënten in regressie-analyse met categorische onafhankelijke variabelen (Verhelst & Eggen, 1989).

Op grond van de schattingen van de effectparameters voor de verschillende categorieën van een achtergrondvariabele kunnen uitspraken gedaan worden over het verschil in gemiddelde vaardigheid tussen deze categorieën. De Z-scores die ontstaan door de schattingen van de effectparameters door hun standaardfout te delen, maken het mogelijk te toetsen of de verschillen in gemiddelde vaardigheid significant zijn. Door de effectschattingen te delen door σ, de binnengroepstandaardafwijking van de vaardigheid, ontstaat een maat die de gelegenheid biedt om uitspraken te doen over de grootte van de effecten. Als richtlijn voor het bepalen van de grootte van een effect geeft Cohen (1977) de waarden 0,2, 0,5 en 0,8. Bij een waarde van 0,2 is sprake van een klein effect, bij een waarde van 0,5 van een middelmatig effect en bij een waarde van 0,8 van een groot effect.

Onderzoek naar beoordelaarsovereenstemming

Bij de kort-open-antwoord-items leesvaardigheid voor Nederlands, de kort-open-antwoord-items schrijfvaardigheid voor Engels en bij alle items voor wiskunde is onderzoek gedaan naar de objectieve scoorbaarheid. Daartoe is voor de betreffende items de beoordelaarsovereenstemming onderzocht. In dat kader zijn zowel voor Nederlands als Engels op aselecte wijze zestig door leerlingen ingevulde toetsboekjes geselecteerd. Het betrof hier twee maal 30 boekjes uit een en dezelfde klasset, zodat het aantal per item te beoordelen

leerlingresponsen 30 bedroeg. Deze toetsboekjes zijn vervolgens gekopieerd en ter correctie voorgelegd aan respectievelijk acht docenten Nederlands en acht docenten Engels.

Bij wiskunde zijn van ieder van de zes toetsboekjes op aselecte wijze 30 ingevulde exemplaren geselecteerd. Vervolgens zijn deze toetsboekjes gekopieerd en in verschillende combinaties van paren ter correctie aangeboden aan steeds drie van in totaal achttien docenten wiskunde. De groepen van drie docenten kregen respectievelijk steeds 30 kopieën van de boekjes 1 en 2; 2 en 3; 3 en 4; 4 en 5; 5 en 6 en 6 en 1. Iedere docent kreeg dus zestig toetsboekjes ter correctie voorgelegd en ieder leerlingantwoord is door zes docenten gescoord. Door deze opzet zijn de beoordelaarsovereenstemmingscoëfficiënten voor alle items met elkaar te vergelijken.

Voor ieder item uit de drie genoemde itemverzamelingen is de beoordelaarsovereenstemmingscoëfficiënt, ρ^2, geschat; $\hat{\rho}^2$ wordt gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + (\hat{\sigma}_b^2 + \hat{\sigma}_{res}^2) / k}, \qquad (2.5)$$

waarbij k staat voor het aantal beoordelaars dat bij de beoordeling is ingeschakeld, en $\hat{\sigma}_p^2$, $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ de variantiecomponenten representeren van respectievelijk de beoordeelde personen, de beoordelaars en het residu. De variantiecomponent voor beoordelaars in uitdrukking (2.5), $\hat{\sigma}_b^2$, geeft aan in welke mate beoordelaarsgemiddelden verschillen. Hoe groter de variantiecomponenten $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ zijn in verhouding tot $\hat{\sigma}_p^2$, des te lager is de overeenstemming. Bij perfecte overeenstemming tussen de beoordelaars zijn $\hat{\sigma}_b^2$ en $\hat{\sigma}_{res}^2$ gelijk aan nul en is de coëfficiënt gelijk aan 1. Bij volledig gebrek aan overeenstemming heeft de coëfficiënt de waarde nul. De coëfficiënt kan geïnterpreteerd worden als een schatting van de mate van overeenstemming tussen de gemiddelde oordelen van k willekeurig gekozen beoordelaars en de gemiddelde oordelen van k andere, eveneens willekeurig gekozen beoordelaars. Indien $k = 1$, dan is de coëfficiënt een schatting van de overeenstemming tussen de oordelen van één willekeurig gekozen beoordelaar en de oordelen van één andere, willekeurig gekozen beoordelaar. Omdat in de praktijk nooit meer dan één docent de antwoorden van leerlingen op de

items zal corrigeren, is bij het berekenen van de waarde van de beoordelaarsovereenstemmingscoëfficiënt voor ieder item $k = 1$ gesteld.

2.2.2 De resultaten voor het vak Nederlands

De volgens de hiervoor beschreven opzet en procedures verzamelde en verwerkte leerlingantwoorden voor het vak Nederlands zijn na de proefafname geanalyseerd. OPLM-analyses zijn uitgevoerd en voor de kort-open-antwoord-items is de beoordelaarsovereenstemming onderzocht. De gegevens die door docenten verstrekt zijn in de docentvragenlijst en op het verslagformulier zijn eveneens verwerkt en geanalyseerd.

Resultaten van de OPLM-analyses

De OPLM-analyses die per itemtype zijn uitgevoerd, laten zien dat de items van hetzelfde type in de regel ook dezelfde vaardigheid meten. Van de cloze-items bleken er slechts drie van de 51 niet binnen het model te passen. Van de kort-open-antwoord-items leesvaardigheid pasten er acht van de 45 niet en bij de meerkeuze-items pasten alle 28 items.

Vervolgens is onderzocht of het mogelijk was de items uit de drie verschillende itemverzamelingen voor Nederlands gezamenlijk te schalen. Bij de betreffende OPLM-analyse zijn de items buiten beschouwing gelaten die in de oorspronkelijke OPLM-analyses een gebrek aan modelfit vertoonden. De OPLM-analyse voor de items Nederlands als geheel liet voor geen enkele van de resterende 113 items een gebrek aan modelfit zien. Daarom mag aangenomen worden dat de drie soorten items een en dezelfde vaardigheid meten.

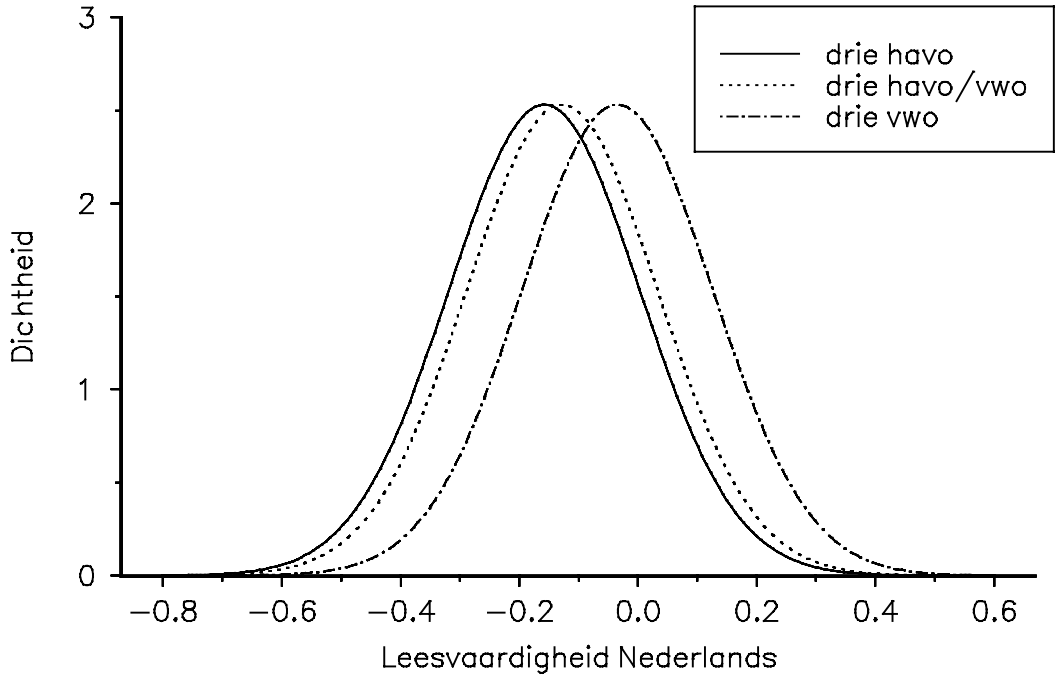
Tabel 2.5
Resultaten van de SAUL-analyse ter bepaling van de verschillen tussen drie havo, drie havo/vwo en drie vwo voor leesvaardigheid Nederlands

Klastype	Effect	SE	n	Z	Effectgrootte
Drie havo	0	0	836		
Drie havo/vwo	0,086	0,039	314	2,204	0,184
Drie vwo	0,358	0,026	1367	13,963	0,765
Contrast					
Drie vwo-drie havo/vwo	0,272	0,037		7,387	0,581

Om de verschillen in vaardigheid tussen de klastypen drie havo, drie havo/vwo en drie vwo te bepalen op de leesvaardigheidschaal die ontstond

na analyse van alle items Nederlands tezamen, is een SAUL-analyse uitgevoerd. Tabel 2.5 bevat de resultaten van deze SAUL-analyse. De effectschatting voor het klastype drie havo is in deze tabel op nul gesteld. Het geschatte algemeen gemiddelde μ op de gestandaardiseerde vaardigheidsschaal is -0,4769 met een standaardfout van 0,0203. De geschatte binnengroepvariantie σ^2 is 0,2193 met een standaardfout van 0,0099.

De kolom met het opschrift 'Effect' bevat de eerder beschreven effectschattingen voor de verschillende klastypen. In de kolom met het opschrift 'SE' staan de standaardfouten van de effectschattingen. De kolom met het opschrift 'n' geeft de aantallen leerlingen in de verschillende klastypen weer. De kolom met het opschrift 'Z' bevat de Z-scores die ontstaan door de effectschattingen door hun standaardfouten te delen. De kolom met het opschrift 'Effectgrootte' bevat het quotiënt van de effecten en de binnengroepstandaardafwijking van de vaardigheid. In de regel onder de term 'contrast' staan de gegevens die betrekking hebben op het verschil in vaardigheid tussen de klastypen drie vwo en drie havo/vwo.



Figuur 2.1

Kansdichtheidsfuncties voor de drie klastypen onder aanname van normaliteit

voor leesvaardigheid Nederlands

Figuur 2.1 bevat de afbeeldingen van de kansdichtheidsfuncties van de leesvaardigheid binnen de drie klastypen op de originele vaardigheidsschaal. Het geometrisch gemiddelde van de discriminatieparameters van de items is 2,970. De gemiddelde vaardigheid binnen de havo-groep op de originele vaardigheidsschaal is -0,161. De gemiddelde vaardigheid binnen de havo/vwo-groep op deze schaal bedraagt -0,132 en de gemiddelde vaardigheid binnen de vwo-groep is -0,040. De binnengroepvariantie is 0,0250.

Tabel 2.5 en figuur 2.1 maken duidelijk dat het verschil in leesvaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-havo/vwo-klassen klein is. Leerlingen uit drie-havo/vwo-klassen zijn gemiddeld wel iets leesvaardiger dan leerlingen uit drie-havo-klassen, maar het effect is klein. Leerlingen uit drie-havo/vwo-klassen zijn op hun beurt gemiddeld weer minder leesvaardig dan leer-lingen uit drie-vwo-klassen en het verschil tussen deze twee groepen is ook groter. Er is sprake van een klein tot middelmatig effect. De tabel en figuur laten zien dat het verschil in vaardigheid tussen leerlingen uit drie-havo- en drie-vwo-klassen relatief het grootst is. Het betreffende effect is middelmatig tot groot.

Beoordelaarsovereenstemming

Bij 26 van de 45 items was de beoordelaarsovereenstemmingscoëfficiënt groter dan 0,80. Bij drie van deze 26 items was de overeenstemming perfect. De overeenstemming tussen docenten bij deze 26 items is goed te noemen. Van de resterende negentien items lag bij negen items de waarde van de coëfficiënt van overeenstemming tussen 0,60 en 0,80. Dat wil zeggen dat de correctievoorschriften van deze negen items redelijk te noemen zijn. Bij de overige tien items was de coëf-ficiënt van overeenstemming kleiner dan 0,60. Bij deze items is sprake van een slechte overeenstemming. Ze mogen dan ook geen deel gaan uitmaken van de te construeren toetsen.

Docentvragenlijst en verslagformulier

Uit de inhoud van de door docenten geretourneerde docentvragenlijsten valt af te leiden dat de docenten over het algemeen tevreden zijn over de items. Een aantal docenten vond de cloze-items moeilijk. Verder moet worden opgemerkt dat uit de reacties van docenten bleek dat ze de correctie van de kort-open-antwoord-items teveel tijd vonden kosten.

2.2.3 De resultaten voor het vak Engels

De volgens de hiervoor beschreven opzet en procedures verzamelde en verwerkte leerlingantwoorden voor het vak Engels zijn na de proefafname geanalyseerd. OPLM-analyses zijn uitgevoerd en voor de items schrijfvaardigheid is de beoordelaarsovereenstemming onderzocht. De gegevens die door docenten verstrekt zijn in de docentvragenlijst en op het verslagformulier zijn eveneens verwerkt en geanalyseerd.

Resultaten OPLM-analyses

De OPLM-analyses die per itemtype zijn uitgevoerd laten zien dat de items van hetzelfde type in de regel ook dezelfde vaardigheid meten. Van de cloze-items bleken er slechts vijf van de 49 niet binnen het model te passen. Van de meerkeuze-items leesvaardigheid pasten er zes van de 50 niet. Al deze items hadden betrekking op het kunnen afleiden van de betekenis van een woord met behulp van een gegeven verwant woord. Bij schrijfvaardigheid bleken alle 46 items te passen.

Vervolgens is onderzocht of het mogelijk was de items uit de drie verschillende itemverzamelingen voor Engels gezamenlijk te schalen. Bij de betreffende OPLM-analyse zijn de items buiten beschouwing gelaten die in de oorspronkelijke OPLM-analyses een gebrek aan modelfit vertoonden. De OPLM-analyse voor de items Engels als geheel liet voor geen enkele van de resterende 134 items een gebrek aan modelfit zien. Daarom mag aangenomen worden dat de drie soorten items een en dezelfde vaardigheid meten.

Tabel 2.6 bevat de resultaten van de SAUL-analyse die is uitgevoerd om de verschillen in vaardigheid tussen de klastypen drie havo, drie havo/vwo en drie vwo te bepalen op de schaal die ontstond na analyse van alle items Engels tezamen. Het geschatte algemeen gemiddelde μ op de gestandaardiseerde vaardigheidsschaal is 0,2499 met een standaardfout van 0,0208. De geschatte binnengroepvariantie σ^2 is 0,5195 met een standaardfout van 0,0161. Voor een nadere toelichting op de opbouw van tabel 2.6 wordt verwezen naar de toelichting bij tabel 2.5 in paragraaf 2.2.2.

Tabel 2.6

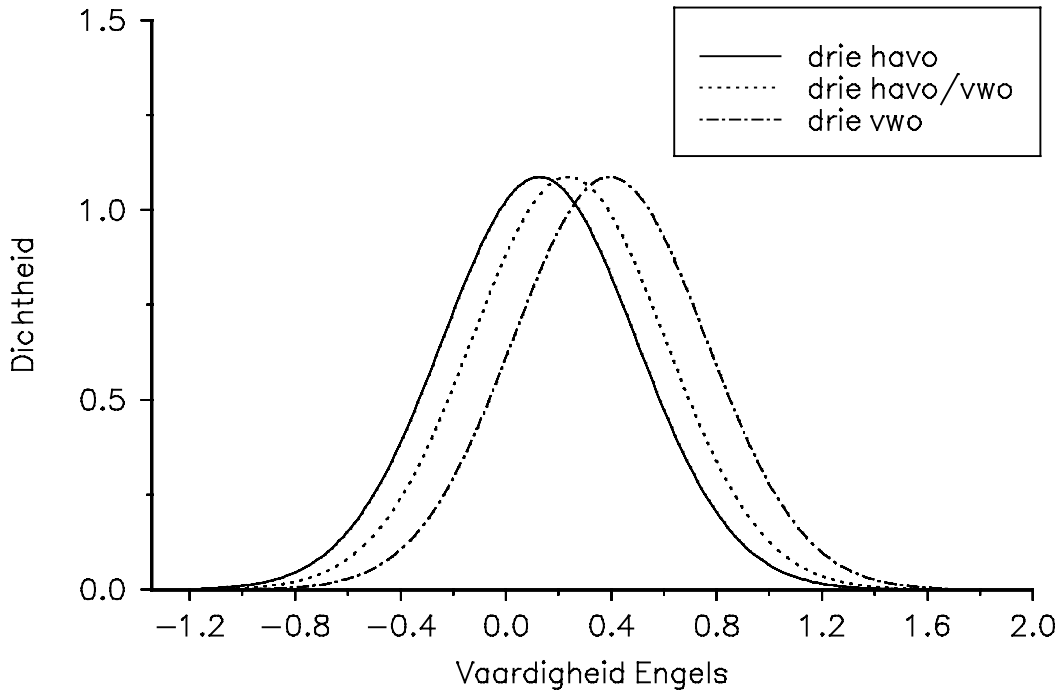
Resultaten van de SAUL-analyse ter bepaling van de verschillen in vaardigheid

tussen drie havo, drie havo/vwo en drie vwo op de schaal voor Engels

Klastype	Effect	SE	n	Z	Effectgrootte
Drie havo	0	0	1489		
Drie havo/vwo	0,216	0,043	464	5,049	0,300
Drie vwo	0,523	0,029	1640	18,116	0,726
Contrast					
Drie vwo-drie havo/vwo	0,307	0,042		7,229	0,426

Figuur 2.2 bevat voor de drie klastypen de afbeeldingen op de originele vaardig-heidsschaal van de kansdichtheidsfuncties van de vaardigheid die gemeten wordt door de voor het vak Engels ontwikkelde items. Het geometrisch gemiddelde van de discriminatieparameters van de items is 1,965. De gemiddelde vaardigheid binnen de havo-groep op de originele vaardigheidsschaal is 0,127. De gemiddelde vaardigheid binnen de havo/vwo-groep op deze schaal bedraagt 0,237 en de gemiddelde vaardigheid binnen de vwo-groep is 0,393. De binnengroepvariantie is 0,1347.

Tabel 2.6 en figuur 2.2 maken duidelijk dat het verschil in de betreffende vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-havo/vwo-klassen klein is. Leerlingen uit drie-havo/vwo-klassen zijn gemiddeld wel iets vaardiger dan leerlingen uit drie-havo-klassen, maar er is sprake van niet meer dan een klein effect. Leerlingen uit drie-vwo-klassen zijn op hun beurt gemiddeld vaardiger dan leerlingen uit drie-havo/vwo-klassen. Het verschil tussen deze twee groepen is groter; de bijbehorende effectgrootte is klein tot middelmatig. De tabel en figuur laten zien dat het verschil in vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-vwo-klassen relatief het grootst is. De bijbehorende effectgrootte is middelmatig tot groot.



Figuur 2.2

Kansdichtheidsfuncties voor de drie klastypen onder aanname van normaliteit voor de vaardigheidsschaal voor Engels

Een vergelijking tussen de proefafnames voor de vakken Nederlands en Engels laat zien dat het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo- en drie-havo/vwo-klassen bij de proefafname voor Engels wat groter is dan bij de proefafname voor Nederlands. Het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo/vwo en drie-vwo-klassen is bij de proefafname voor Nederlands weer wat groter dan bij de proefafname voor Engels. Het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-vwo-klassen in de proefafnames voor beide vakken is vergelijkbaar.

Beoordelaarsovereenstemming

De overeenstemming tussen docenten bij 21 van de 46 items is goed. De beoordelaarsovereenstemmingscoëfficiënt was bij deze items groter dan 0,80. Bij één van deze items was de overeenstemming perfect. Van de resterende 25 items was er bij 22 sprake van een redelijke overeenstemming. De coëfficiënt van overeenstemming lag bij deze items tussen 0,60 en de 0,80. Bij de overige drie items was de coëfficiënt kleiner dan 0,60. De overeenstemming tussen docenten bij deze items is zo slecht dat ze geen deel mogen gaan uitmaken van de te construeren toetsen.

Docentvragenlijst en verslagformulier

Analyse van de reacties van docenten op de docentvragenlijst en het verslagformulier laat zien dat docenten niet veel commentaar hadden op de meerkeuze-items leesvaardigheid. Het weinige commentaar betrof vooral de afnametijd van de toetsboekjes. In veel gevallen waren de leerlingen snel klaar. Een enkele keer zijn er opmerkingen gemaakt over de moeilijkheid van de items. De meeste en de meest uitgebreide reacties zijn gekomen op de items schrijfvaardigheid, zowel in het algemeen, als met betrekking tot afzonderlijke items.

De meeste docenten vonden de correctie van de items schrijfvaardigheid te tijdrovend. Er is weinig kritiek geuit op de aard van de items. De items die zich richtten op het gebruik van het woordenboek ondervonden weerstand vanwege het praktische probleem van het ontbreken van woordenboeken en vanwege de geringe vertrouwdheid van leerlingen met dit soort items. Ook gaven docenten regelmatig kritiek op de correctievoorschriften.

2.2.4 De resultaten voor het vak wiskunde

De volgens de hiervoor beschreven opzet en procedures verzamelde en verwerkte leerlingantwoorden voor het vak wiskunde zijn na de proefafname geanalyseerd. OPLM-analyses zijn uitgevoerd en voor alle items is de beoordelaarsovereenstemming onderzocht. De gegevens die door docenten

verstrekt zijn in de docentvragenlijst en op het verslagformulier zijn eveneens verwerkt en geanalyseerd.

Resultaten OPLM-analyses

De OPLM-analyse liet in eerste instantie zien dat bij een aantal items enkele scorecategorieën niet of nauwelijks gevuld waren. Daarom zijn allereerst de scorecategorieën bij de betreffende items herzien en zijn de antwoordmodellen van deze items dienovereenkomstig aangepast. De OPLM-analyse na hercategorisering laat zien dat 83 van de 93 items dezelfde onderliggende vaardigheid meten. Tien items bleken niet binnen het model te passen.

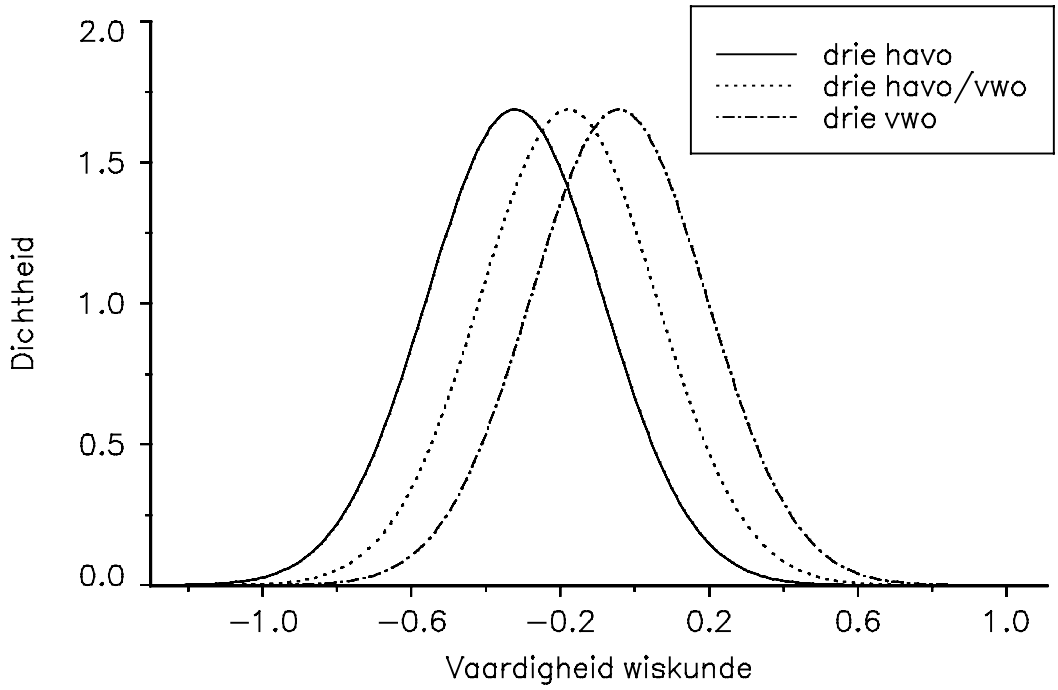
Tabel 2.7 bevat de resultaten van de SAUL-analyse die is uitgevoerd om de verschillen in vaardigheid te bepalen tussen de klastypen drie havo, drie havo/vwo en drie vwo op de vaardigheidsschaal voor wiskunde. Het geschatte algemeen gemiddelde μ op de gestandaardiseerde vaardigheidsschaal is -0,9826 met een standaardfout van 0,0243. De geschatte binnengroepvariantie σ^2 is 0,5217 met een standaardfout van 0,0182. Voor een nadere toelichting op de opzet van tabel 2.7 wordt verwezen naar de toelichting bij tabel 2.5 in paragraaf 2.2.2.

Tabel 2.7

Resultaten van de SAUL-analyse ter bepaling van de verschillen in vaardigheid
tussen drie havo, drie havo/vwo en drie vwo op de schaal voor wiskunde

Klastype	Effect	SE	n	Z	Effectgrootte
Havo	0	0	1217		
Havo/vwo	0,437	0,048	405	9,174	0,605
Vwo	0,852	0,032	1548	26,691	1,179
Contrast					
Vwo - Havo/vwo	0,415	0,046		9,011	0,574

Figuur 2.3 bevat voor de drie klastypen de afbeeldingen op de originele vaardigheidsschaal van de kansdichtheidsfuncties van de vaardigheid die gemeten wordt door de voor het vak wiskunde ontwikkelde items. Het geometrisch gemiddelde van de discriminatieparameters van de items is 3,054. De gemiddelde vaardigheid binnen de havo-groep op de originele vaardigheidsschaal is -0,322. De gemiddelde vaardigheid binnen de havo/vwo-groep op deze schaal bedraagt -0,179 en de gemiddelde vaardigheid binnen de vwo-groep is -0,043. De binnengroepvariantie is 0,0557.



Figuur 2.3

Kansdichtheidsfuncties voor de drie klastypen onder aanname van normaliteit

voor de vaardigheidsschaal voor wiskunde

Tabel 2.7 en figuur 2.3 maken duidelijk dat het verschil in de betreffende vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-havo/vwo- klassen tamelijk groot is. Leerlingen uit drie-havo/vwo-klassen zijn gemiddeld vaardiger dan leerlingen uit drie-havo-klassen. Er is sprake van

een klein tot middelmatig effect. Leerlingen uit drie-havo/vwo-klassen zijn op hun beurt gemiddeld weer minder vaardig dan leerlingen uit drie-vwo-klassen. Het verschil tussen deze twee groepen is iets kleiner, maar ook in dit geval kan gesproken worden van een klein tot middelmatig effect. De tabel en figuur laten zien dat het verschil in vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-vwo-klassen aanzienlijk is. De bijbehorende effectgrootte is hier zelfs zodanig dat gesproken moet worden van een groot effect.

Een vergelijking tussen de proefafnames voor de vakken Nederlands, Engels en wiskunde laat zien dat het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-havo/vwo-klassen bij de proefafname voor wiskunde een stuk groter is dan bij de proefafname voor Engels en Nederlands. Het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo/vwo-klassen en leerlingen uit drie-vwo-klassen is bij de proefafname voor wiskunde vergelijkbaar met de proefafname voor Nederlands en wat groter dan bij de proefafname voor Engels. Het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo-klassen en leerlingen uit drie-vwo-klassen is bij de proefafname wiskunde veel groter dan bij de proefafnames voor de twee andere vakken.

Beoordelaarsovereenstemming

Bij 38 van de 93 items was de beoordelaarsovereenstemmingscoëfficiënt groter dan 0,80. Bij twee van deze items was de overeenstemming perfect. De overeenstemming tussen docenten bij deze 38 items is goed. Bij de resterende 55 items waren er 44 waarbij de coëfficiënt van overeenstemming tussen 0,60 en 0,80 lag. Dat wil zeggen dat de correctievoorschriften bij deze items redelijk zijn. Bij de overige elf items was de coëfficiënt van overeenstemming kleiner dan 0,60. Bij deze items is sprake van een slechte overeenstemming. Ze mogen dan ook niet in de te construeren toetsen worden opgenomen.

Docentvragenlijst en verslagformulier

Uit de antwoorden van docenten valt af te leiden dat docenten veel van de items uit de toetsboekjes redelijk bij hun onderwijs vonden passen. Dat neemt niet weg dat bij sommige context-items enkele docenten opmerkten dat ze geen relevante vaardigheden in de items weerspiegeld zagen. Over het correctievoorschrift zijn per item uiteenlopende opvattingen naar voren gebracht.

2.3 Het samenstellen en normeren van de toetsen

Na afronding van de analyses van de in de proefafnames verzamelde informatie moesten de vakmedewerkers toetsen samenstellen voor de vakken Nederlands, Engels en wiskunde. Hoe zij daarbij te werk dienden te gaan en hoe de geconstrueerde toetsen vervolgens zijn genormeerd is het onderwerp van deze paragraaf.

2.3.1 Het samenstellen van de toetsen

De samen te stellen toetsen moesten aan de volgende specificaties voldoen:
- alle items in de toets passen binnen het OPLM;
- de toets bevat het aantal items dat door de leerlingen in de ter beschikking staande toetstijd van twee lesuren redelijkerwijs gemaakt kan worden;
- de toets maakt een zo goed mogelijk onderscheid tussen vaardige en minder vaardige leerlingen aan het einde van het derde leerjaar in het onderwijstype waar de toets voor bedoeld is (drie havo of drie vwo);
- de in de toets opgenomen items vertonen geen feilen en gebreken wat betreft formulering, duidelijkheid en lay-out;
- de in de toets opgenomen items zijn door een zo groot mogelijk aantal docenten geschikt bevonden voor opname in de toets;
- de correctievoorschriften bij de kort-open-antwoord-items zijn duidelijk en de beoordelaarsovereenstemming bij de in de toetsen opgenomen kort-open-antwoord-items ligt boven een aanvaardbaar minimum.

Zowel bij het vak Nederlands als bij het vak Engels was sprake van drie verzamelingen met verschillende soorten items. Hiervoor is aangegeven dat bij beide vakken een groot deel van de items uit de drie itemverzamelingen gezamenlijk schaalbaar was. Bij het samenstellen van de toetsen voor Nederlands en Engels is gebruik gemaakt van de itemparameters bij de vaardigheidsschalen die ontstonden door de items uit de drie itemverzamelingen gezamenlijk te schalen.

Een aanvullende eis voor de toetsen wiskunde was nog dat in beide toetsen items moesten voorkomen die betrekking hebben op de onderwerpen:

- getalsmatige verhoudingen en verbanden;
- wiskundige formuleringen, formules en notaties en (soorten) functies en relaties;
- meetkundige (vlakke en ruimtelijke) figuren;
- statistiek en combinatoriek;
- kans.

2.3.2 De methode van normeren

Het was de wens van het ministerie van Onderwijs, Cultuur en Wetenschappen dat de instrumentaria zo snel mogelijk ter beschikking van de scholen zouden komen. Daarom is er vanaf gezien voor de te ontwikkelen instrumenten een apart normeringsonderzoek uit te voeren. De items die in de plaatsingstoetsen zijn opgenomen, waren echter afkomstig uit verschillende toetsboekjes. Voor geen van de drie vakken gold dat er leerlingen waren aan wie alle items waren voorgelegd die uiteindelijk zijn opgenomen in de plaatsingstoetsen. Het was dus niet mogelijk de scores van leerlingen op de geconstrueerde plaatsingstoetsen af te leiden uit hun scores op de items. Omdat items uitsluitend voor opname in de toetsen in aanmerking kwamen indien zij volgens het OPLM geschaald waren, kon echter voor iedere toets de scoreverdeling als volgt afgeleid worden.

Stel dat op een toets *t*, die bestaat uit *n* gecalibreerde items, de scores 0, 1, ..., M_t behaald kunnen worden, waarbij M_t de maximumscore op de betreffende toets is. De conditionele kans op een score S_t op deze toets, gegeven een vaardigheid θ, is de som van de kansen op het optreden van de

verschillende responspatronen die leiden tot de score s_t, gegeven deze vaardigheid:

$$p(S_t|\theta) = \sum_{\{\mathbf{x}: r(\mathbf{x})=s_t\}} \prod_i f_{ij}(\theta). \tag{2.5}$$

Hierbij heeft $f_{ij}(\theta)$ dezelfde vorm als uitdrukking (2.1) of, indien sprake is van dichotome items, als uitdrukking (2.2). Verder staat $\mathbf{x}$ voor een vector van itemscores en geldt dat $r(\mathbf{x}) = \sum_i a_i x_i$.

De marginale kans op score S_t wordt gegeven door:

$$p(S_t) = \int_{-\infty}^{\infty} p(S_t|\theta)\, d\, G(\theta), \tag{2.6}$$

waarbij $G(\theta)$ de vaardigheidsverdeling is. Uitdrukking (2.6) kan niet rechtstreeks berekend worden. Een goede numerieke benadering ervan is echter wel mogelijk en deze is geïmplementeerd in het programma OPTAL (Verstralen, 1997). Binnen OPTAL wordt aangenomen dat $\theta \sim N(\mu,\sigma^2)$. Met behulp van uitdrukking (2.6) zijn cumulatieve scoreverdelingen te berekenen bij een toets *t*. De kans dat op de toets *t* een score S_t behaald wordt van r_t of lager is:

$$P(S_t \leq r_t) = \sum_{S_t \leq r_t} p(s_t). \tag{2.7}$$

Met behulp van OPTAL zijn bij de drie ontwikkelde toetsen voor de havo en de drie ontwikkelde toetsen voor het vwo cumulatieve scoreverdelingen berekend. Daarbij is gebruik gemaakt van de μ's en σ's die eerder in de SAUL-analyses voor de leerlingen uit drie havo en drie vwo bepaald zijn. Om eenvoudig interpreteerbare normtabellen te verkrijgen zijn voor iedere toets de scores met behulp van de door OPTAL berekende cumulatieve scoreverdeling omgezet naar decielscores. Deze normtabellen staan in bijlage A.

2.3.3 Beschrijving van de toetsen

Op grond van de eerder gegeven specificaties zijn voor leerlingen uit drie havo en leerlingen uit drie vwo plaatsingstoetsen opgesteld voor de vakken Nederlands, Engels en wiskunde.

De toets Nederlands voor drie havo bestaat uit 38 items. In de toets zijn alle drie de itemsoorten vertegenwoordigd die bij de proefafname onderscheiden zijn. De eerste negen items zijn kort-open-antwoord-items die betrekking hebben op een tekst. De volgende twaalf items zijn cloze-items en de laatste zeventien zijn meerkeuze-items die elk betrekking hebben op een korte tekst. De maximumscore op de toets is 115.

De toets Nederlands die voor drie vwo is ontwikkeld, is identiek aan deze toets. Daarvoor waren drie redenen. Allereerst was het verschil in leesvaardigheid tussen havo- en vwo-leerlingen niet erg groot. Bovendien moesten er uit het oogpunt van efficiëntie bij de cloze-items en de kort-open-antwoord-items leesvaardigheid teksten met bijbehorende items geselecteerd worden in plaats van individuele items. En ook was het aantal meerkeuze-items beperkt dat voor opname in aanmerking kwam.

De toets Engels die voor drie havo bestemd is, bevat 43 items. De eerste tien items zijn cloze-items. De volgende dertien items vragen leerlingen de betekenis te bepalen van een niet bestaand woord in een korte tekst. De volgende acht items zijn meerkeuze-items over een Engelse tekst. De laatste twaalf items zijn wederom cloze-items. De toets heeft een maximumscore van 81.

De voor drie vwo ontwikkelde toets Engels bestaat uit 47 items. De eerste twaalf items zijn cloze-items. De volgende dertien items vragen leerlingen de betekenis te bepalen van een niet bestaand woord in een korte tekst. De laatste 22 items zijn kort-open-antwoord-items die bepaalde aspecten van de schrijfvaardigheid van leerlingen meten. De maximaal te behalen score op de toets is 97.

De toets wiskunde die voor drie havo is samengesteld bestaat uit 27 kort-open-antwoord-items. De maximumscore op deze toets is 57. Voor drie vwo is een toets opgesteld bestaande uit 28 items. De toets heeft een maximumscore van 60.

2.4 Kort resumé

In het voorafgaande is aangegeven langs welke weg de plaatsingstoetsen voor Nederlands, Engels en wiskunde voor drie havo en drie vwo tot stand gekomen zijn. Allereerst zijn de functie en de gewenste kenmerken van de toetsen gespecificeerd. Vervolgens is de opzet en uitvoering van de proefafnames beschreven en is aangegeven van welke procedures gebruik gemaakt is om de kwaliteit van de ontwikkelde items te onderzoeken. Ten slotte is geschetst hoe te werk gegaan is bij het samenstellen van de toetsen en het normeren van de toetsscores. Daarmee is het proces van toetsconstructie nog niet afgerond. Een volgende belangrijke stap is het onderzoek naar de kwaliteit van de geconstrueerde toetsen. Dit vormt het onderwerp van de hoofdstukken drie en vier. Hoofdstuk drie heeft betrekking op de meetnauwkeurigheid van de toetsen en hoofdstuk vier op hun validiteit.

3 De meetnauwkeurigheid van de plaatsingstoetsen

Om een uitspraak te kunnen doen over de kwaliteit van een toets moet onderzocht worden in welke mate deze voldoet aan twee fundamentele eisen. De eerste eis is die van meetnauwkeurigheid. Een toets voldoet beter aan deze eis, naarmate de toetsscores vrijer zijn van meetfouten. De tweede eis is die van validiteit. Validiteit is een overkoepelend begrip dat betrekking heeft op de betekenis, bruikbaarheid en juistheid van de conclusies die getrokken kunnen worden uit de scores op een toets. In dit hoofdstuk komt het onderzoek aan de orde dat is uitgevoerd naar de meetnauwkeurigheid van de ontwikkelde toetsen. De validiteit van de ontwikkelde toetsen is het onderwerp van hoofdstuk vier.

3.1 Betrouwbaarheid en toetsinformatie

De klassieke toetstheorie (KTT; Lord en Novick, 1968), hanteert het begrip betrouwbaarheid om de meetnauwkeurigheid van een toets aan te duiden. In de KTT valt af te leiden dat de variantie van de waargenomen toetsscores X gelijk is aan de som van de varianties van de ware scores T en de meetfouten E in de waargenomen scores:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{3.1}$$

De betrouwbaarheid van een toets wordt in de KTT gedefinieerd als de gekwadrateerde correlatie van ware en waargenomen score. Deze blijkt onder de aannamen van het testmodel waar de KTT vanuit gaat gelijk te zijn aan de verhouding van de varianties van de ware en de waargenomen scores:

$$\rho_{XT}^2 = \sigma_T^2/\sigma_X^2. \tag{3.2}$$

Deze klassieke maat voor de meetnauwkeurigheid is populatie-afhankelijk: ze beschrijft de meetnauwkeurigheid van een toets in relatie tot een bepaalde populatie. De betrouwbaarheid van een en dezelfde toets zal hoger zijn naarmate de spreiding van de vaardigheid groter is in de groep personen aan wie de toets is voorgelegd. In het extreme geval dat alle leden van een populatie dezelfde vaardigheid en dus dezelfde ware score hebben, is de betrouwbaarheid van een toets voor deze populatie gelijk aan nul (zie bijvoorbeeld Fischer, 1974; Samejima, 1994). Verder geldt dat de betrouwbaarheid en de daarvan afgeleide standaardmeetfout constant zijn voor het hele scorebereik van een toets. Dat is echter geen realistische veronderstelling. Er zal altijd sprake zijn van lokale verschillen in meetnauwkeurigheid. Dat wordt direct duidelijk bij het beschouwen van de minimum- en maximumscore op een toets. Omdat ware scores hoger dan de maximumscore of lager dan de minimumscore niet mogelijk zijn, zou een toets voor deze beide scores een andere meetnauwkeurigheid moeten hebben dan voor de overige scores. Binnen de KTT is overigens een aantal procedures ontwikkeld om lokale standaardmeetfouten te bepalen. Een overzicht daarvan is te vinden in Feldt, Steffen en Gupta (1985).

De maat die de IRT gebruikt voor de meetnauwkeurigheid van een toets kent de twee genoemde eigenschappen - populatie-afhankelijkheid en constantheid voor het hele scorebereik - niet. Om de lokale meetnauwkeurigheid van een toets te specificeren wordt in de IRT gebruik gemaakt van het statistische begrip informatie (zie bijvoorbeeld Lindgren, 1976, p. 248). De IRT maakt het mogelijk bij ieder item een zogeheten iteminformatiefunctie $I_i(\theta)$ op te stellen. Bij het gebruik van het OPLM geldt voor de iteminformatiefunctie in het geval van polytome items:

$$I_i(\theta) = a_i^2 \sum_{j=0}^{m_i} f_{ij}(\theta)\left[j - \sum_{h=0}^{m_i} h f_{ih}(\theta) \right]^2,$$

waarbij $f_{ij}(\theta)$ en $f_{ih}(\theta)$ dezelfde vorm hebben als in uitdrukking (2.1). In het geval van dichotome items geldt bij het gebruik van het OPLM voor de iteminformatiefunctie:

$$I_i(\theta) = a_i^2 f_i(\theta)[1 - f_i(\theta)],$$

waarbij $f_i(\theta)$ dezelfde vorm heeft als in uitdrukking (2.2). De iteminformatiefunctie geeft aan welke bijdrage een item levert aan de meetnauwkeurigheid voor ieder punt op de vaardigheidsschaal waar het item betrekking op heeft. De iteminformatiefunctie van een dichotoom item bereikt zijn maximum wanneer $\theta = \beta_i$ en wordt kleiner naarmate het verschil tussen θ en β_i toeneemt.

Met behulp van de informatiefuncties van de items in een toets *t* kan een toetsinformatiefunctie opgesteld worden. Voor de toetsinformatiefunctie $I_t(\theta)$ geldt:

$$I_t(\theta) = \sum_{i \in t} I_i(\theta). \qquad (3.3)$$

De waarde die (3.3) aanneemt voor een bepaald punt op de vaardigheidsschaal is afhankelijk van de waarden van de β_i's van de items in de betreffende toets. De waarde van de toetsinformatiefunctie is hoger op een bepaald punt, naarmate de β_i's meer geconcentreerd zijn rond dit punt.

Met behulp van de toetsinformatiefunctie valt de standaardfout te schatten van de vaardigheidsschattingen die op grond van de toetsscores gemaakt kunnen worden. Voor de standaardfout *SE* van een met behulp van een toets *t* geschatte vaardigheid geldt namelijk:

$\lim_{n \to \infty} SE_t(\hat{\theta}) = \sqrt{1/I_t(\theta)}$, waarbij *n* het aantal items is.

Daardoor is het mogelijk een asymptotisch (1-α)-procents-betrouwbaarheidsinterval op te stellen rond iedere vaardigheidsschatting:

$$\hat{\theta} - Z_{\alpha/2} \cdot SE_t(\hat{\theta}) \leq \theta \leq \hat{\theta} + Z_{\alpha/2} \cdot SE_t(\hat{\theta}).$$

Hierbij is α de waarschijnlijkheid dat het betrouwbaarheidsinterval de werkelijke waarde van θ niet bevat en *Z* de met de betreffende α corresponderende waarde uit de standaardnormale verdeling.

Het feit dat toetsinformatie een populatie-onafhankelijke maat is voor de meetnauwkeurigheid van een toets is een van de grote voordelen van de IRT

(zie bijvoorbeeld Fischer, 1974; Hambleton & van der Linden, 1982; Van der Linden & Hambleton, 1997).

Hoewel de concepten betrouwbaarheid en toetsinformatie betrekking hebben op verschillende aspecten van meetnauwkeurigheid, zijn ze toch nauw aan elkaar verwant (Mellenbergh, 1996). Toetsinformatie is een conditionele maat voor de meetnauwkeurigheid van een toets, namelijk voor een persoon met een specifieke vaardigheid. Toetsinformatie heeft derhalve betrekking op de lokale meetnauwkeurigheid van een toets. Betrouwbaarheid is de niet-conditionele pendant van het begrip toetsinformatie en geeft een algemeen beeld van de meetnauwkeurigheid van een toets voor een populatie van personen. Betrouwbaarheid is een maat voor de globale meetnauwkeurigheid van een toets.

3.2 Het bepalen van de meetnauwkeurigheid van de toetsen

In het voorafgaande is betoogd dat betrouwbaarheid en toetsinformatie concepten zijn die elkaar aanvullen. Om een volledige beschrijving te geven van de meetnauwkeurigheid van een toets moet derhalve voor beide concepten aandacht zijn. In de twee volgende paragrafen wordt beschreven welke procedures gehanteerd zijn om de lokale en globale meetnauwkeurigheid te bepalen van de plaatsingstoetsen die ontwikkeld zijn voor de vakken Nederlands, Engels en wiskunde. Deze twee paragrafen hebben een technisch karakter en verschillen van aard met de rest van dit hoofdstuk.

3.2.1 Het bepalen van de lokale meetnauwkeurigheid

De lokale meetnauwkeurigheid van een toets kan beschreven worden door bij iedere mogelijke toetsscore de vaardigheid te schatten en bij iedere schatting een betrouwbaarheidsinterval te geven. Hoe kleiner een betrouwbaarheidsinterval bij een bepaalde schatting, des te nauwkeuriger meet de toets op het betreffende punt op de vaardigheidsschaal. Ook is het mogelijk de toetsinformatie of de standaardfout als functie van de vaardigheid in een

grafiek af te beelden. Voor de gebruikers van de toetsen zijn beide vormen van informatie echter moeilijk te interpreteren.

Om een eenvoudige interpretatie van de toetsscores mogelijk te maken, zijn bij de plaatsingstoetsen normtabellen opgesteld. In deze normtabellen zijn de toetsscores onderverdeeld in decielen. Er is voor gekozen om de gebruikers aan de hand van deze decielen informatie te verschaffen over de lokale meetnauwkeurigheid van de toetsen. Daartoe zijn zogeheten betrouwbaarheidsmatrices opgesteld met behulp van het programma OPTAL (Verstralen, 1997).

Om een betrouwbaarheidsmatrix voor een toets op te kunnen stellen dient allereerst de vaardigheidsschaal bij de betreffende toets onderverdeeld te worden in K ($k = 1, 2, ..., K$) geordende klassen. Iedere klasse heeft betrekking op een bepaald interval op een schaal. Tezamen bestrijken de klassen het volledige bereik van deze schaal. Iedere cel b_{ij} in een betrouwbaarheidsmatrix $\boldsymbol{B}$ geeft de kans dat iemand toegewezen wordt aan klasse i, gegeven dat de betreffende persoon in werkelijkheid deel uitmaakt van klasse j, of vice versa. De indices i en j kunnen betrekking hebben op identieke definities van klassen, maar ook op verschillende. Voor de betrouwbaarheidsmatrices bij de plaatsingstoetsen geldt dat $K = 10$. Deze betrouwbaarheidsmatrices geven dus aan hoe groot de kans is dat een leerling met een score die valt in deciel i in werkelijkheid deel uitmaakt van deciel j, waarbij $i, j = 1, 2, ..., 10$.

De waarden op de vaardigheidsschaal die corresponderen met de gehanteerde indeling in decielen zijn met behulp van de uitdrukking $\theta_k = G^{-1}(P_k)$ bepaald, waarbij G^{-1} de inverse is van de cumulatieve verdelingsfunctie G en P_k de kans is dat een leerling deel uitmaakt van een klasse met een index lager dan of gelijk aan k. Bij decielen is $P_k = k/10$ ($k = 1, 2, ..., 10$). Voor de ondergrens op de vaardigheidsschaal voor het eerste deciel en de bovengrens op de vaardigheidsschaal voor het tiende deciel geldt hierbij respectievelijk dat $\theta_0 \rightarrow -\infty$ en $\theta_{10} \rightarrow \infty$.

Bij het bepalen van de waarden van de scores die corresponderen met de grenzen van de decielen doet zich een probleem voor, omdat deze waarden in de regel geen gehele getallen zijn. Om dit probleem op te lossen wordt de

cumulatieve score-verdeling *P(s)* (*s = 1, 2, ..., M*) van de discrete variabele *s* benaderd door een cumulatieve scoreverdeling *D(u)* van een continue variabele *u*. Voor *D(u)* geldt:

$$D(u + 0{,}5) = P(u) \qquad (\forall\, u \in \{0,\ 1,\ ...,\ M\});$$
$$D(u + 0{,}5) = P(\lfloor u) + (u - \lfloor u)\,(P(\lceil u) - P(\lfloor u)) \qquad (\forall\, u \notin \{0,\ 1,\ ...,\ M\}).$$

Hierbij staat $\lfloor u$ voor het eerste gehele getal dat kleiner is dan *u* en $\lceil u$ voor het eerste gehele getal dat groter is dan *u*. Indien *u* geen integer is, wordt de waarde van de cumulatieve scoreverdeling *D(u)* dus bepaald via lineaire interpolatie.

De waarden van *u* die corresponderen met de indeling in decielen kunnen nu met behulp van de uitdrukking $u_k = D^{-1}(P_k)$ bepaald worden. Hierbij is D^{-1} de inverse van de cumulatieve verdelingsfunctie *D*. Voor de minimumscore in het eerste deciel en de maximumscore in het tiende deciel geldt hierbij respectievelijk dat $u_0 = 0{,}5$ en $u_{10} = M + 0{,}5$, zodat $D(u_0) = 0$ en $D(u_{10}) = 1$.

Van scores die vallen in de intervallen $s - 0{,}5 < u_k \leq s + 0{,}5$ is het onduidelijk tot welk deciel zij behoren. Indien leerlingen met dergelijke scores alle aan een en hetzelfde deciel zouden worden toegewezen, zou dat leiden tot een over- of onderschatting van de kansen in de betrouwbaarheidsmatrix. Dit probleem wordt opgelost door ervoor te zorgen dat dergelijke leerlingen met een kans $u_k - (s - 0{,}5)$ worden toegewezen aan deciel $k - 1$ en met een kans van $1 - u_k + (s - 0{,}5)$ aan deciel *k*.

Alvorens de betrouwbaarheidsmatrix opgesteld kan worden moet eerst nog de conditionele vaardigheidsverdeling $G(\theta|s)$ continu in s gemaakt worden. Daartoe wordt $H(\theta|u)$ gedefinieerd als:

$$H(\theta|u) = G(\theta|u) \qquad (\forall\, u \in \{0,\ 1,\ ...,\ M\});$$
$$H(\theta|u) = G(\theta|u) + (u - \lfloor u)\,(G(\theta|\lceil u) - G(\theta|\lfloor u)) \qquad (\forall\, u \notin \{0,\ 1,\ ...,\ M\}).$$

Voor iedere cel b_{ij} in de betrouwbaarheidsmatrix ***B*** kan nu de conditionele kans op het lidmaatschap van het vaardigheidsinterval *j*, gegeven een geobserveerde score die valt in klasse *i*, berekend worden als:

$$b_{ij} = \int_{u_{i-1}}^{u_i} \left\{ H(\theta_j | u) - H(\theta_{j-1} | u) \right\} dD(u) \qquad (i, j = 1, 2, ..., 10)$$

Voor het evalueren van deze integraal wordt gebruik gemaakt van het Simpson algoritme, met dertig stappen in ieder interval $[\theta_j, \theta_{j+1}]$. De ondergrens op de vaardigheidsschaal voor het eerste deciel en de bovengrens op de vaardigheidsschaal voor het tiende deciel, θ_0 en θ_{10}, krijgen hierbij respectievelijk een waarde die zeven standaardafwijkingen lager en zeven standaardafwijkingen hoger ligt dan het gemiddelde van de vaardigheidsverdeling.

3.2.2 Het bepalen van de betrouwbaarheid van de toetsen

In hoofdstuk twee is reeds toegelicht dat het niet mogelijk was de toetsen in hun uiteindelijke vorm in een apart normeringsonderzoek aan leerlingen voor te leggen. Daarom viel hun betrouwbaarheid niet rechtstreeks te bepalen. De leerlingen die deelgenomen hebben aan de proefafname die ten grondslag lag aan de ontwikkeling van een toets hebben immers zonder uitzondering slechts een deel gemaakt van de items die uiteindelijk in de betreffende toets opgenomen zijn. Hieronder wordt duidelijk gemaakt dat het echter wel mogelijk is de betrouwbaarheid van iedere toets te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de betreffende toets OPLM-geschaald zijn.

Voor de ware score $\tau_i(\theta)$ - de verwachte waarde van de geobserveerde score gegeven θ - op een OPLM-gecalibreerd item i met J_i scorecategorieën en itemscore X_i geldt:

$$\tau_i(\theta) \doteq \mathscr{E}(X_i|\theta) = \sum_{j=1}^{J_i} j f_{ij}(\theta),$$

waarbij $f_{ij}(\theta)$ gegeven is door uitdrukking (2.1). Voor het tweede niet-centrale moment van de geobserveerde score, gegeven θ, geldt:

$$m_{2_i}(\theta) \doteq \mathscr{E}(X_i^2|\theta) = \sum_{j=1}^{J_i} j^2 f_{ij}(\theta).$$

Bij een vaste waarde van θ is de variantie van de ware score op een item per definitie gelijk aan nul. Daarom is de variantie van de geobserveerde score, gegeven θ, gelijk aan de foutenvariantie en geldt:

$$\sigma^2_{X_i}(\theta) = \sigma^2_{E_i}(\theta) = m_{2_i}(\theta) - \tau_i(\theta)^2.$$

Voor de verwachte marginale geobserveerde score op een item, gegeven een populatie met vaardigheidsverdeling $G(\theta)$, geldt:

$$\tau_i \doteq \mathscr{E}(X_i) = \int \tau_i(\theta)\, dG(\theta).$$

Voor het tweede niet-centrale moment van de marginale geobserveerde score geldt:

$$m_{2_i} \doteq \mathscr{E}(X_i^2) = \int m_{2_i}(\theta)\, dG(\theta).$$

Voor de variantie van de marginale geobserveerde score geldt dus:

$$\sigma^2_{X_i} = m_{2_i} - \tau_i^2. \qquad (3.4)$$

Voor de verwachte marginale ware score op een item, gegeven een populatie met vaardigheidsverdeling $G(\theta)$, geldt:

$$\mathscr{E}(T_i) = \mathscr{E}(X_i) = \tau_i.$$

Voor het tweede niet centrale moment van de marginale ware score geldt:

$$\mu_{2_i} = \mathscr{E}(T_i^2) = \int \tau_i(\theta)^2 \, dG(\theta).$$

Derhalve geldt voor de variantie van de marginale ware scores:

$$\sigma_{\tau_i}^2 = \mu_{2_i} - \tau_i^2. \tag{3.5}$$

Voor de marginale foutenvariantie geldt:

$$\sigma_{E_i}^2 = \int \sigma_{E_i}^2(\theta) \, dG(\theta) = m_{2_i} - \mu_{2_i}. \tag{3.6}$$

Met behulp van uitdrukking (3.4) is het mogelijk de variantie van de gewogen geobserveerde scores te schatten op een toets *t* die is samengesteld uit een verzameling OPLM-gecalibreerde items:

$$\sigma_{X_t}^2 = \sum_{i \in t} a_i^2 \sigma_{X_i}^2. \tag{3.7}$$

Uitdrukking (3.5) maakt het mogelijk de variantie van de gewogen ware scores op toets *t* te schatten:

$$\sigma_{\tau_t}^2 = \sum_{i \in t} a_i^2 \sigma_{\tau_i}^2. \tag{3.8}$$

Uitdrukking (3.6) biedt de gelegenheid de variantie van de meetfouten in de gewogen geobserveerde scores op toets *t* te schatten:

$$\sigma_{E_t}^2 = \sum_{i \in t} a_i^2 \sigma_{E_i}^2.$$

In deze drie laatste uitdrukkingen zijn de a_i steeds de discriminatie-indices van de items die deel uitmaken van toets *t.*

Met behulp van de uitdrukkingen (3.2), (3.7) en (3.8) kan voor een populatie met vaardigheidsverdeling $G(\theta)$ de betrouwbaarheid geschat worden van elke toets t die samengesteld is uit een verzameling OPLM-gecalibreerde items:

$$\hat{\rho}^2_{X_t T_t} = \frac{\sigma^2_{\tau_t}}{\sigma^2_{X_t}} = \frac{\sum_{i \in t} a_i^2 \sigma^2_{\tau_i}}{\sum_{i \in t} a_i^2 \sigma^2_{X_i}} = \frac{\sum_{i \in t} a_i^2 (\mu_{2_i} - \tau_i^2)}{\sum_{i \in t} a_i^2 (m_{2_i} - \tau_i^2)}. \qquad (3.9)$$

In het eerder genoemde programma OPTAL is een procedure geïmplementeerd die op de bovenbeschreven wijze de betrouwbaarheid schat van toetsen die zijn samengesteld uit OPLM-gecalibreerde items. OPTAL berekent tevens een 95-procents betrouwbaarheidsinterval rond de schatting van de betrouwbaarheid met behulp van de standaardfout van de standaardafwijking van de vaardigheidsverdeling. Uiteraard wordt de schatting van de betrouwbaarheid ook beïnvloed door fouten in de schattingen van itemparameters en de gemiddelde vaardigheid. De grootte van de standaardfout van de standaardafwijking heeft echter een veel grotere invloed op de nauwkeurigheid van de schatting van de betrouwbaarheid dan deze andere schattingsfouten (Verstralen, 1997).

3.3 Resultaten

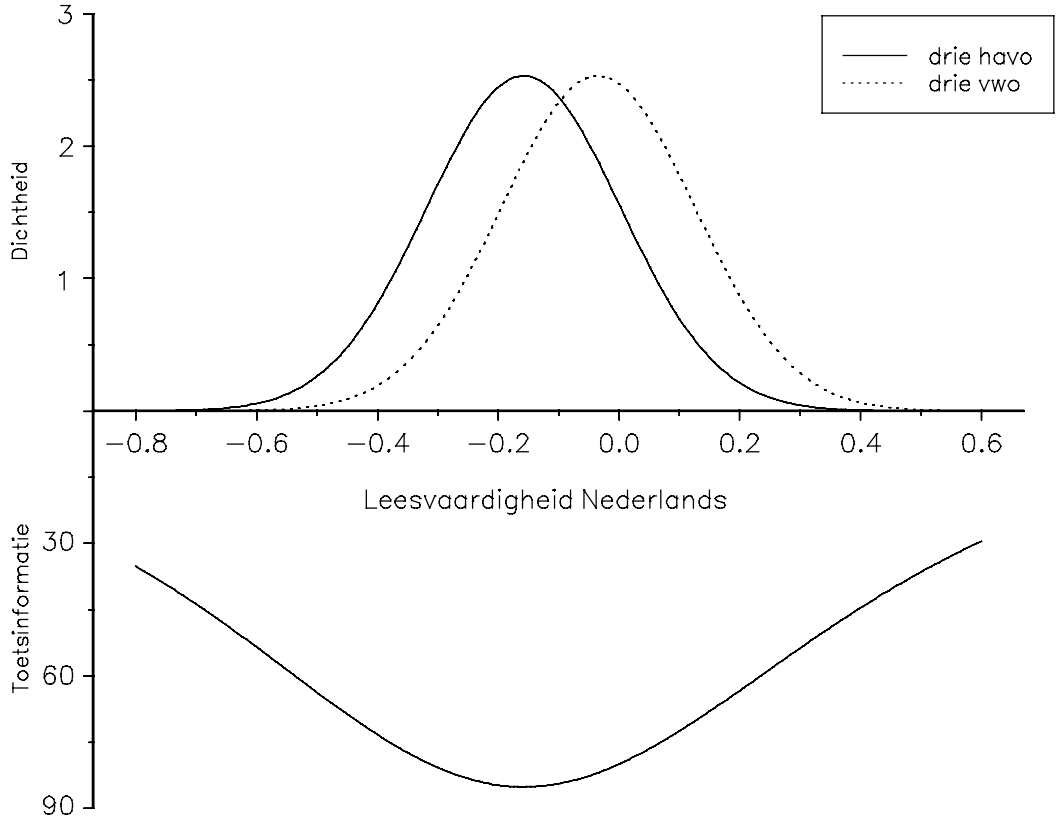
De informatiefuncties van de toetsen die voor de vakken Nederlands, Engels en wiskunde ontwikkeld zijn, staan afgebeeld in de figuren 3.1, 3.2 en 3.3. De figuren maken in een oogopslag duidelijk hoe het gesteld is met de meetnauwkeurigheid van de toetsen voor ieder punt op de vaardigheidsschaal waar zij betrekking op hebben.

De informatie die de figuren leveren is verder geconcretiseerd door betrouwbaarheidsmatrices bij de toetsen op te stellen en klassieke kengetallen bij de toetsen te berekenen met behulp van de procedures die respectievelijk in de paragrafen 3.2.1 en 3.2.2 beschreven zijn. Deze drie verschillende vormen van informatie over de meetnauwkeurigheid van de toetsen komen in de drie volgende paragrafen aan de

orde. Tezamen geven ze een gedetailleerde beschrijving van de meetnauwkeurigheid van de toetsen.

3.3.1 De toetsinformatiefuncties

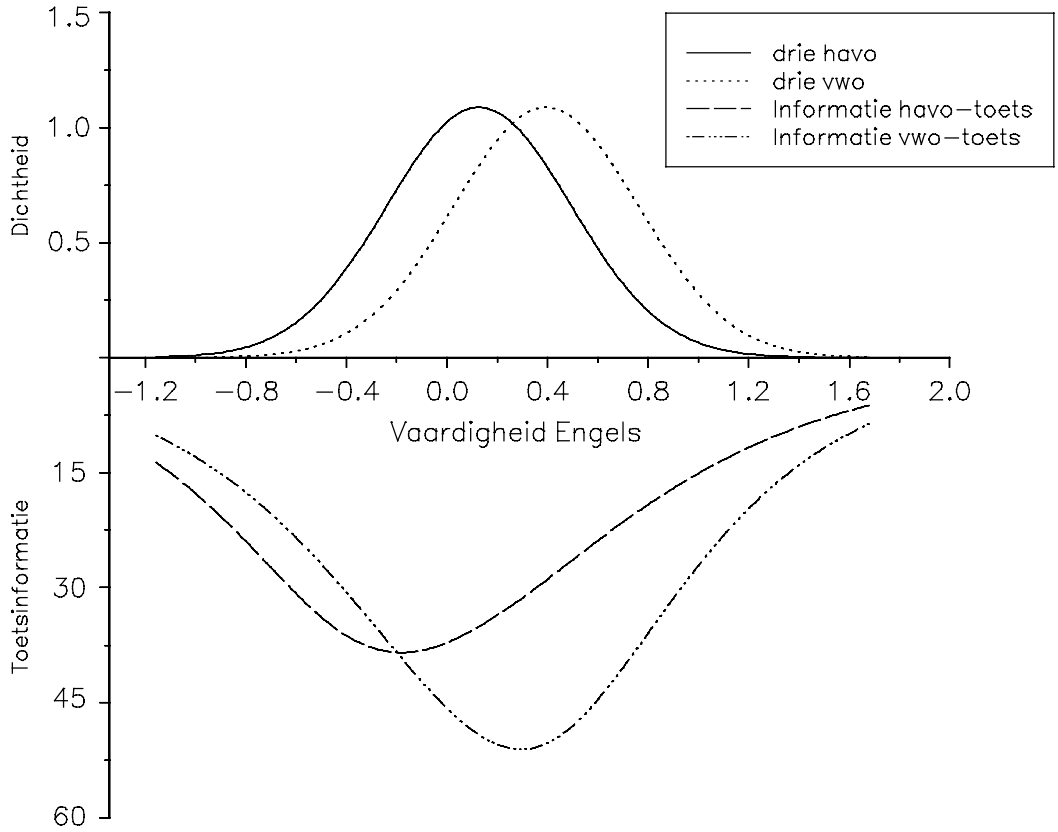
In de figuren 3.1, 3.2 en 3.3 worden naast de informatiefuncties van de toetsen ook de kansdichtheidsfuncties weergegeven die respectievelijk stonden afgebeeld in de figuren 2.1, 2.2 en 2.3. De kansdichtheidsfuncties staan steeds in het bovenste deel van de figuur en de toetsinformatiefuncties in het onderste deel. De figuren als geheel laten niet alleen zien hoe nauwkeurig de toetsen op ieder punt van de vaardigheidsschaal meten, maar ook in hoeverre de toetsen binnen de populaties onderscheid maken tussen leerlingen van verschillende vaardigheidsniveaus. De figuren geven dus zowel informatie over de lokale- als over de globale meetnauwkeurigheid van de toetsen.



Figuur 3.1

Kansdichtheidsfuncties van leesvaardigheid Nederlands voor drie havo en drie vwo en de toetsinformatiefunctie van de bijbehorende toets

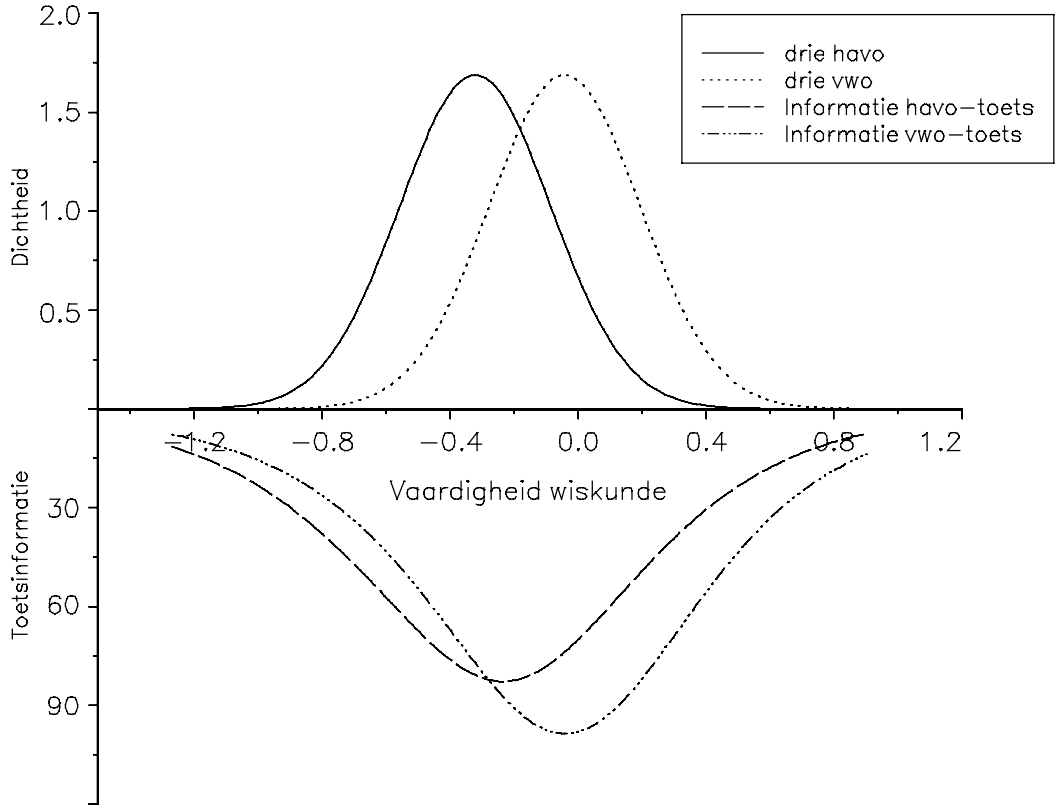
Figuur 3.1 bevat de kansdichtheidsfuncties van leesvaardigheid Nederlands voor de leerlingen uit drie havo en drie vwo. In hoofdstuk twee is al aangegeven dat voor beide groepen een identieke toets Nederlands ontwikkeld is. Het onderste deel van de figuur bevat de informatiefunctie van deze toets. Een vergelijking van de posities van de kansdichtheidsfuncties enerzijds en de toetsinformatiefunctie anderzijds leert dat de toets die is ontwikkeld voor drie havo het nauwkeurigst meet bij leerlingen met een gemiddelde vaardigheid. De voor drie vwo ontwikkelde toets meet daarentegen het nauwkeurigst bij de wat minder vaardige leerlingen.



Figuur 3.2

Kansdichtheidsfuncties voor drie havo en drie vwo van de vaardigheid die de toetsen Engels meten en de toetsinformatiefuncties van deze toetsen

In figuur 3.2 staan de kansdichtheidsfuncties van de vaardigheid die de toetsen Engels meten voor de leerlingen uit drie havo en drie vwo en de informatiefuncties van deze toetsen. Op grond van de posities van de kansdichtheidsfuncties enerzijds en de toetsinformatiefuncties anderzijds is vast te stellen dat de toets die bestemd is voor drie havo de grootste meetnauwkeurigheid heeft bij de wat minder vaardige leerlingen. De voor drie vwo bestemde toets meet eveneens het nauwkeurigst bij de wat minder vaardige leerlingen, zij het in mindere mate.



Figuur 3.3

Kansdichtheidsfuncties voor drie havo en drie vwo van de vaardigheid die de toetsen wiskunde meten en de toetsinformatiefuncties van deze toetsen

In figuur 3.3 worden de kansdichtheidsfuncties afgebeeld van de vaardigheid die de toetsen wiskunde meten voor de leerlingen uit drie havo en drie vwo

tezamen met de informatiefuncties van deze toetsen. Uit de ligging van de kansdichtheidsfunctie ten opzichte van de toetsinformatiefuncties blijkt dat de voor drie havo bedoelde toets het nauwkeurigst meet bij de betere leerlingen. De toets die bedoeld is voor de drie vwo meet daarentegen het nauwkeurigst bij leerlingen met een gemiddelde vaardigheid.

3.3.2 Betrouwbaarheidsmatrices bij de toetsen

De tabellen 3.1 en 3.2 bevatten, respectievelijk voor de leerlingen uit drie havo en drie vwo, de betrouwbaarheidsmatrices voor de toets Nederlands. In de eerste kolom van iedere tabel wordt tussen haakjes aangegeven welke scores op de toets bij elk deciel behoren. In de tabellen valt per deciel te zien hoe groot de kans is dat leerlingen een ware score hebben die in hetzelfde deciel valt. Voornoemde kansen zijn in de tabel vet gedrukt. Zo maakt tabel 3.1 duidelijk dat leerlingen uit drie havo met een score die valt in deciel 1 een kans hebben van 0,573 dat hun ware score op de toets Nederlands ook binnen dit deciel valt. Verder wordt per deciel aangegeven hoe groot de kans is dat leerlingen een ware score hebben die binnen een ander deciel valt. Zo laat tabel 3.1 zien dat leerlingen uit drie havo met een score die valt in het eerste deciel een kans hebben van 0,235 dat hun ware score op de toets Nederlands in het tweede deciel valt.

Tabel 3.1

Betrouwbaarheidsmatrix voor de toets Nederlands voor leerlingen uit drie havo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0- 36)	**0,573**	0,235	0,108	0,049	0,022	0,009	0,003	0,001	0,000	0,000
2	(37- 42)	0,239	**0,276**	0,204	0,133	0,079	0,042	0,019	0,007	0,002	0,000
3	(43- 48)	0,109	0,207	**0,212**	0,178	0,131	0,086	0,049	0,022	0,007	0,001
4	(49- 52)	0,049	0,135	0,179	**0,185**	0,165	0,129	0,088	0,049	0,019	0,003
5	(53- 56)	0,021	0,079	0,133	0,166	**0,175**	0,162	0,130	0,087	0,041	0,008
6	(57- 60)	0,008	0,041	0,087	0,130	0,161	**0,175**	0,166	0,133	0,079	0,020
7	(61- 65)	0,003	0,019	0,049	0,088	0,129	0,165	**0,185**	0,179	0,135	0,048
8	(66- 70)	0,001	0,007	0,022	0,049	0,086	0,131	0,178	**0,212**	0,207	0,108
9	(71- 77)	0,000	0,002	0,007	0,019	0,041	0,078	0,133	0,205	**0,276**	0,239
10	(78-115)	0,000	0,000	0,001	0,003	0,009	0,021	0,049	0,108	0,235	**0,574**

Een onderlinge vergelijking van beide tabellen maakt duidelijk dat de kansen op een al dan niet correcte classificatie in de verschillende decielen voor de beide groepen leerlingen niet veel afwijken. De kansen op een correcte classificatie bij de vwo-toets Nederlands zijn ten opzichte van de havo-toets iets groter in de decielen 1 tot en met 3, gelijk in deciel 4 en iets kleiner in de decielen 5 tot en met 10. Ook laten de tabellen zien dat bij beide populaties de kansen op een correcte classificatie in de meer extreme decielen groter zijn dan in de middelste.

Tabel 3.2

Betrouwbaarheidsmatrix voor de toets Nederlands voor leerlingen uit drie vwo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0- 45)	**0,587**	0,233	0,104	0,046	0,020	0,008	0,003	0,001	0,000	0,000
2	(46- 52)	0,235	**0,280**	0,206	0,133	0,078	0,041	0,019	0,007	0,002	0,000
3	(53- 57)	0,103	0,207	**0,213**	0,179	0,132	0,086	0,049	0,023	0,007	0,001
4	(58- 61)	0,045	0,133	0,178	**0,185**	0,165	0,130	0,089	0,051	0,020	0,003
5	(62- 65)	0,019	0,078	0,131	0,164	**0,173**	0,161	0,130	0,089	0,045	0,009
6	(66- 69)	0,008	0,041	0,086	0,129	0,159	**0,172**	0,164	0,134	0,083	0,024
7	(70- 74)	0,003	0,019	0,049	0,088	0,128	0,162	**0,182**	0,177	0,138	0,054
8	(75- 78)	0,001	0,007	0,023	0,051	0,087	0,131	0,174	**0,206**	0,204	0,116
9	(79- 85)	0,000	0,002	0,008	0,021	0,045	0,082	0,134	0,200	**0,266**	0,242
10	(86-115)	0,000	0,000	0,001	0,004	0,011	0,026	0,056	0,115	0,236	**0,552**

In de tabellen 3.3 en 3.4 staan de betrouwbaarheidsmatrices voor de toetsen Engels. Ook deze twee tabellen laten zien dat de kansen op een correcte classificatie in de meer extreme decielen hoger zijn dan in de middelste decielen. De kansen op een correcte classificatie zijn bij de vwo-toets voor Engels in alle decielen iets groter dan bij de havo-toets. Een vergelijking van de tabellen 3.3 en 3.4 met de twee voorafgaande maakt duidelijk dat de matrices in deze laatste twee tabellen meer door hun diagonalen worden gedomineerd dan de matrices in de twee voorafgaande tabellen. Leerlingen

hebben bij de beide toetsen voor Engels een wat hogere kans op een correcte classificatie dan bij de toets voor Nederlands. De kans is bijvoorbeeld 0,709 dat leerlingen op de havo-toets voor Engels een ware score hebben die in het eerste deciel valt, gegeven dat hun geobserveerde score eveneens in het eerste deciel valt. De corresponderende kans bij de havo-toets voor Nederlands is 0,573.

Tabel 3.3

Betrouwbaarheidsmatrix voor de toets Engels voor leerlingen uit drie havo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0-28)	**0,709**	0,221	0,055	0,012	0,002	0,000	0,000	0,000	0,000	0,000
2	(29-35)	0,223	**0,386**	0,239	0,104	0,036	0,010	0,002	0,000	0,000	0,000
3	(36-40)	0,054	0,240	**0,295**	0,220	0,121	0,051	0,016	0,003	0,000	0,000
4	(41-44)	0,012	0,104	0,219	**0,254**	0,207	0,127	0,057	0,018	0,003	0,000
5	(45-46)	0,002	0,036	0,121	0,205	**0,236**	0,201	0,128	0,056	0,014	0,001
6	(48-51)	0,000	0,010	0,051	0,124	0,198	**0,232**	0,204	0,129	0,048	0,005
7	(52-54)	0,000	0,002	0,017	0,058	0,126	0,200	**0,241**	0,215	0,120	0,022
8	(56-58)	0,000	0,000	0,004	0,019	0,057	0,124	0,209	**0,270**	0,238	0,079
9	(59-63)	0,000	0,000	0,001	0,004	0,015	0,048	0,117	0,229	**0,342**	0,244
10	(64-81)	0,000	0,000	0,000	0,000	0,001	0,007	0,025	0,081	0,236	**0,650**

Tabel 3.4

Betrouwbaarheidsmatrix voor de toets Engels voor leerlingen uit drie vwo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0-31)	**0,728**	0,218	0,046	0,008	0,001	0,000	0,000	0,000	0,000	0,000
2	(32-39)	0,221	**0,418**	0,243	0,089	0,023	0,004	0,001	0,000	0,000	0,000
3	(40-45)	0,044	0,248	**0,327**	0,231	0,108	0,034	0,007	0,001	0,000	0,000
4	(46-51)	0,007	0,090	0,233	**0,287**	0,221	0,115	0,039	0,008	0,001	0,000
5	(52-56)	0,001	0,023	0,108	0,222	**0,270**	0,218	0,116	0,037	0,005	0,000
6	(57-61)	0,000	0,004	0,034	0,114	0,216	**0,267**	0,221	0,113	0,028	0,001
7	(62-66)	0,000	0,001	0,007	0,039	0,115	0,217	**0,279**	0,232	0,100	0,010
8	(67-71)	0,000	0,000	0,001	0,008	0,038	0,111	0,225	**0,311**	0,249	0,057
9	(72-78)	0,000	0,000	0,000	0,001	0,006	0,029	0,098	0,239	**0,388**	0,238
10	(79-97)	0,000	0,000	0,000	0,000	0,000	0,002	0,013	0,061	0,230	**0,694**

Tabel 3.5

Betrouwbaarheidsmatrix voor de toets wiskunde voor leerlingen uit drie havo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0-11)	**0,717**	0,229	0,047	0,006	0,001	0,000	0,000	0,000	0,000	0,000
2	(12-16)	0,237	**0,428**	0,244	0,076	0,014	0,001	0,000	0,000	0,000	0,000
3	(17-20)	0,043	0,257	**0,360**	0,236	0,085	0,017	0,002	0,000	0,000	0,000
4	(21-24)	0,004	0,075	0,248	**0,335**	0,234	0,087	0,016	0,001	0,000	0,000
5	(25-28)	0,000	0,011	0,086	0,242	**0,329**	0,235	0,083	0,012	0,000	0,000
6	(29-32)	0,000	0,001	0,014	0,087	0,239	**0,336**	0,240	0,076	0,007	0,000
7	(33-36)	0,000	0,000	0,001	0,015	0,084	0,239	**0,354**	0,246	0,059	0,002
8	(37-41)	0,000	0,000	0,000	0,001	0,012	0,075	0,241	**0,391**	0,251	0,028
9	(42-46)	0,000	0,000	0,000	0,000	0,001	0,008	0,060	0,244	**0,473**	0,214
10	(47-57)	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,030	0,211	**0,757**

Tabel 3.6

Betrouwbaarheidsmatrix voor de toets wiskunde voor leerlingen uit drie vwo

		Deciel waarin de ware score valt									
		1	2	3	4	5	6	7	8	9	10
Deciel											
1	(0-11)	**0,794**	0,190	0,016	0,001	0,000	0,000	0,000	0,000	0,000	0,000
2	(12-16)	0,194	**0,530**	0,234	0,040	0,003	0,000	0,000	0,000	0,000	0,000
3	(17-20)	0,014	0,239	**0,437**	0,243	0,066	0,007	0,000	0,000	0,000	0,000
4	(21-24)	0,000	0,039	0,245	**0,385**	0,243	0,075	0,012	0,001	0,000	0,000
5	(25-28)	0,000	0,003	0,061	0,242	**0,354**	0,242	0,084	0,014	0,001	0,000
6	(29-32)	0,000	0,000	0,008	0,075	0,236	**0,336**	0,244	0,088	0,013	0,000
7	(33-36)	0,000	0,000	0,001	0,013	0,083	0,232	**0,331**	0,249	0,085	0,007
8	(37-41)	0,000	0,000	0,000	0,001	0,017	0,088	0,235	**0,346**	0,258	0,056
9	(42-46)	0,000	0,000	0,000	0,000	0,002	0,017	0,084	0,244	**0,404**	0,249
10	(47-57)	0,000	0,000	0,000	0,000	0,000	0,001	0,010	0,059	0,241	**0,690**

De tabellen 3.5 en 3.6 bevatten de betrouwbaarheidsmatrices voor de toetsen voor wiskunde. Ook uit deze tabellen blijkt weer dat de kansen op een correcte classificatie in de meer extreme decielen hoger zijn dan in de middelste. Een onderlinge vergelijking van beide tabellen leert dat de kans op een correcte classificatie bij de vwo-toets voor wiskunde iets groter is in de decielen 1 tot en met 6 en bij de havo-toets iets groter in de decielen 7 tot en met 10. Vergelijking met de twee andere vakken laat zien dat kans op een correcte classificatie bij de twee toetsen voor wiskunde groter is dan bij de toets Nederlands. Verder is de kans op een correcte classificatie bij de havo-toets voor wiskunde iets groter dan bij de havo-toets voor Engels. En bij de vwo-toets voor wiskunde blijkt de kans op een correcte classificatie in alle decielen met uitzondering van deciel 10 groter te zijn dan bij de vwo-toets voor Engels.

3.3.3 De geschatte betrouwbaarheid van de toetsen

Tabel 3.7 geeft voor elk van de zes samengestelde toetsen een aantal met behulp van OPTAL berekende kengetallen. De eerste kolom geeft de naam van de toets; 'Ned.' staat voor Nederlands, 'Eng.' voor Engels en 'Wis.' voor wiskunde. De tweede kolom geeft de maximumscore bij iedere toets. Bij de toetsen voor Nederlands en Engels betreft het hier de gewogen maximumscore ($\Sigma a_i J_i$) en bij wiskunde de ongewogen maximumscore (ΣJ_i). De derde kolom ($\mathscr{E}(X)$) geeft de gemiddelde score per toets. De vierde kolom bevat de geschatte standaardafwijking ($\hat{\sigma}_X$) van de scores op de toets. De geschatte standaardmeetfout ($\hat{\sigma}_E$) van iedere toets staat in de vijfde kolom, terwijl de laatste kolom gereserveerd is voor de schatting van de betrouwbaarheid ($\hat{\rho}^2_{XT}$). Deze kolom bevat tussen haakjes het 95-procents betrouwbaarheidsinterval rond de schatting van de betrouwbaarheid. De betrouwbaarheid van alle toetsen is geschat met behulp van uitdrukking (3.9). Bij de toetsen wiskunde zijn de waarden van de discriminatie-indices echter gelijk aan één gesteld.

De totaalscore op de toetsen voor Nederlands en Engels komt tot stand door de scores op de items te wegen met de bijbehorende discriminatie-indices. - Omdat bij deze toetsen zonder uitzondering sprake is van dichotome items blijft het vaststellen van de totaalscore voor de gebruikers echter relatief

simpel. Gebruikers kunnen totaalscores eenvoudigweg bepalen door de discriminatie-indices, die in het correctievoorschrift als score bij een correct antwoord vermeld staan, te sommeren. De items in de toetsen wiskunde zijn daarentegen veelal polytoom. Daarom is bij de toetsen wiskunde gekozen voor een ongewogen totaalscore. Omdat de scores op de items niet gewogen hoeven te worden met de discriminatie-indices, kunnen gebruikers ook hier volstaan met het sommeren van de op de items behaalde scores om een totaalscore te bepalen. Bij de toetsen wiskunde kunnen de gewogen scores zonder bezwaar door de ongewogen scores vervangen worden. De correlatie tussen de gewogen en de ongewogen somscores voor de toetsen wiskunde bedraagt namelijk 0,980 voor de toets voor havo en 0,978 voor de toets voor vwo.

Tabel 3.7
Kengetallen voor de plaatsingstoetsen

Toets	*Max.*	$\mathscr{E}(X)$	$\hat{\sigma}_X$	$\hat{\sigma}_E$	$\hat{\rho}^2_{XT}$
Ned. havo	115	55,844	15,527	8,963	0,667 (0,647-0,685)
Ned. vwo	115	65,443	15,186	8,840	0,661 (0,641-0,679)
Eng. havo	81	46,504	13,266	5,696	0,816 (0,807-0,824)
Eng. vwo	97	55,414	17,396	6,604	0,856 (0,849-0,862)
Wis. havo	57	27,068	13,183	4,158	0,901 (0,896-0,905)
Wis. vwo	60	37,958	13,897	4,093	0,913 (0,909-0,917)

3.3.4 De drie soorten informatie over meetnauwkeurigheid vergeleken

De figuren in paragraaf 3.3.1. geven een totaalindruk van de meetnauwkeurigheid van de toetsen op de vaardigheidsschalen waar zij betrekking op hebben. Bovendien laten de figuren zien hoe de meetnauwkeurigheid van de toetsen binnen de verschillende populaties verloopt. De betrouwbaarheidsmatrices geven een concreet en gedetailleerd overzicht van de meetnauwkeurigheid van de toetsen op de normschalen. De geschatte betrouwbaarheid van de toetsen geeft in één getal weer hoe het gesteld is met de globale meetnauwkeurigheid van de toetsen binnen de verschillende populaties.

Uit de betrouwbaarheidsmatrices blijkt dat bij alle toetsen de kansen op een correcte classificatie groter zijn in de meer extreme decielen dan in de middelste decielen. De figuren met de toetsinformatiefuncties geven echter aan dat de meetnauwkeurigheid van de toetsen juist het kleinst is voor de meer extreme delen van de vaardigheidsschalen. De oorzaak van deze schijnbare tegenstelling is gelegen in het feit dat de kans op een correcte classificatie niet alleen afhankelijk is van de meetnauwkeurigheid van de toets op specifieke punten van de vaardigheidsschaal, maar ook van de breedte van het vaardigheidsinterval dat bij ieder deciel hoort. Indien de grenzen van de decielen in de figuren in paragraaf 3.3.1. zouden worden afgebeeld, zou blijken dat de vaardigheidsintervallen bij de meer extreme decielen aanmerkelijk breder zijn dan bij de middelste. Dit laatste komt ook tot uitdrukking in het ruwe-score bereik van de decielen. De vaardigheidsintervallen bij de meer extreme decielen zijn klaarblijkelijk zelfs zoveel breder dat de geringere meetnauwkeurigheid van de toetsen binnen deze intervallen volstrekt niet tot uitdrukking kan komen. Hoewel de exacte positie van de leerlingen op de vaardigheidsschalen in de meer extreme decielen minder goed is te bepalen, is de kans groter dat hun vaardigheid wel valt in het bij het betreffende deciel behorende vaardigheidsinterval, omdat dit interval relatief breder is.

Een vergelijking van de betrouwbaarheidsmatrices bij de toetsen met hun geschatte betrouwbaarheid toont aan dat deze twee vormen van informatie elkaar aanvullen en bevestigen. De conclusie dat er bij de toets Nederlands slechts kleine verschillen zijn in meetnauwkeurigheid tussen de havo-populatie en de vwo-populatie wordt bevestigd door het kleine verschil in geschatte betrouwbaarheid van de toets voor deze twee populaties. De betrouwbaarheidsmatrices bevatten echter wel rijkere informatie. Ze laten zien dat bij de toets Nederlands in de lagere decielen de kans op een classificatie in het juiste deciel bij de havo-populatie iets lager is dan bij de vwo-populatie, maar dat in de hogere decielen het omgekeerde het geval is. Verder zien we het feit dat de kansen op een correcte classificatie hoger zijn bij de vwo-toets Engels dan bij de havo-toets Engels weerspiegeld in de geschatte betrouwbaarheid van deze toetsen. Bij de toetsen voor wiskunde geldt, net als bij de toets Nederlands, dat de betrouwbaarheidsmatrices informatie leveren die in de geschatte betrouwbaarheid van de toetsen niet tot uiting kan komen. De geschatte betrouwbaarheid van de vwo-toets

wiskunde is hoger dan die van de havo-toets, maar uit de betrouwbaarheidsmatrices blijkt dat in de lagere decielen de kans op een correcte classificatie bij de havo-toets lager is dan bij de vwo-toets en dat in de hogere decielen het omgekeerde het geval is.

Ook de constatering dat de kansen op een correcte classificatie bij de toetsen Engels hoger liggen dan bij de toets Nederlands strookt met de schattingen van de betrouwbaarheid van de betreffende toetsen. Het feit dat de kansen op een correcte classificatie hoger liggen bij de toetsen wiskunde dan bij de toets Nederlands komt eveneens tot uitdrukking in de geschatte betrouwbaarheid van deze toetsen. Ten slotte leert een vergelijking van de geschatte betrouwbaarheden en de betrouwbaarheidsmatrices voor de toetsen Engels enerzijds en de toetsen wiskunde anderzijds dat deze gegevens elkaar eveneens bevestigen. De kanttekening die hier te maken valt is dat bij de vwo-toets wiskunde de kans op een correcte classificatie in deciel 10 groter is dan de corresponderende kans bij de vwo-toets Engels, terwijl de geschatte betrouwbaarheid van de vwo-toets wiskunde groter is dan die van de vwo-toets Engels.

Het voorafgaande maakt duidelijk dat toetsinformatiefuncties, betrouwbaarheidsmatrices en betrouwbaarheidscoëfficiënten elk op zich een onvolledig beeld geven van de meetnauwkeurigheid van een toets. De figuren met toetsinformatiefuncties en kansdichtheden geven een goede totaalindruk, maar geven geen concreet beeld van de specifieke meetnauwkeurigheid van de toets op de normschalen. Dit laatste geldt ook voor de geschatte betrouwbaarheid en de andere klassieke kengetallen in tabel 3.7. De betrouwbaarheidsmatrices maken op hun beurt wel duidelijk hoe het gesteld is met de nauwkeurigheid waarmee leerlingen in decielen geclassificeerd zijn, maar geven weer geen informatie over de meetnauwkeurigheid van de toetsen op de vaardigheidsschalen waar ze betrekking op hebben. Verder geldt voor de gepresenteerde figuren en de betrouwbaarheidsmatrices dat ze de meetnauwkeurigheid van de toetsen niet in een eenvoudig communiceerbare vorm uitdrukken. Dat gebeurt bij de betrouwbaarheidscoëfficiënt wel, maar gebruik van deze maat alleen is te weinig informatief. Alleen door de verschillende vormen van informatie tezamen te presenteren ontstaat een afdoende beschrijving van de meetnauwkeurigheid van de toetsen.

4 De validiteit van de plaatsingstoetsen

Validiteit is de tweede fundamentele eis waar toetsen aan moeten voldoen. De opvattingen over validiteit zijn in de loop der jaren sterk gewijzigd. Daarom volgt hieronder, voorafgaand aan de behandeling van de resultaten van het valideringsonderzoek, een korte bespreking van de ontwikkeling van het begrip en de huidige standpunten. Deze bespreking bevat een kritische beschouwing van de vraag of een valideringsonderzoek zich ook behoort te richten op de mogelijke gevolgen van het gebruik van een test. Het zal nodig blijken een onderscheid aan te brengen tussen de begrippen test en toets om een genuanceerd antwoord op deze vraag te kunnen geven. In het onderstaande wordt de term test echter vooralsnog in generieke zin gebruikt.

4.1 Het begrip validiteit

Oorspronkelijk werd validiteit gedefinieerd als de mate waarin een test meet wat men ermee beoogt te meten: '*The validity of a test is the extent to which ... it measures what it purports to measure*' (Garret, 1937, p.324, geciteerd in Angoff, 1988). Verder was het tot omstreeks 1950 gebruikelijk de correlatie tussen de score op een test en een criteriumvariabele te rapporteren als indicatie voor de validiteit van deze test (Angoff, 1988; Shepard, 1993).

Sinds 1954 zijn er verschillende publicaties verschenen die de opvattingen over validiteit in een bepaalde periode codificeren (American Psychological Association, 1954; American Psychological Association, 1966; American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1974; American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985). Deze publicaties geven een duidelijk beeld van de veranderingen die zich de laatste decennia hebben voorgedaan in deze opvattingen.

Een meer theoretisch georiënteerde zienswijze, waarin verschillende soorten validiteit een plaats innamen, verdrong langzamerhand de oorspronkelijke, strikt operationele opvatting van het begrip validiteit. Deze verschillende soorten validiteit hoorden aanvankelijk ook bij verschillende typen tests. Zo moesten tests die zich richtten op het bepalen van de prestaties van een persoon op een welomschreven domein van kennis of vaardigheden, zoals bijvoorbeeld een bepaald leerstofgebied, inhoudsvalide zijn. Waren tests bedoeld om uitspraken te doen over psychologische kenmerken van personen, dan dienden zij constructvalide te zijn. Moesten tests toekomstige prestaties of kenmerken van personen voorspellen, dan behoorden zij predictief valide te zijn. En hadden tests betrekking op actuele prestaties of kenmerken van personen dan dienden zij gelijktijdige ('concurrent') validiteit te bezitten. Al spoedig raakte echter de term criterium gerelateerde validiteit in zwang om te verwijzen naar de twee laatstgenoemde soorten validiteit.

Hoewel zich vanaf de jaren vijftig een ware wildgroei van begrippen heeft voltrokken waarin de term validiteit voorkwam (zie bijvoorbeeld van Berkel, 1984), was tot halverwege de jaren tachtig de als canon geaccepteerde indeling die naar inhouds-, construct- en criterium gerelateerde validiteit. De inhoudsvaliditeit van een test had betrekking op de mate waarin de items in de test een welomschreven en afgebakend universum representeren van mogelijk in de test op te nemen items. De constructvaliditeit van een test had betrekking op de mate waarin testscores zijn toe te schrijven aan verklarende concepten en constructen die deel uitmaken van het theoretisch kader dat aan de ontwikkeling van de test ten grondslag ligt. De criterium gerelateerde validiteit van een test ten slotte, was de correlatie tussen de scores op de test en de scores op een variabele die een operationalisatie is van een actueel of toekomstig gegeven waar de test betrekking op moet hebben.

In de loop der jaren groeide het besef dat de gegevens die men verzamelde om de inhouds-, construct- of criterium gerelateerde validiteit van een test te bepalen alle bijdragen aan een juiste interpretatie en een correct gebruik van de scores op deze test. Tegenwoordig is de klassieke driedeling dan ook minder populair. Veel theoretici, onder aanvoering van Messick (1988, 1989, 1990, 1995), zijn van mening dat de drie genoemde soorten validiteit verschillende aspecten zijn van een en hetzelfde overkoepelende begrip. Volgens de huidige opvattingen dient een valideringsonderzoek zich te

richten op '*... several types of evidence, which span all three of the traditional categories.*' (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985, p. 9).

Het huidige standpunt is dat: '*... validity is a unified though faceted concept and ... validation is scientific inquiry.*' (Messick, 1989, p. 14). Valideren zou een continu proces moeten zijn waarin men op gezette tijden op grond van empirische gegevens en theoretische overwegingen bepaalt in hoeverre het geoorloofd is om de scores op een test op een bepaalde manier te interpreteren en te gebruiken. Strikt gezien is het dan ook niet de test die gevalideerd moet worden als wel de interpretatie van de testscores en het gebruik van de test. Desondanks is het nog steeds gebruikelijk te spreken van een 'valide test', hoewel duidelijk is dat validiteit geen intrinsieke eigenschap van een test is.

Validiteit en de consequenties van testgebruik

In een valideringsonderzoek moet volgens Messick (1990) niet alleen aandacht zijn voor de betekenis van testscores, maar ook voor de consequenties van het gebruik van een test: '*... test validation is empirical evaluation of the meaning and consequences of measurement.*' (Messick, 1990, p. 2). Cronbach (1988), Shepard (1993, 1997) en Linn (1997) stellen zich eveneens op dit standpunt. Het is in hun ogen niet voldoende om uitsluitend te onderzoeken in hoeverre het gewettigd is tot bepaalde uitspraken over eigenschappen van personen te komen. Ook de beslissingen die uit deze uitspraken voortvloeien, dienen onderwerp van onderzoek te zijn. Een valideringsonderzoek zal dan ook aandacht moeten besteden aan de gevolgen van het gebruik van een test: '*... since validity depends on the uses to which results are put, it needs to include an evaluation of the consequences of those uses.*' (Linn, 1997, p. 14).

Met deze laatste stelling zijn Popham (1997) en Mehrens (1997) het niet eens. Zij vinden dat valideringsonderzoek zich zou moeten concentreren op de interpretatie van testscores. Hoewel het in kaart brengen van de gevolgen van

het gebruik van een test belangrijk is, zou dit geen deel mogen uitmaken van het valideringsonderzoek. Zij voeren hiervoor twee argumenten aan.

Hun eerste argument is ingegeven door de vrees dat een valideringsonderzoek dat ook betrekking heeft op de consequenties van het gebruik van een test in sommige gevallen een onterecht beeld zal geven van de kwaliteit van deze test als meetinstrument. Zij vrezen dat een test het stigma 'niet valide' zou kunnen krijgen, terwijl met deze test wel degelijk uitspraken mogelijk zijn over de mate waarin de geteste personen over de eigenschap beschikken waar de test betrekking op heeft: '*The problem with tying consequences of a specific use to the notion of validity is that a test score therefore may be considered* ***invalid*** *for making an inference about a construct when, in fact, the inference would be accurate.*' (Mehrens, 1997, p. 17).

Tegen het argument dat een test onterecht het stigma 'niet valide' zou kunnen krijgen, zijn twee bedenkingen in te brengen. In de eerste plaats is het goed mogelijk de kwaliteit van een test als operationalisatie van een bepaald construct en de gevolgen van het gebruik van een test als twee aparte zaken te beschouwen. Wanneer een valideringsonderzoek heeft aangetoond dat gebruik van een test bij het nemen van een bepaalde beslissing ongewenste consequenties heeft, moet de conclusie luiden dat de test niet valide is voor het nemen van de betreffende beslissing. Deze conclusie doet echter op geen enkele wijze afbreuk aan de kwaliteit van de test als meetinstrument. Het feit dat een test niet valide is voor het nemen van een zekere beslissing impliceert niet dat de betreffende test ook onbruikbaar zou zijn voor het nemen van andere beslissingen.

In de tweede plaats laten beschrijvingen van de ontwikkeling en de functie van de eerste psychologische tests, zoals die te vinden zijn in Du Bois (1970), Gould (1981) en Van der Linden (1983), zien dat tests al vanaf het prille begin bedoeld waren om met behulp van de testresultaten beslissingen over personen te nemen. Het feit dat tests welbeschouwd geen meetinstrumenten, maar beslisinstrumenten zijn, vindt zowel in de klassieke als in de huidige opvattingen over validiteit een duidelijke weerklank. Hoewel de opvattingen over validiteit in de loop der jaren sterk gewijzigd zijn, heeft validiteit vrijwel vanaf de introductie van het begrip niet alleen betrekking gehad op de vraag of een test meet wat deze beoogt te meten, maar ook op de vraag of een test

aan zijn doel beantwoordt. Een valideringsonderzoek dat zich beperkt tot de vraag wat een test meet, is derhalve onvolledig.

Beantwoording van de vraag of een test meet wat deze beoogt te meten, kan nooit het einddoel van een valideringsonderzoek zijn (Linn, 1997; Shepard, 1997). Een positief antwoord op deze vraag is een noodzakelijke, maar geen voldoende, voorwaarde voor een test om een rol te kunnen spelen bij het nemen van een bepaalde beslissing. Indien een valideringsonderzoek zich zou beperken tot de vraag of een test meet wat deze beoogt te meten, is het niet mogelijk een uitspraak te doen over de geschiktheid van deze test voor het nemen van een bepaalde beslissing. Het afnemen van een test met meten als enig doel kan zich voordoen in wetenschappelijk onderzoek waarin een testscore de rol speelt van afhankelijke of onafhankelijke variabele. Maar ook dan is het uiteindelijke doel van het afnemen van de test het nemen van een beslissing, namelijk het al of niet verwerpen van een onderzoekshypothese.

Het tweede argument van Popham en Mehrens komt voort uit hun opvatting dat gebruikers vaak niet meer dan een vage notie van het begrip validiteit hebben en testscores veelal zien als absoluut en onfeilbaar. Popham en Mehrens stellen dat het begrip validiteit begrijpelijker zou zijn voor testgebruikers, indien het uitsluitend betrekking heeft op de kwaliteit van een test als meetinstrument. Het presenteren van validiteit op deze manier zou een middel zijn om ervoor te zorgen dat gebruikers: '*... focus their attention on the adequacy of the* ***evidence*** *that was used to support a score-based inference about a student's status.*' (Popham, 1997, p. 11).

Voor een adequate reactie op het argument dat het begrip validiteit begrijpelijker zou zijn voor testgebruikers, indien het uitsluitend betrekking zou hebben op de kwaliteit van een test als meetinstrument, dienen we een onderscheid te maken tussen de begrippen test en toets. In het eerste hoofdstuk van dit proefschrift is het begrip test gedefinieerd als een gestandaardiseerde en systematische meetprocedure die het mogelijk maakt kwantitatieve uitspraken te doen over eigenschappen van personen, met het doel om beslissingen te nemen. Deze definitie is verder gespeci-ficeerd door aan te geven dat een toets een instrument is dat door onderwijs en studie verworven kennis, inzicht of vaardigheden meet, terwijl het begrip test in de

niet-generieke zin staat voor een instrument dat eigenschappen meet die niet door intentioneel onderwijs en studie verworven zijn.

Door een onderscheid aan te brengen tussen de begrippen test en toets valt allereerst duidelijk te maken dat de stelling maar ten dele houdbaar is dat gebruikers van dergelijke instrumenten slechts een vage notie hebben van het begrip validiteit. Inderdaad geldt dat *toets*gebruikers - docenten - in de regel niet goed bekend zullen zijn met het begrip validiteit. Voor *test*gebruikers geldt dat laatste in veel mindere mate. Het gebruik van tests vindt immers in de regel plaats binnen professionele kaders. Van *test*gebruikers, zoals beroepskeuze-adviseurs en testpsychologen, mag men verwachten dat zij weten wat het begrip validiteit inhoudt.

Verder biedt het aanbrengen van een onderscheid tussen de begrippen test en toets de gelegenheid de stelling te nuanceren dat het nodig is de gevolgen van het gebruik van een test in kaart te brengen. Het valideringsonderzoek van *toetsen* hoort zonder meer aandacht te besteden aan de mogelijke gevolgen van het toetsgebruik. Indien docenten niet exact weten wat het gebruiksdoel is van een toets en in hoeverre deze toets geschikt is voor het gebruiksdoel, kan dat leiden tot onoordeelkundig en verkeerd gebruik, met alle gevolgen vandien. Omdat toetsen veelal ontwikkeld worden met een specifiek gebruiksdoel voor ogen, is het voor toetsontwikkelaars ook zeer wel mogelijk het beoogde gebruik van de toets te rechtvaardigen. Zij zijn hiertoe zelfs verplicht, omdat het niet reëel is te verwachten dat de gebruikers van de toets - scholen en docenten - deze taak op zich kunnen nemen.

Bij een test kan het veel moeilijker zijn in het valideringsonderzoek aandacht te besteden aan de mogelijke gevolgen van het gebruik. Dit is het geval wanneer de test een theoretisch psychologisch construct meet, zoals bijvoorbeeld intelligentie. Een test van dit type kan in principe een rol spelen bij iedere situatie waarin een beslisser of beslissende instantie de informatie die de test levert van belang acht. Het is voor de ontwikkelaars van dergelijke tests niet goed mogelijk om te anticiperen op alle mogelijke toepassingen. Dat ontslaat ze volgens Shepard (1997) echter niet van de verplichting om tenminste voor één specifiek gebruiksdoel te onderzoeken wat de gevolgen van het gebruik van de test zijn. Het is dan vervolgens aan de professionele gebruikers van de test om aan te tonen dat de test geschikt

is voor de specifieke beslissing die zij voor ogen hebben: '*When users appropriate tests for purposes not sanctioned and studied by the test developers, users become responsible for the needed validity investigation.*' (Shepard, 1997, p. 8). Terzijde zij opgemerkt dat toetsgebruikers uiteraard zèlf de verantwoording dienen te dragen voor het uitvoeren van een valideringsonderzoek, indien zij een toets gebruiken voor een ander doel dan waarvoor deze ontwikkeld is.

Richtlijnen voor validering

De opvattingen over validiteit in dit proefschrift sluiten aan op de meest recente versie van de 'Standards for Educational and Psychological Testing' (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985) en stroken met de recente opvat-tingen van de meeste theoretici, zoals Messick (1988, 1989, 1990, 1995), Cron-bach (1988), Shepard (1993, 1997) en Linn (1997). Toetsen worden in dit proefschrift beschouwd als beslisinstrumenten. Verder wordt het aantonen van de kwali-teit van een toets als meetinstrument gezien als een noodzakelijke, maar niet voldoende, voorwaarde voor het doen van de uitspraak dat de toets valide is voor het nemen van een bepaalde beslissing.

Het aantal eisen ten aanzien van validiteit, zoals die met name door Messick (1988, 1989, 1990, 1995) verwoord zijn, is uitgebreid. De relevantie van de verschillende eisen varieert met het specifieke doel van een test. Het werk van Messick biedt echter nauwelijks criteria met behulp waarvan de mate van relevantie van al deze eisen, gegeven een specifiek gebruiksdoel, valt te bepalen (Cronbach, 1988; Kane, 1992; Shepard, 1993). De afwezigheid van concrete richtlijnen voor het opzetten van een valideringsonderzoek maakt het moeilijk om onderzoeksvragen te prioriteren. Een bijkomend probleem is dat er veel bronnen van informatie zijn waar onderzoekers uit kunnen putten om na te gaan of een bepaalde interpretatie van een score op een test of toets juist is, of om een bepaalde vorm van gebruik te rechtvaardigen. In de meest recente versie van de 'Standards for Educational and Psychological Testing' staat dat '*Resources should be invested in obtaining the combination of evidence that optimally reflects the value of a test for an intended purpose.*'

(American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985. p. 9). De twee voornoemde problemen tezamen zorgen er echter voor dat niet expliciet vast te stellen is of in een valideringsonderzoek sprake is van een optimale of zelfs goede 'combination of evidence'.

Om te kunnen bepalen waar het valideringsonderzoek zich met name op moet richten, stelt Kane (1992) daarom voor de opzet van dit onderzoek te baseren op een zo concreet mogelijke omschrijving van het doel van een test. Vervolgens moet dan zo expliciet mogelijk aangegeven worden *waarom* de test gebruikt kan worden als bedoeld. Soortgelijke uitgangspunten zijn ook bij Cronbach (1988) en Shepard (1993) te vinden. Argumenten die het gebruik als bedoeld rechtvaardigen en eventuele argumenten tegen het gebruik als bedoeld kunnen vervolgens richting geven aan de vorm en opzet van het valideringsonderzoek. Validering komt dan neer op het aandragen van empirische gegevens die, eventueel ondersteund door logische beweringen of gevestigde theorieën, de argumenten plausibel maken en eventuele tegenargumenten ontkrachten.

Ontwikkelaars van tests dienen, met andere woorden, in een steekhoudend betoog duidelijk te maken dat hun instrumenten geschikt zijn voor het gebruiksdoel waar ze voor ontwikkeld zijn. Dit betoog moet aan drie kenmerken voldoen (Kane, 1992). Allereerst dienen de ontwikkelaars het gebruiksdoel van het instrument helder te verwoorden en duidelijk geformuleerde argumenten te hanteren om de stelling te onderbouwen dat het instrument voor dit doel te gebruiken is. In de tweede plaats moet hun betoog coherent zijn, wat betekent dat de conclusies redelijkerwijze uit de aannamen moeten volgen. En ten derde dienen de aannamen van de ontwikkelaars plausibel te zijn, of moet het mogelijk zijn hun geldigheid aan te tonen.

De rest van dit hoofdstuk is gereserveerd voor een betoog dat voldoet aan de voornoemde kenmerken en aannemelijk moet maken dat scholen de ontwikkelde plaatsingstoetsen kunnen gebruiken als bedoeld. In dit betoog komen de drie aspecten van validiteit aan de orde die in de klassieke opvattingen bekend stonden als inhouds-, construct- en criterium gerelateerde validiteit. Het vierde aspect - de consequenties van het gebruik van de

toetsen - komt kort aan bod in hoofdstuk zes, wanneer de vraag aan de orde komt hoe scholen de plaatsingstoetsen kunnen gebruiken bij het nemen van doorstroombeslissingen.

4.2 Het bepalen van de validiteit van de toetsen

In hoofdstuk twee is al ingegaan op het doel en de functie van de toetsen. Het doel van de toetsen en de checklist voor studievaardigheid is daar omschreven als het bieden van *ondersteuning* bij het nemen van beslissingen over de doorstroming van leerlingen na het derde leerjaar van de havo en het derde leerjaar van het vwo. Een nadere precisering van de vorm van ondersteuning die de toetsen en de checklist bieden, komt aan de orde in hoofdstuk zes.

In hoofdstuk twee zijn eisen geformuleerd waar de toetsen aan zouden moeten voldoen en zijn kenmerken beschreven waar de toetsen over zouden moeten beschikken om bruikbaar te zijn voor het ondersteunen van doorstroombeslissingen aan het einde van het derde leerjaar van de havo en het vwo. In het valideringsonderzoek is onderzocht in hoeverre de toetsen aan deze eisen voldoen en in hoeverre ze over deze kenmerken beschikken. Het onderzoek levert duidelijke aanwijzingen voor de geldigheid van de stelling dat de toetsen bruikbaar zijn voor het doel waarvoor ze ontwikkeld zijn.

In hoofdstuk twee is als eerste eis gesteld dat de toetsen curriculumonafhankelijk moeten zijn. Ze mogen uitsluitend items bevatten die een beroep doen op kennis of vaardigheden die onderwezen zijn aan alle leerlingen die in aanmerking komen om de toetsen te maken. De tweede eis was dat de potentiële toetsgebruikers de indruk moeten hebben dat de toetsen dienst kunnen doen bij het ondersteunen van doorstroombeslissingen. In hoofdstuk twee is verder gesteld dat de toetsen het voorafgaande onderwijsaanbod niet volledig hoeven te dekken. Wanneer de toetsen betrekking hebben op onderdelen van de leerstof waarin de te toetsen vaardigheid aan bod komt, is dat afdoende.

De kenmerken van het universum van mogelijk in een plaatsingstoets op te nemen items zijn af te leiden uit voornoemde toetseigenschappen en de omschrijving van de vaardigheid waar de toets zich op moet richten. Zijn de in een plaatsingstoets opgenomen items representatief voor dit universum, dan is dit een eerste duidelijke aanwijzing voor de validiteit van deze toets. De toetsen zijn dan volgens de klassieke opvatting over validiteit inhoudsvalide.

In hoofdstuk twee is als tweede eis gesteld dat de toetsen betrekking dienen te hebben op vaardigheden waarvan de beheersing van belang is voor het succesvol vervolgen van de havo- of vwo-opleiding. Indien blijkt dat de toetsen de beoogde constructen daadwerkelijk representeren, is dat een belangrijke positieve aanwijzing voor hun validiteit. Het beantwoorden van de vraag of de toetsen inderdaad betrekking hebben op de beoogde vaardigheden valt onder de noemer van wat in de klassieke opvatting over validiteit constructvaliditeit genoemd wordt. Ditzelfde geldt voor het beantwoorden van de vraag naar de onderlinge samenhang tussen de beoogde vaardigheden.

Centraal in het valideringsonderzoek ten slotte, staat de beantwoording van de vraag in welke mate de toetsen het mogelijk maken de toekomstige prestaties van leerlingen te voorspellen. Deze vraag heeft betrekking op wat in de klassieke opvatting over validiteit onder criterium gerelateerde validiteit verstaan werd. Een positief antwoord op de vraag naar het voorspellend vermogen van de plaatsingstoetsen levert de belangrijkste aanwijzing voor de bruikbaarheid van de toetsen als hulpmiddel bij het nemen van doorstroombeslissingen.

De termen inhoudsrepresentativiteit, begripsrepresentativiteit en voorspellend vermogen zullen in de rest van dit proefschrift de termen inhouds-, construct- en criterium gerelateerde validiteit vervangen om aan te geven dat steeds sprake is van de drie met elkaar samenhangende aspecten van het begrip validiteit zoals Messick (1989) dat opvat. De term begripsrepresentativiteit mag daarbij niet verward worden met het door Embretson (1983) geïntroduceerde begrip ‘construct representation’, dat betrekking heeft op de cognitieve processen die responsen op opgaven kunnen verklaren.

4.3 De inhoudsrepresentativiteit van de toetsen

Een item dat representatief is voor het universum aan mogelijk in een plaatsingstoets op te nemen items heeft drie kenmerken. Het is allereerst curriculumonafhankelijk. Een item is curriculumonafhankelijk, indien de leerstof waar het item betrekking op heeft, onderwezen is aan alle leerlingen voor wie de toets bestemd is. In de tweede plaats moeten docenten van mening zijn dat het item geschikt is om in een plaatsingstoets op te nemen. In de derde plaats moet het plausibel zijn dat het item een beroep doet op de vaardigheid waar de toets betrekking op moet hebben.

De voor ieder vak gehanteerde procedure bij de constructie van de items maakt aannemelijk dat de items daadwerkelijk een beroep doen op de vaardigheden waar de toetsen betrekking op moeten hebben. De constructie-opdrachten voor ieder vak gaven een duidelijke omschrijving van de verschillende aspecten van de vaardigheid waar de items betrekking op moesten hebben. Bij de vakken Engels en wiskunde zijn de items ontwikkeld in samenwerking met een groep docenten met ruime ervaring in de boven- en onderbouw van havo en vwo. Bij het vak Nederlands was dit bij twee van de drie soorten items ook het geval. Een derde soort items was reeds eerder ontwikkeld en beproefd. Daarnaast is een deel van de ontwikkelde items naar aanleiding van besprekingen binnen de constructieteams bijgesteld of verworpen. Omdat de voor ieder vak ontwikkelde items betrekking hebben op verschillende aspecten van de beoogde vaardigheid, is voorkomen dat slechts onderdelen van de beoogde constructen zijn geoperationaliseerd.

In hoeverre items aan de eisen van curriculumonafhankelijkheid en gepercipieerde geschiktheid voldeden, is onderzocht in de proefafnames. Een vragenlijst bood docenten de gelegenheid aan te geven welke items ze niet geschikt vonden voor opname in een toets en welke items een beroep deden op leerstof die niet behandeld was.

De twee voornoemde eisen speelden een rol bij de toetsconstructie voor ieder vak. Omdat blijkens de reacties van de docenten veruit de meeste van de ontwikkelde items curriculumonafhankelijk waren en geschikt om in een

plaatsingstoets op te nemen, was het mogelijk toetsen te construeren met een goede inhoudsrepresentativiteit. Dit houdt allereerst in dat de in de toetsen opgenomen items door docenten nooit of slechts incidenteel als niet geschikt bestempeld zijn. En bovendien betekent het dat bij ieder opgenomen item vrijwel alle docenten aangegeven hebben dat de leerstof waar het betrekking op heeft, behandeld is.

4.4 De begripsrepresentativiteit van de toetsen

Een eerste belangrijke aanwijzing voor de begripsrepresentativiteit van de toetsen is dat de items die voor ieder vak in de toetsen zijn opgenomen zonder uitzondering OPLM-schaalbaar waren. Dat wil zeggen dat voor iedere toets geldt dat de items betrekking hebben op een en hetzelfde construct. Een tweede duidelijke aanwijzing voor de begripsrepresentativiteit van de toetsen is al gepresenteerd in hoofdstuk twee. In dit hoofdstuk is namelijk beschreven dat voor ieder vak de leerlingen uit drie havo in de regel de vaardigheden waar de toetsen betrekking op hadden minder beheersten dan de leerlingen uit drie havo/vwo en dat deze op hun beurt de vaardigheden waar de toetsen betrekking op hadden weer minder beheersten dan de leerlingen uit drie vwo.

Een derde positieve aanwijzing voor de begripsrepresentativiteit van de toetsen is te vinden in de samenhang van de vaardigheden waar ze betrekking op hebben met gegevens waarvan bekend is dat ze een rol spelen bij doorstroombeslissingen, zoals oordelen van docenten en rapportcijfers. In de proefafnames zijn deze gegevens dan ook verzameld.

Aan de vakdocenten van de deelnemende leerlingen is gevraagd het afgeronde rapportcijfer van deze leerlingen te verstrekken voor het eindrapport van het derde leerjaar. Dit leverde voor 617 leerlingen uit drie havo en voor 1088 leerlingen uit drie vwo een cijfer op voor Nederlands; voor 1049 leerlingen uit drie havo en voor 1278 leerlingen uit drie vwo een cijfer voor Engels en voor 987 leerlingen uit drie havo en voor 1356 leerlingen uit drie vwo een cijfer voor wiskunde. Per vak zijn de leerlingen op grond van

hun rapportcijfers in vier categorieën ingedeeld. De eerste categorie bestond uit de leerlingen met een afgerond rapportcijfer van 5 of lager. De tweede categorie bevatte de leerlingen met het afgeronde rapportcijfer 6 en de derde categorie de leerlingen met het afgeronde rapportcijfer 7. De vierde categorie bestond uit de leerlingen met een afgerond rapportcijfer van 8 of hoger.

Verder is aan de docenten gevraagd het toekomstig prestatieniveau van de leerlingen die over zouden gaan te voorspellen door ze in drie categorieën in te delen. De eerste categorie bestond uit de leerlingen van wie docenten verwachtten dat ze wel over zouden gaan, maar dat ze in het volgende leerjaar tot de 'slechtere' leerlingen (de onderste 25 procent) zouden gaan behoren. De tweede categorie bevatte de leerlingen van wie docenten voorspelden dat ze in het volgende leerjaar tot de 'gemiddelde' leerlingen zouden gaan behoren. De derde categorie bestond uit de leerlingen van wie docenten aangaven dat ze in het volgende leerjaar tot de 'betere' leerlingen (de bovenste 25 procent) zouden gaan behoren. Bij Nederlands leverde dit verzoek docentoordelen op voor 487 leerlingen uit drie havo en 921 leerlingen uit drie vwo; bij Engels docentoordelen voor 1011 leerlingen uit drie havo en 1055 leerlingen uit drie vwo en bij wiskunde docentoordelen voor 852 leerlingen uit drie havo en 1193 leerlingen uit drie vwo.

Zowel voor de leerlingen uit drie havo als de leerlingen uit drie vwo is voor de hierboven beschreven variabelen 'rapportcijfer' en 'docentoordeel' per vak onderzocht in welke mate leerlingen uit de beschreven categorieën in vaardigheid verschilden. Dit is gedaan door het uitvoeren van analyses met behulp van het al in hoofdstuk twee beschreven programma SAUL (Verhelst & Verstralen, 1996). SAUL schat de gemiddelde vaardigheid in verschillende subgroepen door het uitvoeren van regressie-analyses met de latente vaardigheid als afhankelijke variabele, en achtergrondvariabelen, zoals in dit geval 'rapportcijfer' en 'docentoordeel', als onafhankelijke variabelen. Om effecten eenduidig schatbaar te maken stelt SAUL het effect van de eerste categorie van een variabele gelijk aan 0. Verder neemt SAUL aan dat binnen de subpopulaties die gevormd kunnen worden op basis van de achtergrondvariabelen de vaardigheid normaal verdeeld is. Bovendien neemt SAUL aan dat de binnengroepvariantie van de vaardigheid voor iedere subpopulatie gelijk is. Door de effectschattingen te delen door de binnengroepstandaardafwijking van de vaardigheid ontstaat een maat voor de effectgrootte van iedere

categorie. Door de verschillen in effectgrootte tussen de onderscheiden categorieën te bepalen, wordt het mogelijk de verschillen in gemiddelde vaardigheid tussen leerlingen uit de verschillende categorieën te interpreteren.

Tabel 4.1 geeft per vaardigheid en onderwijstype een overzicht van de effectgroottes voor de opeenvolgende categorieën van de variabele rapportcijfer. Tabel 4.2 doet hetzelfde voor de variabele docentoordeel. De effectgroottes voor niet-opeenvolgende categorieën worden in deze tabellen niet vermeld. Ze zijn echter eenvoudig te berekenen door de effectgroottes voor de opeenvolgende categorieën te sommeren. Voor meer uitgebreide overzichten van de resultaten van de SAUL-analyses voor deze twee variabelen, met informatie over de effectschattingen zelf en de aantallen leerlingen in iedere categorie, wordt verwezen naar Sluijter (1998).

In tabel 4.1 staat in de kolom met het opschrift ‘contrasten rapportcijfer’ bij welke twee opeenvolgende categorieën van de variabele rapportcijfer de effectgrootte hoort die in de volgende kolommen vermeld staat. Voor het verschil in leesvaardigheid in drie havo tussen de groep leerlingen met een afgerond cijfer van 5 of lager en de groep leerlingen met het afgeronde cijfer 6 bedraagt de effectgrootte bijvoorbeeld 0,341. Cohen (1977) spreekt dan van een klein tot middelmatig effect. Dat wil zeggen dat leerlingen die in drie havo een afgerond rapportcijfer van 5 of lager voor het vak Nederlands hebben wat minder leesvaardig zijn dan leerlingen met het afgeronde cijfer 6. Uit de tabel is verder bijvoorbeeld af te leiden dat sprake is van een groot effect bij het verschil in leesvaardigheid tussen de groep leerlingen met een afgerond cijfer van 5 of lager en de groep leerlingen met een afgerond cijfer van 8 of hoger. De grootte van het effect bedraagt namelijk (0,341 + 0,291 + 0,568 =) 1,200. Leerlingen die in drie havo een afgerond rapportcijfer van 5 of lager hebben voor het vak Nederlands, blijken veel minder leesvaardig dan leerlingen met een afgerond cijfer van 8 of hoger.

Tabel 4.1

Effectgrootte voor opeenvolgende categorieën van de variabele rapportcijfer voor de vaardigheidsschalen voor Nederlands, Engels en wiskunde, voor de leerlingen uit drie havo en drie vwo

Contrasten rapportcijfer	Effectgrootte*					
	Nederlands		Engels		wiskunde	
	drie havo	drie vwo	drie havo	drie vwo	drie havo	drie vwo
≤ 5 - 6	0,341	0,232	0,603	0,758	0,563	0,573
6 - 7	0,291	0,361	0,536	0,474	0,600	0,381
7 - ≥ 8	0,568	0,674	0,491	0,553	0,793	0,570

* 0,2: klein effect; 0,5: middelmatig effect; 0,8: groot effect (Cohen, 1977)

De gegevens in tabel 4.1 laten zien dat voor iedere vaardigheid en voor beide onderwijstypen zonder uitzondering geldt dat leerlingen, naarmate zij een hoger rapportcijfer hebben, de vaardigheden waar de toetsen betrekking op hebben ook beter beheersen. De leerlingen met een afgerond rapportcijfer van 5 of lager blijken het minst vaardig, gevolgd door de leerlingen met een afgerond cijfer van 6. De leerlingen met een afgerond cijfer van 7 blijken weer vaardiger dan de leerlingen met een afgerond cijfer van 6. De leerlingen met een afgerond cijfer van 8 of hoger blijken het vaardigst.

In tabel 4.2 staat in de kolom met het opschrift ‘contrasten docentoordeel’ bij welke twee opeenvolgende categorieën van de variabele docentoordeel de effectgrootte hoort die in de volgende kolommen vermeld staat. Voor het verschil in leesvaardigheid in drie havo tussen de ‘slechtere’ leerlingen en de ‘gemiddelde’ leerlingen bedraagt de effectgrootte bijvoorbeeld 0,541. Bij deze effectgrootte is sprake van een middelmatig tot groot effect. Dat wil zeggen dat leerlingen van wie docenten in drie havo voorspelden dat zij in vier havo slecht zouden presteren een stuk minder leesvaardig blijken te zijn dan leerlingen van wie docenten voorspelden dat zij in vier havo gemiddeld zouden presteren. Verder valt bijvoorbeeld uit de tabel af te leiden dat bij het verschil in leesvaardigheid in drie havo tussen de ‘slechtere’ leerlingen en de ‘betere’ leerlingen een effectgrootte hoort van (0,541 + 0,279 =) 0,820. Hier is sprake van een groot effect: leerlingen van wie docenten in drie havo

voorspelden dat zij in vier havo slecht zouden presteren, blijken veel minder leesvaardig dan leerlingen van wie docenten in drie havo voorspelden dat zij in vier havo goed zouden presteren.

De gegevens in tabel 4.2 maken duidelijk dat voor iedere vaardigheid en voor beide onderwijstypen zonder uitzondering geldt dat naarmate docenten een gunstiger voorspelling doen, leerlingen de vaardigheden waar de toetsen betrekking op hebben ook beter beheersen. De leerlingen van wie de docenten aangeven dat zij in het volgende leerjaar tot de slechtere leerlingen zullen behoren, blijken minder vaardig dan de leerlingen van wie de docenten aangeven dat zij tot de gemiddelde leerlingen zullen behoren. En deze laatste groep leerlingen blijkt op zijn beurt weer minder vaardig dan de leerlingen van wie de docenten aangeven dat zij in het volgende leerjaar tot de betere leerlingen zullen behoren.

Tabel 4.2

Effectgrootte voor opeenvolgende categorieën van de variabele docentoordeel

voor de vaardigheidsschalen voor Nederlands, Engels en wiskunde, voor leerlingen uit drie havo en drie vwo

Contrasten docentoordeel	Effectgrootte*					
	Nederlands		Engels		wiskunde	
	drie havo	drie vwo	drie havo	drie vwo	drie havo	drie vwo
slechter - gemiddeld	0,541	0,580	0,745	0,875	0,711	0,610
gemiddeld - beter	0,279	0,647	0,772	0,779	0,737	1,021

* 0,2: klein effect; 0,5: middelmatig effect; 0,8: groot effect (Cohen, 1977)

Uit de gegevens in de tabellen 4.1 en 4.2 blijkt dat de voor de drie vakken ontwikkelde toetsen elk betrekking hebben op een vaardigheid die relevant is. Rapportcijfers en de oordelen van docenten over prestaties in het volgende leerjaar zijn gegevens die van belang zijn bij het nemen van doorstroombeslissingen. Verder geldt voor ieder vak dat vaardiger leerlingen

hogere rapportcijfers krijgen en dat docenten over leerlingen die vaardiger zijn ook positiever oordelen voor wat betreft hun toekomstig presteren.

Een vierde positieve aanwijzing voor de begripsrepresentativiteit van de toetsen is te vinden in de schoolloopbaan van de leerlingen die in het schooljaar 1993-1994 aan het einde van drie havo en vier havo hebben deelgenomen aan de proefafnames. In 1995 zijn gegevens verzameld die een beeld geven van het studiesucces van deze leerlingen. Deze gegevens maakten het mogelijk vast te stellen welke verschillen in beheersing er bestonden tussen succesvolle en niet succesvolle leerlingen voor de verschillende vaardigheden waar de toetsen betrekking op hebben.

De verzamelde gegevens hadden onder meer betrekking op het onderwijstype en het leerjaar waarin de leerlingen zich respectievelijk in het schooljaar 1994-1995 en het schooljaar 1995-1996 bevonden. In totaal zijn gegevens verkregen van 1452 leerlingen die zich ten tijde van de proefafnames in drie havo bevonden en van 2110 leerlingen die zich destijds in drie vwo bevonden. Op grond van deze gegevens zijn leerlingen in twee categorieën ingedeeld. De ene categorie bevat de 'succesvolle' leerlingen: leerlingen die zich in het leerjaar 1995-1996 in vijf havo of vijf vwo bevonden en dus zonder vertraging zijn doorgestroomd. De andere groep bevat de 'niet succesvolle' leerlingen: leerlingen die zich in het schooljaar 1995-1996 niet in vijf havo of vijf vwo bevonden. Het betreft hier leerlingen die niet naar behoren hebben gepresteerd: ze zijn blijven zitten, zijn afgestroomd naar een ander onderwijstype of hebben hun opleiding inmiddels afgebroken.

Om de verschillen tussen succesvolle en niet succesvolle leerlingen te bepalen voor wat betreft de mate van beheersing van de vaardigheden waar de toetsen betrekking op hebben, is per onderwijstype en per vak een SAUL-analyse uitgevoerd. Tabel 4.3 geeft per vaardigheid en onderwijstype een overzicht van het verschil in effectgrootte voor deze twee groepen leerlingen. Zo bedraagt het verschil in effectgrootte tussen niet succesvolle en succesvolle havo-leerlingen voor leesvaardigheid 0,530. Bij deze effectgrootte spreken we van een middelmatig effect. Dat wil zeggen dat succesvolle havo-leerlingen een stuk leesvaardiger zijn dan niet succesvolle havo-leerlingen. Voor meer uitgebreide overzichten van de resultaten van de SAUL-analyses voor deze twee variabelen, met informatie over de effectschattingen zelf en

de aantallen leerlingen in iedere categorie, wordt verwezen naar Sluijter (1998).

De gegevens in tabel 4.3 laten zien dat voor iedere vaardigheid en voor beide onderwijstypen geldt dat leerlingen die de vaardigheden waar de toetsen betrekking op hebben beter beheersen ook meer succesvol zijn. Deze gegevens bevestigen eens te meer de begripsrepresentativiteit van de plaatsingstoetsen. Ze maken duidelijk dat de toetsen betrekking hebben op vaardigheden die relevant zijn voor het nemen van doorstroombeslissingen.

Tabel 4.3.

Effectgrootte voor het verschil tussen succesvolle leerlingen en niet-succesvolle leerlingen op de vaardigheidsschalen voor Nederlands, Engels en wiskunde, voor leerlingen uit drie havo en drie vwo

Effectgrootte*					
Nederlands		Engels		wiskunde	
drie havo	drie vwo	drie havo	drie vwo	drie havo	drie vwo
0,530	0,472	0,194	0,313	0,570	0,549

* 0,2: klein effect; 0,5: middelmatig effect; 0,8: groot effect (Cohen, 1977)

Voornoemde gegevens hebben betrekking op de begripsrepresentativiteit van iedere plaatsingstoets afzonderlijk. De gegevens die hierna aan de orde komen, hebben te maken met de toetsen als geheel. Een uitgangspunt bij de keuze voor het ontwikkelen van de toetsen voor de vakken Nederlands, Engels en wiskunde en de checklist studievaardigheid was dat de verschillende te meten vaardigheden niet al te sterk mochten samenhangen. De tabellen 4.4 en 4.5 geven respectievelijk voor leerlingen uit drie havo en leerlingen uit drie vwo per vak een overzicht van een aantal correlaties. Het betreft hier de correlaties tussen de geschatte vaardigheden van de leerlingen, de correlaties tussen de geschatte vaardigheden enerzijds en de scores op de checklist studievaardigheid anderzijds en de correlaties tussen de drie scores op de checklisten. Tussen haakjes staat steeds het aantal waarnemingen waarop de correlaties gebaseerd zijn.

De tabellen 4.4 en 4.5 laten zien dat zowel voor leerlingen uit drie havo als voor leerlingen uit drie vwo geldt dat de vaardigheden waar de verschillende toetsen betrekking op hebben geen al te grote samenhang vertonen. Ditzelfde geldt in iets mindere mate voor de samenhang tussen de scores op de checklisten onderling en de samenhang tussen de vaardigheden enerzijds en de scores op de checklisten anderzijds. Dit betekent dat te verwachten valt dat zowel de ontwikkelde toetsen

Tabel 4.4

Correlaties tussen de per vak geschatte vaardigheden, correlaties tussen geschatte vaardigheden en checklistscores en correlaties tussen de checklist-sco-
res voor leerlingen uit drie havo. Tussen haakjes staan de aantallen observaties

	Nederlands	Engels	Wiskunde	Checklist Nederlands	Checklist Engels
Engels	0,25 (449)				
Wiskunde	0,19 (433)	0,17 (612)			
Checklist Nederlands	0,11 (242)	*	0,11 (145)		
Checklist Engels	*	0,17 (437)	0,12 (199)	0,30 (118)	
Checklist wiskunde	*	*	0,33 (320)	0,27 (114)	0,26 (184)

* correlaties zijn niet significant

Tabel 4.5

Correlaties tussen de per vak geschatte vaardigheden, correlaties tussen geschatte vaardigheden en checklistscores en correlaties tussen de checklist-sco-
res voor leerlingen uit drie vwo. Tussen haakjes staan de aantallen observaties

	Nederlands	Engels	Wiskunde	Checklist Nederlands	Checklist Engels
Engels	0,40 (710)				
Wiskunde	0,28 (658)	0,21 (988)			
Checklist Nederlands	0,17 (386)	0,14 (187)	0,30 (199)		
Checklist Engels	0,26 (250)	0,31 (540)	0,22 (327)	0,31 (165)	
Checklist wiskunde	*	*	0,32 (434)	0,50 (145)	0,34 (271)

* correlaties zijn niet significant

als de checklisten elk een deels unieke bijdrage zullen leveren aan het voorspellend vermogen van het plaatsingsinstrument als geheel. Opvallend is nog dat de correlaties bij de leerlingen uit drie vwo groter zijn dan bij de leerlingen uit drie havo. De inhoud van de drie toetsen die ontwikkeld zijn voor beide groepen stemt overeen en de inhoud van de checklist is zelfs volledig identiek. Daarom mocht bij een gelijke heterogeniteit van beide leerlinggroepen verwacht worden dat de correlaties overeenkomstig zouden zijn. Blijkbaar is de groep leerlingen uit drie havo meer homogeen. De oorzaak hiervan kan mogelijkerwijs liggen in het feit dat de leerlingen uit de drie-vwo-groep afkomstig zijn van scholen met een ongedeeld vwo, van scholen met een gedeeld vwo en van categoriale gymnasia.

4.5 Het voorspellend vermogen van de toetsen

Alvorens het mogelijk was het voorspellend vermogen te bepalen van de op grond van de proefafnames ontwikkelde instrumenten, moesten er eerst twee problemen opgelost worden.

Het eerste probleem was dat het niet mogelijk was, zoals al in hoofdstuk twee is toegelicht, een apart onderzoek uit te voeren waarin een nieuwe groep leerlingen de toetsen in hun definitieve vorm kon maken en de docenten Nederlands, Engels en wiskunde van de betreffende leerlingen de checklist in zijn definitieve vorm konden invullen. Voor het bepalen van het voorspellend vermogen van de instrumenten moest dus gebruik gemaakt worden van de scores van de leerlingen op de door hen gemaakte toetsboekjes en hun scores op het prototype van de checklist. De items in de definitieve versie van de checklist vormden een deelverzameling van de items in het prototype (zie De Wit, Heuvelmans & Sluijter, 1995). Daarom was het eenvoudig de scores van de leerlingen op de definitieve checklist te bepalen. De items die in de plaatsingstoetsen opgenomen zijn, waren echter afkomstig uit verschillende toetsboekjes. Voor geen van de drie vakken gold dat er leerlingen waren aan wie alle items waren voorgelegd die uiteindelijk zijn opgenomen in de plaatsingstoetsen. Zoals al in hoofdstuk twee werd opgemerkt, waren er dus geen leerlingen van wie de scores op de geconstrueerde plaatsingstoetsen direct vielen te bepalen. Daarom zijn voor ieder vak en onderwijstype met behulp van het programma OPTAL (Verstralen, 1996) de verwachte scores van de leerlingen, gegeven hun geschatte vaardigheid, bepaald voor de betreffende plaatsingstoets. Zoals in hoofdstuk drie aangegeven, betrof het bij de toetsen voor Nederlands en Engels de met de discriminatie-indices gewogen scores en bij de toetsen voor wiskunde de ruwe scores.

Om het voorspellend vermogen van de plaatsingstoetsen en de checklist studievaardigheid te bepalen, is gekozen voor het in paragraaf 4.4 beschreven criterium voor studiesucces. In welke mate het studiesucces van leerlingen uit drie havo en drie vwo te voorspellen is met behulp van hun scores op deze instrumenten, is onderzocht door het uitvoeren van logistische regressie-analyses (Hosmer & Lemeshow, 1989). Met behulp van een logistische regressie-analyse kan voor iedere onafhankelijke variabele X_v uit een reeks **X** van v ($v = 1, 2, ..., V$) variabelen het optimale gewicht bepaald worden ten aanzien van een te voorspellen discrete afhankelijke variabele Y, die de waarden 0 en 1 aan kan nemen. Het logistisch regressiemodel bepaalt voor een persoon i de verwachte score op de discrete afhankelijke variabele, gegeven de scores op de onafhankelijke variabelen, als:

$$\mathcal{E}(Y_i|\boldsymbol{x}_i) = \frac{\exp[g(\boldsymbol{x}_i)]}{1+\exp[g(\boldsymbol{x}_i)]}. \qquad (4.1)$$

In uitdrukking (4.1) is

$$g(\boldsymbol{x}_i) = B_0 + \sum_{v=1}^{V} B_v x_{iv},$$

waarbij B_0 een constante is, en de B_v de niet-gestandaardiseerde regressiegewichten van de onafhankelijke variabelen zijn.

In ons geval is Y een dichotome variabele die wel of geen succes aangeeft. Het model (4.1) schat dan ook voor iedere leerling de kans dat deze deel uitmaakt van de groep die succes heeft. Het model maakt het mogelijk voor leerlingen uit drie havo en drie vwo te bepalen hoe groot de kans is dat zij, gegeven hun scores op de verschillende instrumenten, zonder vertraging in respectievelijk vijf havo en vijf vwo zullen belanden.

Voordat het bepalen van de verwachte scores met behulp van OPTAL daadwerkelijk plaats kon vinden, moest eerst het tweede probleem opgelost worden. De logistische regressie-analyses vereisten namelijk dat de records in de gegevensbestanden volledig waren. Er waren echter maar weinig records die naast de criteriumscore ook drie vaardigheidsschattingen en drie checklistscores bevatten. Hiervan was slechts sprake bij 32 van de 1452 leerlingen uit drie havo en bij 62 van de 2110 leerlingen uit drie vwo. Voor de overige leerlingen varieerde het aantal ontbrekende gegevens van één tot vijf. Het percentage ontbrekende gegevens bij drie-havo-leerlingen was 57,9 en bij leerlingen uit drie vwo 43,4. De logistische regressie-analyses zouden geen generaliseerbare regressiegewichten opleveren, indien ze uitsluitend voor de volledige records uitgevoerd zouden worden. Daarom is ervoor gekozen de ontbrekende vaardigheidsschattingen en checklistscores te imputeren. Daarvoor is gebruik gemaakt van een methode die vergelijkbaar is met de procedure die in Fahrmeier & Tutz (1994, hoofdstuk acht) beschreven staat (Kamphuis, 1998, persoonlijke mededeling).

Tabel 4.6 geeft de resultaten van drie likelihood ratio tests die zijn uitgevoerd om voor de leerlingen uit drie havo en drie vwo de passing van verschillende logistische regressiemodellen te onderzoeken. Het eerste model - het volledige model - bevat alle zes onafhankelijke variabelen, het tweede model

bevat alleen de drie toetsscores en het derde model uitsluitend de drie checklistscores. In de tabel staat in de tweede kolom aangegeven op welk model de gegevens in de verschillende rijen betrekking hebben. De letter 'V' staat voor het volledige model. De letter 'T' maakt duidelijk dat de gegevens betrekking hebben op het model waarin alleen de drie toetsscores zijn opgenomen. De letter 'C' geeft aan dat de gegevens gelden voor het model waarin alleen de drie scores op de checklist zijn opgenomen. De kolom met het opschrift 'Modelfit' in deze tabel bevat informatie over de passing van het model op de data. De grootheid die gebruikt wordt om de passing aan te duiden, ontstaat door van het quotiënt van de likelihood van het model dat uitsluitend de constante B_0 bevat en het model dat in de tweede kolom omschreven staat de logaritme te nemen en deze met een factor -2 te vermenigvuldigen. De resulterende passingsmaat is asymptotisch chi-kwadraat verdeeld (zie bijvoorbeeld Bishop, Fienberg & Holland, 1975). In de vierde kolom van de tabel staat het aantal vrijheidsgraden van de bij de statistische toetsing gebruikte chi-kwadraat verdelingen en in de vijfde kolom de overschrijdingskans voor iedere likelihood-ratio test.

De overschrijdingskansen voor de likelihood ratio tests bij de verschillende modellen laten een goede passing van de modellen zien, wat gezien het aantal waarnemingen niet zo wonderlijk is. Een vergelijking van de waarden die de passingsmaten aannemen voor de verschillende modellen maakt duidelijk dat zowel voor de leerlingen uit drie havo als voor de leerlingen uit drie vwo het model met de drie checklistscores beter past dan het model met de drie toetsscores. De passing van het model waarin zowel de toets- als de checklistscores zijn opgenomen blijkt iets hoger te liggen dan de passing van het model waarin alleen de drie checklistscores zijn opgenomen.

Tabel 4.6

Resultaten voor leerlingen uit drie havo en drie vwo van de likelihood-ratio tests

voor het model dat de zes onafhankelijke variabelen bevat (V), het model dat

de drie toetsscores bevat (T) en het model dat de drie checklistscores bevat (C)

Klastype	Model	Modelfit	Vrijheids-graden	Overschrijdings-kans
	V	399,803	6	0,0000
Drie havo	**T**	156,523	3	0,0000
	C	319,288	3	0,0000
	V	569,425	6	0,0000
Drie vwo	**T**	234,142	3	0,0000
	C	511,726	3	0,0000

Om een uitspraak te kunnen doen over de grootte van het voorspellend vermogen van ieder model is Nagelkerke's (1991) $\tilde{R}^2$ berekend. Deze maat loopt van 0 tot 1 en is vergelijkbaar met het percentage verklaarde variantie bij lineaire regressie-analyse. Verder is voor alle modellen de zogeheten C-statistic berekend. Deze geeft in ons geval de mate aan waarin de modellen onderscheid maken tussen leerlingen die zonder vertraging doorstromen en leerlingen bij wie dit niet het geval is. Om de C-statistic te bepalen zijn alle paren leerlingen in beschouwing genomen waarvoor geldt dat de ene leerling succesvol is en de andere niet. Vervolgens zijn voor ieder leerlingpaar de op grond van het logistische regressiemodel berekende kansen op succes met elkaar vergeleken. De C-statistic geeft de proportie van de betreffende leerlingparen waarbij de berekende kans op succes voor de succesvolle leerling groter is dan de berekende kans op succes voor de niet-succesvolle leerling. Deze grootheid kan variëren van 0,5 tot 1. De C-statistic heeft de waarde 1 als het model alle succesvolle leerlingen hogere kansen op succes toedicht dan alle niet succesvolle leerlingen. Heeft de C-statistic een waarde van 0,5, dan wil dat zeggen dat het model geen enkel voorspellend vermogen heeft.

Tabel 4.7

Waarden van $\tilde{R}^2$ en de C-statistic voor het model met de zes

onafhankelijke variabelen (V), het model met de toetsscores (T) en het model

met de checklistscores (C) voor leerlingen uit drie havo en drie vwo

Klastype	Model	$\tilde{R}^2$	C-statistic
	V	0,241	0,82
Drie havo	**T**	0,102	0,70
	C	0,197	0,79
	V	0,237	0,83
Drie vwo	**T**	0,105	0,72
	C	0,215	0,81

In tabel 4.7 staan de waarden van $\tilde{R}^2$ en de C-statistic voor de verschillende modellen. Het volledige model wordt in deze tabel weer aangeduid met de letter 'V'; het model met de drie toetsscores door de letter 'T' en het model met de drie checklistscores door de letter 'C'. De tabel laat zien dat zowel voor leerlingen uit drie havo als voor leerlingen uit drie vwo het voorspellend vermogen van het volledige model beter is dan dat van de modellen waarin alleen de drie toetsscores of de drie checklistscores opgenomen zijn. Ook van de volledige modellen is het voorspellend vermogen echter beperkt. Bij alle modellen is de waarde van $\tilde{R}^2$ relatief laag. De waarden van de C-statistic maken duidelijk dat de verschillende modellen bij 70 tot 83 procent van alle mogelijke leerlingparen de succesvolle leerlingen een grotere kans op succes toedichten dan de niet succesvolle leerlingen.

Meer gedetailleerde informatie over de resultaten van de logistische regressie-analyses voor de drie onderscheiden modellen staat in de tabellen 4.8 en 4.9. De eerste tabel bevat de resultaten voor de leerlingen uit drie havo; de tweede de resultaten voor de leerlingen uit drie vwo. Ook in deze tabel worden de modellen weer aangeduid met de letters 'V', 'T' en 'C'. In de kolom met het opschrift 'B' staan de geschatte niet-gestandaardiseerde regressiegewichten van de onafhankelijke variabelen voor het betreffende model. In de vierde kolom staan de standaardfouten van deze schattingen. In

de vijfde kolom staat de Wald-statistic. Dit is een chi-kwadraat verdeelde grootheid waarmee statistisch getoetst kan worden of de regressiegewichten van de verschillende onafhankelijke variabelen significant van nul afwijken. In de zesde kolom staan de uitkomsten van deze statistische toetsing. Het aantal vrijheidsgraden van de bij de toetsing gebruikte chi-kwadraat verdelingen is bij alle modellen 1 en is niet in de tabel opgenomen. De 'R-statistic' in de zevende kolom van beide tabellen geeft de partiële correlatie tussen het criterium en iedere onafhankelijke variabele afzonderlijk.

Tabel 4.8 laat zien dat, met uitzondering van de score op de toets Engels, alle onafhankelijke variabelen een unieke bijdrage leveren aan het voorspellend vermogen van het volledige model ten aanzien van het studiesucces van leerlingen uit drie havo. Het regressiegewicht van de betreffende toetsscore wijkt niet significant van nul af en de waarde van de R-statistic is nul. In het model dat alleen de drie toetsscores bevat, levert de score op de toets Engels een unieke, zij het bescheiden, bijdrage. Bij dit laatste model wijkt het regressiegewicht van de toetsscore wèl significant af van nul en heeft de R-statistic de waarde 0,0519.

De eerder geformuleerde verwachting dat ieder instrument een unieke bijdrage zou leveren aan het voorspellend vermogen van het instrumentarium komt voor de leerlingen uit drie havo dus niet uit. Een mogelijke verklaring is dat er maar een klein verschil bestaat tussen succesvolle en niet succesvolle leerlingen voor wat betreft de vaardigheid die de toets Engels voor drie havo meet, zoals tabel 4.3 laat zien. Bovendien hebben de toetsen Nederlands en Engels beide betrekking op leesvaardigheid en laat tabel 4.4 zien dat de samenhang tussen de prestaties op deze twee toetsen het minst laag is van de correlaties tussen de drie toetsen. Deze feiten zouden er in combinatie toe kunnen leiden dat de toetsscore Engels niets meer bijdraagt aan het voorspellend vermogen van het volledige model, wanneer de toetsscore Nederlands er deel van uitmaakt. Een logistische regressie-analyse met daarin opgenomen de drie checklistscores en de scores op de toetsen voor Engels en wiskunde bevestigt deze verklaring. Bij dit model heeft de score op de toets voor Engels een regressiegewicht van 0,0131. De bijbehorende overschrijdingskans is 0,0411 en de R-statistic heeft de waarde 0,0357.

Tabel 4.8

*Resultaten van de logistische regressie-analyses voor leerlingen uit drie havo voor het model met zes onafhankelijke variabelen (**V**), het model met de drie toetsscores (**T**) en het model met de drie checklistscores (**C**)*

Variabele	Model	B	Standaard-fout	Wald-statistic	Overschrij-dingskans	R-statistic
Toets	**V**	0,0334	0,0060	31,4323	0,0000	0,1315
Nederlands	**T**	0,0237	0,0052	20,7356	0,0000	0,1049
Toets	**V**	0,0051	0,0067	0,5758	0,4480	0,0000
Engels	**T**	0,0150	0,0058	6,5914	0,0102	0,0519
Toets	**V**	0,0297	0,0069	18,5505	0,0000	0,0986
wiskunde	**T**	0,0498	0,0059	72,5091	0,0000	0,2036
Checklist	**V**	0,0480	0,0081	35,3624	0,0000	0,1400
Nederlands	**C**	0,0510	0,0077	43,5362	0,0000	0,1562
Checklist	**V**	0,0551	0,0070	61,4265	0,0000	0,1869
Engels	**C**	0,0510	0,0066	64,9940	0,0000	0,1924
Checklist	**V**	0,0552	0,0084	42,8245	0,0000	0,1549
wiskunde	**C**	0,0633	0,0075	70,3919	0,0000	0,2005
	V	-10,4086	0,7371	199,4117	0,0000	
Constante	**T**	- 2,3691	0,3532	44,9938	0,0000	
	C	- 8,0292	0,5877	186,6283	0,0000	

Tabel 4.9 maakt duidelijk dat twee van de zes onafhankelijke variabelen een niet of nauwelijks unieke bijdrage leveren aan het voorspellend vermogen van het volledig model ten aanzien van het studiesucces van leerlingen uit drie vwo. Zowel de score op de toets Engels als de score op de checklist studievaardigheid voor wiskunde dragen weinig bij aan de mate waarin het volledige model onderscheid maakt tussen leerlingen die zonder vertraging zijn doorgestroomd naar vijf vwo en leerlingen bij wie dit niet het geval is. De regressiegewichten van beide variabelen wijken niet significant van nul af en de waarde van de R-statistic is in beide gevallen nul. De resultaten voor het model met de drie toetsscores leert dat de score op de toets Engels wel degelijk een unieke bijdrage levert, indien alleen de toetsen beschouwd worden. Bij dit laatste model wijkt het betreffende regressiegewicht wèl

significant af van nul en heeft de R-statistic de waarde 0,0944. Verder laten de resultaten voor het model met de drie checklistscores zien dat de checklistscore voor wiskunde ook geen unieke bijdrage levert, indien alleen de scores op de checklisten beschouwd worden. Ook bij dit laatste model wijkt het regressiegewicht van de checklistscore voor wiskunde niet significant van nul af.

Tabel 4.9

*Resultaten van de logistische regressie-analyses voor leerlingen uit drie vwo voor het model met zes onafhankelijke variabelen (**V**), het model met de drie toetsscores (**T**) en het model met de drie checklistscores (**C**)*

Variabele	Model	B	Standaard-fout	Wald-statistic	Overschrij-dingskans	R-statistic
Toets	**V**	0,0190	0,0062	9,5566	0,0000	0,0572
Nederlands	**T**	0,0237	0,0054	19,0839	0,0000	0,0861
Toets	**V**	0,0061	0,0044	1,9656	0,1609	0,0000
Engels	**T**	0,0182	0,0038	22,5426	0,0000	0,0944
Toets	**V**	0,0234	0,0054	18,5833	0,0000	0,0848
wiskunde	**T**	0,0378	0,0046	68,1847	0,0000	0,1694
Checklist	**V**	0,0640	0,0079	65,6625	0,0000	0,1662
Nederlands	**C**	0,0701	0,0077	82,3394	0,0000	0,1867
Checklist	**V**	0,0714	0,0070	102,9205	0,0000	0,2092
Engels	**C**	0,0826	0,0066	157,7984	0,0000	0,2599
Checklist	**V**	0,0106	0,0076	1,9369	0,1640	0,0000
wiskunde	**C**	0,0122	0,0070	2,9824	0,0842	0,0206
	V	-9,6288	0,7371	281,1860	0,0000	
Constante	**T**	-2,7512	0,3009	83,5787	0,0000	
	C	-8,2802	0,4941	280,8692	0,0000	

Ook voor de leerlingen uit drie vwo komt de verwachting niet uit dat ieder instrument een unieke bijdrage zou leveren aan het voorspellend vermogen van het instrumentarium. Net als bij het instrumentarium voor drie havo blijkt de score op de toets Engels niets meer bij te dragen aan het voorspel-

lend vermogen van het volledige model, wanneer de toetsscore Nederlands er deel van uitmaakt. Bij het model met de drie checklistscores en de toetsscores voor Engels en wiskunde heeft de toetsscore Engels een regressiegewicht van 0,0123 met een overschrijdingskans van 0,0016 en heeft de R-statistic de waarde 0,0588.

Verder inzicht in het voorspellend vermogen van het instrumentarium als geheel valt te verkrijgen door bij iedere leerling een uitspraak te doen over de vraag of deze al of niet succesvol zal zijn op grond van de met behulp van de logistische regressie-analyse bepaalde verwachte kans op studiesucces, $\mathscr{E}(Y_i|\boldsymbol{x}_i)$. Deze uitspraken kunnen vervolgens afgezet worden tegen de score van iedere leerling op het dichotome criterium voor studiesucces. Daarbij zijn de volgende vier uitkomsten mogelijk.

1/1 Leerlingen van wie voorspeld wordt dat ze succesvol zullen zijn, blijken ook daadwerkelijk succesvol.
0/0 Leerlingen van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken ook daadwerkelijk **niet** succesvol.
1/0 Leerlingen van wie voorspeld wordt dat ze succesvol zullen zijn, blijken **niet** succesvol.
0/1 Leerlingen van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken wel succesvol.

Bij de uitkomsten **1/1** en **0/0** zijn de voorspellingen correct en bij de uitkomsten **1/0** en **0/1** niet. Hoe hoger het percentage correcte voorspellingen, des te sterker is het voorspellend vermogen van de instrumenten. De frequentie van de verschillende uitkomsten is echter ook afhankelijk van $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i)$, de verwachte kans op studiesucces die minimaal vereist is voor de voorspelling dat een leerling succesvol zal zijn.

Tabel 4.10 geeft voor een aantal verschillende waarden van de minimaal vereiste verwachte kans op studiesucces aan hoe vaak de onderscheiden uitkomsten voorkomen bij leerlingen uit drie havo en hoe groot het percentage correcte voorspellingen is. Tabel 4.11 doet hetzelfde voor de leerlingen uit drie vwo. Verder laten de tabellen voor de vermelde waarden van $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i)$ zien hoe groot het aantal leerlingen is dat binnen iedere

uitkomst valt. In beide tabellen staat tussen haakjes vermeld welk percentage van de leerlingen het hier betreft.

Tabel 4.10 laat bijvoorbeeld zien dat het percentage correcte voorspellingen op basis van de toets- en de checklistscores 77,9 bedraagt, indien $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i) = 0{,}5$ voor leerlingen uit drie havo. Bij deze minimaal vereiste verwachte kans blijken 964 leerlingen (66,4%) van wie voorspeld wordt dat ze succesvol zullen zijn ook daadwerkelijk succesvol (uitkomst **1/1**). En 167 leerlingen (11,5%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken ook daadwerkelijk **niet** succesvol (uitkomst **0/0**). Verder blijken er 229 leerlingen (15,8%) van wie voorspeld wordt dat ze succesvol zullen zijn **niet** succesvol (uitkomst **1/0**). En ten slotte blijken er 92 leerlingen (6,3%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn wel succesvol (uitkomst **0/1**).

Tabel 4.10

Aantallen correcte en niet correcte voorspellingen en percentage correct voorspeld bij verschillende waarden van de minimaal vereiste verwachte kans op studiesucces voor de leerlingen uit drie havo. Tussen haakjes staat het percentage leerlingen dat binnen iedere uitkomst valt

$\mathscr{E}_{\min}(Y_i\|\boldsymbol{x}_i)$	Uitkomsten				Percentage correct voorspeld
	Correct		Niet correct		
	1/1	**0/0**	**1/0**	**0/1**	
0,5	964 (66,4%)	167 (11,5%)	229 (15,8%)	92 (6,3%)	77,9%
0,6	896 (61,7%)	232 (16,0%)	164 (11,3%)	160 (11,0,%)	77,7%
0,7	799 (55,0%)	282 (19,4%)	114 (7,9%)	257 (17,7%)	74,4%
0,8	673 (46,3%)	338 (23,3%)	58 (4,0%)	383 (26,4%)	69,6%

Tabel 4.10 maakt verder duidelijk in hoeverre de verschillende uitkomsten afhankelijk zijn van de waarde van $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i)$ die men hanteert. Wordt bijvoorbeeld een minimale verwachte kans op studiesucces van 0,8 vereist voor de voorspelling dat een leerling uit drie havo succesvol zal zijn, dan blijken 673 leerlingen (46,3%) van wie voorspeld wordt dat ze succesvol zullen zijn ook daadwerkelijk succesvol (uitkomst **1/1**). En 338 leerlingen (23,3%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken ook daadwerkelijk **niet** succesvol (uitkomst **0/0**). Verder blijken er 58 leerlingen (4,0%) van wie voorspeld wordt dat ze succesvol zullen zijn **niet** succesvol (uitkomst **1/0**). En 383 leerlingen (26,4%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken wel succesvol (uitkomst **0/1**).

Tabel 4.11

Aantallen correcte en niet correcte voorspellingen en percentage correct voorspeld bij verschillende waarden van de minimaal vereiste verwachte kans op studiesucces voor de leerlingen uit drie vwo. Tussen haakjes staat het percentage leerlingen dat binnen iedere uitkomst valt

$\mathscr{E}_{\min}(Y_i\|\boldsymbol{x}_i)$	Uitkomsten				Percentage correct voorspeld
	Correct		Niet correct		
	1/1	**0/0**	**1/0**	**0/1**	
0,5	1515 (71,8%)	210 (10,0%)	288 (13,7%)	97 (4,0%)	81,8%
0,6	1430 (67,8%)	255 (12,1%)	243 (11,5%)	182 (8,6%)	79,9%
0,7	1322 (62,7%)	330 (15,6%)	168 (8,0%)	290 (13,7%)	78,3%
0,8	1117 (52,9%)	394 (18,7%)	104 (4,9%)	495 (23,2%)	71,6%

Tabel 4.11 laat zien dat het percentage correcte voorspellingen op basis van de toets- en de checklistscores voor leerlingen uit drie vwo 81,8 bedraagt, indien $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i)$ = 0,5. Bij het hanteren van deze minimaal vereiste kans op studiesucces blijken 1515 leerlingen (71,8%) van wie voorspeld wordt dat ze succesvol zullen zijn werkelijk succesvol (uitkomst **1/1**). En 210 leerlingen (10,0%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn, blijken ook werkelijk **niet** succesvol (uitkomst **0/0**). Verder blijken er 288 leerlingen (13,7%) van wie voorspeld wordt dat ze succesvol zullen zijn **niet** succesvol (uitkomst **1/0**) en blijken 97 leerlingen (4,0%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn wel succesvol (uitkomst **0/1**).

Verder laat tabel 4.11 ook zien in hoeverre de verschillende uitkomsten afhankelijk zijn van de waarde van $\mathscr{E}_{\min}(Y_i|\boldsymbol{x}_i)$ die men hanteert. Wordt bijvoorbeeld een minimaal verwachte kans van 0,8 vereist voor de voorspelling dat een leerling uit drie vwo succesvol zal zijn, dan blijken 1117 leerlingen (52,9%) van wie voorspeld wordt dat ze succesvol zullen zijn ook daadwerkelijk succesvol (uitkomst **1/1**). En 394 leerlingen (18,7%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn blijken ook daadwerkelijk **niet** succesvol (uitkomst **0/0**). Verder blijken er 104 leerlingen (4,9%) van wie voorspeld wordt dat ze succesvol zullen zijn **niet** succesvol (uitkomst **1/0**). En 495 leerlingen (23,2%) van wie voorspeld wordt dat ze **niet** succesvol zullen zijn blijken wel succesvol (uitkomst **0/1**).

Zowel uit tabel 4.10 als tabel 4.11 blijkt dat het hanteren van een hogere minimaal vereiste verwachte kans op studiesucces ertoe leidt dat het aantal leerlingen toeneemt van wie correct voorspeld wordt dat ze **niet** succesvol zullen zijn. Uiteraard blijkt dit ten koste te gaan van het aantal leerlingen van wie correct voorspeld wordt dat ze succesvol zullen zijn. In tabel 4.10 neemt het percentage leerlingen bij uitkomst **0/0** toe van 11,5 naar 23,3, terwijl het percentage leerlingen bij uitkomst **1/1** afneemt van 66,4 naar 46,3. In tabel 4.11 neemt het percentage leerlingen bij uitkomst **0/0** toe van 10,0 naar 18,7, terwijl het percentage leerlingen bij uitkomst **1/1** afneemt van 71,8 naar 52,2.

Voor de leerlingen die deelnamen aan de proefafnames is een totaalscore bepaald. Deze totaalscore kwam tot stand door de scores van de leerlingen uit drie havo en drie vwo op de zes instrumenten te vermenigvuldigen met gewichten die gebaseerd waren op de volledige logistische regressiemodellen

die opgesteld waren voor beide klastypen. Deze gewichten waren de geschatte regressiegewichten vermenigvuldigd met 100 en afgerond op één decimaal. Vervolgens zijn de zes resulterende gewogen scores per leerling gesommeerd.

Zowel de leerlingen uit drie havo als de leerlingen uit drie vwo zijn op grond van de aldus bepaalde totaalscores gesorteerd en over decielen verdeeld. Elk deciel bevat tien procent van de leerlingen. In het eerste deciel bevinden zich de leerlingen met de tien procent laagste totaalscores. Het tweede deciel bevat de leerlingen met de tien procent van de totaalscores die volgen op de laagste scores, enzovoorts. De leerlingen met de tien procent hoogste totaalscores maken deel uit van het tiende deciel. Vervolgens is een zogeheten verwachtingstabel opgesteld waarin aangegeven wordt welk percentage van de leerlingen in ieder deciel succesvol is en welk percentage niet. Scholen kunnen de informatie in de verwachtings-tabellen gebruiken bij het nemen van doorstroombeslissingen over nieuwe populaties leerlingen in drie havo en drie vwo.

Tabel 4.12

Percentages succesvolle en niet succesvolle leerlingen uit drie havo en drie vwo

bij een decielverdeling van de berekende totaalscores

Leerlingen uit drie havo			Leerlingen uit drie vwo		
Deciel	Percentage niet succesvol	Percentage succesvol	Deciel	Percentage niet succesvol	Percentage succesvol
1 (≤ 587)	73,5	26,5	**1** (≤ 562)	74,0	26,0
2 (588- 665)	51,4	48,6	**2** (563- 632)	43,8	56,2
3 (666- 700)	45,5	54,4	**3** (633- 684)	40,5	59,5
4 (701- 741)	32,9	67,1	**4** (685- 727)	23,4	76,7
5 (742- 783)	30,3	69,7	**5** (728- 767)	20,8	79,2
6 (784- 822)	15,1	84,9	**6** (768- 807)	12,4	87,6
7 (823- 857)	9,7	90,3	**7** (808- 847)	7,6	92,4
8 (858- 905)	6,9	93,1	**8** (848- 885)	6,8	93,2
9 (906- 970)	5,5	94,5	**9** (886- 938)	4,2	95,8
10 (≥ 971)	0,7	99,3	**10** (≥ 939)	1,5	98,5
Normgroep als geheel	27,3	72,7		23,6	76,4

Tabel 4.12 bevat de verwachtingstabellen voor de leerlingen uit drie havo en drie vwo. In de tabel is in de kolommen met decielaanduidingen tevens tussen haakjes het bijbehorende interval van de totaalscores vermeld. In de onderste regel van de tabel met het opschrift ‘normgroep als geheel’ staat welk percentage van de leerlingen uit drie havo en drie vwo als succesvol bestempeld kon worden en welk percentage niet.

De geldigheid van de informatie in deze tabel hangt af van twee aannamen. De eerste aanname is dat de relevante kenmerken van nieuwe populaties leerlingen overeenkomen met de kenmerken van de populaties waaruit de leerlingen afkomstig zijn die aan de proefafnames hebben deelgenomen. De

informatie in de tabel is bijvoorbeeld niet meer geldig, indien het curriculum in de leerjaren voorafgaand aan het afnemen van de instrumenten zich wijzigt. En de informatie in de tabel verliest ook zijn geldigheid, indien het curriculum in de tweede fase van het voortgezet onderwijs verandert.

De tweede aanname bij het verstrekken van dergelijke tabellen is dat de kans op succes voor een leerling niet afhankelijk is van de school die deze leerling bezoekt. Scholen moeten als uitwisselbaar beschouwd kunnen worden. Door het uitvoeren van een multi-niveau-logistische regressie-analyse (zie bijvoorbeeld Goldstein, 1995, hoofdstuk zeven) is onderzocht in hoeverre deze aanname gerechtvaardigd is. Deze multi-niveau-analyse gaat uit van twee niveaus: leerlingen en scholen. Het derde mogelijke niveau, dat van de klas, is buiten beschouwing gelaten, omdat veel scholen slechts met één klas vertegenwoordigd zijn en het aantal deelnemende klassen binnen scholen slechts bij hoge uitzondering groter dan twee was.

De verwachte kans op succes voor leerling *i* op school *k* is in een multi-niveau- logistisch regressiemodel met twee niveaus gedefinieerd als:

$$\mathscr{E}(Y_{ik}|\boldsymbol{x}_{ik}) = \frac{\exp[g(\boldsymbol{x}_{ik})]}{1+\exp[g(\boldsymbol{x}_{ik})]},$$

waarbij

$$g(\boldsymbol{x}_{ik}) = B_{0k} + \sum_{v=1}^{V} B_{vk} x_{vik}.$$

Het subscript *k* maakt hier duidelijk dat het intercept, B_{0k}, en de overige regressie-gewichten, B_{vk}, van school tot school kunnen variëren.

Zowel voor het intercept als de overige regressiegewichten geldt:

$$B_{.k} = B_{.} + \upsilon_k,$$

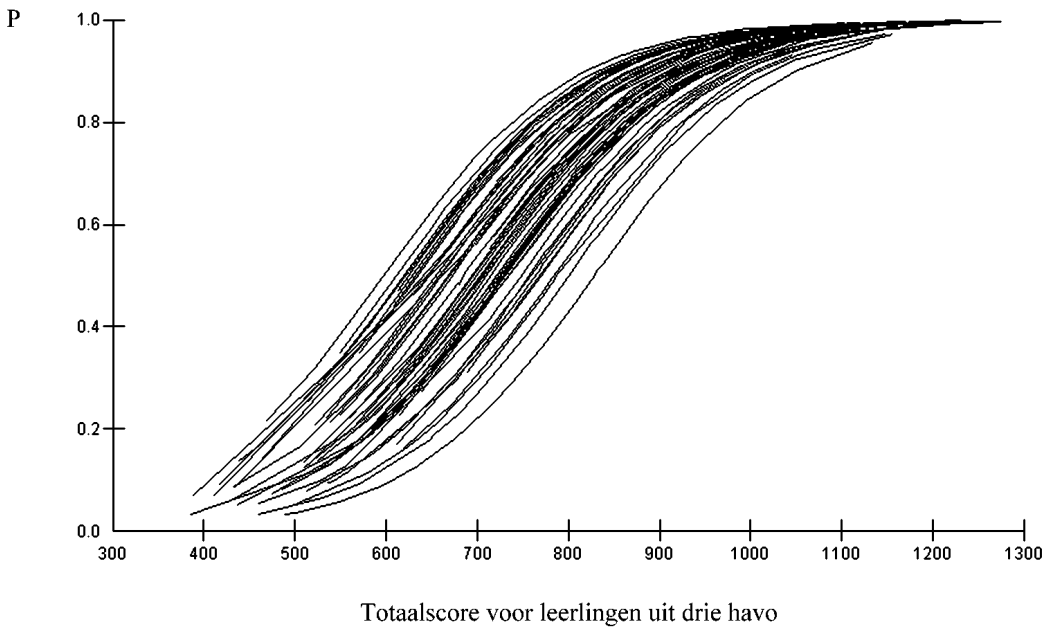
waarbij $B_{.}$ constant is over scholen en υ_k een normaal verdeelde variabele is met een gemiddelde gelijk aan nul en variantie σ^2_{υ}.

Allereerst zijn voor de leerlingen uit drie havo en drie vwo modellen opgesteld waarin alleen het intercept B_0 over scholen varieerde. De

regressiegewichten van de zes instrumenten werden in dit model als constant over scholen gespecificeerd. Het betreffende model voor leerlingen uit drie havo toont aan dat er duidelijke verschillen zijn tussen scholen voor wat betreft het intercept. De variabele υ_{0k} heeft een standaarddeviatie van 0,475 met een standaardfout van 0,143. De kans op succes voor een leerling uit drie havo, gegeven zijn of haar scores op de zes instrumenten, blijkt tevens afhankelijk te zijn van de school die de leerling bezoekt. Vervolgens zijn modellen opgesteld waarin ook de regressiegewichten van de scores op de instrumenten van school tot school mochten variëren. De regressiegewichten van de scores op de instrumenten bleken echter constant te zijn.

Het voor leerlingen uit drie vwo opgestelde model maakt duidelijk dat ook hier sprake is van duidelijke verschillen tussen scholen. De variabele υ_{0k} heeft bij de leerlingen uit drie vwo een standaarddeviatie van 0,623 met een standaardfout van 0,151. Voor deze leerlingen geldt eveneens dat hun kansen op succes, gegeven de behaalde scores op de zes instrumenten, afhankelijk zijn van de school die zij bezoeken. Ook voor drie-vwo-leerlingen zijn vervolgens modellen opgesteld waarin de regressiegewichten van school tot school uiteen mochten lopen. Bij deze leerlingen bleken de regressiegewichten van de instrumentscores eveneens niet tussen scholen te verschillen.

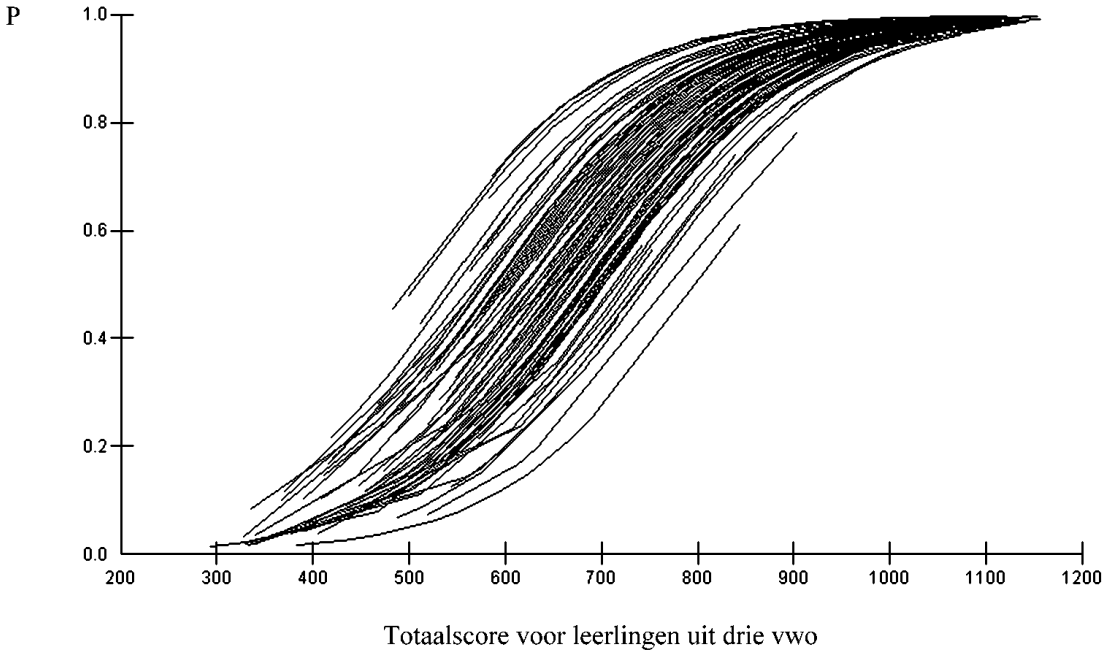
De discrepanties tussen scholen die voor beide klastypen in de multi-niveau-analyses geconstateerd zijn voor wat betreft het intercept worden naar alle waarschijnlijkheid voor een deel veroorzaakt doordat scholen verschillen in strengheid bij het beoordelen van leerlingen in het vierde leerjaar. De geconstateerde discrepanties zullen echter ook voor een deel het gevolg zijn van verschillen tussen scholen voor wat betreft de kwaliteit van de havo- en vwo-opleidingen in het vierde leerjaar. De in tabel 4.12 gepresenteerde informatie over de kansen op succes voor drie havo- en drie vwo-leerlingen is minder representatief voor een specifieke school, naarmate deze een intercept heeft dat verder afwijkt van het gemiddelde van de intercepten van de scholen die deelgenomen hebben aan de proefafnames.



Figuur 4.1

Relatie tussen de via het logistisch regressiemodel met één niveau bepaalde kans

op succes voor de leerlingen uit drie havo en hun totaalscores



Figuur 4.2

Relatie tussen de via het logistisch regressiemodel met één niveau bepaalde kans op succes voor de leerlingen uit drie vwo en hun totaalscores

Het schooleffect is weer te geven door in een figuur de kansen op succes die voor de leerlingen bepaald zijn via het logistische regressiemodel met één niveau af te zetten tegen hun totaalscores. De punten die behoren bij leerlingen die afkomstig zijn uit dezelfde school zijn in de figuur met elkaar verbonden. Daardoor ontstaat voor iedere school een regressiekromme die aangeeft hoe de totaalscores van de leerlingen die de school bezoeken zich verhouden tot hun kans op succes. Figuur 4.1 bevat de schoolregressiekrommen voor de leerlingen uit drie havo en figuur 4.2 de schoolregressiekrommen voor de leerlingen uit drie vwo. De figuren maken duidelijk dat de kans op succes voor een leerling, gegeven zijn of haar scores op de instrumenten, afhankelijk is van de bezochte school.

Figuur 4.1 laat bijvoorbeeld zien dat de kans op succes voor een leerling met een totaalscore rond de 700 op het voor drie havo ontwikkelde instrumentarium varieert van om en nabij de 0,2 tot bijna 0,8. En blijkens figuur 4.2 kan de kans op succes voor een leerling met een totaalscore rond de 700 op het

instrumentarium dat voor drie vwo ontwikkeld is uiteenlopen van iets meer dan 0,2 tot ongeveer 0,9.

Omdat de multi-niveau-modellen rekening houden met de verschillen tussen scholen, is hun voorspellend vermogen groter dan dat van de logistische regressiemodellen met maar één niveau. Om de winst in voorspellend vermogen te bepalen, is weer per leerling een uitspraak gedaan over de vraag of deze succesvol zal zijn of niet. Deze uitspraken zijn nu echter gebaseerd op de met behulp van de multi-niveau-modellen bepaalde verwachte kans op studiesucces. Vervolgens is onderzocht hoe groot het percentage correcte voorspellingen was bij de vier minimaal vereiste kansen op studiesucces die ook bij het opstellen van tabel 4.10 en 4.11 gehanteerd zijn.

Bij de leerlingen uit drie havo neemt het percentage correcte voorspellingen bij een minimaal vereiste kans van 0,5 toe van 77,9 naar 81,0. Bij een minimaal vereiste kans van 0,6 gaat het percentage correcte voorspellingen van 77,7 naar 80,4; bij een minimaal vereiste kans van 0,7 stijgt het van 74,4 naar 77,6 procent en is de minimaal vereiste kans 0,8, dan neemt het toe van 69,6 tot 73,5 procent. Bij de leerlingen uit drie vwo stijgt bij een minimaal vereiste kans van 0,5 het percentage correcte voorspellingen van 81,8 naar 83,5. Bij een minimaal vereiste kans van 0,6 neemt het percentage correcte voorspellingen toe van 79,9 naar 83,7; bij een minimaal vereiste kans van 0,7 stijgt het van 78,3 naar 81,5 procent en bij een minimaal vereiste kans van 0,8 van 71,6 naar 75,6 procent.

Om de vraag te beantwoorden hoe gevoelig de in tabel 4.12 gepresenteerde informatie is voor de geconstateerde verschillen tussen scholen, zijn voor de leerlingen uit drie havo en drie vwo nieuwe totaalscores op het instrumentarium bepaald. Deze nieuwe totaalscores zijn berekend door eerst voor iedere klas de waarde te bepalen van B_{0k}, het intercept voor de betreffende klas. De waarde van ieder intercept is vervolgens met 100 vermenigvuldigd en afgerond op één decimaal. Daarna zijn de aldus voor iedere klas bepaalde waarden opgeteld bij de oorspronkelijke niet afgeronde totaalscore van iedere leerling in de betreffende klas en afgerond op een geheel getal.

Tabel 4.13

Percentages succesvolle en niet succesvolle leerlingen uit drie havo en drie vwo
bij een decielverdeling op grond van de berekende multi-niveau-totaalscores

Leerlingen uit drie havo			Leerlingen uit drie vwo		
Deciel	Percentage niet succesvol	Percentage succesvol	Deciel	Percentage niet succesvol	Percentage succesvol
1 (≤ 570)	79,9	20,1	**1** (≤ 552)	79,1	20,9
2 (571- 648)	61,6	38,4	**2** (553- 626)	59,7	40,3
3 (649- 797)	42,7	57,3	**3** (627- 685)	36,4	63,6
4 (698- 737)	36,3	63,7	**4** (686- 727)	22,4	77,6
5 (738- 783)	29,7	70,3	**5** (728- 776)	15,2	84,8
6 (784- 824)	9,1	90,9	**6** (777- 815)	11,4	88,6
7 (825- 866)	7,5	92,5	**7** (816- 858)	8,0	92,0
8 (867- 914)	2,0	98,0	**8** (859- 907)	2,9	97,1
9 (915- 988)	4,8	95,2	**9** (908- 962)	1,9	98,1
10 (≥ 989)	0,0	100,0	**10** (≥ 963)	0,5	99,5
Normgroep als geheel	27,3	72,7		23,6	76,4

De leerlingen zijn op grond van de aldus vastgestelde ‘multi-niveau’-totaalscores weer gesorteerd en over decielen verdeeld. Tabel 4.13 laat opnieuw voor beide instrumentaria per deciel het percentage succesvolle en niet succesvolle leerlingen zien. In de kolommen met decielaanduidingen staat tussen haakjes weer het bij ieder deciel behorende interval van de multi-niveau-totaalscores vermeld.

Een vergelijking tussen tabel 4.12 en 4.13 laat zien dat de score-intervallen bij de corresponderende decielen niet sterk verschillen. Verder blijken ook de kansen op succes in de onderscheiden decielen niet veel af te wijken. De verwachtingstabellen veranderen dus niet sterk, wanneer er rekening

gehouden wordt met verschillen tussen scholen voor wat betreft de strengheid van beoordeling en/of de kwaliteit van de opleiding.

Door de oorspronkelijke decielscores van de leerlingen te vergelijken met hun 'multi-niveau'-decielscores ontstaat een goed beeld van de gevoeligheid van de informatie in tabel 4.12 voor de verschillen tussen scholen. De tabellen 4.14 en 4.15 bevatten de resultaten van deze vergelijking voor respectievelijk leerlingen uit drie havo en drie vwo. Aan de buitendiagonale cellen is te zien welk percentage van de leerlingen een andere decielscore krijgt, wanneer rekening gehouden wordt met de geconstateerde verschillen tussen scholen. Op de diagonaal staat steeds het percentage leerlingen dat bij de oorspronkelijke totaalscores en de multi-niveau- totaalscore dezelfde decielscore krijgt. Deze percentages zijn vet gedrukt.

De gegevens in tabel 4.14 maken duidelijk dat de oorspronkelijke decielscores van een redelijk groot aantal leerlingen uit drie havo veranderen, wanneer rekening gehouden wordt met de verschillen tussen scholen. Van de leerlingen met een oorspronkelijke decielscore van 1 krijgt bijvoorbeeld 74,6 procent wederom een decielscore van 1. Van de resterende leerlingen krijgt 22,5 procent een decielscore van 2 en 2,8 procent een decielscore van 3. De verschuivingen in de middelste decielen zijn groter dan in de buitenste. Van de leerlingen met een oorspronkelijke decielscore van 7 krijgt bijvoorbeeld slechts 20,7 procent wederom een decielscore van 7, wanneer met schoolverschillen rekening gehouden wordt. Van de overige leerlingen met een oorspronkelijke decielscore van 7 krijgt 42,8 procent een lagere en 36,6 procent een hogere decielscore.

Tabel 4.14

Vergelijking van de oorspronkelijke decielindeling met de multi-niveau-indeling

voor leerlingen uit drie havo

Multi-niveau-decielindeling	Oorspronkelijke decielindeling									
	1	2	3	4	5	6	7	8	9	10
1	**74,6%**	22,3%	3,5%							
2	22,5%	**43,9%**	21,5%	9,2%	3,3%					
3	2,8%	24,3%	**36,1%**	23,9%	8,7%	1,4%	1,4%			
4		8,8%	26,4%	**24,6%**	32,0%	6,3%	1,4%	0,7%		
5		0,7%	12,5%	29,6%	**23,3%**	26,6%	5,5%	1,4%	0,7%	
6				12,7%	18,0%	**25,9%**	34,5%	6,8%	0,7%	
7					14,7%	23,8%	**20,7%**	30,6%	8,9%	1,4%
8						16,1%	29,7%	**33,3%**	19,9%	2,8%
9							6,9%	27,2%	**45,2%**	20,7%
10									24,7%	**75,2%**

Tabel 4.15 laat zien dat ook bij een tamelijk groot deel van de leerlingen uit drie vwo de oorspronkelijk decielscores veranderen, wanneer rekening gehouden wordt met de verschillen tussen scholen. Van de leerlingen met een oorspronkelijke decielscore van 1 krijgt bijvoorbeeld 71,9 procent wederom een decielscore van 1. Van de overige leerlingen krijgt 22,9 procent een decielscore van 2, 4,3 procent een decielscore van 3 en 1,0 procent een decielscore van 4. Ook bij drie vwo-leerlingen zijn de verschuivingen in de middelste decielen omvangrijker dan in de buitenste. Van de leerlingen met een oorspronkelijke decielscore van 5 krijgt bijvoorbeeld niet meer dan 21,9 procent wederom een decielscore van 5, wanneer voor schoolverschillen gecontroleerd wordt. Van de overige leerlingen met een oorspronkelijke decielscore van 5 krijgt 42,0 procent een lagere en 36,2 procent een hogere decielscore.

De tabellen 4.14 en 4.15 tonen aan dat de informatie die de verwachtingstabellen in tabel 4.12 bevatten gevoelig is voor de verschillen tussen scholen die in de multi-niveau-analyses geconstateerd zijn. Als rekening gehouden zou

worden met het schooleffect, zou voor een deel van de leerlingen de verwachte kans op succes hoger of lager liggen.

Tabel 4.15

Vergelijking van de oorspronkelijke decielindeling met de multi-niveau-indeling

voor leerlingen uit drie vwo

Multi-niveau-decielindeling	Oorspronkelijke decielindeling									
	1	2	3	4	5	6	7	8	9	10
1	**71,9%**	24,6%	1,4%							
2	22,9%	**41,7%**	31,0%	3,8%	0,5%	0,5%				
3	4,3%	21,8%	**32,4%**	30,2%	11,0%	1,9%				
4	1,0%	6,6%	23,8%	**25,9%**	30,5%	10,7%	0,9%	0,5%		
5		5,2%	9,0%	20,3%	**21,9%**	30,1%	11,6%	2,4%		
6			1,9%	9,9%	22,9%	**24,3%**	25,5%	14,3%	1,4%	
7			0,5%	7,5%	9,0%	21,8%	**22,2%**	23,3%	15,4%	0,9%
8				2,4%	4,3%	7,8%	22,7%	**29,5%**	25,2%	7,1%
9						2,9%	13,9%	22,4%	**33,6%**	27,5%
10							3,2%	7,6%	24,3%	**64,5%**

4.6 Conclusies

Voor iedere plaatsingstoets geldt dat de items die erin zijn opgenomen dezelfde vaardigheid meten. De toetsen zijn curriculumonafhankelijk en de items in de toetsen zijn door vrijwel alle docenten die deelnamen aan de proefafnames geschikt bevonden om deel uit te maken van de plaatsingstoetsen. De inhoudsrepresentativiteit van de ontwikkelde toetsen is goed.

Verder meten de toetsen, zoals verwacht, vaardigheden waarvan de beheersing van belang is voor het succesvol vervolgen van een havo- of vwo-opleiding. Voor iedere vaardigheid geldt dat leerlingen, naarmate zij de

betreffende vaardigheid beter beheersen ook hogere afgeronde cijfers hebben op het eindrapport voor het betreffende vak. Bovendien geldt voor iedere vaardigheid dat naarmate leerlingen de betreffende vaardigheid beter beheersen, docenten ook een positiever beeld hebben van hun toekomstige prestaties. Ook geldt voor ieder vak dat leerlingen die hun havo- of vwo-opleiding zonder vertraging vervolgen vaardiger zijn dan de leerlingen die vertraging oplopen, afstromen of het onderwijs verlaten. Ten slotte bestrijken de toetsen, gezien hun onderlinge samenhang, een breed scala aan relevante vaardigheden. De begripsrepresentativiteit van de ontwikkelde toetsen is goed.

De resultaten van de logistische regressie-analyses maken duidelijk dat het voorspellend vermogen van het voor drie havo en drie vwo ontwikkelde instrumentarium beperkt is. Desondanks valt op grond van de scores op de toetsen en de checklist redelijk te voorspellen of leerlingen succesvol zullen zijn. Hoge scores geven aan dat leerlingen naar alle waarschijnlijkheid zonder vertraging door zullen stromen naar vijf havo of vijf vwo. Met behulp van de instrumenten valt echter minder goed te voorspellen of leerlingen niet succesvol zullen zijn. Lage scores op de instrumenten leiden niet zonder meer tot de conclusie dat leerlingen vertraging zullen oplopen of zullen afstromen. Leerlingen kunnen blijkbaar het gedeeltelijk ontbreken van de vaardigheden die in de toetsen en de checklisten aan de orde komen zodanig compenseren dat zij toch succesvol zijn.

Gezien het feit dat de toetsen voor Engels vrijwel geen unieke bijdrage leveren aan het voorspellend vermogen van het instrumentarium als geheel, lijkt de vraag gewettigd of niet beter volstaan had kunnen worden met de ontwikkeling van toetsen voor twee in plaats van drie vakken. Deze vraag is echter niet zonder meer bevestigend te beantwoorden, omdat de toetsen Engels wel degelijk studiesucces voorspellen. De analyses voor de modellen waarin uitsluitend de drie toetsscores zijn opgenomen, laten zien dat de toetsen Engels een unieke bijdrage leveren aan het voorspellend vermogen van de verzameling toetsen. Pas als ook de checklistscores in de modellen opgenomen worden, voegen de toetsen niets meer toe aan het voorspellend vermogen van het instrumentarium. De conclusie moet dan ook luiden dat het aantal van zes instrumenten bij nader inzien te ruim gekozen is.

De resultaten van de multi-niveau-analyse maken duidelijk dat voor beide instrumentaria geldt dat de kans op studiesucces van leerlingen, gegeven hun scores op de instrumenten, ook afhankelijk is van de specifieke school die bezocht wordt. Het eerste nadelige gevolg van dit schooleffect is dat het voorspellend vermogen van de instrumentaria erdoor afneemt. Het tweede nadelige gevolg van het schooleffect is dat de informatie in tabel 4.12 minder bruikbaar is voor een school naarmate deze een intercept heeft dat sterker afwijkt van het gemiddelde intercept.

Het bestaan van een schooleffect impliceert echter geenszins dat de ontwikkelde instrumenten ongeschikt zouden zijn voor het ondersteunen van doorstroombeslissingen. De verwachtingstabellen geven immers een indicatie van de kans op succes van een leerling met een bepaalde totaalscore op een ‘gemiddelde’ school. Als zodanig bieden ze scholen een referentiekader waar deze wel degelijk gebruik van kunnen maken bij het nemen van doorstroom-beslissingen aan het einde van het derde leerjaar van de havo en het vwo.

5 De rol van toetsen bij doorstroombeslissingen

Binnen de psychologische besliskunde is sprake van een normatieve en een descriptieve benadering. Deze twee benaderingen bieden de mogelijkheid om doorstroombeslissingen te optimaliseren, te analyseren en om aan te geven welke rol tests of toetsen bij het nemen van doorstroombeslissingen zouden kunnen spelen.

De normatieve besliskunde heeft tot doel methoden te ontwikkelen met behulp waarvan personen, groepen en instanties beslissingen kunnen nemen die vanuit ra-tioneel standpunt bezien optimaal zijn. Het eerste deel van dit hoofdstuk beschrijft kort hoe de normatieve besliskunde het nemen van beslissingen met tests kan opti-maliseren. Ter afronding van dit deel wordt gedemonstreerd op welke wijze met behulp van de normatieve besliskunde cesuren voor de totaalscores op de twee ontwikkelde instrumentaria te bepalen zijn die rationeel gezien optimaal zijn.

De descriptieve besliskunde heeft tot doel te onderzoeken op welke wijze feitelijke beslissingen, oordelen en keuzes van personen, groepen en organisaties zijn te begrijpen, te verklaren en zo mogelijk te voorspellen. In het tweede deel van dit hoofdstuk wordt verslag gedaan van onderzoek dat is uitgevoerd om antwoord te geven op de vraag welke rol toetsen in de praktijk kunnen spelen bij doorstroombeslissingen in het voortgezet onderwijs.

5.1 Beslissen met tests

Cronbach en Gleser (1965) braken in hun klassieke boek *Psychological Tests and Personnel Decisions* als eersten een lans voor de ‘psychometrische besliskunde’: het toepassen van de normatieve besliskunde bij het nemen van beslissingen met behulp van tests. In de psychometrische besliskunde is een

indeling in vier hoofdsoorten beslissingen gebruikelijk, namelijk selectie, plaatsing, classificatie en beheersing. De eerste drie soorten psychometrische beslissingen werden al door Cronbach en Gleser onderscheiden. Later is door Van der Linden (1985) de beheersingsbeslissing aan deze indeling toegevoegd.

Bij de verschillende soorten psychometrische beslissingen kunnen drie elementen onderscheiden worden:
- de test die de informatie levert op basis waarvan de beslissing genomen wordt;
- de 'behandeling' of 'behandelingen' waaruit gekozen moet worden; bijvoorbeeld een opleiding waarvoor kandidaten kunnen worden aangenomen of afgewezen, of een reeks verschillende cursussen;
- het criterium of de criteria waarmee het welslagen van de behandeling(en) te meten valt.

Bij een selectiebeslissing gaat het om het selecteren van personen voor een bepaalde behandeling op grond van hun testresultaten. Een voorbeeld van een selectiebeslissing is het al dan niet aannemen van personen voor een bepaalde opleiding op grond van hun resultaten op een toelatingsexamen.

Bij een plaatsingsbeslissing worden personen aan kwalitatief verschillende behandelingen toegewezen op grond van hun testresultaten en stelt men het welslagen van iedere behandeling vast met behulp van één en hetzelfde criterium. Van een plaatsingsbeslissing is bijvoorbeeld sprake, wanneer men besluit personen op grond van hun toetsprestaties individueel of klassikaal onderwijs te laten volgen en na afronding van het onderwijs aan alle personen hetzelfde afsluitende examen geeft.

Bij een classificatiebeslissing wijst men eveneens personen toe aan kwalitatief verschillende behandelingen op grond van hun testresultaten. Alleen wordt bij een classificatiebeslissing het succes van de onderscheiden behandelingen vastgesteld met behulp van verschillende criteria. Regionale opleidingscentra bieden bijvoorbeeld cursussen aan die betrekking hebben op dezelfde vaardigheid, maar die verschillen van niveau en met een ander examen worden afgerond. Wanneer personen die een cursus willen volgen bij een regionaal opleidingscentrum op grond van hun prestaties op een toets aan

verschillende niveaus van de cursus toegewezen worden, is sprake van een classificatiebeslissing.

Deze drie soorten beslissingen verschillen in een belangrijk opzicht van de beheersingsbeslissing. Bij selectie-, classificatie- en plaatsingsbeslissingen is het criterium extern en bij beheersingsbeslissingen niet. Bij de beheersingsbeslissing wordt met behulp van een test vastgesteld of een behandeling voldoende effect heeft gesorteerd. Test en criterium vallen samen, in die zin dat de testscore een met meetfouten behepte weergave van de criteriumscore is. Het criterium bij de beheersingsbeslissing is vanuit het standpunt van de klassieke testtheorie bezien de ware score op de test en vanuit het perspectief van de itemresponstheorie de latente vaardigheid. Een voorbeeld van een beheersingsbeslissing is het vaststellen of personen een opleiding al dan niet met succes hebben afgerond op grond van hun scores op een afsluitend examen.

Bovenstaande indeling in soorten beslissingen maakt duidelijk dat vanuit de psychometrische besliskunde bezien de term ‘plaatsingstoets’ een niet geheel juiste benaming is voor de toetsen die het onderwerp zijn van dit proefschrift. Een juistere term zou ‘selectietoetsen’ zijn. Bij de naamgeving van de toetsen is echter niet uitgegaan van de wijze waarop plaatsing in de psychometrische besliskunde wordt omschreven. Het woord plaatsing in de term plaatsingstoets heeft een meer generieke betekenis (vgl. Nederlands Instituut van Psychologen, 1988). Zo is in het verleden de term plaatsingstoets ook gebruikt voor toetsen die dienden ter ondersteuning van determinatiebeslissingen (Sluijter, Boertien, De Klijn & Van Roosmalen, 1991), hoewel voor dit laatste type toets de term ‘classificatietoets’ volgens de psychometrische besliskunde een betere benaming zou zijn.

Psychometrische besliskunde

Uit de tabellen 4.10 en 4.11 in het voorgaande hoofdstuk bleek dat de frequenties van de vier mogelijke uitkomsten van doorstroombeslissingen aan het einde van drie havo en drie vwo afhankelijk zijn van de met behulp van de instrumentaria bepaalde verwachte kans op succes die minimaal vereist is. De minimaal vereiste verwachte kans op succes kan vertaald worden naar

een minimaal vereiste totaal-score op het instrumentarium. Er bestaat immers een door het logistische regressiemodel gegeven directe relatie tussen de verwachte kans op studiesucces en de totaalscore op het instrumentarium. Naarmate de minimaal vereiste totaalscore - die gewoonlijk met de term *cesuur* aangeduid wordt - hoger ligt, neemt het aantal leerlingen toe van wie op grond van de totaalscore op het instrumentarium correct voorspeld wordt dat ze **niet** succesvol zullen zijn. Maar dit gaat natuurlijk ten koste van het aantal leerlingen van wie correct voorspeld wordt dat ze succesvol zullen zijn. Verder neemt het aantal leerlingen af van wie **onterecht** voorspeld wordt dat ze succesvol zullen zijn. Maar dit gaat uiteraard weer ten koste van het aantal leerlingen van wie **onterecht** voorspeld wordt dat ze **niet** succesvol zullen zijn.

De uitkomsten van doorstroombeslissingen zijn afhankelijk van de voorkeuren die men heeft voor de verschillende beslissingsuitkomsten. Wanneer een school zich zou baseren op de scores van leerlingen op de ontwikkelde instrumentaria bij het nemen van doorstroombeslissingen, dan zou de voorkeur van de school zo goed mogelijk tot uiting moeten komen in de cesuur die men kiest. Heeft een school het streven te voorkomen dat leerlingen na drie havo of drie vwo zullen afstromen, blijven zitten, of afhaken, dan zal deze school voor een relatief hoge cesuur moeten kiezen. Hecht een school er meer waarde aan om ook zwakkere leerlingen in drie havo of vwo de kans te bieden om hun opleiding zonder vertraging af te ronden, dan zal de betreffende school daarentegen voor een relatief lage cesuur moeten kiezen. De psychometrische besliskunde biedt een formele methode om cesuren op tests zodanig te kiezen dat deze de voorkeuren van de ‘beslisser’ - een rationeel handelende persoon, groep of instantie - optimaal weerspiegelen.

Om psychometrische beslissingen te kunnen optimaliseren, is het nodig de voorkeuren van de beslisser voor de mogelijke beslissingsuitkomsten te kwantificeren. Bij bepaalde soorten beslissingen kunnen voorkeuren in een objectieve eenheid, zoals geld, worden uitgedrukt. Maar bij psychometrische beslissingen is dat niet het geval. Om de voorkeuren van beslissers te representeren, maakt de psychometrische besliskunde gebruik van het begrip ‘utiliteit’. Dit begrip geeft aan welke subjectieve waarde de beslisser aan een uitkomst hecht. Bij psychometrische beslissingen worden de subjectieve

waarden die de beslisser, gegeven een bepaalde actie, hecht aan de criteriumscores vastgelegd in een zogeheten utiliteitsfunctie.

Bij psychometrische beslissingen heeft de beslisser de keuze uit twee of meer acties. Iedere actie heeft een aantal uitkomsten. Welke uitkomst een actie bij een bepaald persoon heeft, is afhankelijk van de criteriumscore van de betreffende persoon. Over de uitkomst van de acties waaruit de beslisser kan kiezen, bestaat onzekerheid, omdat tests niet meer dan een feilbare indicatie geven van de prestaties van personen op het criterium. Wanneer de utiliteitsfuncties voor de mogelijke acties en de simultane kansverdeling van test- en criteriumscores bekend zijn, kan voor iedere mogelijke cesuur de verwachte utiliteit berekend worden. Een rationele beslisser dient dan te kiezen voor de cesuur die de hoogste verwachte utiliteit oplevert.

Utiliteitsfuncties

Er bestaan verschillende typen utiliteitsfuncties. De simpelste is de drempelutiliteitsfunctie (Hambleton & Novick, 1973). Deze veronderstelt dat de beslisser een constante subjectieve waarde hecht aan elk van de mogelijke beslissingsuitkomsten. Drempelutiliteitsfuncties zijn discontinu en daarom veelal weinig realistisch. Bij een selectiebeslissing bijvoorbeeld, kent de utiliteit van de actie ‘aannemen’ maar twee waarden. De eerste (lagere) waarde geldt voor alle criteriumscores die lager zijn dan de score op het criterium die minimaal benodigd is om de actie ‘aannemen’ terecht te vinden. Bij deze minimaal benodigde criteriumscore wijzigt de waarde van de utiliteit zich en deze blijft vervolgens constant voor alle criteriumscores die hoger zijn dan de minimaal vereiste.

Een meer realistische aanname bij een selectiebeslissing is dat er bij de actie ‘aannemen’ een monotoon stijgend verband bestaat tussen de waarde van de utiliteit en de criteriumscore en bij de actie ‘afwijzen’ een monotoon dalend verband. Een beter alternatief voor de drempelutiliteitsfunctie is daarom de lineaire (Van der Linden & Mellenbergh 1977). Bij een lineaire utiliteitsfunctie bestaat er bij iedere behandeling een lineair verband tussen de utiliteit en criteriumscores. Bij een selectiebeslissing bijvoorbeeld, stijgt de utiliteit van de actie ‘aannemen’ naarmate de criteriumscore toeneemt, terwijl de utiliteit

van de actie ‘afwijzen’ daalt. Bij de lineaire utiliteitsfunctie is dan ook geen sprake van discontinuïteit, zoals bij drempelutiliteit.

Een nog realistischer functie is de normaalogief (Novick & Lindley, 1978). Bij een normaalogieve utiliteitsfunctie bestaat er een kromlijnig verband tussen de utiliteit en de criteriumscores. Bij een selectiebeslissing bijvoorbeeld, stijgt de utiliteit van de actie ‘aannemen’ met het toenemen van de criteriumscore eerst zeer langzaam, maar klimt steeds sneller, naarmate criteriumscores minder extreem worden. Worden criteriumscores vervolgens weer extremer, dan neemt de utiliteit langzamer toe, om uiteindelijk zeer langzaam zijn maximum te naderen. Voor de utiliteit van de actie ‘afwijzen’ geldt exact het omgekeerde. De utiliteit neemt eerst zeer langzaam af met de criteriumscore, maar daalt daarna steeds sneller om vervolgens weer langzamer af te nemen en tenslotte zeer langzaam zijn minimum te naderen.

Een bezwaar tegen drempel-, lineaire en normaalogieve utiliteitsfuncties is dat onduidelijk is in hoeverre ze de subjectieve waarden die beslissers hechten aan beslissingsuitkomsten werkelijk representeren. Empirische utiliteitsfuncties die met behulp van een elicitatieprocedure aan beslissers ontlokt zijn, kennen dit bezwaar niet. Onderzoek op dit terrein (Vrijhof, Mellenbergh & Van den Brink, 1982; Van der Gaag, Mellenbergh & Van den Brink, 1987; Van der Gaag, 1990; Lamers, Van der Gaag & Mellenbergh, 1992) laat zien dat empirische utiliteitsfuncties in de regel de subjectieve waardering van beslissers voor verschillende uitkomsten goed representeren. Verder maken deze onderzoeken duidelijk dat lineaire en normaalogieve utiliteitsfuncties een goede benadering geven van empirische utiliteitsfuncties. Een bezwaar tegen het gebruik van empirische utiliteitsfuncties is echter dat het ontlokken ervan een complex en tijdrovend proces is.

Complexere vormen van psychometrische beslissingen

Naast de vier beschreven soorten psychometrische beslissingen zijn er uiteraard ook meer complexe psychometrische beslissingen mogelijk. In de eerste plaats kunnen beslissingen gebaseerd zijn op meer dan één test. Het ligt natuurlijk voor de hand om beslissingen over personen te baseren op hun resultaten op verschillende tests die elk betrekking hebben op een voor de

beslissing relevante vaardigheid. In de tweede plaats kan er sprake zijn van een reeks criteria die ieder betrekking hebben op het welslagen van een ander aspect van een behandeling. In de derde plaats kan een psychometrische beslissing betrekking hebben op personen die afkomstig zijn uit verschillende populaties. Zijn personen afkomstig uit een meerderheid- en een minderheid-populatie, dan speelt het probleem van de 'cultuur-eerlijkheid'. In dat geval kunnen de utiliteitsfuncties per populatie verschillen en kunnen er andere beslisregels gelden voor de verschillende populaties. Terzijde zij opgemerkt dat alleen de psychometrische besliskunde een acceptabele methode biedt voor het oplossen van dit probleem (Petersen & Novick, 1976). Gross en Su (1975) en Mellenbergh en Van der Linden (1981) laten zien hoe de psychometrische besliskunde toegepast kan worden bij cultuur-eerlijke selectie. In de vierde plaats kan sprake zijn van quota-restricties: beperkingen voor wat betreft het aantal aan een behandeling toe te wijzen personen.

Ook kunnen de vier beschreven hoofdsoorten psychometrische beslissingen in combinaties voorkomen. Zo is bij een plaatsingsbeslissing waarbij een deel van de leerlingen voor geen van de behandelingen in aanmerking komt, formeel sprake van een combinatie van een plaatsing- en selectiebeslissing. Wanneer succes bepaald wordt aan de hand van een grenswaarde op een criterium, zal bij selectie-, plaatsing- en classificatiebeslissingen welbeschouwd vaak sprake zijn van een combinatie van deze beslissingen met een beheersingsbeslissing (Van der Linden, 1990). Normaal gesproken zal het criterium namelijk meetfouten bevatten en zal derhalve na de behandeling formeel gezien een beheersingsbeslissing plaats moeten vinden. Ten slotte kan ook sprake zijn, bijvoorbeeld bij individuele studiesystemen, van reeksen al dan niet verschillende psychometrische beslissingen die elkaar in de tijd opvolgen. Van der Linden (1998) geeft bijvoorbeeld aan hoe optimale cesuren kunnen worden vastgesteld bij een gecombineerde plaatsings- en beheersingsbeslissing. Vos (1994) geeft aan op welke wijze optimale beslisregels voor reeksen van psychometrische beslissingen te vinden zijn.

Uitwerking voor de ontwikkelde instrumentaria

Ook voor de voor drie havo en drie vwo ontwikkelde instrumentaria is het mogelijk om optimale cesuren te bepalen bij de totaalscores. Deze cesuren gelden uiteraard alleen indien scholen bij het nemen van doorstromingsbeslissingen *uitsluitend* gebruik zouden maken van deze instrumentaria. In dat geval zou in psychometrisch-besliskundige termen sprake zijn van selectiebeslissingen op grond van meer dan een test. Van der Linden (1990) geeft een voorbeeld van een toepassing van de psychometrische besliskunde bij beslissingen op grond van meer dan één test. In zijn voorbeeld is sprake van conjuncte beslisregels. Dit houdt in dat personen op iedere test boven een bepaalde cesuur moeten scoren om aan een bepaalde behandeling toegewezen te worden. Deze benadering is niet gevolgd bij de instrumentaria die ontwikkeld zijn ter ondersteuning van doorstroombeslissingen aan het einde van drie havo en drie vwo. In ons geval zijn, zoals in hoofdstuk vier beschreven is, de scores op de verschillende instrumenten via een gewogen sommering omgezet naar een totaalscore. Dat impliceert dat er van uitgegaan is dat leerlingen een lage score op het ene instrument kunnen compenseren met een hogere score op een ander instrument. Formeel beschouwd is daarom in dit geval sprake van een selectiebeslissing op grond van één test, wat, gezien de aard van de beslissing, een realistische opvatting is.

Het criterium voor studiesucces is in ons geval dichotoom. Het criterium geeft aan of leerlingen zich in het schooljaar 1995-1996 al dan niet in het vijfde leerjaar van havo of vwo bevonden. Omdat het criterium dichotoom is, moet gekozen worden voor een drempelutiliteitsfunctie. Voor de verwachte utiliteit $\mathcal{E}(u(t_c))$ bij een cesuur t_c op de totaalscore voor het instrumentarium geldt bij het hanteren van een drempelutiliteitsfunctie:

$$\mathcal{E}(u(t_c)) = P_{00}u_{00} + P_{01}u_{01} + P_{10}u_{10} + P_{11}u_{11}, \qquad (5.1)$$

waarbij
P_{00} de kans is op het terecht afwijzen van een leerling en u_{00} de utiliteit die bij deze beslissingsuitkomst hoort;
P_{01} de kans is op het *onterecht* afwijzen van een leerling en u_{01} de utiliteit die bij deze beslissingsuitkomst hoort;
P_{10} de kans is op het *onterecht* aannemen van een leerling en u_{10} de utiliteit die bij deze beslissingsuitkomst hoort;

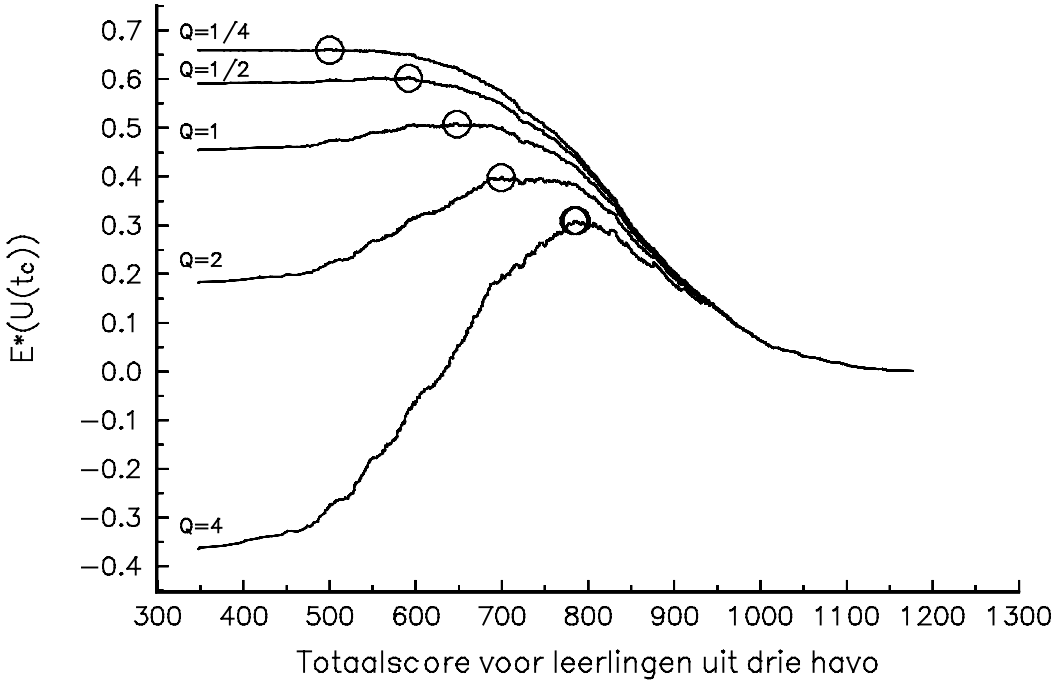
P_{11} de kans is op het terecht aannemen van een leerling en u_{11} de utiliteit die bij deze beslissingsuitkomst hoort;

Het is niet nodig dat de waarden van de afzonderlijke utiliteiten bekend zijn om het maximum van uitdrukking (5.1) te bepalen. Het valt aan te tonen (zie bijvoorbeeld Mellenbergh, 1993) dat het maximaliseren van uitdrukking (5.1) hetzelfde resultaat oplevert als het maximaliseren van de alternatieve uitdrukking:

$$\mathscr{E}^*(u(t_c)) = \frac{-(u_{00} - u_{10})}{u_{11} - u_{01}} P_{10} + P_{11} = -Q P_{10} + P_{11}, \qquad (u_{11} > u_{01}).$$

De waarde van u_{00} - u_{10} is het verschil in utiliteit tussen het afwijzen en laten doorstromen van een ongeschikte leerling. De waarde van u_{11} - u_{01} is het verschil in utiliteit tussen het laten doorstromen en afwijzen van een geschikte leerling. Als Q gelijk aan 1 is, dan zijn beide verschillen gelijk aan elkaar. Vindt een school het relatief belangrijker om ongeschikte leerlingen af te wijzen dan om geschikte leerlingen te laten doorstromen, dan is Q voor deze school groter dan 1. Vindt een school het daarentegen belangrijker om geschikte leerlingen door te laten stromen dan om ongeschikte leerlingen af te wijzen, dan is Q voor deze school kleiner dan 1. Hoe strenger een school is, des te groter is Q en des te hoger is de cesuur.

Figuur 5.1 laat voor de totaalscores van de leerlingen uit drie havo zien welke waarden $\mathscr{E}^*(u(t_c))$ kan aannemen voor vijf waarden van Q. Figuur 5.2 doet hetzelfde voor de totaalscores van de leerlingen uit drie vwo. Bij Q = 1/4 is de optimale cesuur voor de leerlingen uit drie havo de score 500. Is Q = 1/2 dan ligt de optimale cesuur bij de score 592 en heeft Q de waarde 1, dan ligt de optimale cesuur bij de score 648. Is Q = 2, dan is de optimale cesuur de score 700. Bij Q = 4 bereikt $\mathscr{E}^*(u(t_c))$ de maximale waarde bij de scores 784 en 787. De maxima van $\mathscr{E}^*(u(t_c))$ voor de verschillende waarden van Q zijn in figuur 5.1 omcirkeld.



Figuur 5.1;

Waarden van $\mathscr{E}^*(u(t_c))$ voor de totaalscores van leerlingen uit drie havo voor

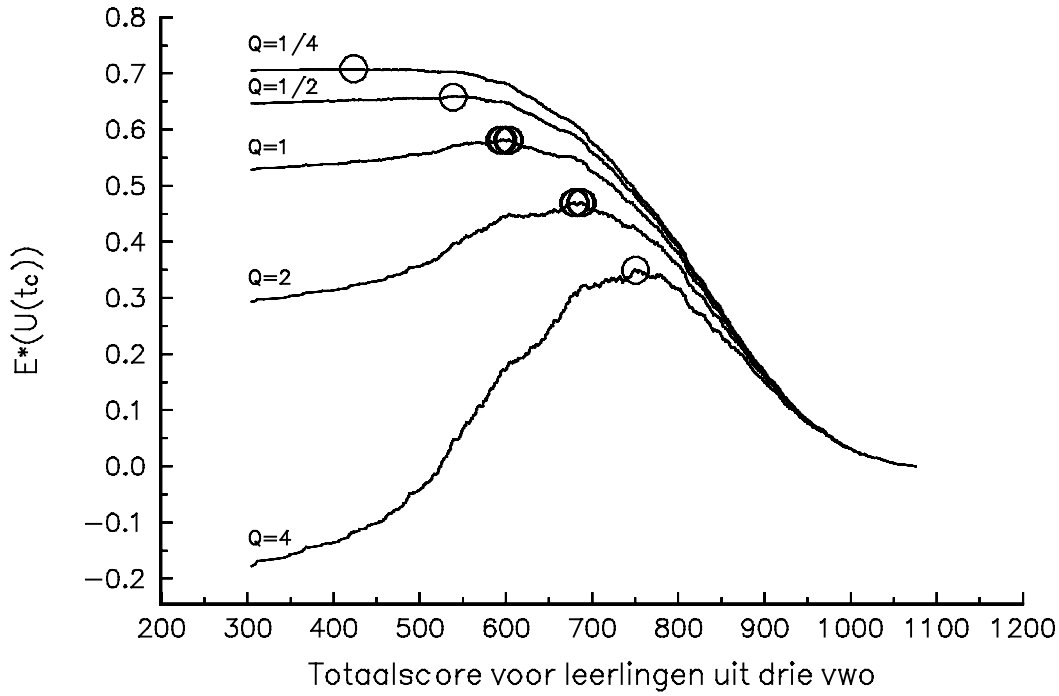
Q = 1/4, 1/2, 1, 2 en 4. De maxima voor de verschillende waarden van Q zijn

omcirkeld.

Voor de leerlingen uit drie vwo ligt de optimale cesuur op het betreffende instrumentarium voor Q = 1/4 bij de score 423. Is Q = 1/2, dan ligt de optimale cesuur bij de score 539. Is Q = 1, dan bereikt $\mathscr{E}^*(u(t_c))$ de grootste waarde bij maar liefst vier scores: 594, 597, 603 en 604. Bij Q = 2 zijn er wederom vier scores waarvoor $\mathscr{E}^*(u(t_c))$ maximaal is. Het betreft hier de scores 679, 680, 686 en 688. Bij Q = 4 ligt de optimale cesuur bij de score 750. Ook in figuur 5.2 zijn de maxima van $\mathscr{E}^*(u(t_c))$ voor de verschillende waarden van Q omcirkeld.

Beide figuren laten zien dat de optimale cesuur bij een hogere score komt te liggen naarmate Q toeneemt. Verder maken ze duidelijk dat er relatief kleine verschillen zijn tussen de waarden van $\mathscr{E}^*(u(t_c))$ bij verschillende totaalsco-

res. Bij een aantal waarden van *Q* is zelfs sprake van meer dan één maximum. Door de relatief kleine verschillen heeft het weinig zin om optimale cesuren te rapporteren. De gevonden
waarden van de optimale cesuren zijn namelijk, vanwege de toevallige samenstelling van de onderzoeksgroepen, niet generaliseerbaar. Er is een grote kans dat er bij een iets andere samenstelling van de onderzoeksgroepen andere optimale cesuren gevonden zouden worden. De figuren laten zien dat, naarmate de waarde



Figuur 5.2;

Waarden van $\mathscr{E}^*(u(t_c))$ voor de totaalscores van leerlingen uit drie vwo voor

Q = 1/4, 1/2, 1, 2 en 4. De maxima voor de verschillende waarden van Q zijn

omcirkeld.

van *Q* lager is, het gebrek aan generaliseerbaarheid van de optimale cesuren sterker is. Het scorebereik waarbinnen sprake is van relatief kleine

verschillen neemt af naarmate de waarde van Q toeneemt. Door het gebrek aan generaliseerbaarheid is er vanaf gezien om in de handleidingen bij de instrumentaria informatie over cesuren op te nemen.

5.2 De rol van toetsen bij docentbeslissingen

In de voorgaande paragraaf is beschreven op welke wijze voor de totaalscores op de ontwikkelde instrumentaria cesuren zijn te vinden die vanuit normatief beslis-kundig oogpunt bezien optimaal zijn. De rest van dit hoofdstuk heeft betrekking op de algemene vraag hoe docenten in de praktijk te werk gaan bij het nemen van doorstroombeslissingen en welke rol toetsen daarbij kunnen spelen. Deze paragraaf bevat een kort overzicht van onderzoek waarin de rol van toetsen bij de totstand-koming van docentbeslissingen ter sprake komt. Ook komt de vraag aan de orde of uit de onderzoeksresultaten conclusies zijn te trekken omtrent de rol die de ontwikkelde instrumentaria in de praktijk zouden kunnen spelen bij doorstroombeslissingen.

Onderzoeken naar de invloed van gestandaardiseerde toetsen op docentbeslissingen rapporteren dat docenten toetsinformatie gebruiken om al gevormde oordelen mee te vergelijken. Docenten schijnen hun leerlingen het ‘voordeel van de twijfel’ te gunnen, wanneer zij van toetsresultaten gebruik maken om oordelen te verifiëren. Dit houdt in dat zij hun oordeel niet veranderen, indien de toetsuitslag lager uitvalt dan verwacht. Valt de toetsuitslag echter hoger uit dan verwacht, dan bestaat de kans dat docenten hun oordeel in positieve richting herzien (Salmon-Cox, 1982; Kellaghan, Madaus & Airasian, 1982; Janssens, 1986). Verder blijken toetsresultaten geen belangrijke rol te spelen bij het nemen van onderwijskundige beslissingen (Kellaghan, Madaus & Airasian, 1982; Janssens, 1986).

Het is echter de vraag in hoeverre de conclusies uit onderzoek naar de invloed van gestandaardiseerde toetsen gewettigd zijn. Bij veel van de betreffende onderzoeken is voor het verzamelen van gegevens namelijk gebruik gemaakt van vragenlijsten en al dan niet gestructureerde interviews. Besliskundig onderzoek heeft echter aangetoond dat personen meestal niet

goed in staat zijn om aan te geven in hoeverre verschillende gegevens hun beslissingen beïnvloeden (Slovic & Lichtenstein, 1971; Brehmer & Brehmer, 1988). Beslissers blijken het aantal gegevens dat daadwerkelijk een rol speelt, te overschatten en zijn in de regel niet in staat om aan te geven hoe sterk verschillende gegevens meespelen. Onderzoek via vragenlijsten en interviews is daarom niet toereikend voor het trekken van valide conclusies over de vorm en inhoud van beslisprocessen. Op grond van de resultaten van dergelijke onderzoeken valt geen antwoord te geven op de vraag welke rol de ontwikkelde instrumentaria zouden kunnen spelen bij doorstroombeslissingen aan het einde van drie havo en drie vwo.

De kritiek die op onderzoek via vragenlijsten en interviews te geven is, geldt niet voor het grootschalige onderzoek dat in Ierland gedaan is naar het effect van het gebruik van gestandaardiseerde toetsen op docentbeslissingen (Airasian, Kellaghan, Madaus & Pedulla, 1977; Kellaghan, Madaus & Airasian, 1982). Dit onderzoek neemt een uitzonderlijke plaats in tussen al het andere onderzoek naar het effect van toetsen, omdat in Ierland docenten en leerlingen geen enkele ervaring met het gebruik van gestandaardiseerde toetsen hadden vóór de introductie ervan. Daardoor was men in de gelegenheid om goed gecontroleerd, bijkans experimenteel, onderzoek te doen naar het effect van het gebruik van toetsen op verschillende soorten beslissingen. De onderzoekers concluderen dat de invloed van toetsen op de beslissingen van docenten niet groot is. Géén van de beslissingen die ter sprake komen heeft echter tot doel te bepalen voor welk onderwijstraject leerlingen het meest geschikt zijn. Ook de resultaten van dit onderzoek maken het daarom niet mogelijk een duidelijke uitspraak te doen over de invloed die de ontwikkelde instrumentaria zouden kunnen uitoefenen op doorstroombeslissingen aan het einde van drie havo en drie vwo.

Een binnen de descriptieve besliskunde veel gehanteerde onderzoeksbenadering richt zich op het in kaart brengen van de samenhang tussen aan personen gepresenteerde gegevens en de beslissingen of oordelen die daaruit voortvloeien. Deze zogeheten structuurgerichte benadering heeft zijn oorsprong in het werk van Brunswik (1952; 1955), Meehl (1954) en Hammond (1955). De kritiek op onderzoek via vragenlijsten en interviews geldt niet voor besliskundig onderzoek dat gebruik maakt van deze benadering.

Besliskundig structuurgericht onderzoek heeft tot doel vergelijkingen op te stellen met behulp van statistische analysetechnieken. Deze vergelijkingen zijn in de regel lineair van aard en beschrijven de samenhang tussen de gegevens die ter beschikking staan en de genomen beslissingen. In gevallen waar een objectief criterium bestaat, is het mogelijk om de juistheid van een beslissing te bepalen en om vergelijkingen op te stellen die de stand van zaken in de werkelijkheid beschrijven. Het model dat de samenhang beschrijft tussen gegevens en oordelen of beslissingen enerzijds en dezelfde gegevens en het objectieve criterium anderzijds, noemt men het 'lensmodel' (Brunswik, 1955; Hammond, Stewart, Brehmer & Steinman, 1975). Omdat objectieve criteria veelal ontbreken, kent het lensmodel maar weinig praktische toepassingen. In de praktijk heeft veel onderzoek binnen de structuurgerichte benadering daarom het karakter van beslissings- of beoordelingsanalyse. Voor dit type onderzoek wordt in de Engelstalige literatuur vaak de term 'policy capturing' (Hammond, 1955; Hoffman, 1960) gebruikt, omdat de beoordelingsanalyse de strategie van een beslisser als het ware 'vangt' in een vergelijking.

Structuurgerichte lineaire modellen hebben in de regel een goede passing op de onderzoeksgegevens. De reden hiervan zou hun robuustheid kunnen zijn. Een lineair model geeft namelijk ook een goede benadering van de uitkomsten van een proces van besluitvorming, wanneer sprake is van relaties die niet strikt lineair zijn (Dawes & Corrigan, 1974). De reden zou echter ook kunnen zijn dat bij veel beslisprocessen daadwerkelijk sprake is van lineaire relaties tussen beslissingen en de gegevens waar die beslissingen op gebaseerd zijn. (Hoffman, 1960; Slovic & Lichtenstein, 1971; Brehmer, 1980; Brehmer & Brehmer, 1988; Brehmer, 1994). Er zijn echter aanwijzingen dat deze laatste veronderstelling niet juist is (Ganzach, 1995; Ganzach & Kzaczkes, 1995).

Op de bruikbaarheid van structurele modellen voor het doen van onderzoek naar beslissingen binnen het onderwijs is meermaals gewezen (Snow, 1968; Shulman & Elstein, 1975; Cooksey & Freebody, 1986; Shavelson, Webb & Burstein, 1986). Desondanks maken overzichten van Shavelson en Stern (1981) en Cooksey (1988) duidelijk dat er niet veel onderwijskundig onderzoek bestaat dat gebruik maakt van structurele modellen. Het onderzoek dat is gedaan, heeft bovendien, met enkele uitzonderingen

(Maniscalco, Doherty & Ullman, 1980; Johnson & Doherty, 1983), geen betrekking op beslissingen ten aanzien van doorstroming van leerlingen of studenten.

In de literatuur is slechts één relevant voorbeeld gevonden van structuurgericht onderzoek waarin expliciet aandacht is voor de rol die toetsen spelen bij doorstroombeslissingen (Hoogstraten & Mellenbergh, 1978). Het betreft een onderzoek naar doorstroombeslissingen van docenten bij de overgang van hun leerlingen van het basisonderwijs naar het voortgezet onderwijs. Het betreffende artikel doet verslag van een experiment waarin 40 onderwijzers van een aantal fictieve leer-lingen aan moesten geven of deze het meest geschikt waren voor vwo, havo, mavo, of lavo/lbo. In dit onderzoek is de invloed onderzocht op het oordeel van de onderwijzers van vier factoren, waaronder de score op een toets. De eigenschappen van de leerling, omschreven als ' *werklust, doorzettingsvermogen, omgang met klasgenoten, intellectuele capaciteiten en schoolprestaties*' bleken de oordelen van de docenten het meest te beïnvloeden. Toetsprestaties bleken na de eigenschappen van de leerling de meeste invloed uit te oefenen. De 'omstandigheden thuis' bleken de op twee na belangrijkste factor, terwijl het 'sociale niveau' van de school er het minst toe deed. Hoogstraten en Mellenbergh rapporteren dat de gevonden resultaten niet stroken met de resultaten van een bevraging van dezelfde docenten. In de bevraging gaven de onderwijzers namelijk toetsresultaten een veel minder prominente rol dan in het experiment. De door Hoogstraten en Mellenbergh geconstateerde discrepantie is mogelijk een illustratie van het al eerder gememoreerde feit dat mensen niet goed in staat zijn aan te geven in hoeverre verschillende gegevens hun beslissingen beïnvloeden (Slovic & Lichtenstein, 1971; Brehmer & Brehmer, 1988).

Het bovenstaande leert dat de literatuur geen duidelijke conclusie toelaat over de rol die de ontwikkelde instrumentaria zouden kunnen spelen bij doorstroombeslissingen aan het einde van drie havo en drie vwo. Onderzoek met behulp van vra-genlijsten en gestructureerde interviews geeft aan dat toetsresultaten geen belangrijke rol spelen, maar het is de vraag of de conclusies uit dergelijk onderzoek valide zijn. Over structuurgericht besliskundig onderzoek naar de rol van toetsen bij doorstroombeslissingen dat wel tot valide conclusies leidt, wordt nauwelijks gerapporteerd.

Er bestaat echter onderzoek waarover nog niet gerapporteerd is en dat wat meer uitsluitsel kan geven over de rol die de ontwikkelde instrumentaria in de praktijk zouden kunnen spelen. Dit onderzoek is een aantal jaren geleden verricht rond de door het Cito (1993a, 1993b, 1993c, 1993d, 1993e, 1993f) ontwikkelde plaatsingstoetsen voor het einde van het eerste leerjaar van het voortgezet onderwijs. Dit onderzoek komt in de komende paragraaf aan de orde. Het richt zich op doorstroombeslissingen aan het einde van het eerste leerjaar van het voortgezet onderwijs en de rol die toetsen daarbij spelen.

5.3 Beoordelingsanalyse van determinatiebeslissingen

Aan het onderzoek naar doorstroombeslissingen aan het einde van het eerste leerjaar ging een kleinschalig vooronderzoek vooraf. In dit vooronderzoek is geïnventariseerd welke kenmerken van leerlingen docenten van belang achten bij de determinatie. Het vooronderzoek geeft tevens een beeld van de overeenstemming tussen docenten, wanneer deze onafhankelijk van elkaar aan het einde van de brugperiode leerlingen moeten determineren.

Het eigenlijke onderzoek had een drieledig doel. In de eerste plaats geeft ook dit onderzoek een overzicht van de mate van overeenstemming tussen docenten, wanneer deze onafhankelijk van elkaar leerlingen moeten determineren. In de tweede plaats beschrijft het onderzoek in welke mate bepaalde leerlingkenmerken de uitkomsten beïnvloeden van de determinatiebeslissingen van iedere docent die deelnam. En in de derde plaats brengt het onderzoek in kaart welke rol de prestaties van leerlingen op plaatsingstoetsen spelen bij het tot stand komen van determinatiebeslissingen. Het onderzoek richt zich op beslissingen van individuele docenten, omdat de structurele modellen voor verschillende personen veelal een andere vorm hebben. Het uitvoeren van analyses op groepsniveau is pas zinvol, wanneer is vastgesteld in welke mate er sprake is van verschillen tussen individuen (vgl. Borko & Cadwell, 1982; Cadwell, 1980; Cooksey, 1988).

5.3.1 Het vooronderzoek

Bij het nemen van doorstroombeslissingen kan een scala aan gegevens een rol spelen. Docenten zullen beslissingen nemen op basis van verzamelingen van al dan niet expliciet geformuleerde criteria die van docent tot docent kunnen verschillen. Om meer duidelijkheid te krijgen over de gegevens die een rol spelen bij doorstroombeslissingen werd daarom een vooronderzoek gehouden op een drietal scholen (Van Dijk & Van 't Land 1990). Het betrof hier twee scholengemeenschappen voor mavo, havo en vwo en een scholengemeenschap voor vbo en mavo. Hier zal alleen over de resultaten op de eerste twee scholen gerapporteerd worden. De eerste taak die iedere deelnemende docent moest uitvoeren, was het kiezen van de meest geschikte categoriale vorm van onderwijs voor een aantal van zijn of haar leerlingen. De tweede taak betrof het aangeven van de leerlingkenmerken die de docent voor het nemen van deze beslissingen van belang vond. De derde taak van iedere docent bestond uit het toekennen van scores aan de betreffende leerlingen op de zes kenmerken die hij of zij het meest relevant vond.

Opzet en uitvoering van het vooronderzoek

De drie hierboven beschreven taken waren in de tijd gescheiden. Tussen het verzamelen van de determinatiebeslissingen van de docenten en het bepalen van de kenmerken die iedere docent belangrijk vond bij het determineren lag een periode van een week. En tussen deze tweede taak en het scoren van leerlingen op de kenmerken lag een periode van een maand. Door het in de tijd gescheiden houden van het nemen van beslissingen en het scoren van leerlingen op relevant geachte kenmerken werd voorkomen dat docenten op oneigenlijke wijze een hoge consistentie konden bereiken tussen hun beslissingen over leerlingen en de scores die zij aan deze leerlingen op de kenmerken toekenden. Voor de aangebrachte fasering was ook nog een praktische reden: het in één keer uitvoeren van alle taken zou veel inspanning van de deelnemers hebben gevergd.

In totaal voerden op elk van beide scholen negen docenten de drie taken volledig uit. Op de ene school bedroeg het totaal aantal leerlingen over wie docenten uitspraken konden doen 47, verdeeld over twee klassen met respectievelijk 23 en 24 leerlingen. Drie van de negen docenten gaven les aan beide klassen; drie docenten gaven alleen aan de ene klas les en eveneens

drie alleen aan de andere. Op de andere school konden docenten uitspraken doen over maximaal 33 leerlingen, afkomstig uit een en dezelfde klas.

Bij de eerste taak kregen de docenten een boekje voorgelegd met daarin een reeks namen van leerlingen. De namen van de leerlingen waren in een eerder stadium via de schooladministratie verzameld en in de boekjes afgedrukt. De docenten moesten in de boekjes aangeven voor welke vorm van categoriaal onderwijs elke leerling het meest geschikt was. De docenten konden daarbij kiezen uit de onderwijsvormen mavo, havo en vwo.

Bij de tweede taak ontvingen de docenten een boekje waarin 21 mogelijk relevante leerlingkenmerken genoemd werden en kort omschreven. Deze reeks leerlingkenmerken was samengesteld op grond van een literatuurstudie (Van Dijk & Van 't Land, 1990). De docenten moesten in het boekje aangeven welke van de 21 leerlingkenmerken een rol speelden bij het tot stand komen van hun eigen beslissingen. Tevens moesten zij aan ieder geselecteerd kenmerk een rangnummer geven dat de mate van belang ervan weerspiegelde. Om te bewerkstelligen dat docenten deze opdracht onafhankelijk van elkaar zouden verrichten, is deze op iedere school onder supervisie van twee proefleiders uitgevoerd. Uiteraard bood het boekje docenten ook de mogelijkheid om leerlingkenmerken in te vullen en te omschrijven die niet in de reeks voorkwamen. Tabel 5.1 geeft een overzicht van de 21 kenmerken en de daarbij behorende omschrijvingen.

Tabel 5.1
Overzicht van de aan docenten voorgelegde leerlingkenmerken

Leerlingkenmerk	Omschrijving
Begeleiding door de ouders	mate waarin de ouders van de leerling ondersteuning kunnen verlenen bij het huiswerk maken
Beroepsinteresse	uitgesproken wens van de leerling voor het uitoefenen van een bepaald beroep
Capaciteiten vrienden en vriendinnen	schoolprestaties van vrienden en vriendinnen van de leerling

Leerlingkenmerk	Omschrijving
Begeleiding door de ouders	mate waarin de ouders van de leerling ondersteuning kunnen verlenen bij het huiswerk maken
Concentratievermogen	het vermogen van de leerling om zijn/haar aandacht op de leerstof te richten
Doorzettingsvermogen	bereidheid om zich met ijver voor de leerstof in te zetten ondanks tegenslag; na het behalen van een onvoldoende kunnen leerlingen verschillend reageren: de een geeft de moed op, de ander zal juist extra zijn/haar schouders er onder zetten.
Eerdere schoolervaringen	schoolprestaties van de leerling in voorgaande schooljaren
Emotionele stabiliteit	mate van emotioneel (on)evenwichtig gedrag van de leerling; zo kan de ene leerling kritiek of een slecht cijfer verdragen zonder de moed op te geven, terwijl de andere leerling hierdoor gemakkelijk overstuur raakt.

Tabel 5.1
Vervolg

Leerlingkenmerk	Omschrijving
Etniciteit	land van herkomst van de ouders van de leerling
Gezondheid	lichamelijke gesteldheid van de leerling
Interesse	aantoonbare belangstelling voor de leerstof
Karakteristieke eigenschappen	in het oog springende, individuele kenmerken van de leerling; bijvoorbeeld afwezigheid tijdens de les, verlegenheid, voortdurend te laat komen in de les, brutaal gedrag, etc.
Leer- en studievaardigheden	het vermogen van de leerling om de leerstof op effectieve wijze te analyseren, te verwerken en te reproduceren; zo kan de ene leerling beter hoofd- en bijzaken in de leerstof onderscheiden dan de andere leerling.
Motivatie	mate waarin de leerling openstaat voor de leerstof, wat onder andere tot uiting komt in actieve deelname tijdens de les
Rapportcijfers	rapportcijfers behaald tijdens de brugperiode
Sexe	geslacht (m/v)
Specifieke prestaties	goede resultaten in enkele specifieke gebieden; bijvoorbeeld goede mondelinge uitdrukkingsvaardigheid, hoger dan gemiddelde prestaties in exacte vakken, expressie vakken of andere vakken, etc.

Leerlingkenmerk	Omschrijving
Etniciteit	land van herkomst van de ouders van de leerling
Studiemogelijkheden thuis	mogelijkheden voor de leerling om thuis huiswerk te kunnen maken in een gunstige omgeving.
Verzuim	afwezigheid door familie-omstandigheden, ziekte, spijbelen, of andere omstandigheden
Voorkeur leerling voor schooltype	schooltype dat de leerling prefereert; een leerling heeft bijvoorbeeld een uitgesproken voorkeur voor mavo
Voorkeur ouders voor schooltype	schooltype dat de ouders prefereren voor hun kind
Zelfperceptie	het beeld dat de leerling heeft van zijn/haar capaciteiten; een leerling heeft bijvoorbeeld een negatief zelfbeeld ondanks voldoende resultaten op school, of overschat juist zijn/haar capaciteiten.

Het aantal kenmerken dat docenten selecteerden liep sterk uiteen. Het hoogste aantal geselecteerde kenmerken was zestien. Drie docenten kwamen tot dit aantal. Vier docenten vonden dertien kenmerken belangrijk. Zeven docenten bestempelden twaalf kenmerken als relevant. Twee docenten kozen voor tien kenmerken en eveneens twee docenten noemden er acht. De achttien docenten brachten in totaal slechts vier kenmerken naar voren die niet in de lijst van mogelijk relevante kenmerken voorkwamen. Deze kenmerken zijn kort te omschrijven als *verzorging van het werk*, *sociale vaardigheid*, *leeftijd* en *probleemoplossend vermogen*.

Zes kenmerken, namelijk *motivatie, concentratievermogen, leer- en studievaardigheden, interesse, doorzettingsvermogen* en *rapportcijfers*, zijn door vrijwel alle docenten geselecteerd en behoorden in het algemeen ook tot de verzameling kenmerken die zij het meest belangrijk vonden. Bij zes van de achttien docenten waren de zes genoemde kenmerken ook de zes meest belangrijke. Bij tien docenten behoorden vijf van de zes kenmerken tot de zes meest belangrijke, terwijl zeven van deze tien docenten het zesde kenmerk eveneens selecteerden. Bij twee docenten behoorden vier van de zes kenmerken tot de belangrijkste zes kenmerken. Beide docenten selecteerden overigens de twee overige kenmerken wel. Anders gezegd: vijftien van de

achttien docenten vonden alle zes voornoemde kenmerken relevant en drie docenten vonden vijf van de zes kenmerken relevant.

In tabel 5.2 staat hoe vaak ieder kenmerk uit de lijst van 21 kenmerken geselecteerd is door de achttien docenten. In de rechterkolom van deze tabel is aangegeven hoe vaak ieder kenmerk - blijkens de door docenten toegewezen rangnummers - behoorde tot de zes meest belangrijk geachte kenmerken. Voor de volledigheid wordt hier ook vermeld dat de vier door docenten zelf genoemde kenmerken op één uitzondering na behoorden tot de zes meest belangrijke kenmerken. De uitzondering was het kenmerk *verzorging van het werk*.

Bij de derde taak moesten de docenten scores toekennen aan de leerlingen op een reeks kenmerken. Niet alle door een docent geselecteerde kenmerken werden in de voor hem of haar bedoelde reeks opgenomen. Het aantal kenmerken is voor iedere docent beperkt tot de zes kenmerken die de betreffende docent het meest belangrijk vond. Bij tien docenten ontbrak in de aldus gevormde reeks één van de kenmerken *motivatie, concentratievermogen, leer- en studievaardigheden*, *interesse*, *doorzettingsvermogen* en *rapportcijfers* en bij twee docenten ontbraken er twee. Indien de genoemde kenmerken geen deel uitmaakten van een reeks, zijn ze er aan toegevoegd.

Deze reductie in kenmerken zorgde ervoor dat de taak voor de docenten geen al te tijdrovend karakter kreeg. Het beperken van het aantal kenmerken kan zonder

Tabel 5.2

Overzicht van door docenten geselecteerde leerlingkenmerken en het aantal malen dat ieder kenmerk als een van de zes belangrijkste werd aangemerkt

Leerlingkenmerk	Aantal malen gekozen	Aantal malen bij zes belangrijkste kenmerken
Motivatie	18	18
Concentratievermogen	18	16
Leer- en studievaardigheden	18	16
Interesse	17	17
Doorzettingsvermogen	17	15
Rapportcijfers	17	12
Specifieke prestaties	16	5
Karakteristieke eigenschappen	15	5
Emotionele stabiliteit	15	2
Eerdere schoolervaringen	11	5
Zelfperceptie	9	2
Voorkeur leerling voor schooltype	8	0
Gezondheid	8	2
Verzuim	7	2
Studiemogelijkheden thuis	7	1
Begeleiding ouders	5	3
Beroepsinteresse	4	2
Voorkeur ouders voor schooltype	3	0
Capaciteiten van vrienden en vriendinnen	1	1
Etniciteit	1	0
Sexe	0	0

bezwaar gebeuren. Besliskundig onderzoek leert immers dat beslissers slechts een beperkt aantal kenmerken daadwerkelijk gebruiken. Brehmer en Brehmer (1988) melden dat het aantal kenmerken dat daadwerkelijk een rol

speelt varieert van één tot elf. Zij geven echter ook aan dat er te weinig goede onderzoeken zijn gedaan om tot harde conclusies te komen. De twee methodologische meest verantwoorde onderzoeken rapporteren het gebruik van tussen de zes en negen kenmerken uit een verzameling van 64 (Roose & Doherty, 1976) en het gebruik van tussen de één en zes kenmerken uit een verzameling van negentien (Ullman & Doherty, 1984).

Ten behoeve van het uitvoeren van de scoringstaak kreeg iedere docent een verzameling materiaal toegezonden, bestaande uit een toelichting, beschrijvingen van de voor de docent geselecteerde leerlingkenmerken en per leerling een scoringsformulier. Ieder formulier bevatte de naam van een leerling en de reeks van zes, zeven of acht kenmerken die voor de betreffende docent geselecteerd waren. Ieder leerlingkenmerk ging vergezeld van een vijfpuntsschaal, waarop de docent de score van de betreffende leerling voor het kenmerk kon aangeven. Het punt 1 op de vijfpuntsschaal droeg het label ‘zeer laag’; punt 2 het label ‘laag’; punt 3 het label ‘gemiddeld’; punt 4 het label ‘hoog’ en punt 5 het label ‘zeer hoog’.

Resultaten van het vooronderzoek

Voor beide scholen is, waar mogelijk, voor ieder paar van docenten de overeenstemming tussen oordelen berekend via Cohen’s kappa. In het algemeen beschouwt men de overeenstemming tussen beoordelaars als ‘acceptabel’ bij een kappa tussen de 0,60 en 0,80 en als ‘goed’ bij een kappa hoger dan 0,80 (Heuvelmans & Sanders, 1993).

Eerder is al opgemerkt dat op een van de twee scholen de te beoordelen leerlingen uit twee klassen afkomstig waren en dat een deel der docenten aan slechts een van beide klassen lesgaf. Het was uiteraard alleen mogelijk kappa te bepalen voor paren docenten die uitspraken deden over een verzameling leerlingen die tenminste voor een deel identiek was. Op de eerste school kon kappa voor 27 van de 36 te vormen paren docenten bepaald worden. Op de tweede school is kappa voor alle 36 mogelijke docentparen bepaald.

De overeenstemming tussen docenten varieerde sterk. Op de eerste school liep Cohen’s kappa voor de 27 paren beoordelaars van 0 tot 1. Bij slechts

twee beoor-delaarsparen was kappa groter dan 0,80 en kon de overeenstemming dus goed genoemd worden. Bij vijftien beoordelaarsparen lag kappa tussen 0,60 en de 0,80 en kon de overeenstemming derhalve als acceptabel beschouwd worden. Kappa was bij de overige tien paren kleiner dan 0,60 en moest daarom als onvoldoende bestempeld worden. Op de tweede school varieerde kappa van 0,17 tot 0,88. Bij slechts twee van de 36 beoordelaarsparen was sprake van een goede overeenstemming. Bij veertien paren was de overeenstemming acceptabel en bij de overige twintig paren onvoldoende. De gemiddelde kappa op de eerste school bedroeg 0,39 en op de tweede school 0,48. Voor beide scholen lijkt de conclusie gewettigd dat de overeenstemming tussen docenten te wensen overlaat, indien zij onafhankelijk van elkaar doorstroombeslissingen moeten nemen.

Verder bleek er, zoals tabel 5.2 laat zien, een duidelijke overeenstemming te zijn tussen docenten over de vraag welke leerlingkenmerken van belang zijn bij het determineren. Vrijwel alle docenten vonden dat de leerlingkenmerken *motivatie*, *concentratievermogen*, *leer- en studievaardigheden*, *interesse*, *doorzettingsvermogen* en *rapportcijfers* in beschouwing moeten worden genomen bij het determineren van leerlingen.

Omdat bij veel docenten substantiële correlaties bleken te bestaan tussen de scores die zij aan leerlingen toekenden op de verschillende kenmerken, werden voor de achttien deelnemende docenten geen structurele modellen opgesteld. De gewichten die sterk samenhangende kenmerken krijgen in structurele modellen geven namelijk geen goede indicatie van de invloed die de kenmerken hebben op het tot stand komen van beslissingen. Een tweede reden om af te zien van het opstellen van structurele modellen was dat het aantal observaties in verhouding tot het aantal kenmerken zo klein was dat de generaliseerbaarheid van de opgestelde modellen te wensen zou overlaten. De vraag of de gebrekkige overeenstemming tussen de docenten veroorzaakt wordt door interactie tussen docenten en kenmerken was in dit onderzoek dan ook niet te beantwoorden.

5.3.2 De invloed van leerlingkenmerken op determinatiebeslissingen

Om wel de invloed te kunnen bepalen die resultaten op plaatsingstoetsen en andere leerlingkenmerken hebben op de determinatiebeslissingen van docenten is een tweede onderzoek uitgevoerd. De opzet van dit onderzoek was zodanig dat het probleem van de hoge onderlinge correlaties tussen leerlingkenmerken kon worden ondervangen. Ook het probleem van het geringe aantal observaties in verhouding tot het aantal kenmerken is in dit onderzoek opgelost. Aan het onderzoek is deelgenomen door 60 docenten uit de onderbouw van het voortgezet onderwijs.

Opzet van het onderzoek

De interactie tussen docenten en leerlingkenmerken bij determinatiebeslissingen is pas goed te onderzoeken wanneer de beslissituatie gestandaardiseerd is. In de eerste plaats dient de verzameling leerlingen over wie docenten uitspraken moeten doen voor alle docenten gelijk te zijn. In de tweede plaats dient de verzameling gegevens op grond waarvan docenten uitspraken doen voor alle docenten identiek te zijn. Aan deze twee voorwaarden werd voldaan door zestig fictieve leerlingprofielen op te stellen en aan de docenten ter beoordeling voor te leggen.

Ieder profiel gaf een beschrijving van een fictieve leerling uit het eerste leerjaar van een scholengemeenschap voor mavo, havo en vwo. In de profielen werden vier kenmerken opgenomen. Drie daarvan waren gebaseerd op de kenmerken die de docenten in het vooronderzoek belangrijk vonden: *motivatie, concentratievermogen, leer- en studievaardigheid, interesse, doorzettingsvermogen* en *rapportcijfers. Leer- en studievaardigheid* en *rapportcijfers* zijn als op zichzelf staande kenmerken in de profielen opgenomen. De kenmerken *motivatie, concentratievermogen, interesse,* en *doorzettingsvermogen* zijn gezamenlijk gepresenteerd onder de noemer *leerinstelling.* Als vierde is aan de profielen het kenmerk *prestaties op Cito-plaatsingstoetsen* toegevoegd.

Het eerste kenmerk in de profielen, *leer- en studievaardigheden* (*LST*), is omschreven als het vermogen van een leerling om de leerstof op effectieve wijze te analyseren, te verwerken en te reproduceren. Het tweede kenmerk, *gemiddeld cijfer op het eindrapport* (*GCE*), is omschreven als het gemiddelde

van de cijfers die een leerling heeft op het eindrapport van het brugjaar voor de vakken die relevant zijn voor beslissingen omtrent de overgang. Het derde kenmerk, *leerinstelling* (*LIN*), is omschreven als een groep leerlingkenmerken die onderling sterk samenhangen, namelijk de motivatie, de interesse, het concentratie- en het doorzettingsvermogen van een leerling. Motivatie is omschreven als de mate waarin de leerling openstaat voor de leerstof, wat onder andere tot uiting komt in actieve deelname aan de lessen. Interesse is omschreven als het tonen van belangstelling voor de leerstof. Concentratievermogen is omschreven als het vermogen van de leerling om de aandacht op de leerstof te richten en doorzettingsvermogen als de bereidheid om zich met ijver voor het beheersen van de leerstof in te zetten. Het vierde kenmerk, *resultaat op Cito-plaatsingstoetsen* (*RCP*), is omschreven als de scoregroep waar de leerling aan kan worden toegewezen op grond van zijn of haar prestaties op een verzameling van drie plaatsingstoetsen voor het einde van het eerste leerjaar.

Posities van leerlingen op de kenmerken werden in de profielen gerepresenteerd door scores die liepen van 1 tot en met 9. Deze scores zijn aselect getrokken uit een uniforme verdeling. Daarom komen alle scores ongeveer even vaak aan bod en zijn de scores op de verschillende kenmerken niet gecorreleerd. Door het standaardiseren van de beslissituatie en het opstellen van profielen met niet correlerende kenmerken is een zuivere vergelijking mogelijk tussen de voor de verschillende docenten opgestelde modellen.

Zowel de *LST*-schaal als de *LIN*-schaal liepen van slecht (1) tot uitstekend (9); de punten daartussen kregen geen label. De *GCE*-schaal was gebaseerd op een verdeling van rapportcijfers met een bereik van 4,5 tot 8,5 met een gemiddelde van 6,5. De normtabellen voor de totaalscores op de plaatsingstoetsen aan het einde van het eerste leerjaar bevatten stanines. De score 1 op de *RCP*-schaal staat dan ook voor een totaalscore die behoort bij de vier procent laagste totaalscores in de populatie. De score 2 representeert de daarop volgende zeven procent totaalscores en scoregroep 3 de twaalf procent die daarop volgt. De scoregroepen 4 tot en met 8 vertegenwoordigen respectievelijk de volgende zeventien, twintig, zeventien, twaalf en zeven procent van de totaalscores. Score 9 representeert de bovenste vier procent van de totaalscores.

Bij het boekje met profielen was een overzicht gevoegd van de proportie leerlingen uit iedere scoregroep die met succes het tweede leerjaar van de havo en het vwo kon afronden. Omdat ten tijde van het onderzoek de werkelijke kansen op succes voor iedere scoregroep, gegeven de scores op de plaatsingstoetsen, nog niet bekend waren, betrof het fictieve proporties. Bij het bepalen van de fictieve proporties is ervan uitgegaan dat de totaalscores normaal verdeeld zouden zijn en dat de gemiddelde totaalscore van de vwo-leerlingen een halve standaardafwijking hoger zou zijn dan de gemiddelde totaalscore van de havo-leerlingen. Bij het bepalen van de fictieve proportie succesvolle leerlingen in iedere scoregroep is ook uitgegaan van een stanineverdeling. De proporties succesvolle leerlingen in de scoregroepen 1 tot en met 9 waren voor de vwo-leerlingen respectievelijk 0,11; 0,23; 0,40; 0,60; 0,77; 0,89; 0,96, 1,00 en 1,00. Bij de havo-leerlingen bedroegen de fictieve proporties succesvolle leerlingen in de scoregroepen 1 tot en met 9 respectievelijk 0,00; 0,04; 0,11; 0,23; 0,40; 0,60; 0,77; 0,89 en 0,96.

Ieder profiel werd voorzien van een fictieve leerlingnaam. De zestig profielen zijn tot een boekje gebundeld. Boekjes werden aan de zestig docenten aangeboden met de vraag bij iedere leerling aan te geven voor welk onderwijstype de leerling in kwestie het meest geschikt zou zijn.

In veel structuurgericht onderzoek moeten beslissers uitspraken doen op een schaal met een intervalkarakter. Bij determinatie zijn de uitspraken discreet en hoogstens ordinaal van aard. Er zijn aanwijzingen dat zowel de vorm van structurele modellen als de uitkomsten van beslisprocessen gevoelig zijn voor het gevraagde type respons (Westenberg & Koele, 1990; Westenberg, 1991; Westenberg & Koele 1992). Om de invloed van de responswijze te kunnen onderzoeken, zijn docenten aselect aan twee groepen toegewezen. Aan docenten uit de ene groep is gevraagd een oordeel te geven over de geschiktheid van leerlingen door een score op een negenpuntsschaal te kiezen. De score 1 op deze schaal gaf aan dat de betreffende leerling op de mavo thuishoorde en het in de andere twee onderwijstypen zeker niet zou redden. De score 9 op de schaal gaf aan dat de betreffende leerling op het vwo thuishoorde en het daar zeker goed zou doen. De overige punten werden niet omschreven. Deze groep docenten zal in het vervolg met de term '*intervalconditie*' worden aangeduid. De docenten uit de andere groep moesten aangeven voor welk type onderwijs iedere leerling in hun ogen het

meest geschikt was: mavo, havo of vwo. Deze groep zal in het vervolg met de naam '*nominale conditie*' aangeduid worden.

De zestig profielen zijn met een tussenperiode van enige weken tweemaal aan de docenten aangeboden. Bij de eerste aanbieding bevatte ieder profiel de vier genoemde kenmerken. De tweede keer ontbrak het kenmerk *resultaat op Cito-plaatsingstoetsen*. Vergelijking van de oordelen van iedere docent voor de overeenkomstige profielen op beide meetmomenten geeft de mogelijkheid uitspraken te doen over de rol die prestaties op plaatsingstoetsen voor het einde van het eerste leerjaar zouden kunnen spelen bij determinatiebeslissingen.

De relatieve gewichten van de kenmerken in de intervalconditie zijn bepaald met behulp van multiple regressie-analyse. Voor het berekenen van de relatieve gewichten van de kenmerken in de nominale conditie is gebruik gemaakt van discriminantanalyse. Beide technieken zijn speciale gevallen van canonische correlatie-analyse. Variabelen worden gecombineerd in een lineaire vergelijking om een zekere doelfunctie te maximaliseren. Er zijn geen fundamentele verschillen tussen de beide technieken. Daarom zijn eventuele verschillende resultaten in beide condities toe te schrijven aan de invloed van de responswijze.

Docenten ontvingen de boekjes met de profielen en de bijbehorende instructies per post. Zij moesten hun taak binnen twee weken uitvoeren en vervolgens de beoordeelde profielen terugsturen. Enkele weken later ontvingen ze dan het tweede boekje met profielen. De schaalscores voor de profielen in dit tweede boekje waren identiek aan die in het eerste boekje, maar in het tweede boekje ontbrak steeds de *RCP*-schaal.

Onderzoeksresultaten

In de intervalconditie retourneerden 28 docenten beide boekjes, terwijl één docent alleen het eerste boekje ingevuld terugstuurde. De correlaties tussen de beslissingen van de verschillende docenten bij de profielen met vier kenmerken varieerden van 0,17 tot 0,96 met een gemiddelde van 0,66 en een standaarddeviatie van 0,20. Bij de profielen zonder *RCP*-schaal liepen de

correlaties uiteen van 0,08 tot 0,97 met een gemiddelde van 0,71 en een standaarddeviatie van 0,19.

De respons in de nominale conditie was identiek. Ook hier retourneerden 28 docenten beide boekjes en stuurde één docent alleen het eerste boekje terug. Cohen's kappa varieerde bij de profielen met vier kenmerken van -0,18 tot 0,87. Hanteren we het uitgangspunt dat kappa's groter dan 0,80 goed zijn en kappa's tussen de 0,60 en 0,80 acceptabel, dan is de overeenstemming bij slechts één van de in totaal 406 te vormen paren docenten goed, bij 51 docentparen acceptabel en bij 354 paren onder de maat. Bij de profielen met drie kenmerken liep Cohen's kappa uiteen van -0,05 tot 0,89. Van de 378 te vormen paren docenten vertoonden er zeven een goede overeenstemming. Bij 128 paren was de overeenstemming acceptabel en bij 243 onvoldoende. Ook in deze gestandaardiseerde situatie laat de overeenstemming tussen docenten nog steeds te wensen over.

Bijlage B bevat een tabel met de resultaten van de regressie-analyses die uitgevoerd zijn in de intervalconditie voor de oordelen die de docenten gegeven hebben bij profielen met vier kenmerken. Kenmerken worden in deze tabel aangeduid met de reeds beschreven labels *LST*, *GCE*, *LIN* en *RCP*. Bij ieder kenmerk wordt het gestandaardiseerde regressiegewicht β gegeven. Het kwadraat van de multiple correlatie, R^2, varieert van 0,65 tot 0,95. Dit houdt in dat de lineaire modellen voor de 29 betreffende docenten een goede tot zeer goede passing hebben.

Omdat de kenmerken onderling niet samenhingen, is het gestandaardiseerde regressiegewicht van ieder kenmerk in een model een goede maat voor het relatieve belang van dit kenmerk voor de betreffende docent. Bij 21 van de 29 docenten was het gemiddelde cijfer op het eindrapport het meest belangrijke kenmerk. Bij vijf docenten legden prestaties op de toetsen het meeste gewicht in de schaal en voor drie docenten was het kenmerk leer- en studievaardigheden het meest relevant. De standaardfouten bij de regressiegewichten worden in de tabel niet vermeld, maar variëren van 0,04 tot 0,07.

Bijlage B bevat tevens een tabel met de resultaten van de discriminantanalyses die voor de leerlingprofielen met vier kenmerken zijn uitgevoerd in de nominale conditie. Ook in deze tabel worden de eerder beschreven labels

gebruikt. De tabel bevat de gestandaardiseerde discriminantgewichten van de vier kenmerken voor de eerste discriminantfunctie. Omdat de kenmerken niet gecorreleerd zijn, kunnen de discriminantgewichten beschouwd worden als een goede maat voor het relatieve belang van de kenmerken. Op één uitzondering na is de passing van de opgestelde modellen in orde. De canonische correlatie r^* varieert bij de passende modellen van 0,70 tot 0,90 en het aantal correct geclassificeerde leerlingen (N_{cor}) loopt van 43 (71,7%) tot 57 (95%). Bij één docent (nr. 25) is de passing slecht; de canonische correlatie bedraagt slechts 0,26 en er zijn niet meer dan 29 leerlingen correct geclassificeerd. Bij 23 docenten was het gemiddelde rapportcijfer het meest belangrijke kenmerk. De prestaties op de toetsen en leer- en studievaardigheid hadden beide bij twee docenten het meeste gewicht. Leerinstelling bleek bij één docent het belangrijkste kenmerk.

Bij een aantal docenten in de nominale conditie waren de overschrijdingskansen voor het onderscheidend vermogen van de tweede discriminantfunctie ook significant. Omdat de relatieve sterkte van de tweede discriminantfunctie ten opzichte van de eerste steeds verwaarloosbaar was, zijn deze echter buiten beschouwing gelaten.

Om het effect van het ontbreken van de *RCP*-schaal te bepalen, werden de 30 profielen geselecteerd waar de score op de *RCP*-schaal en de gemiddelde scores op de andere drie kenmerken de grootste discrepantie vertoonden. Deze selectie vond plaats, omdat te verwachten viel dat docenten eerder geneigd zouden zijn een beslissing bij de profielen zonder *RCP*-schaal te herzien, naarmate de discrepantie tussen scores op deze schaal en scores op de andere kenmerken groter was. Vijftien van de geselecteerde profielen kenden een hoge score op de *RCP*-schaal en een lage gemiddelde score op de drie andere kenmerken; de andere vijftien een lage score op de *RCP*-schaal en een hoge gemiddelde score op de drie andere kenmerken.

Docenten die hun leerlingen het voordeel van de twijfel geven, herzien hun oordeel in positieve zin, indien toetsresultaten beter uitvallen dan verwacht, maar herzien hun oordeel niet, indien toetsresultaten slechter uitvallen dan verwacht. Bij een docent die leerlingen het voordeel van de twijfel geeft, mochten daarom bij de 15 profielen met een hoge score op de *RCP*-schaal in het eerste boekje positievere oordelen verwacht worden dan in het boekje

waarin de *RCP*-schaal ontbrak. Bovendien mocht verwacht worden dat bij de profielen met een lage score op de *RCP*-schaal de oordelen van een dergelijke docent in het eerste boekje niet wezenlijk zouden afwijken van de oordelen in het boekje waarin de *RCP*-schaal ontbrak.

Bijlage C bevat een tabel die betrekking heeft op de oordelen van de 28 docenten in de intervalconditie die zowel het eerste als het tweede boekje retourneerden. In deze tabel staan voor beide boekjes de gemiddelde oordelen van de docenten voor de 30 geselecteerde profielen. De tabel vermeldt ook voor iedere docent het verschil tussen het gemiddelde oordeel voor het eerste boekje en het boekje waar de *RCP*-schaal aan ontbrak. Bijlage C bevat ook een tabel met gegevens over de oordelen van de 28 docenten in de nominale conditie die zowel het eerste als het tweede boekje terugzonden. De tabel laat voor iedere docent zien hoe de oordelen in het eerste boekje zich verhouden tot de overeenkomstige oordelen in het boekje waarin de *RCP*-schaal ontbrak.

Omdat er in beide condities sprake was van interacties tussen docenten en kenmerken, vond de statistische toetsing van de verschillen tussen oordelen in het eerste en het tweede boekje voor iedere docent afzonderlijk plaats. In de intervalconditie werden per docent twee t-toetsen uitgevoerd op zijn of haar gepaarde oordelen. De eerste t-toets had betrekking op de oordelen voor de vijftien profielen met een hoge *RCP*-score en een gemiddeld lage score op de drie andere kenmerken. De tweede t-toets had betrekking op de oordelen voor de vijftien profielen met een lage *RCP*-score en een gemiddeld hoge score op de drie andere kenmerken. Bij elke toetsing was het significantieniveau 0,05. In de nominale conditie zijn voor iedere docent afzonderlijk twee tekentoetsen uitgevoerd voor de twee eerder beschreven verzamelingen van vijftien profielen. Ook hier was bij iedere toetsing het significantieniveau 0,05.

Tabel 5.3

Aantallen significante resultaten voor de in de intervalconditie uitgevoerde t-toetsen op verschillen in gemiddelde oordelen en de in de nominale conditie uitgevoerde tekentoetsen. Het significantieniveau is 0,05.

	Aantallen significante resultaten bij de profielen met een hoge RCP-score			Aantallen significante resultaten bij de profielen met een lage RCP-score		
Conditie	Met RCP hoger	Geen verschil	Met RCP lager	Met RCP hoger	Geen verschil	Met RCP lager
Interval	15	13	0	0	7	21
Nominaal	11	17	0	0	19	9

Tabel 5.3 geeft een overzicht van de uitkomsten van de statistische toetsingen in beide condities. Bij de profielen met een *hoge RCP*-score is bij dertien docenten in de intervalconditie geen sprake van significante verschillen tussen gemiddelde oordelen. Bij vijftien docenten in de intervalconditie is het gemiddelde oordeel voor profielen die de *RCP*-schaal bevatten significant hoger. In de nominale conditie veranderen bij de profielen met een *hoge RCP*-score de oordelen van zeventien docenten niet significant. Bij elf docenten is sprake van veranderingen in het oordeel. Hun oordelen zijn positiever voor profielen die de *RCP*-schaal bevatten. In geen van beide condities is sprake van significant lagere gemiddelde oordelen c.q. negatievere oordelen voor profielen die de *RCP*-schaal bevatten. De conclusie is dan ook dat toetsprestaties die positief afsteken tegen andere relevante kenmerken kunnen leiden tot een meer positieve determinatiebeslissing van een docent.

Bij de profielen met een *lage RCP*-score is bij zeven docenten in de intervalconditie *geen* sprake van significante verschillen tussen gemiddelde oordelen. Bij 21 docenten in de intervalconditie is het gemiddelde oordeel voor profielen mèt de *RCP*-schaal significant lager. In de nominale conditie is bij de profielen met een *lage RCP*-score bij negentien docenten geen sprake van significante veranderingen in oordeel van profielen mèt *RCP*-schaal naar profielen waarbij deze schaal ontbrak. Bij negen docenten is sprake van veranderingen in het oordeel. Hun oordelen zijn negatiever voor

de profielen die de *RCP*-schaal bevatten. In geen van beide condities is sprake van significant hogere gemiddelde oordelen c.q. positievere oordelen voor profielen die de *RCP*-schaal bevatten. De conclusie is dat toetsprestaties die negatief afsteken tegen andere kenmerken kunnen leiden tot een meer negatieve beslissing.

5.3.3 Discussie en conclusies

Tussen de 18 docenten die hebben deelgenomen aan het in paragraaf 5.3.1 beschreven vooronderzoek bleek vrij veel overeenstemming te zijn over de vraag welke kenmerken van leerlingen relevant zijn bij het nemen van determinatiebeslissingen. De overeenstemming over het voor leerlingen meest geschikte onderwijstype zelf laat echter veel te wensen over. Voor maar vier van de in totaal 63 onderzochte docentparen is Cohen's kappa zodanig dat gesproken mag worden van een goede overeenstemming. De overeenstemming tussen 29 paren kan nog voor acceptabel doorgaan, maar voor de overige 30 docentparen is kappa onder de maat.

Aan dit gebrek aan overeenstemming tussen docenten kunnen drie elkaar verster-kende oorzaken ten grondslag liggen. In de eerste plaats kunnen docenten verschillende kenmerken van leerlingen in ogenschouw nemen bij het bepalen van het meest geschikte onderwijstype. In de tweede plaats kan sprake zijn van interactie tussen docenten en kenmerken. Docenten die wel dezelfde kenmerken in ogen-schouw nemen, betrekken deze kenmerken dan op verschillende wijzen in hun be-sluitvorming. Een kenmerk dat bij de ene docent veel gewicht krijgt, kan bij een andere docent nauwelijks meetellen. In de derde plaats kunnen docenten, wanneer zij van dezelfde kenmerken uitgaan en deze ook op dezelfde manier in hun oor-deelsvorming verdisconteren, nog verschillen voor wat betreft het inschatten van de positie van hun leerlingen op deze kenmerken. Zo kan de ene docent een bepaalde leerling goed gemotiveerd vinden, terwijl een collega van mening kan zijn dat dezelfde leerling maar matig gemotiveerd is.

Het vooronderzoek was te beperkt van opzet om te kunnen bepalen wat de oorzaak was van het gebrek aan overeenstemming tussen de docenten. Het lijkt echter niet aannemelijk dat het waargenomen gebrek aan overeenstem-

ming ontstaan is doordat verschillende docenten geheel andere verzamelingen kenmerken van leerlingen in ogenschouw nemen bij de determinatie. Er is sprake van een duidelijk begrensde verzameling leerlingkenmerken die alle docenten relevant vinden en die zij ook daadwerkelijk in de besluitvorming betrekken. Het in het vooronderzoek geconstateerde gebrek aan overeenstemming tussen docenten komt naar alle waarschijnlijkheid voort uit verschillen in invloed van dezelfde kenmerken. Een soortgelijke conclusie wordt in veel structuurgericht besliskundig onderzoek getrokken. Het gebrek aan overeenstemming dat door interactie tussen docenten en kenmerken ontstaat, kan overigens nog versterkt worden door verschillen in opvatting over de positie van leerlingen op identieke kenmerken.

In het eigenlijke onderzoek kregen docenten fictieve profielen van leerlingen voorgelegd. Deze profielen waren voor alle docenten identiek. Verschillen in oordeel konden dan ook niet veroorzaakt worden doordat docenten andere kenmerken in ogenschouw namen, of identieke leerlingen anders scoorden op dezelfde kenmerken. Gebrek aan overeenstemming kon alleen nog maar voortkomen uit verschillen in invloed van identieke kenmerken. Ook in deze gestandaardiseerde en relatief eenvoudige situatie bleek nog een duidelijk gebrek aan overeenstemming te bestaan. In de intervalconditie bedroeg de gemiddelde correlatie tussen oordelen van docenten bij de profielen met vier kenmerken 0,66 en bij de profielen met drie kenmerken 0,71. In de nominale conditie was bij de profielen met vier kenmerken de overeenstemming volgens Cohen's kappa bij slechts één van de in totaal 406 te vormen paren docenten goed (groter dan 0,80), bij 51 acceptabel (tussen 0,60 en 0,80) en bij 354 paren onder de maat (kleiner dan 0,60). Bij de profielen met drie kenmerken vertonen slechts zeven van de 378 te vormen paren docenten een goede overeenstemming. Bij 128 paren is de overeenstemming acceptabel en bij 243 onvoldoende.

De twee beschreven onderzoeken leiden tot de conclusie dat de kans op verschillen van mening over de geschiktheid van leerlingen voor verschillende onderwijstypen aanzienlijk is, wanneer docenten daarover aan het einde van de brugperiode onafhankelijk van elkaar beslissingen moeten nemen. Een belangrijke oorzaak voor dit gebrek aan overeenstemming is dat het beslisproces van docent tot docent kan verschillen. Docenten wijken wat dat betreft in niets af van beslissers op andere terreinen. Overeenkomstige

kenmerken van leerlingen oefenen een uiteenlopende invloed uit op de beslissingen van verschillende docenten. Het is echter niet zo dat de beslisprocessen van alle docenten van elkaar afwijken. Er zijn docenten te onderscheiden van wie de beslisprocessen niet opvallend verschillen.

Het feit dat de invloed van identieke kenmerken op docentbeslissingen van docent tot docent kan variëren, geldt natuurlijk ook voor de invloed van de prestaties van leerlingen op gestandaardiseerde toetsen. Aan het begin van paragraaf 5.2 is aangegeven dat uit de literatuur bekend is dat docenten geneigd zijn leerlingen het voordeel van de twijfel te geven, indien zij toetsresultaten in hun beslissingen betrekken. De uitkomsten van het hiervoor beschreven onderzoek stroken niet met de in de literatuur gevonden resultaten. Er is geen sprake van dat leerlingen het voordeel van de twijfel krijgen. Dit geldt zowel voor de nominale als de intervalconditie. Wanneer toetsresultaten positief afsteken ten opzichte van de andere kenmerken leiden ze tot een positiever oordeel, maar het omgekeerde is evenzeer het geval. Wanneer toetsresultaten negatief afsteken ten opzichte van de andere kenmerken worden ze niet - zoals in de literatuur aangegeven - bij het nemen van een beslissing buiten beschouwing gelaten. Deze discrepantie kan verklaard worden door het feit dat de conclusies in de literatuur gebaseerd zijn op vragenlijstonderzoek of onderzoek met behulp van al dan niet gestructureerde interviews, terwijl de besliskunde leert dat de beschrijvingen die mensen van hun beslisprocessen geven vaak verschillen van de wijze waarop hun beslissingen daadwerkelijk tot stand komen.

Aan het begin van paragraaf 5.2 is ook aangegeven dat uit de literatuur blijkt dat toetsprestaties geen belangrijke rol spelen bij beslissingen van docenten. Bij een deel van de docenten die deelnamen aan het hiervoor beschreven onderzoek blijken toetsprestaties inderdaad geen belangrijke rol te spelen. Er zijn echter ook docenten van wie de beslissingen wel degelijk door de toetsprestaties van leerlingen worden beïnvloed.

Een mogelijk punt van kritiek op het beschreven onderzoek is dat gebruik gemaakt is van *paper cases*. Iedere docent kreeg een reeks fictieve profielen voorgelegd en werd niet geconfronteerd met *echte* leerlingen. Het is de vraag of de verkregen resultaten generaliseerbaar zijn, omdat docenten in de praktijk ook andere informatie kunnen en zullen gebruiken bij het vellen van

een oordeel. Er zijn echter aanwijzingen dat, wanneer sprake is van een taak waar de beoordelaars mee vertrouwd zijn, het gebruik van *paper cases* op zich niet tot belangrijke vertekeningen in de opgestelde structurele modellen hoeft te leiden (Brehmer & Brehmer, 1988). Dat laatste is in het beschreven onderzoek zeker het geval. Docenten moeten ieder jaar voor grote aantallen leerlingen doorstroombeslissingen nemen. Bovendien laat het vooronderzoek zien dat de profielen kenmerken bevatten die docenten in de regel relevant vinden en ook daadwerkelijk gebruiken.

Een tweede mogelijk punt van kritiek op het onderzoek is dat structurele modellen niets meer doen dan weergeven hoe het gesteld is met de samenhang tussen de gegevens die beslissers tot hun beschikking hebben en hun uiteindelijke besluit. Het is niet duidelijk in hoeverre structurele modellen een correcte weergave leveren van het proces dat werkelijk aan het totstandkomen van beslissingen ten grondslag ligt (Payne, 1976; Einhorn, Kleinmuntz & Kleinmuntz, 1979). Een techniek die wel een juiste weergave van dit proces geeft, is de analyse van verbale protocollen (zie bijvoorbeeld Svenson; 1979 en Borko & Cadwell; 1982). De analyse van verbale protocollen kan dienen om te bepalen of de beoordelingsstrategieën van docenten werkelijk goed weergegeven worden door structurele modellen.

Het in dit hoofdstuk beschreven onderzoek had uitsluitend tot doel modellen op te stellen die aangaven in welke mate de beslissingen van docenten bepaald werden door verschillende kenmerken van leerlingen. Een op zichzelf interessante vraag is echter of de mentale processen van docenten in dit onderzoek door structurele modellen beschreven kunnen worden. Daarom is aan een groep van tien docenten gevraagd de profielen met vier kenmerken van de 60 fictieve leerlingen hardop denkend te beoordelen. De resultaten van de analyses van de verbale protocollen worden elders (Harte, 1995; Harte & Koele, 1995) uitgebreid beschreven. Hier wordt volstaan met de mededeling dat de protocolanalyses duidelijk maakten dat eenvoudige lineaire modellen inderdaad een goede weergave leveren van de mentale processen die zich afspelen bij docenten die moeten beslissen over het voor hun leerlingen meest geschikte onderwijstype.

Door de verschillen in beslisprocessen van individuele docenten is het niet mogelijk om algemeen geldende uitspraken te doen over de invloed van test-

en toetsresultaten op doorstroombeslissingen van docenten. Het lijkt er echter niet op dat het gebruik van de test- en toetsresultaten bij het nemen van doorstroombeslissingen ertoe zal leiden dat docenten hun leerlingen massaal het voordeel van de twijfel zullen geven, of juist zullen laten doubleren of afstromen. Het valt dan ook niet te verwachten dat het ter beschikking stellen van de instrumentaria die voor drie havo en drie vwo ontwikkeld zijn grote gevolgen zal hebben voor de instroom in de tweede fase van het voortgezet onderwijs.

6 Samenvatting en conclusies

In dit proefschrift stonden twee thema's centraal. Het eerste thema was de ontwikkeling van plaatsingstoetsen voor het ondersteunen van doorstroombeslissingen aan het einde van het derde leerjaar van de havo en het vwo. Het tweede onderwerp dat in dit proefschrift aan de orde kwam, was de besliskundige analyse van doorstroombeslissingen en de rol die tests bij het nemen van doorstroombeslissingen spelen.

De hoofdstukken twee, drie en vier van dit proefschrift hebben betrekking op het eerste thema. Hoofdstuk twee geeft aan op welke wijze de plaatsingstoetsen voor de vakken Nederlands, Engels en wiskunde tot stand gekomen zijn. Het hoofdstuk bevat een overzicht van de uitgangspunten die gehanteerd zijn bij de ontwikkeling van iedere toets. De opzet, uitvoering en resultaten van de proefafname voor ieder vak worden uitgebreid beschreven. Ten slotte wordt geschetst hoe te werk gegaan is bij het samenstellen en normeren van de toetsen.

Hoofdstuk drie heeft betrekking op verschillende aspecten van de meetnauwkeurigheid van de ontwikkelde toetsen. Het betreffende hoofdstuk beschrijft op welke wijze de klassieke betrouwbaarheid van de toetsen af te leiden is uit de proefafnamegegevens. Ook laat het hoofdstuk zien hoe het gesteld is met de lokale meetnauwkeurigheid van de ontwikkelde toetsen. Dit gebeurt op twee manieren. De lokale meetnauwkeurigheid van iedere toets wordt grafisch weergegeven door voor iedere toets de toetsinformatiefunctie en de dichtheidsfunctie van de vaardigheid voor de doelpopulatie af te beelden. De lokale meetnauwkeurigheid wordt ook beschreven met behulp van betrouwbaarheidsmatrices. Voor iedere toets zijn de scores getransformeerd naar decielen. De betrouwbaarheidsmatrices laten voor ieder deciel zien hoe groot de proportie leerlingen is die een ware score heeft die in het betreffende deciel valt en hoe groot de proporties leerlingen zijn binnen ieder deciel die ware scores hebben die in andere dan het betreffende deciel vallen. Aldus ontstaat een gedetailleerd beeld van de meetnauwkeurigheid van de ontwik-

kelde toetsen. Geconcludeerd mag worden dat de meetnauwkeurigheid van de toetsen in orde is.

Hoofdstuk vier bevat het verslag van het valideringsonderzoek. In het hoofdstuk wordt beargumenteerd dat de toetsen een goede inhoudsrepresentativiteit en constructrepresentativiteit hebben. In het valideringsonderzoek is voor leerlingen uit drie havo en drie vwo het voorspellend vermogen van zes variabelen onderzocht. Het betrof de scores op de drie plaatsingstoetsen en drie scores op een checklist studievaardigheid, bepaald door docenten Nederlands, Engels en wiskunde. Deze checklist is naast de twee reeksen plaatsingstoetsen ontwikkeld voor het ondersteunen van doorstroombeslissingen aan het einde van drie havo en drie vwo. Het gehanteerde criterium was een simpele dichotome variabele. Leerlingen die zich in het schooljaar 1995-1996 in het vijfde leerjaar van de havo of het vwo bevonden, werden beschouwd als succesvol. Leerlingen die waren blijven zitten, waren afgestroomd, of geen onderwijs meer volgden, werden beschouwd als niet succesvol. De uitgevoerde logistische regressie-analyses laten zien dat het voorspellend vermogen van de toets- en de checklistscores ten aanzien van het betreffende criterium beperkt is. De scores op de twee toetsen voor Engels bleken zelfs nauwelijks een unieke bijdrage te leveren aan het voorspellend vermogen van de instrumentaria. Het feit dat beide instrumentaria een beperkt voorspellend vermogen hebben, is echter geen reden om de kwaliteit van de instrumentaria in twijfel te trekken. Het voorspellend vermogen van tests ten aanzien van studiesucces blijkt steevast gering.

In het valideringsonderzoek zijn ook multi-niveau logistische regressie-analyses uitgevoerd. Deze maken duidelijk dat zowel voor leerlingen uit drie havo als voor leerlingen uit drie vwo sprake is van een schooleffect. De kans op succes voor een leerling, gegeven zijn of haar scores op de instrumenten, blijkt mede afhankelijk te zijn van de school die de betreffende leerling bezoekt. De verwachtingstabellen bij de ontwikkelde instrumentaria zijn hierdoor voor scholen minder bruikbaar naarmate zij verder afwijken van wat in hoofdstuk vier een ‘gemiddelde’ school genoemd is.

Het optreden van een schooleffect is mogelijk te voorkomen door gebruik te maken van een criterium dat niet gevoelig is voor de interne standaard die scholen hanteren bij het beoordelen van de prestaties van leerlingen. De

prestaties van leerlingen op het eindexamen kunnen de basis vormen van een dergelijk criterium. Omdat de instrumenten zo snel mogelijk ter beschikking van de scholen moesten komen, behoorde het ontwikkelen van een criterium op grond van de eindexamenresultaten van leerlingen echter niet tot de mogelijkheden. Een tweede reden dat op voorhand niet gekozen is voor het ontwikkelen van een criterium van dit type is dat prestaties minder goed te voorspellen zijn naarmate het criterium verder in de toekomst ligt.

Het gesignaleerde schooleffect zou ook veroorzaakt kunnen zijn door verschillen in gemotiveerdheid tussen de leerlingen in de deelnemende klassen tijdens de proefafnames. De toetsscores van minder gemotiveerde leerlingen geven immers een negatief beeld van hun werkelijke vaardigheid, waardoor de kans op studiesucces voor dergelijke leerlingen onderschat wordt. Bij iedere proefafname is docenten gevraagd een zogeheten 'afnameprotocol' in te vullen. Op dit afnameprotocol konden de docenten onder meer aangeven of tijdens de proefafname sprake was van afwijkingen van de omstandigheden die normaal gelden bij het maken van een proefwerk. Analyse van de afnameprotocollen gaf geen aanleiding om klassen van analyses uit te sluiten. Omdat de proefafnames echter niet onder volledig gecontroleerde omstandigheden hebben plaatsgevonden, is deze alternatieve verklaring voor het schooleffect niet volledig uit te sluiten.

Hoofdstuk vijf heeft betrekking op het tweede thema en gaat met name in op de rol van tests bij het nemen van doorstroombeslissingen in het voortgezet onderwijs. In het betreffende hoofdstuk wordt kort beschreven hoe de normatieve besliskunde het nemen van beslissingen met tests kan optimaliseren. Tevens wordt gedemonstreerd op welke wijze met behulp van de psychometrische besliskunde cesuren voor de totaalscores op beide instrumentaria te vinden zijn die vanuit rationeel standpunt bezien optimaal zijn. Omdat bleek dat de gevonden cesuren sterk afhankelijk waren van de toevallige samenstelling van de onderzoeksgroepen is ervan afgezien informatie over de optimale cesuren op te nemen in de handleidingen bij de instrumentaria.

Het praktisch toepassen van de psychometrische besliskunde bij het nemen van doorstroombeslissingen in het voortgezet onderwijs is overigens in het algemeen problematisch. In de eerste plaats kan er, zoals bij de ontwikkelde

instrumenten, sprake zijn van een schooleffect. In dat geval zullen de gevonden optimale cesuren de voorkeuren van een school, naarmate deze verder afwijkt van een 'gemiddelde' school, minder goed weergeven. In de tweede plaats zullen de ontwikkelde instrumenten altijd betrekking hebben op niet meer dan een deel van vaardigheden, kenmerken en eigenschappen van leerlingen die bepalend zijn voor studiesucces. Het is dan ook de vraag in hoeverre het gebruik van de ontwikkelde instrumenten acceptabel zal zijn voor alle betrokken docenten. In de derde plaats zijn docenten volledig onbekend met het begrip utiliteit en toepassingen van de psychometrische besliskunde.

Hoofdstuk vijf bevat verder een bespreking van de resultaten van onderzoek naar de rol van gestandaardiseerde toetsen bij beslissingen van docenten. De betreffende onderzoeksresultaten geven aan dat toetsen geen belangrijke rol spelen. Bovendien maken de onderzoeksresultaten duidelijk dat docenten geneigd zijn hun leerlingen het voordeel van de twijfel geven, indien zij toetsresultaten in hun oordelen betrekken. Zij blijken hun oordeel niet te veranderen wanneer het resultaat op de toets lager is dan verwacht. Is het toetsresultaat daarentegen hoger dan verwacht, dan zijn docenten geneigd hun oordeel te herzien in positieve richting.

De resultaten van een onderzoek naar de rol van toetsen bij determinatiebeslissingen maken duidelijk dat niet verwacht mag worden dat het ter beschikking stellen van de instrumentaria grote invloed uit zal oefenen op doorstroombeslissingen aan het einde van het derde leerjaar van havo en vwo. Bovendien geeft het onderzoek aan dat scholen bij het gebruik van de instrumentaria naar alle waarschijnlijkheid hun leerlingen niet massaal het voordeel van de twijfel zullen geven, of juist zullen laten doubleren of afstromen.

Voordelen van het gebruik van de ontwikkelde instrumentaria

Het gebruiken van tests bij het nemen van doorstroombeslissingen kan een aantal voordelen hebben. Zo kan het gebruik van tests het rendement van doorstroombeslissingen verhogen. Verder bieden tests scholen de gelegenheid om een goed onderbouwd advies te geven over de meest geschikte voortzet-

ting van een leerweg. Ook maken tests het mogelijk om verschillen van mening te beslechten, zowel tussen docenten onderling als tussen docenten enerzijds en leerlingen of hun ouders anderzijds. Het is de vraag in hoeverre de ontwikkelde instrumentaria de genoemde voordelen bieden.

Er zijn verschillende redenen om aan te nemen dat het ter beschikking stellen van de instrumentaria geen grote gevolgen zal hebben voor het rendement van de bovenbouw van havo en vwo. De eerste reden is dat het niet voor de hand ligt dat scholen ervoor zullen kiezen de informatie die de instrumentaria leveren te gebruiken bij het nemen van doorstroombeslissingen over *alle* leerlingen. Bij een belangrijk deel van de leerlingen is de informatie die scholen tot hun beschikking hebben zo eenduidig dat het evident is dat deze leerlingen geschikt zijn voor het vervolgen van hun havo- of vwo-opleiding. Van zulke leerlingen valt aan te nemen dat zij, calamiteiten daargelaten, hun opleiding zonder vertraging zullen voltooien. Het heeft weinig zin om de geschiktheid van deze leerlingen met behulp van de instrumentaria te bepalen. Het valideringsonderzoek maakt duidelijk dat veel van deze leerlingen een hoge totaalscore zouden behalen op de instrumentaria. Bovendien is het onwaarschijnlijk dat lage toetsscores van deze leerlingen ertoe zullen leiden dat docenten hun oordelen over deze leerlingen herzien. Het inzetten van de instrumentaria bij leerlingen die gezien hun prestaties overduidelijk niet geschikt zijn om hun havo- of vwo-opleiding te vervolgen, lijkt evenmin zinvol. Ook voor deze leerlingen geldt dat het onwaarschijnlijk is dat docenten hun oordelen zouden herzien, indien zij hoge toetsscores zouden behalen.

In de handleidingen bij de ontwikkelde instrumentaria is dan ook het advies aan scholen opgenomen om de instrumentaria alleen maar te gebruiken, indien veel docenten bij een specifieke leerling twijfelen over de beste voortzetting van een leerweg. Ook het bestaan van verschillen van mening, hetzij tussen docenten onderling, hetzij tussen docenten enerzijds en leerlingen of hun ouders anderzijds, kan een reden zijn om de instrumentaria te gebruiken.

Een tweede reden om te verwachten dat het ter beschikking stellen van de instrumentaria weinig invloed zal hebben op het rendement van de bovenbouw van havo en vwo is dat het voorspellend vermogen van de

instrumenten beperkt is. Een derde reden is dat niet valt te verwachten, zoals hoofdstuk vijf heeft duidelijk gemaakt, dat de leerlingen bij wie de instrumentaria worden ingezet op grote schaal zullen worden afgewezen of het voordeel van de twijfel zullen krijgen.

Het ter beschikking stellen van de instrumenten biedt scholen wel de mogelijkheid om doorstroombeslissingen beter te onderbouwen. Het valideringsonderzoek heeft immers aangetoond dat de instrumenten betrekking hebben op vaardigheden waarvan de beheersing van belang is voor het succesvol vervolgen van een havo- of vwo-opleiding. De verwachtingstabellen bij de instrumentaria geven informatie over de kans die een leerling heeft op wat in hoofdstuk vier een 'gemiddelde' school wordt genoemd om het vijfde leerjaar van de havo of het vwo zonder vertraging te bereiken. Docenten op een specifieke school kunnen in gezamenlijk overleg kiezen voor een mildere of strengere interpretatie van de totaalscores van leerlingen op het instrumentarium, afhankelijk van het niveau dat een school nastreeft of het pedagogisch-didactische klimaat dat er heerst.

Indien een school besluit om de instrumentaria bij een aantal leerlingen in te zetten, dan zijn er twee manieren waarop de school de informatie die de instrumentaria leveren, kan betrekken bij het nemen van doorstroombeslissingen. De school zou ervoor kunnen kiezen de informatie die de instrumentaria leveren te gebruiken als *aanvullend gegeven* naast de informatie die de school al ter beschikking heeft. De school kan er echter ook voor kiezen om de informatie die de instrumentaria leveren te gebruiken als enig en *absoluut besliscriterium*. In dat geval zou de school doorstroombeslissingen uitsluitend baseren op de totaalscores van leerlingen op het instrumentarium. Het kiezen voor deze laatste mogelijkheid zou docenten de nodige tijd en inspanning besparen. Bovendien zou de wijze waarop doorstroombeslissingen tot stand komen in het geval van twijfel of verschillen van mening volledig transparant zijn voor de ouders en hun leerlingen. Desondanks is het niet waarschijnlijk dat veel scholen voor deze mogelijkheid zullen kiezen. Zoals eerder aangegeven, dekken de verschillende instrumenten immers lang niet alle vaardigheden, kenmerken en eigenschappen van leerlingen die bepalend zijn voor studiesucces. Voor docenten die lesgeven in andere vakken dan Nederlands, Engels en wiskunde zal de keuze van deze optie - het voorspel-

lend vermogen van de instrumentaria in aanmerking genomen - waarschijnlijk niet acceptabel zijn.

Resumerend kan gesteld worden dat de ontwikkelde instrumentaria een goede kwa-liteit hebben. Het valt echter niet te verwachten dat het ter beschikking stellen van de instrumentaria het rendement in de bovenbouw van havo en vwo zal beïnvloeden. De instrumentaria bieden echter wel de mogelijkheid om, indien dat nodig is, studie-adviezen aan het einde van het derde leerjaar van de havo en het vwo beter te onderbouwen.

Summary

Tests and Decision Making:
Making Placement Decisions in Secondary Education

The Dutch secondary education system consists of four different school types. Two of these prepare students for higher education. Senior general secondary education ('havo') is a five-year program that prepares students for professional schools and pre-university education ('vwo') is a six-year program that prepares students for admission to university. In 1993, the Dutch Ministry of Education requested Cito, the National Institute for Educational Measurement, to develop instruments that could help havo and vwo schools to improve their procedures for admitting stu-dents to the second stage of secondary education (the fourth and fifth years of havo and the fourth to sixth years of vwo).

A set of instruments was developed for both havo and vwo. Each set consisted of three 'placement' tests and a checklist for general study skills. The first part of this thesis gives an account of the construction process of the placement tests and reports on their reliability and validity. The construction of the study skills checklist is described in De Wit, Heuvelmans, and Sluijter (1995). The second part of this thesis focuses on the role the placement tests and the checklist can play in procedures for admission to the second stage of secondary education.

The tests were constructed for three major subjects taught in havo and vwo: Dutch, English, and mathematics. The tests for Dutch were aimed at measuring reading ability. The English tests measured reading ability and basic writing skills. The tests for mathematics were focused on measuring mathematical abilities and skills which were considered prerequisites for success in mathematics and related subjects in the second stage of secondary education.

In 1994, pretests for each subject were conducted at the end of the third year of havo and vwo. In each pretest, data were collected to determine the psychometric properties of a large number of items. In addition, teachers of the participating students assessed the suitability of each item for inclusion in a placement test. Following each pretest, a calibration study using the OPLM (One-Parameter Logistic Model; Verhelst & Glas, 1996) was conducted. Next, tests were constructed for each subject using the items that fitted the OPLM. A detailed account of the design of each pretest and the pretest results is given in chapter two of this thesis.

Because the instruments had to be available as soon as possible, it was decided to forgo separate norm studies for the tests and the checklist and to establish the properties of each instrument from the pretest data. In chapter three of this thesis, it is shown how - using the fact that all items fit the OPLM - the classical reliability of each test was determined from the pretest data. The local reliability of each test is described by graphically displaying its test information function and the ability density function of the target population. In addition, 'reliability matrices' that contain local reliability information are presented. Reliability matrices were constructed by dividing the score scale of each test into deciles and by determining for each decile the proportion of students with a true score within the same decile, as well as the distinctive proportions of students with true scores within all other deciles. It is concluded that all tests have a good measurement precision.

In this thesis, test validity is considered to be a unified though faceted concept and test validation is seen as empirical evaluation of the meaning and consequences of measurement. In the validation study, presented in chapter four, an argument-based approach (Kane, 1992) was used to underpin the validity of the placement tests that were developed. It is demonstrated that each test has a good content representativeness and it is argued that each test has a good construct representativeness.

Data were collected in the school year 1995-1996 on the school success of the students who participated in the pretests. The criterion for success was a simple dichotomous variable. Both in havo and vwo, students who were in the fifth year were considered to be successful, while students who were still in the fourth year, had dropped out, or had left havo or vwo for a lower level

type of education were considered to be unsuccessful. Logistic regression analyses were performed for both school types to examine the predictive power of six variables: three placement test scores and three checklist scores, provided by teachers of Dutch, English, and mathematics. These analyses showed that the predictive power of the three placement test scores and the three checklist scores was limited. Because the predictive power of tests for study success is never high, this result was not unexpected. The scores on the placement tests for English turned out to be only marginally predictive in the presence of the other predictors.

The results from the logistic regression analyses were used to determine the total scores on both sets of instruments. These total scores were transformed to deciles, and expectation tables were constructed that gave the proportions of successful and unsuccessful students in each decile and the overall proportion of successful and unsuccessful students. To examine the generalizability of the expectation tables for havo and vwo, multi-level analyses were conducted. These analyses showed that the odds of being successful at both the havo and vwo are partly influenced by the specific school students attend. Schools therefore have to interpret the total scores of their students with caution.

The next topic that is considered, in chapter five, is the practical use that can be made of the instruments that were developed. A short description is given of how normative decision theory can be employed to optimize test-based decision making. It is also demonstrated how normative decision theory can be applied to find cutting scores on both sets of instruments that are optimal for rational decision makers. The cutting scores turn out to be considerably sample-dependent.

The results from research on the role standardized tests play in teachers' decision making are discussed next. These results indicate that tests have little influence on teachers' decisions and that teachers tend to give their students the benefit of the doubt when they employ standardized tests to make educational decisions. That is, if test results are higher than expected, teachers are predisposed to revise their decisions in a positive way. If test results are lower than expected, however, they tend to be disregarded.

A structural modeling study is presented that examined the placement decisions of teachers at the end of the first year of secondary education. In the light of the results of this study, it can be assumed that total scores on the instruments that were developed will probably not greatly influence procedures for admission to the second stage of secondary education. Moreover, it can be assumed that when the instruments are used in these procedures, the benefit-of-the-doubt phenomenon will not occur.

In the final chapter of this thesis the results from the previous chapters are summarized. It is concluded that it is not to be expected that introducing the instruments will significantly reduce the number of unsuccessful students in the final stage of secondary education. Nevertheless, using the instruments can im-prove procedures for admitting students to the second stage of secondary education. For instance, when teachers disagree about the best placement for a student, the instruments offer a way to solve these disagree-ments in a more objective manner. And when there are differences of opinion between teachers, on the one hand, and students or their parents, on the other hand, these differences in opinion can be settled in a way that is transparent to all involved.

Literatuur

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W.H. (1988). Validity: an evolving concept. In: H. Wainer, & H.I. Braun (Eds.). *Test validity* (pp. 19-32). Hillsdale: Lawrence Erlbaum.

Airasian, P.W., Kellaghan, T., Madaus, G.F., & Pedulla, J. (1977). Proportion and direction of teacher rating changes of pupil progress attributable to standardized test information. *Journal of Educational Psychology, 69*, 702-709.

Arkes, H.R., & Hammond, K.R. (Eds.). (1986). *Judgment and decision making: an interdisciplinary reader*. Cambridge: Cambridge university press.

Berkel, H.J.M. van. (1984). *De diagnose van toetsvragen*. Proefschrift. Amsterdam: Universiteit van Amsterdam.

Bell, D.E., Raiffa, H., & Tversky, A. (Eds.). (1988). *Decision making: Descriptive, normative and prescriptive interactions*. Cambridge: Cambridge university press.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: MIT press.

Borko, H., & Cadwell, J. (1982). Individual differences in teachers' decision

strategies: An investigation of classroom organization and management decisions. *Journal of Educational Psychology, 74*, 598-610.

Bos, K.T., Cremers-van Wees, L.M.C.M., & Lugthart, E. (1996). *Selectie en verwijzing in de eerste fase voortgezet onderwijs; Deel I: uitkomsten*. Enschede: Universiteit Twente

Brehmer, B. (1980). In one word: not from experience. *Acta Psychologica, 45*, 223-241.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137-154.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In: B. Brehmer, & C.R.B. Joyce (Eds.). *Human Judgment: The SJT view* (pp 75-114). Amsterdam: Elsevier Science.

Brehmer, B., & Joyce, C.R.B. (Eds.). (1988). *Human judgment: The SJT View* (Advances in psychology volume 54). Amsterdam: Elsevier science publishers.

Brunswik, E. (1952). The conceptual framework of psychology. In: *Interna tional Encyclopedia of Unified Science* (Vol. 1, No. 10). Chicago: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193-217.

Cadwell, J. (1980). *Alternative regression models of teacher judgment and decision making*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1980.

Cito Instituut voor Toetsontwikkeling. (1986). *Leestoetsen Nederlands Onderbouw Voortgezet Onderwijs*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993a). *Plaatsingstoetsen Nederlands toets 1; vwo-havo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993b). *Plaatsingstoetsen Nederlands toets 2; havo-mavo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993c). *Plaatsingstoetsen Engels toets 1; vwo-havo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993d). *Plaatsingstoetsen Engels toets 2; havo-mavo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993e). *Plaatsingstoetsen wiskunde toets 1; vwo-havo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1993f). *Plaatsingstoetsen wiskunde toets 2; havo-mavo*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1996a). *Ondersteuning bij studie-adviezen (OSA) 3 havo: plaatsingstoetsen/checklist studievaardigheid*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1996b). *Ondersteuning bij studie-adviezen (OSA) 3 vwo: plaatsingstoetsen/checklist studievaardigheid*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Cito Instituut voor Toetsontwikkeling. (1997). *Eindtoetsbulletin september 1997*.Arnhem: Cito Instituut voor Toetsontwikkeling.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cooksey, R.W. (1988). Social judgment theory in education: current and potential applications. In: B. Brehmer, & C.R.B. Joyce (Eds.). *Human judgment: the SJT View* (Advances in psychology volume 54) (pp. 273-315). Amsterdam: Elsevier science publishers.

Cooksey, R.W., & Freebody, P. (1986). Social judgment theory and cognitive feedback: a general model for analysing educational policies and decisions. *Educational Evaluation and Policy Analysis, 8*, 17-29.

Cronbach, L.J. (1988). Five perspectives on validity argument. In: H. Wainer, & H.I. Braun (Eds.). *Test validity* (pp. 3-17). Hillsdale: Lawrence Erlbaum.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois press.

Dawes, R.M., & Corrigan, B. (1974). Linear Models in Decision Making. *Psychological Bulletin, 81*, 95-106.

Dijk, J. van, & Land, H. van 't (1990). *Plaatsing: analytisch of intuïtief? Een policy capturing onderzoek in het voortgezet onderwijs*. Amsterdam: Universiteit van Amsterdam.

Drenth, P.J.D., & Sijtsma, K. (1990). *Testtheorie: Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu Van Loghum.

Du Bois, P.H. (1970). *A history of psychological testing*. Boston: Allyn and Bacon.

Einhorn, H.J., Kleinmuntz, D.N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*, 465-485.

Evers, A., Vliet-Mulder, J.C. van, & Laak, J. ter (1992). *Documentatie van tests en testresearch in Nederland*. Amsterdam: Nederlands Instituut van Psychologen.

Embretson, S.E. (1983) Construct representation and nomothetic span. *Psychological Bulletin, 93,* 179-197.

Fahrmeier, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models.* New York: Springer.

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests.* Bern: Huber.

Gaag, N.L. van der, Mellenbergh, G.J., & Brink, W.P. van den (1987). Empirical utility functions for pass/fail situations. *Methodika, 2,* 40-52.

Gaag, N.L. van der (1990). *Empirische utiliteiten voor psychometrische beslissingen*. Proefschrift. Universiteit van Amsterdam.

Ganzach, Y. (1995). Nonlinear models of clinical jugment: Meehl's data revisited. *Psychological Bulletin, 118,* 422-429.

Ganzach, Y., & Czaczkes, B. (1995). On detecting nonlinear noncompensato ry judgment strategies: comparison of alternative regression models. *Organizational Behavior and Human Decision Processes, 61,* 168-176.

Garret, H.E. (1937). *Statistics in psychology and education*. New York: Longmans, Green.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Gould, S.J. (1981). *The mismeasure of man*. New York: Norton & Company.

Gross, A.L., & Su, W.H. (1975). Defining a 'fair' or 'unbiased' selection model: a question of utilities. *Journal of Applied Psychology, 60,* 345-351.

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

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion referenced tests. *Journal of Educational Measurement, 10,* 159-170

Hambleton, R.K., & Linden, W.J. van der (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 4,* 373-378

Hammond, K.R. (1955). Probabilistic functioning and the clinical method.

Psychological Review, 62, 255-262.
Hammond, K.R., Stewart, T.R., Brehmer, B., & Steinman D.O. (1975). Social judgment theory. In: M. Kaplan, & S. Schwartz (Eds.). *Human judgment and decision processes* (pp. 271-312). New York: Academic Press.
Harte, J.M. (1995). *Multiattribute evaluation processes: evaluation types and research techniques.* Proefschrift. Universiteit van Amsterdam.
Harte, J.M., & Koele, P. (1995). A comparison of different methods for the elicitation of attribute weights: structural modeling, process tracing and self-reports. *Organizational Behavior and Human Decision Processes, 64,* 49-64.
Heuvelmans, A.P.J.M., & Sanders, P.F. (1993). In: T.J.H.M. Eggen, & P.F. Sanders (red.). *Psychometrie in de praktijk* (pp. 443-469). Arnhem: Cito Instituut voor Toetsontwikkeling.
Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116-131.
Hoogstraten, Joh., & Mellenbergh, G.J. (1978). Relevante variabelen bij het doorverwijzen na de lagere school; een experiment. *Tijdschrift voor Onderwijsresearch, 3,* 161-172.
Hosmer, D.W., & Lemeshow, S, (1989). *Applied logistic regression*. New York: Wiley.
Instituut voor Didactiek en Onderwijspraktijk. (1994). *Voorstudie ten behoeve van de constructie van plaatsingstoetsen Engels voor de bovenbouw van het havo en vwo*. Amsterdam: Instituut voor Didactiek en Onderwijspraktijk van de Vrije Universiteit Amsterdam.
Janssens, F.J.G. (1986). Toetsgebruik in de onderwijspraktijk: stand van zaken.*Tijdschrift voor Onderwijsresearch, 11*, 2-22.
Johnson, W.R., & Doherty, M.E. (1983). Social judgment theory and academic advisement. *Journal of Counselling Psychology, 30,* 271-274
Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge university press.
Kane, M.T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527-535.
Kapel, R.C., & Roessingh, M.J. (1996). Onderwijsmatrix 1996. *Kwartaaltijd-schrift onderwijsstatistieken, 3,* 35-53.
Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1982). *The effects of standardized testing*. Den Haag: Kluwer.

Lamers, L.M., Gaag, N.L. van der, & Mellenbergh, G.J. (1992). Empirical utility functions for selection. *Methodika, 7*, 13-29.

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

Linden, W.J. van der (1985). Decision theoru in educational research and testing.

In: T. Husen, & T.N. Postlethwaite (Eds.). *International encyclopedia of education: Research and studies.* Oxford: Pergamon Press.

Linden, W.J. van der (1990). Applications of decision theory to test based decision making. In: R.K. Hambleton, & J.N. Zaal (1990). *Advances in educational and psychological testing* (pp. 129-156). Boston: Kluwer.

Linden, W.J. van der (1998). A decision theory model for course placement. *Journal of Educational and Behavioral Statistics, 23*, 18-34.

Linden, W.J. van der, & Mellenbergh, G.J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1*, 593-597.

Linden, W.J. van der, & Mellenbergh, G.J. (1978). Coefficients for tests from a decision theoretic point of view. *Applied Psychological Measurement, 2*, 119-34.

Linden, W.J. van der, & Hambleton, R.K. (1997). Item response theory: brief history, common models and extensions. In: W.J. van der Linden, & R.K. Hambleton (Eds.). *Handbook of modern item response theory*. New York: Springer.

Lindgren, B.W. (1976) *Statistical theory* (3rd ed.). New York: Macmillan

Linn, R.L. (1997). Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice, 16*, 2, 14-16.

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

Maniscalco, C.T., Doherty, M.E., & Ullman, D.G. (1980). Assessing discrimination: An application of social judgment technology. *Journal of Applied Psychology, 65*, 284-288

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Meehl, P.E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota press.

Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*, 2, 16-18.

Mellenbergh, G.J. (1993). Beslissen met tests en studietoetsen. In: P. Koele, & J.van der Pligt (red.). *Beslissen en beoordelen* (pp. 96-120). Amsterdam: Boom.

Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91-100.

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*, 293-299.

Mellenbergh, G.J., & Linden, W.J. van der (1981). The linear utility model for optimal selection. *Psychometrika, 46,* 283-294.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In: H. Wainer, & H.I. Braun (Eds.). *Test validity* (pp. 33-45). Hillsdale: Lawrence Erlbaum.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton: Educational Testing Service; Research Report 90-11.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 4, 5-8.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In: R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 335-366). Washington, DC: American Council on Education.

Ministerie van Onderwijs, Cultuur en Wetenschappen (1996). *Voortgezet onderwijs in cijfers*. Den Haag: Sdu.

Nagelkerke, S.J.D. (1991). A note on general definition of the coefficient of determination, *Biometrika, 78*, 691-692.

Nederlands Instituut van Psychologen. (1988). *Richtlijnen voor ontwikkeling en gebruik van psychologische tests en studietoetsen*. Amsterdam: Nederlands Instituut van Psychologen.

Novick, M.R., & Lindley, D.V. (1978). The use of more realistic utility functions in educational applications. *Journal of Educational Measurement, 15*, 181-191.

Payne J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 23*, 86-112.

Petersen, N.S., & Novick, M.R. (1976). An evaluation of some models for

culture-fair selection. *Journal of Educational Measurement, 13,* 3-31

Popham, W.J. (1997). Consequential validity: right concern - wrong concept. *Educational Measurement: Issues and Practice, 16,* 2, 9-13.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish institute of educational research.

Roose, J.E., & Doherty, M.E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior and Human Performance, 16,* 231-249.

Roosmalen, W.M.M. van, & Sluijter, C. (1991). De constructie van toetsen voor classificatie in het voortgezet onderwijs. In: J. Hoogstraten, & W.J. van der Linden (red.). *Methodologie* (pp. 47-56). Amsterdam: SCO.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: what's really happening? *Phi Delta Kappan, 62,* 631-634.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229-244.

Sanders P.F., & Eggen, T.J.H.M. (1993). Inleiding. In: T.J.H.M. Eggen, & P.F. Sanders (red.). *Psychometrie in de praktijk* (pp. 1-16). Arnhem: Cito Instituut voor Toetsontwikkeling.

Shavelson, R.J., & Stern, P. (1981). Research on teachers's pedagogical thoughts, judgments, decisions and behaviors. *Review of Educational Research, 51,* 485-498.

Shavelson, R.J., Webb, N.M., & Burstein, L. (1986). Measurement of teaching.
In: M.C. Wittrock (Ed.). *Handbook of Research on Testing* (3rd. ed.) New York: MacMillan.

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammond (Ed.). *Review of research in education: Vol 19* (pp.405-450). Washington, DC: American Educational Research Association.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity.
Educational Measurement: Issues and Practice, 16, 2, 5-8; 13; 24.

Shulman, L.S., & Elstein, A.S. (1975). Studies of problem solving, judgment and decision making: Implications for educational research. In: F.N. Kerlinger (Ed.). *Review of Research in Education. Vol 3.* Itasca: Peacock.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression

approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649-744.

Sluijter, C. (1988a). *Project eindtoetsen brugklas; resultaten van een behoeftenonderzoek*. (Interne documentatie nr. 272). Arnhem: Cito Instituut voor Toetsontwikkeling.

Sluijter, C. (1988b). *Toetsen als hulpmiddel bij het determineren van leerlin gen; Een besliskundige benadering* (Specialistisch Bulletin nr. 63). Arnhem: Cito Instituut voor Toetsontwikkeling.

Sluijter, C. (1995). De rol van toetsen (en andere meetinstrumenten) bij selectie voor de tweede fase. *Dekanoloog,* 9, 368-373.

Sluijter, C. (1998). *Verantwoording plaatsingstoetsen voor drie havo en drie vwo*. (OPD-memorandum 98-4). Arnhem: Cito Instituut voor Toetsontwikkeling.

Sluijter, C., Boertien, H., Klijn, W.J. de, & Roosmalen, W.M.M. van (1991). *De constructie van plaatsingstoetsen* (Onderzoeksrapporten beginfase voorgezet onderwijs nr. 6). Arnhem: Cito Instituut voor Toetsontwikkeling.

Snow, R.E. (1968). Brunswikian approaches to research on teaching. *American Educational Research Journal, 5,* 475-489.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid; De ontwikkeling van een domeingericht meetinstrument.* Proefschrift. Universiteit Twente.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance, 23,* 86-112.

Ullman, D.G., & Doherty, M.E. (1984). Two determinants of the diagnosis of hyperactivity: the child and the clinician. *Advances in Developmental and Behavioral Pediatrics, 5,* 167-219.

Uiterwijk, J.H. (1994). *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen*. Proefschrift. Katholieke Universiteit Brabant.

Vale, D.C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Verhelst, N.D. (1993) Itemresponstheorie. In: T.J.H.M. Eggen, & P.F. Sanders (red.). *Psychometrie in de praktijk* (pp. 83-178). Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek*. (PPON-rapport, nr. 4). Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D., & Glas, C.A.W. (1995). The one parameter logistic model. In:

G.H. Fischer, & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments, and applications* (pp. 215-239). New York: Springer.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: Computer program and manual.* Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D., & Verstralen, H.H.F.M. (1996). *SAUL handleiding.* Arnhem: Cito Instituut voor Toetsontwikkeling.

Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.J.H.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* Measurement and Research Department Reports 91-10, Arnhem: Cito Instituut voor Toetsontwikkeling.

Verstralen, H.H.F.M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations.* Arnhem: Cito Instituut voor Toetsontwikkeling.

Vos, H.J. (1994). *Simultaneous optimization of test-based decisions in education.* Proefschrift. Universiteit Twente.

Vrijhof, B.J., Mellenbergh, G.J., & Brink, W.P. van den (1983). Assessing and studying utility functions in psychometric decision theory. *Applied Psychological Measurement, 7*, 341-357.

Wainer, H. (1987). *The first four milennia of mental testing; from ancient China to the computer age* (Research report 87-34). Princeton: Educational Testing Service.

Westenberg, M.R.M. (1991). *Response modes and decision strategies.* Proefschrift Universiteit van Amsterdam.

Westenberg, M.R.M., & Koele, P. (1990). Response modes and decision strategies. In: K. Borchering, O.I. Larichev, & D.M. Messick (Eds.). *Contemporary issues in decision making* (pp. 159-170). Amsterdam: North-Holland.

Westenberg, M.R.M., & Koele, P. (1992). Response modes, decision processes and decision outcomes. *Acta Psychologica, 80*, 169-184.

Wit, C.A. de, Heuvelmans A.P.J.M., & Sluijter, C. (1995). *Verantwoording checklist studievaardigheid* (OPD-memorandum 95-1). Arnhem: Cito Instituut voor Toetsontwikkeling.

Bijlagen

Bijlage A
Normtabellen bij de ontwikkelde plaatsingstoetsen

Tabel A.1
Normtabellen bij de toetsen voor Nederlands

Decielscore	Scores toets drie havo	Scores toets drie vwo
1	0 - 36	0 - 45
2	37 - 42	46 - 52
3	43 - 48	53 - 57
4	49 - 52	58 - 61
5	53 - 56	62 - 65
6	57 - 60	66 - 69
7	61 - 65	70 - 74
8	66 - 70	75 - 78
9	71 - 77	79 - 85
10	78 - 115	86 - 115

Tabel A.2
Normtabellen bij de toetsen voor Engels

Decielscore	Scores toets drie havo	Scores toets drie vwo
1	0 - 28	0 - 31
2	29 - 35	32 - 39
3	36 - 40	40 - 45
4	41 - 44	46 - 50
5	45 - 47	52 - 56
6	48 - 51	57 - 61
7	52 - 55	62 - 66
8	56 - 58	67 - 71
9	59 - 63	72 - 78
10	64 - 81	79 - 97

Tabel A.3
Normtabellen bij de toetsen voor wiskunde

Decielscore	Scores toets drie havo	Scores toets drie vwo
1	0 - 11	0 - 18
2	12 - 16	19 - 25
3	17 - 20	26 - 31
4	21 - 24	32 - 36
5	25 - 28	37 - 40
6	29 - 32	41 - 44
7	33 - 36	45 - 48
8	37 - 41	49 - 51
9	42 - 46	52 - 54
10	47 - 57	55 - 60

Bijlage B
Resultaten beoordelingsanalyses van determinatiebeslissingen

Tabel B.1

Resultaten van de regressieanalyses in de intervalconditie met vier kenmerken (GCE: gemiddeld cijfer op het eindrapport; LST: leer- en studievaardigheden; LIN: leerlinginstelling; RCP: resultaat op Cito-plaatsingstoetsen; R^2: percentage door het model verklaarde variantie)

	Gestandaardiseerde regressiegewichten				
Docent	*GCE*	*LST*	*LIN*	*RCP*	R^2
1	0,92	0,15	0,15	0,18	0,87
2	0,92	0,14	0,04	0,12	0,85
3	0,68	0,37	0,30	0,39	0,73
4	0,26	0,08	0,16	0,92	0,90
5	0,86	0,13	0,04	0,34	0,85
6	0,77	0,26	0,17	0,42	0,79
7	0,99	0,16	0,03	0,04	0,95
8	0,89	0,21	0,15	0,07	0,81
9	0,81	0,15	0,46	0,25	0,85
10	0,91	0,29	0,08	0,22	0,91
11	0,25	0,08	0,35	0,76	0,67
12	0,73	0,32	0,30	0,50	0,86
13	0,84	0,24	0,27	0,32	0,85
14	0,40	0,25	0,43	0,70	0,75
15	0,94	0,01	-0,01	0,06	0,89
16	0,93	0,18	0,11	0,09	0,87
17	0,97	0,14	0,10	0,02	0,94
18	0,91	0,11	-0,02	0,14	0,85
19	0,83	0,43	0,15	0,12	0,84
20	0,55	0,47	0,36	0,47	0,74
21	0,90	0,28	0,10	0,15	0,86
22	0,37	0,53	0,23	0,52	0,65
23	0,72	0,30	0,29	0,29	0,67
24	0,95	0,15	0,17	0,17	0,93
25	0,91	0,19	0,18	0,15	0,86
26	0,22	0,27	0,11	0,86	0,83
27	0,23	0,65	0,32	0,42	0,66
28	0,38	0,37	0,44	0,60	0,70
29	0,15	0,91	-0,04	0,15	0,87

Tabel B.2

Resultaten van de discriminantanalyses in de nominale conditie met vier kenmerken (N: aantal beoordeelde leerlingen; r: canonische correlatie; GCE: gemiddeld cijfer op het eindrapport; LST: leer- en studievaardigheden; LIN: leerlinginstelling; RCP: resultaat op Cito-plaatsingstoetsen; N_{cor}: aantal leerlingen bij wie het oordeel correct voorspeld wordt door het model)*

			Gestandaardiseerde discriminantgewichten				
Docent	N	r*	*GCE*	*LST*	*LIN*	*RCP*	N_{cor}
1	60	0,78	0,46	1,05	0,35	0,53	49
2	60	0,77	0,42	1,03	0,33	0,58	50
3	56	0,77	1,05	0,80	0,19	0,29	48
4	60	0,70	0,74	0,76	0,81	0,30	43
5	60	0,85	1,04	0,26	-0,02	0,29	50
6	60	0,86	1,07	0,42	-0,13	0,33	53
7	60	0,86	1,14	0,24	0,31	0,77	51
8	60	0,87	1,14	0,23	-0,03	0,81	53
9	60	0,86	1,03	0,26	0,04	-0,12	55
10	60	0,79	0,70	0,58	0,73	1,06	48
11	60	0,90	1,04	0,18	-0,32	0,13	50
12	60	0,89	1,01	0,09	0,02	0,02	48
13	60	0,77	0,99	0,78	0,59	0,66	50
14	60	0,79	1,08	0,49	0,49	0,06	44
15	60	0,85	1,13	1,03	0,76	0,59	47
16	60	0,85	1,01	0,11	-0,24	-0,14	51
17	60	0,89	1,08	0,30	0,17	0,39	52
18	59	0,89	1,07	0,09	0,41	0,13	49
19	60	0,82	1,11	0,44	0,47	0,51	52
20	59	0,81	0,98	0,43	0,25	0,85	46
21	60	0,86	1,12	0,41	0,48	0,20	49
22	60	0,85	1,13	0,31	0,28	0,75	47
23	60	0,87	1,21	0,64	0,89	0,98	53
24	60	0,81	1,09	0,53	0,17	0,37	49
25	59	0,26*					29
26	60	0,94	1,03	0,01	-0,33	0,01	57
27	60	0,84	0,80	0,86	0,71	1,13	53
28	60	0,90	1,01	0,06	-0,19	0,07	53
29	60	0,89	1,03	0,19	0,13	-0,09	53

* eerste discriminantfunctie niet significant

Bijlage C
Vergelijking van docentoordelen

Tabel C.1
Gemiddeld oordeel van iedere docent in de intervalconditie voor het boekje met RCP-schaal, het boekje zonder RCP-schaal en het verschil tussen beide gemiddelde oordelen, bij de 15 geselecteerde profielen met hoge scores en de 15 geselecteerde profielen met lage scores op de RCP-schaal

	Gemiddeld oordeel voor de 15 profielen met een hoge score op de *RCP*-schaal			Gemiddeld oordeel voor de 15 profielen met een lage score op de *RCP*-schaal		
Docent	Mét RCP	Zonder RCP	Verschil	Mét RCP	Zonder RCP	Verschil
1	5,07	4,33	0,73	5,73	6,27	- 0,53
2	3,87	3,00	0,87	4,27	5,40	- 1,13
3	4,67	3,27	1,40	4,40	6,13	- 1,73
4	6,87	4,13	2,73	2,80	5,73	- 2,93
5	4,47	4,07	0,40	3,80	5,87	- 2,07
6	5,15	4,15	1,00	4,53	5,73	- 1,20
7	4,29	4,21	0,07	5,47	5,40	0,07
8	4,07	4,73	-0,67	5,67	6,60	- 0,93
9	5,00	3,80	1,20	5,60	5,27	0,33
10	4,67	4,13	0,53	5,07	5,87	- 0,80
11	3,76	2,87	0,80	1,80	4,87	- 3,07
12	5,20	3,73	1,47	4,47	6,20	- 1,73
13	4,53	4,00	0,53	4,60	5,40	- 0,80
14	7,14	4,07	3,07	5,00	6,47	- 1,47
15	4,07	4,07	0,00	5,20	5,60	- 0,40
16	4,00	3,53	0,47	5,27	5,40	- 0,13
17	3,67	3,73	-0,67	4,93	5,87	- 0,93
18	3,60	2,40	1,20	3,87	4,67	- 0,80
19	4,53	4,20	0,33	5,76	6,13	- 0,47
20	4,47	3,73	0,73	4,07	5,80	- 1,73
21	3,93	3,67	0,27	4,87	5,67	- 0,80
22	5,60	3,33	2,27	4,20	6,27	- 2,07
23	4,27	4,27	0,00	4,53	6,13	- 1,60
24	4,67	4,47	0,20	5,27	5,53	- 0,27
25	3,50	3,36	0,14	4,13	5,40	- 1,27
26	5,13	3,60	1,53	4,87	5,73	- 0,87
27	4,93	4,33	0,60	3,80	6,67	- 2,87
28	3,73	2,80	0,93	4,07	4,36	- 0,29

Tabel C.2

Veranderingen in oordelen voor iedere docent in de nominale conditie van het boekje met RCP-schaal naar het boekje zonder RCP-schaal voor de 15 geselecteerde profielen met hoge scores en de 15 geselecteerde profielen met lage scores op de RCP-schaal

	Veranderingen voor de 15 profielen met een hoge score op de *RCP*-schaal			Veranderingen voor de 15 profielen met een lage score op de *RCP*-schaal		
Docent	Positief	Negatief	Geen	Positief	Negatief	Geen
1	1	3	11	6	0	9
2	0	6	9	5	2	6
3	0	2	12	4	0	10
4	0	2	13	4	0	10
5	1	2	12	6	0	9
6	0	7	8	8	0	7
7	0	8	7	5	1	9
8	0	1	14	1	3	11
9	0	8	7	8	0	7
10	0	6	9	1	1	13
11	0	3	12	6	0	8
12	2	8	5	4	1	10
13	1	1	13	5	1	9
14	0	7	8	1	4	10
15	1	1	13	3	1	11
16	1	1	13	2	1	12
17	0	5	10	0	4	11
18	0	6	9	2	0	13
19	0	8	6	9	0	6
20	1	1	13	1	1	13
21	0	2	13	9	0	6
22	0	6	9	4	0	11
23	0	1	14	7	0	8
24	3	7	3	7	3	5
25	0	1	14	0	0	15
26	1	10	4	11	0	4
27	0	3	12	1	0	14
28	0	1	14	2	0	13