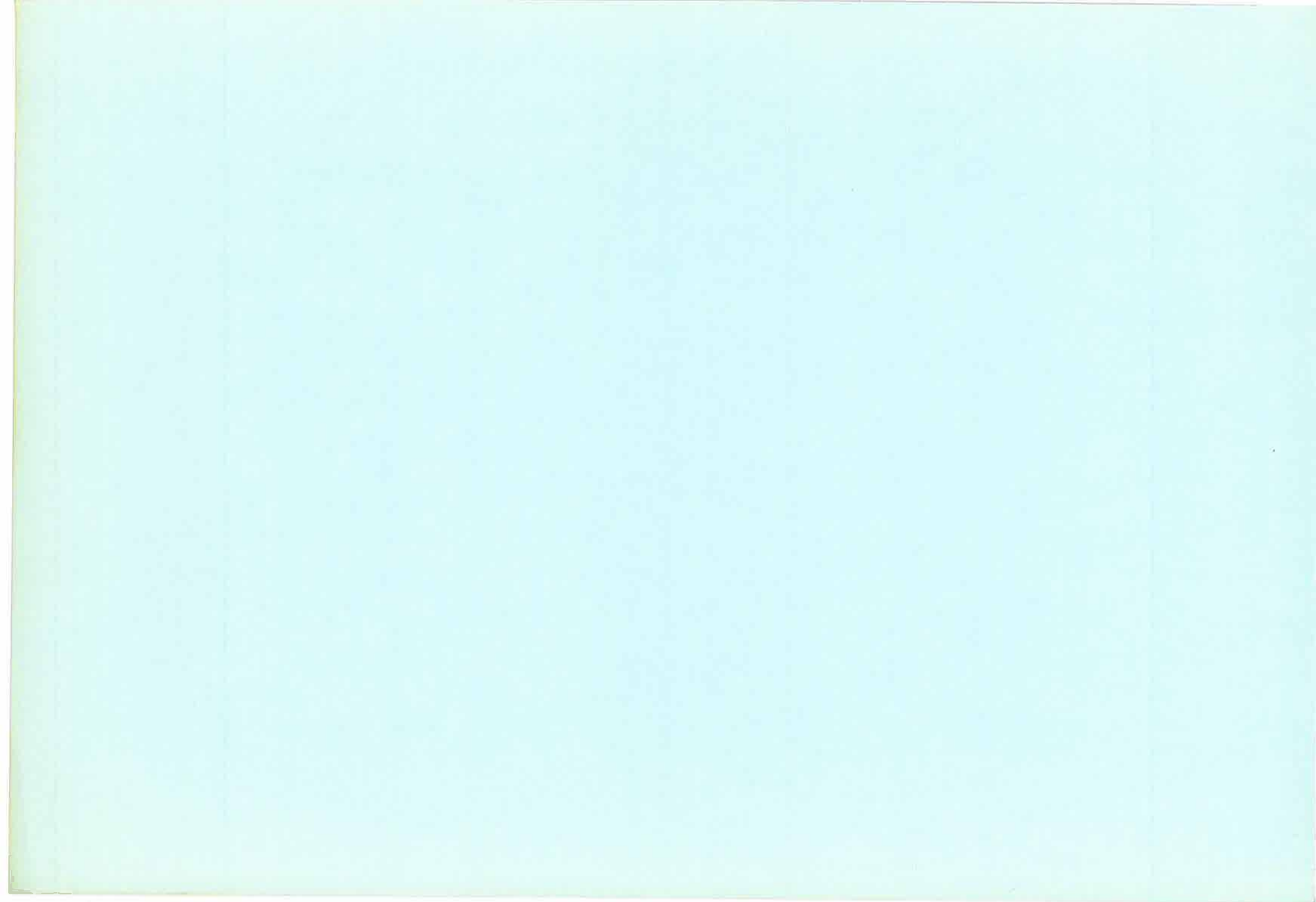


Verantwoording plaatsingstoetsen voor drie havo en drie vwo

C. Sluiter

Cito



OPD Memorandum

98-4

Verantwoording plaatsingstoetsen voor drie havo en drie vwo

C. Sluijter

Cito
Arnhem, september 1998



© Cito Arnhem 1998

Uit deze uitgave mogen zonder toestemming van de auteur korte aanhalingen met volledige bronvermelding worden opgenomen.

Inhoudsopgave

1	De constructie van plaatsingstoetsen voor drie havo en drie vwo	1
1.1	Uitgangspunten	1
1.2	De proefafnames	8
1.2.1	Opzet en uitvoering	8
1.2.2	De resultaten voor het vak Nederlands	21
1.2.3	De resultaten voor het vak Engels	24
1.2.4	De resultaten voor het vak wiskunde	28
1.3	Het samenstellen en normeren van de toetsen	31
1.3.1	Het samenstellen van de toetsen	31
1.3.2	De methode van normeren	32
1.3.3	Beschrijving van de toetsen	33
1.4	Kort resumé	34
2	De meetnauwkeurigheid van de plaatsingstoetsen	35
2.1	Betrouwbaarheid en toetsinformatie	35
2.2	Het bepalen van de meetnauwkeurigheid van de toetsen	38
2.2.1	Het bepalen van de lokale meetnauwkeurigheid	38
2.2.2	Het bepalen van de betrouwbaarheid van de toetsen	41
2.3	Resultaten	44
2.3.1	De toetsinformatiefuncties	44
2.3.2	Betrouwbaarheidsmatrices bij de toetsen	47
2.3.3	De geschatte betrouwbaarheid van de toetsen	51
2.4	De drie soorten informatie over meetnauwkeurigheid vergeleken	52
3	De validiteit van de plaatsingstoetsen	55
3.1	Het begrip validiteit	55
3.2	Het bepalen van de validiteit van de toetsen	62
3.3	De inhoudsrepresentativiteit van de toetsen	64
3.4	De begripsrepresentativiteit van de toetsen	65
3.5	Het voorspellend vermogen van de toetsen	73

4 Samenvatting en conclusies

93

Literatuur

97

Bijlagen

103

1 De constructie van plaatsingstoetsen voor drie havo en drie vwo

Eind 1993 kreeg het Cito van het ministerie van Onderwijs, Cultuur en Wetenschappen de opdracht een instrumentarium te ontwikkelen dat scholen ondersteuning kon 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.

Aan deze opdracht is 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 studievoordigheid ontwikkeld die het docenten mogelijk maakt de studievoordigheid van een leerling systematisch te beoordelen en in een score uit te drukken. Dit memorandum bevat een beschrijving van de ontwikkeling, meetnauwkeurigheid en validiteit van de plaatsingstoetsen. De checklist komt in dit memorandum alleen zijdelings aan de orde. Voor een verslag van de ontwikkeling, normering en validering van de checklist wordt verwezen naar De Wit, Heuvelmans, & Sluiter (1995) en naar de handleidingen bij de instrumentaria die voor drie havo en drie vwo ontwikkeld zijn (Cito Instituut voor Toetsontwikkeling, 1996a, 1996b).

1.1 Uitgangspunten

Aan de keuze voor de ontwikkeling van toetsen voor de vakken Nederlands, Engels en wiskunde en een checklist studievoordigheid 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 memorandum geen aandacht aan de ontwikkeling, normering en validering van deze checklist.

De constructie van toetsen is een complex en tijdrovend proces dat uit een reeks opeenvolgende stappen bestaat (Millman & Greene, 1986; 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 proefafnames;
- 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 wordt ingegaan op de constructie van de items. Vervolgens komen de opzet en de resultaten van de proefafnames aan de orde. 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, 1997). Om hun functie te benadrukken hebben dergelijke 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. In dit kader is tevens besloten in iedere toets een zodanig aantal items op te nemen dat de afnametijd per toets maximaal twee lesuren van vijfenveertig minuten zou bedragen.

Een volgende eis was dat de toetsen curriculumonafhankelijk moesten 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 de toetsen 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 immers 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 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 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 14 betrekking op het kunnen herkennen van het hoofdonderwerp van een tekst en 14 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 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 geraden 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 beter 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. Bovendien moesten voornoemde kennis en vaardigheden binnen en buiten het vak toepassingen hebben die los staan van de context waarin ze oorspronkelijk verworven zijn, omdat een noodzakelijke voorwaarde voor voortgang in de ontwikkeling van de leerlingen is dat ze de verworven kennis en vaardigheden in nieuwe situaties kunnen gebruiken.

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 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 ontwikkelde items komt derhalve ‘praktische’ formele wiskunde aan de orde, zoals die naar verwachting in de toekomst in de onderbouw van de havo en het vwo aan de orde zal komen en in de bovenbouw van de havo en het vwo in de vakken wiskunde A en B door leerlingen beheerst moet 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 ook rekening gehouden met het feit dat in de bovenbouw van de havo en het vwo nu 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.

1.2 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 studievoordigheid 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.

1.2.1 Opzet en uitvoering

In januari 1994 is een aselechte 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 uit de door vaksecties opgegeven klassen een op toeval gebaseerde selectie gemaakt.

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 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 in vier complementaire deelverzamelingen, zogeheten clusters, ondergebracht. 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 1.1 geeft het bij de aanbidding 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 1.1
Proefafnamedesigns voor Nederlands en Engels

Toetsboekje	Cluster				Klasset
	1	2	3	4	
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 kettingsdesign 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 zij alle mogelijke paren gevormd. In totaal konden derhalve negen verschillende boekjesparen aan klassen worden 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 werd 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. Het proefafnamedesign voor Nederlands en Engels leverde hierbij als voordeel dat iedere docent alle items uit twee van de drie itemverzamelingen kon beoordelen.

Het verslagformulier bood docenten de gelegenheid informatie te verstrekken over het verloop van een proefafname. Ze 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 studievoordigheid in te vullen.

Inschrijving en deelname

Tabel 1.2 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.

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 ook al medewerking verleend had aan een proefafname van afsluitingstoetsen basisvorming.

Tabel 1.2

Aantal ingeschreven en daadwerkelijk deelnemende klassen per vak en klastype

		Klastype			Totaal
		Havo	Havo/vwo	Vwo	
Nederlands	Ingeschreven	75	21	101	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

Gegevensverwerking

Na verzameling van al 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 rangordening 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. Dedden 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 van de leerlingen 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.

In tabel 1.3 staat hoeveel bruikbare leerlingrecords er uiteindelijk na de gegevensverwerking overbleven voor elk vak en welk klastype de betreffende leerlingen zich bevonden. Van de 2.517 bruikbare records voor Nederlands zijn er bijvoorbeeld 836 afkomstig van leerlingen uit drie-havo-klassen, 314 uit drie-havo/vwo klassen en 1.367 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.

Tabel 1.3
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

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 5.544 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
1.035		X	X
616	X		
1.153		X	
839			X
Totaal 5.544	2.517	3.593	3.170

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 ($j = 0, 1, \dots, m_i$) aan zal nemen als:

$$f_{ij}(\theta) = \frac{\exp \left[a_i \left(j\theta - \sum_{g=1}^j \beta_{ig} \right) \right]}{1 + \sum_{h=1}^{m_i} \exp \left[a_i \left(h\theta - \sum_{g=1}^h \beta_{ig} \right) \right]}, \quad (j=0, 1, \dots, m_i), \quad (1.1)$$

waarbij de β_{ig} ($g = 1, \dots, m_i$), 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 (1.1) 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 (1.1) dezelfde vorm als het partial credit model (Masters, 1982). Is bovendien $m_i = 1$ voor alle items i , dan heeft uitdrukking (1.1) 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 bij deze itemsoorten gebruikt is bij het analyseren van de leerlingantwoorden beschrijft de kans dat de score X_i op item i van een persoon met vaardigheid θ de waarde 1 aan zal nemen als:

$$f_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]} \quad (1.2)$$

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-schattingmethode is dat het mogelijk is de parameters van de modellen in (1.1) en (1.2) 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 mis-specificatie 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 havo, havo/vwo en vwo aan een nader onderzoek te onderwerpen is gebruik gemaakt van het eveneens op het Cito ontwikkeld 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 dezelfde 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:

$$G = \left[\prod_{i=1}^n a_i \right]^{\frac{1}{n}}, \quad (1.3)$$

waarbij n staat voor het aantal items dat bij een 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:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}. \quad (1.4)$$

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 0 gesteld. 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 (1.4) 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 scorbaarheid. Daartoe is voor de betreffende items de beoordelaarsovereenstemming onderzocht. In dat kader zijn zowel voor Nederlands als Engels op aselechte wijze zestig door leerlingen ingevulde toetsboekjes geselecteerd. Het betrof hier twee maal 30 boekjes uit een en dezelfde klas, 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 aselechte 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 beoordelaars-overeenstemmingscoëfficiënt, ρ^2 , geschat; $\hat{\rho}^2$ wordt gedefinieerd als:

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + (\hat{\sigma}_b^2 + \hat{\sigma}_{res}^2) / k}, \quad (1.5)$$

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 (1.5), $\hat{\sigma}_b^2$, 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 beoordeling van k willekeurig gekozen beoordelaars en de gemiddelde beoordeling 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.

1.2.2 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 betreffende analyse 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. De parameters van de 113 items voor Nederlands staan in bijlage A1.

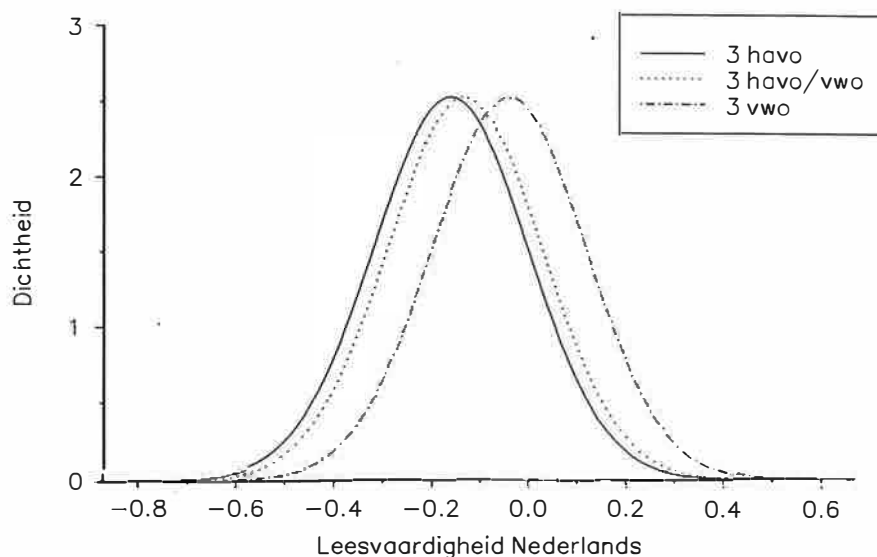
Tabel 1.5

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 1.5 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 vwo en havo/vwo.



Figuur 1.1

Kansdichtheidsfuncties voor de drie klastypen onder aanname van normaliteit voor leesvaardigheid Nederlands

Figuur 1.1 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 1.5 en figuur 1.1 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 leerlingen 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-klassen en leerlingen uit 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 de 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 coëfficiënt 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. Bijlage A2 bevat een gedetailleerd overzicht van de resultaten van het onderzoek naar de beoordelaarsovereenstemming voor de items Nederlands.

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 vindt de cloze-items moeilijk. Ten slotte moet worden opgemerkt dat uit de reacties van docenten bleek dat ze de correctie van de kort-open-antwoord items teveel tijd vonden kosten.

1.2.3 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 beoordelaars-overeenstemming 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 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 betreffende analyse liet voor geen enkele van de resterende 134 items een gebrek aan modelfit zien. Daarom mag aangenomen worden dat de drie soorten items die door de vakmedewerker ontwikkeld zijn een en dezelfde vaardigheid meten. De parameters van de 134 items voor Engels staan in bijlage B1.

Tabel 1.6 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 1.6 wordt verwezen naar de toelichting bij tabel 1.5 in paragraaf 1.2.2.

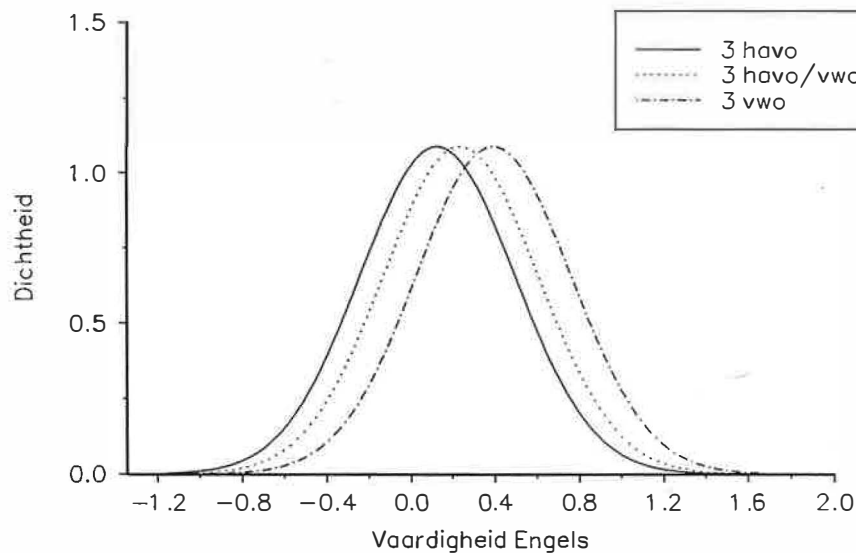
Tabel 1.6

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 1.2 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 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.



Figuur 1.2

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

Tabel 1.6 en figuur 1.2 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 vaardig 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.

Een vergelijking tussen de proefafnames voor de vakken Nederlands en Engels 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 Engels wat groter is dan bij de proefafname voor Nederlands. Het verschil in gemiddelde vaardigheid tussen leerlingen uit drie-havo/vwo-klassen en leerlingen uit drie vwo 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 voor 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. Bijlage B2 bevat een gedetailleerd overzicht van de resultaten van het onderzoek naar de beoordelaarsovereenstemming voor de items Engels.

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 vertrouwdsheid van leerlingen met dit soort items. Ook gaven docenten regelmatig kritiek op de correctievoorschriften.

1.2.4 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 van de 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. De parameters van deze items staan in bijlage C1.

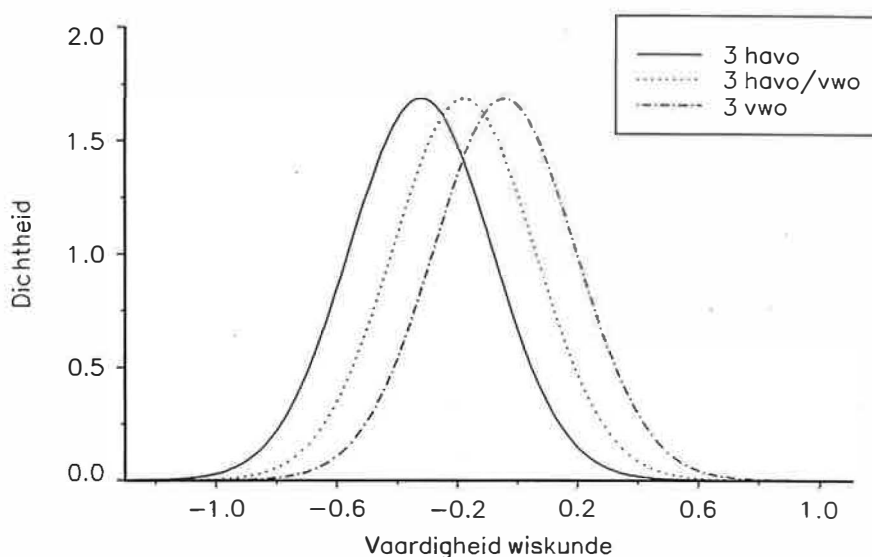
Tabel 1.7 bevat de resultaten van deze SAUL-analyse die is uitgevoerd om de verschillen in vaardigheid tussen de klastypen drie havo, drie havo/vwo en drie vwo te bepalen op de vaardigheidsschaal voor wiskunde. Het geschatte algemeen gemiddelde μ op de gestandaardiseerde vaardigheidsschaal is -0,9826 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 1.7 wordt verwezen naar de toelichting bij tabel 1.5 in paragraaf 1.2.2.

Tabel 1.7

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 1.3 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 1.3

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

Tabel 1.7 en figuur 1.3 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 duidelijk 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 de 0,60 en 0,80. 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. Bijlage C2 bevat een gedetailleerd overzicht van de resultaten van het onderzoek naar de beoordelaarsovereenstemming voor de items wiskunde.

Docentvragenlijst en verslagformulier

Uit de antwoorden van docentvragenlijst 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.

1.3 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.

1.3.1 *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.

1.3.2 De methode van normeren

Het was de wens van het ministerie van Onderwijs, Cultuur en Wetenschappen dat de instrumentarium zo snel mogelijk ter beschikking van de scholen zou 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 $0, 1, \dots, M_t$ 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:

$$p(S_t | \theta) = \sum_{\{\mathbf{x}: r(\mathbf{x}) = s_t\}} \prod_i f_{ij}(\theta). \quad (1.6)$$

Hierbij heeft $f_{ij}(\theta)$ dezelfde vorm als uitdrukking (1.1) of, indien sprake is van dichotome items, uitdrukking (1.2), 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:

$$p(S_t) = \int_{-\infty}^{\infty} p(S_t | \theta) dG(\theta), \quad (1.7)$$

waarbij $G(\theta)$ de vaardigheidsverdeling is. Uitdrukking (1.7) 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 (1.7) 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:

$$P(S_t \leq r_t) = \sum_{S_t \leq r_t} p(s_t). \quad (1.8)$$

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 D.

1.3.3 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 deze 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 Engels 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 de toets is 57. Voor drie vwo is een toets opgesteld bestaande uit 28 items. De toets heeft een maximumscore van 60.

1.4 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 twee en drie. Hoofdstuk twee heeft betrekking op de meetnauwkeurigheid en hoofdstuk drie op de validiteit van de toetsen.

2 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 en de validiteit van de ontwikkelde toetsen. De validiteit van de ontwikkelde toetsen is het onderwerp van hoofdstuk drie.

2.1 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:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \quad (2.1)$$

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:

$$\rho_{XT}^2 = \sigma_T^2 / \sigma_X^2. \quad (2.2)$$

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 levert een kleinere waarde op 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:

$$I_t(\theta) = \sum_{i \in t} I_i(\theta). \quad (2.3)$$

De waarde die (2.3) 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 gedaan kunnen worden. Voor de standaardfout SE van een met behulp van een toets t geschatte vaardigheid geldt namelijk:

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

Daardoor is het mogelijk een asymptotisch $(1 - \alpha)$ -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 met 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.

2.2 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 paragraaf 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.

2.2.1 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 $K(k = 1, 2, \dots, K)$ geordende klassen. Iedere klasse heeft betrekking op een bepaald interval op een schaal. Tezamen bestrijken de klassen het volledige bereik van een schaal. Iedere cel b_{ij} in een betrouwbaarheidsmatrix 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 $i, j = 1, 2, \dots, 10$.

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 $P_k = k/10 (k = 1, 2, \dots, 10)$. 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 $P(s) (s = 1, 2, \dots, M)$ van de discrete variabele s benaderd door een cumulatieve scoreverdeling $D(u)$ van een continue variabele u . Voor $D(u)$ geldt:

$$\begin{aligned} D(u + 0,5) &= P(u) & (\forall u \in \{0, 1, \dots, M\}); \\ D(u + 0,5) &= P(\lfloor u \rfloor) + (u - \lfloor u \rfloor)(P(\lceil u \rceil) - P(\lfloor u \rfloor)) & (\forall u \notin \{0, 1, \dots, M\}). \end{aligned}$$

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:

$$\begin{aligned} H(\theta | u) &= G(\theta | u) & (\forall u \in \{0, 1, \dots, M\}); \\ H(\theta | u) &= G(\theta | u) + (u - \lfloor u \rfloor) (G(\theta | \lceil u \rceil) - G(\theta | \lfloor u \rfloor)) & (\forall u \notin \{0, 1, \dots, M\}). \end{aligned}$$

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:

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

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.

2.2.2 Het bepalen van de betrouwbaarheid van de toetsen

In het eerste hoofdstuk is reeds toegelicht dat het niet mogelijk was de toetsen in hun uiteindelijke vorm in een apart normeringsonderzoek aan leerlingen voor te leggen. Daarom was het niet mogelijk hun betrouwbaarheid 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 er uiteindelijk in 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:

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

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

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

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:

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

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

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

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

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

Voor de variantie van de marginale geobserveerde score geldt dus:

$$\sigma_{X_i}^2 = m_{2_i} - \tau_i^2. \quad (2.4)$$

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

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

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

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

Derhalve geldt voor de variantie van de marginale ware scores:

$$\sigma_{\tau_i}^2 = \mu_{2_i} - \tau_i^2. \quad (2.5)$$

Voor de marginale foutenvariantie geldt:

$$\sigma_{E_i}^2 = \int \sigma_{E_i}^2(\theta) dG(\theta) = m_{2_i} - \mu_{2_i} \quad (2.6)$$

Met behulp van uitdrukking (2.4) 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:

$$\sigma_{X_t}^2 = \sum_{i \in t} a_i^2 \sigma_{X_i}^2. \quad (2.7)$$

Uitdrukking (2.5) maakt het mogelijk de variantie van de gewogen ware scores op toets t te schatten:

$$\sigma_{\tau_t}^2 = \sum_{i \in t} a_i^2 \sigma_{\tau_i}^2. \quad (2.8)$$

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 (2.2), (2.7) en (2.8) 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:

$$\hat{\rho}_{X_t T_t}^2 = \frac{\sigma_{\tau_t}^2}{\sigma_{X_t}^2} = \frac{\sum_{i \in t} a_i^2 \sigma_{\tau_i}^2}{\sum_{i \in t} a_i^2 \sigma_{X_i}^2} = \frac{\sum_{i \in t} a_i^2 (\mu_{2_i} - \tau_i^2)}{\sum_{i \in t} a_i^2 (m_{2_i} - \tau_i^2)}. \quad (2.9)$$

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).

2.3 Resultaten

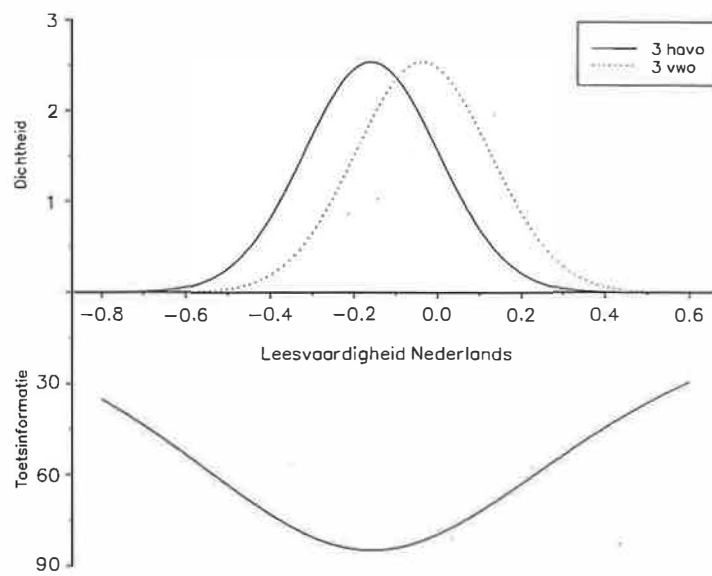
De informatiefuncties van de toetsen die voor de vakken Nederlands, Engels en wiskunde ontwikkelde zijn, staan afgebeeld in de figuren 2.1, 2.2 en 2.2. De figuren maken in een oogopslag duidelijk hoe het gesteld is met de meetnauwkeurigheid van de toetsen voor ieder punt op de vaardigheidsschaal waar 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 2.2.1 en 2.2.2 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.

2.3.1 *De toetsinformatiefuncties*

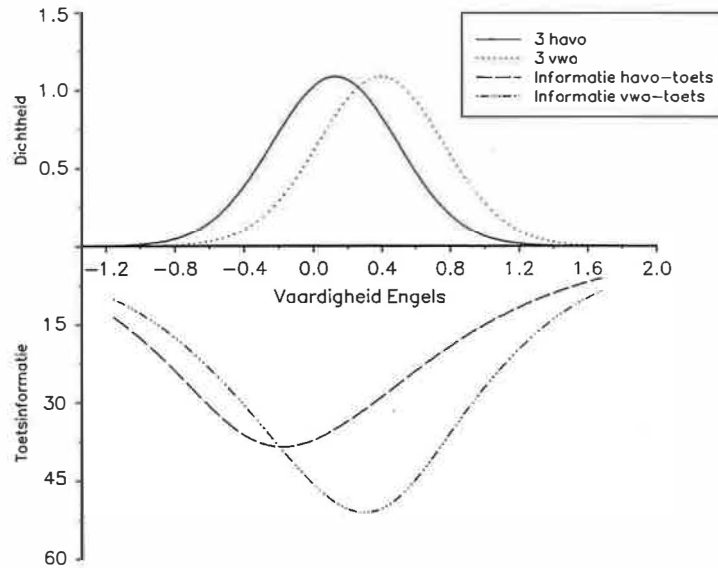
In de figuren 2.1, 2.2 en 2.3 worden naast de informatiefuncties van de toetsen ook de kansdichtheidsfuncties weergegeven die respectievelijk stonden afgebeeld in de figuren 1.1, 1.2 en 1.3. De kansdichtheidsfuncties staan steeds in het bovenste deel van de figuur en de toetsinformatiefuncties in het 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 2.1 bevat de kansdichtheidsfuncties van leesvaardigheid Nederlands voor de leerlingen uit drie havo en drie vwo. In het eerste hoofdstuk 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 2.1

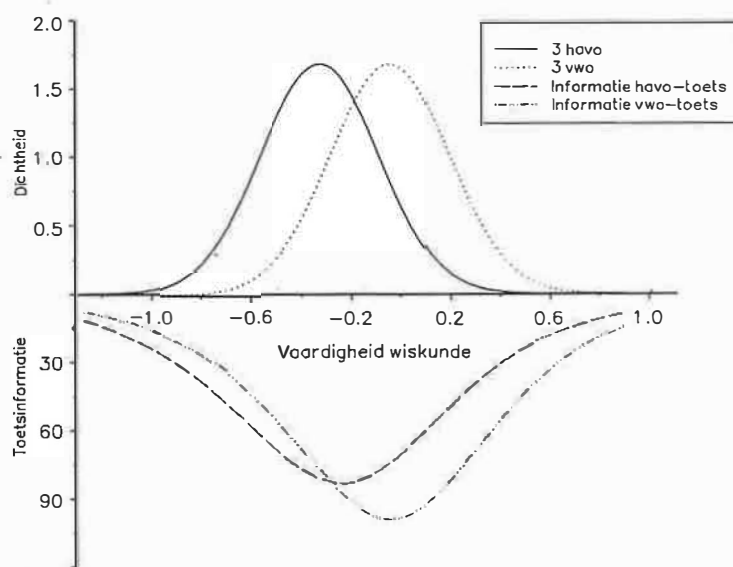
Kansdichtheidsfuncties van leesvaardigheid Nederlands voor drie havo en drie vwo en de toetsinformatiefunctie van de bijbehorende toets



Figuur 2.2

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

Het bovenste deel van figuur 2.2 bevat 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 2.3

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

In figuur 2.3 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 de 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.

2.3.2 Betrouwbaarheidsmatrices bij de toetsen

De tabellen 2.1 en 2.2 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 2.1 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 2.1 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 2.1

Betrouwbaarheidsmatrix voor de toets Nederlands voor leerlingen uit drie havo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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 2.2
Betrouwbaarheidsmatrix voor de toets Nederlands voor 3 vwo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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

Tabel 2.3
Betrouwbaarheidsmatrix voor de toets Engels voor leerlingen uit drie havo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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

In de tabellen 2.3 en 2.4 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.

Tabel 2.4

Betrouwbaarheidsmatrix voor de toets Engels voor leerlingen uit drie vwo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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

Een vergelijking van de tabellen 2.3 en 2.4 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.

De tabellen 2.5 en 2.6 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

Tabel 2.5

Betrouwbaarheidsmatrix voor de toets wiskunde voor leerlingen uit drie havo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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 2.6

Betrouwbaarheidsmatrix voor de toets wiskunde voor leerlingen uit drie vwo

Deciel	Deciel waarin de ware score valt									
	1	2	3	4	5	6	7	8	9	10
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

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 de kans op een correcte classificatie in alle decielen met uitzondering van deciel 10 groter te zijn dan bij de vwo-toets voor Engels.

2.3.3 De geschatte betrouwbaarheid van de toetsen

Tabel 2.7 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 ($\sum a_i J_i$) en bij wiskunde de ongewogen maximumscore ($\sum J_i$). De derde kolom ($\mathcal{E}(X)$) geeft de gemiddelde score per toets. De vierde kolom bevat de geschatte standaardafwijking ($\hat{\sigma}_X$) van de scores op de toets.

Tabel 2.7
Kengetallen voor de plaatsingstoetsen

<i>Toets</i>	<i>Max.</i>	$\mathcal{E}(X)$	$\hat{\sigma}_X$	$\hat{\sigma}_E$	$\hat{\rho}_{XT}^2$
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)

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 ($\hat{\rho}_{XT}$). Deze kolom bevat tevens tussen haakjes het 95-procents betrouwbaarheidsinterval rond de schatting van de betrouwbaarheid. De betrouwbaarheid van alle toetsen is geschat met behulp van uitdrukking (2.9). 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.

2.4 De drie soorten informatie over meetnauwkeurigheid vergeleken

De figuren in paragraaf 2.3.1. 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 bij ieder deciel. Indien de grenzen van de decielen in de figuren in paragraaf 2.3.1. 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 grotendeels 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 grotendeels 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 2.7. 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.

3 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.

3.1 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 behoorden 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 dienden zij predictief valide te zijn. En hadden tests betrekking op actuele prestaties of kenmerken van personen dan moesten zij gelijktijdige ('concurrent') validiteit 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 testcores 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 testcores, 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 testcores. 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 testcores 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 memorandum 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 gespecificeerd door aan te geven dat een toets een instrument is dat door onderwijs en studie verworven kennis, inzicht of vaardigheden meet, terwijl een 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 *toetsgebruikers* - docenten - in de regel niet goed bekend zullen zijn met het begrip validiteit. Voor *testgebruikers* geldt dat laatste in veel mindere mate. Het gebruik van tests vindt immers in de regel plaats binnen professionele kaders. Van *testgebruikers*, 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 één 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 zelf 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 memorandum 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 opvattingen van de meeste theoretici, zoals Messick (1988, 1989, 1990, 1995), Cronbach (1988), Shepard (1993, 1997) en Linn (1997). Toetsen worden in dit memorandum beschouwd als beslisinstrumenten. Verder wordt het aantonen van de kwaliteit van een toets als meetinstrument gezien als een noodzakelijke, maar niet voldoende, voorwaarde om de toets valide te doen zijn 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 of toets. 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 of toets. Vervolgens moet dan zo expliciet mogelijk aangegeven worden *waarom* de test of toets 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 of toetsen 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 pas aan bod in hoofdstuk vier, wanneer de vraag aan de orde komt hoe scholen de plaatsingstoetsen kunnen gebruiken bij het nemen van doorstroombeslissingen.

3.2 Het bepalen van de validiteit van de toetsen

In het eerste hoofdstuk 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 vier.

In het eerste hoofdstuk 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 het eerste hoofdstuk 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 het eerste hoofdstuk 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 het eerste hoofdstuk 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 memorandum 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.

3.3 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 aannemelijk 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 zijn 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 namelijk 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.

3.4 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 het eerste hoofdstuk. 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 het eerste hoofdstuk 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 3.1 geeft per vaardigheid en onderwijstype een overzicht van de effectgroottes voor de opeenvolgende categorieën van de variabele rapportcijfer. Tabel 3.2 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 bijlage E.

In tabel 3.1 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 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 3.1

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 3.1 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 3.2 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 3.2 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 3.2

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 3.1 en 3.2 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 3.3 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 bijlage F.

De gegevens in tabel 3.3 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 toetsen betrekking hebben op vaardigheden die relevant zijn voor het nemen van doorstroombeslissingen.

Tabel 3.3

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

Effectgroottes*					
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 studievoordigheid was dat de verschillende te meten vaardigheden niet al te sterk mochten samenhangen. De tabellen 3.4 en 3.5 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 studievoordigheid anderzijds en de correlaties tussen de drie scores op de checklisten. Tussen haakjes staat steeds het aantal waarnemingen waarop de correlaties gebaseerd zijn.

Tabel 3.4

Correlaties tussen de per vak geschatte vaardigheden, correlaties tussen geschatte vaardigheden en checklistscores en correlaties tussen de checklistscores 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

De tabellen 3.4 en 3.5 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.

Tabel 3.5

Correlaties tussen de per vak geschatte vaardigheden, correlaties tussen geschatte vaardigheden en checklistscores en correlaties tussen de checklistscores 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

Dit betekent dat te verwachten valt dat zowel de ontwikkelde toetsen 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 mogelijk 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.

3.5 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 studievoordigheid 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 $\mathbf{X}$ van v ($v = 1, 2, \dots, V$) 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:

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

In uitdrukking (4.1) is

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

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 3.6 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 grootte 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 overschrijdingskansen 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 3.6

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
Drie havo	V	399,803	6	0,0000
	T	156,523	3	0,0000
	C	319,288	3	0,0000
Drie vwo	V	569,425	6	0,0000
	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) $\bar{R}^2$ 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 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 3.7

Waarden van $\tilde{R}^2$ en de C-statistic voor het model met de zes onafhankelijke variabelen (V), het model met de toetscores (T) en het model met de checklistscores (C) voor leerlingen uit drie havo en drie vwo

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

In tabel 3.7 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 toetscores 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 toetscores 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 3.8 en 3.9. 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 3.8

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	Overschrijdings- kans	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
Constante	V	-10,4086	0,7371	199,4117	0,0000	
	T	-2,3691	0,3532	44,9938	0,0000	
	C	-8,0292	0,5877	186,6283	0,0000	

Tabel 3.8 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 een bescheiden, bijdrage. Bij dit laatste model wijkt het regressiegewicht van de toetsscore wél 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 3.3 laat zien. Bovendien hebben de toetsen Nederlands en Engels beide betrekking op leesvaardigheid en laat tabel 3.4 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 3.9 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 waarden 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 wél 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 3.9

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	Overschrijdings- kans	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	1,93695	0,1640	0,0000
wiskunde	C	0,0122	0,0070	2,98249	0,0842	0,0206
Constante	V	-9,6288	0,7371	281,1860	0,0000	
	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 voorspellend 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, $\mathcal{E}(Y_i | x_i)$. 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 $\mathcal{E}_{\min}(Y_i | x_i)$, de verwachte kans op studiesucces die minimaal vereist is voor de voorspelling dat een leerling succesvol zal zijn.

Tabel 3.10 geeft voor een aantal verschillende waarden van de minimaal vereiste verwachte kans op studiesucces aan hoe vaak de onderscheiden uitkomsten voorkomen bij leerlingen in drie havo en hoe groot het percentage correcte voorspellingen is. Tabel 3.11 doet hetzelfde voor de leerlingen in drie vwo. Verder laten de tabellen voor de vermelde waarden van $\mathcal{E}_{\min}(Y_i | x_i)$ 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 3.10 laat bijvoorbeeld zien dat het percentage correcte voorspellingen op basis van de toets- en de checklistscores 77,9 bedraagt, indien $\mathcal{E}_{\min}(Y_i | x_i) = 0,5$ 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 3.10

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

$\mathcal{E}_{\min}(Y_i x_i)$	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 3.10 maakt verder duidelijk in hoeverre de verschillende uitkomsten afhankelijk zijn van de waarde van $\mathcal{E}_{\min}(Y_i | x_i)$ 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 3.11

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

$\mathcal{E}_{\min}(Y_i \mathbf{x}_i)$	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 3.11 laat zien dat het percentage correcte voorspellingen op basis van de toets- en de checklistscores voor leerlingen uit drie vwo 81,8 bedraagt, indien $\mathcal{E}_{\min}(Y_i | x_i) = 0,5$. 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 3.11 ook zien in hoeverre de verschillende uitkomsten afhankelijk zijn van de waarde van $\mathcal{E}_{\min}(Y_i | x_i)$ 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 3.10 als tabel 3.11 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 3.10 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 3.11 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 verwachtingstabellen gebruiken bij het nemen van doorstroombeslissingen over nieuwe populaties leerlingen in drie havo en drie vwo.

Tabel 3.12

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			Deciel	Percentage	
		niet succesvol	succesvol			niet succesvol	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 3.12 bevat de verwachtingstabellen voor de leerlingen uit drie havo en drie vwo. In de tabel is in de kolommen met deciel aanduidingen 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:

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

waarbij

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

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

Zowel voor het intercept als de overige regressiegewichten geldt:

$$B_{\cdot k} = B_{\cdot} + u_k,$$

waarbij $B_{\cdot}$ constant is over scholen en u_k een normaal verdeelde variabele is met een gemiddelde gelijk aan nul en variantie σ_u^2 .

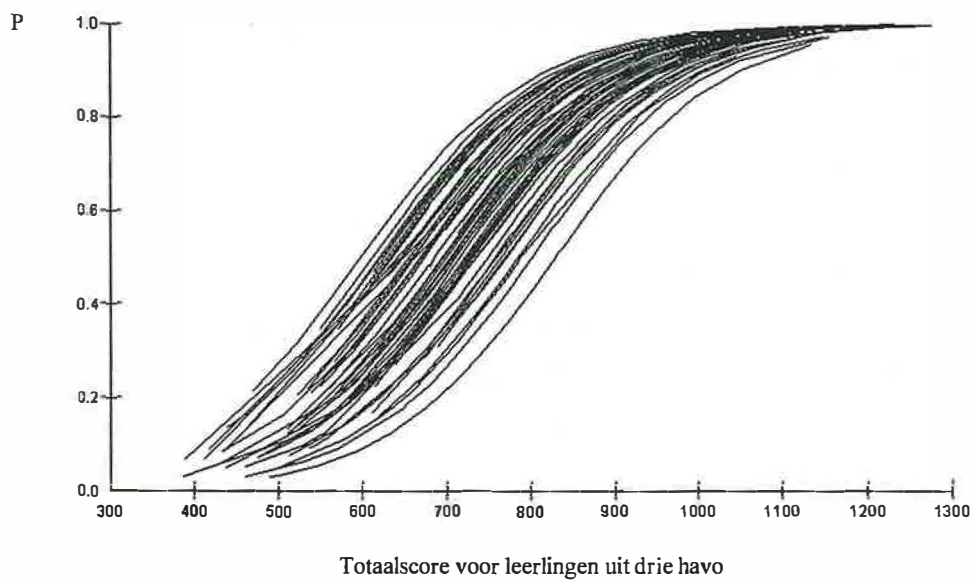
Allereerst zijn voor de leerlingen uit drie havo en drie vwo modellen opgesteld waarin alleen het intercept B_0 over scholen varieerde. De regressiegewichten van de zes instrumenten werden in dit model als constant over scholen gespecificeerd. Het betreffende model voor leerlingen uit drie havo toont aan dat er duidelijke verschillen zijn tussen scholen voor wat betreft het intercept. De variabele u_{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 v_{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.

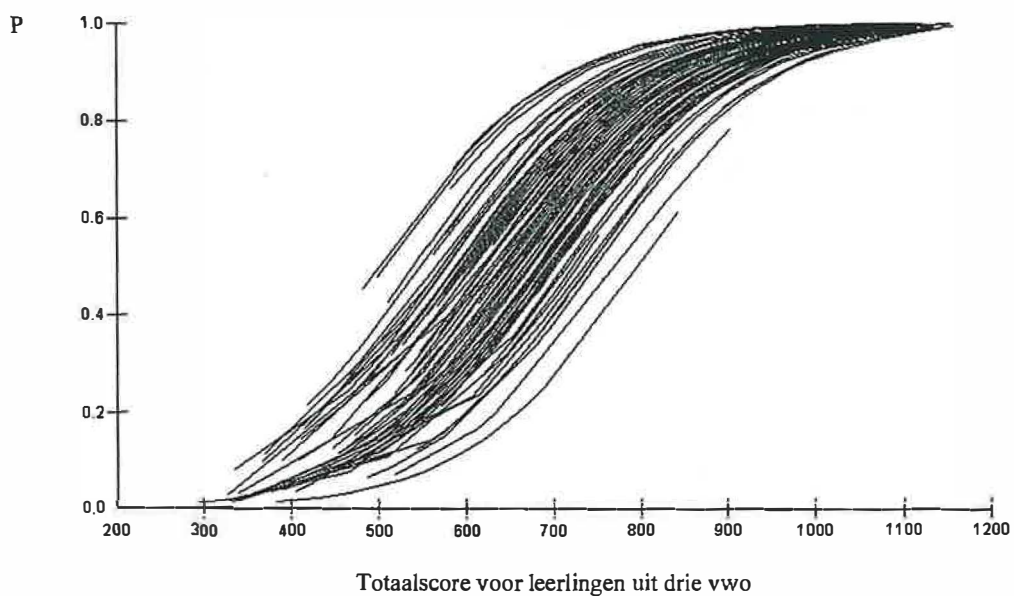
De discrepanties tussen scholen die voor beide klastypen in de multi-niveau analyses geconstateerd zijn voor wat betreft het intercept worden naar alle waarschijnlijkheid voor een deel veroorzaakt doordat scholen verschillen in strengheid bij het beoordelen van leerlingen in het vierde leerjaar. De geconstateerde 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 3.12 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.

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 ontstaan 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

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

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 blijktens 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 3.10 en 3.11 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 3.12 gepresenteerde informatie is voor de geconstateerde verschillen tussen scholen, zijn voor de leerlingen uit drie havo en drie vwo nieuwe totaalscores op het instrumentarium berekend. 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 zo 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.

De leerlingen zijn op grond van de aldus vastgestelde ‘multi-niveau’ totaalscores weer gesorteerd en over decielen verdeeld. Tabel 3.13 laat opnieuw voor beide instrumentaria per deciel het percentage succesvolle en niet succesvolle leerlingen zien. In de kolommen met deciel aanduidingen staat tussen haakjes weer het bij ieder deciel behorende interval van de multi-niveau totaalscores vermeld.

Een vergelijking tussen tabel 3.12 en 3.13 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.

Tabel 3.13

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		Deciel		Percentage	
		niet succesvol	Percentage succesvol			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

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 3.12 voor de verschillen tussen scholen. De tabellen 3.14 en 3.15 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 3.14 maken duidelijk dat de oorspronkelijke decielscores van een redelijk groot aantal leerlingen 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 3.14

Vergelijking van de oorspronkelijke decielindeling met de multi-niveau
indeling voor leerlingen uit drie havo

Multi-niveau decielindeeling	Oorspronkelijke decielindeling									
	1	2	3	4	5	6	7	8	9	10
1	76,6%	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 3.15 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 3.14 en 3.15 tonen aan dat de informatie die de verwachtingstabellen in tabel 3.12 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 3.15

Vergelijking van de oorspronkelijke decielindeling met de multi-niveau indeling voor leerlingen uit drie vwo

Multi-niveau decielindeeling	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 Samenvatting en conclusies

In dit memorandum wordt verslag gedaan van de ontwikkeling, normering en van plaatsingstoetsen voor het derde leerjaar van de havo en het vwo. Het eerste hoofdstuk 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 twee 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 ontwikkelde toetsen. Geconcludeerd mag worden dat de meetnauwkeurigheid van de toetsen in orde is.

Hoofdstuk drie 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.

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.
- Berkel, H.J.M. van. (1984). *De diagnose van toetsvragen*. Proefschrift. Amsterdam: Universiteit van Amsterdam.
- Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: theory and practice*. Cambridge: MIT press.
- 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
- 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 studievvaardigheid*. Arnhem: Cito Instituut voor Toetsontwikkeling.
- Cito Instituut voor Toetsontwikkeling (1996b). *Ondersteuning bij studie-adviezen (OSA) 3 vwo: plaatsingstoetsen/checklist studievvaardigheid*. 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.
- 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.
- Drenth, P.J.D., & Sijsma, 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.
- 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.
- 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.
- Groot, A.D. de, & Naerssen, R.F. van (1969). *Studietoetsen construeren, afnemen en analyseren*. Den Haag: Mouton.
- Hambleton, R.K., & Linden, W.J. van der (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 4, 373-378
- 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.
- 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.
- Kane, M.T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.
- Kapel, R.C., & Roessingh, M.J. (1996). Onderwijsmatrix 1996. *Kwartaaltijdschrift onderwijsstatistiek*, 3, 35-53.
- Linden, W.J. van der (1983). *Van standaardtest naar itembank*. Enschede: Universiteit Twente
- 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.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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 leerlingen; Een beslistkundige 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., Boertien, H., Klijn, W.J. de, & Roosmalen, W.M.M. van (1991). *De constructie van plaatsingstoetsen* (Onderzoeksrapporten beginfase voortgezet onderwijs nr. 6). Arnhem: Cito Instituut voor Toetsontwikkeling.
- Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid; De ontwikkeling van een domeingericht meetinstrument*. Proefschrift. Universiteit Twente.
- 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.
- Wainer, H. (1987). *The first four millennia of mental testing; from ancient China to the computer age* (Research report 87-34). Princeton: Educational Testing Service.
- 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 A1	Parameters voor de items Nederlands	105
Bijlage A2	Resultaten beoordelaarsovereenstemmingsonderzoek Nederlands	107
Bijlage B1	Parameters voor de items Engels	109
Bijlage B2	Resultaten beoordelaarsovereenstemmingsonderzoek Engels	113
Bijlage C1	Parameters voor de items wiskunde	115
Bijlage C2	Resultaten beoordelaarsovereenstemmingsonderzoek wiskunde	117
Bijlage D	Normtabellen bij de ontwikkelde plaatsingstoetsen	119
Bijlage E	Resultaten SAUL-analyses voor rapportcijfer en docentoordeel	121
Bijlage F	Resultaten SAUL-analyses voor het succescriterium	125

Bijlage A1; Parameters voor de items Nederlands

geometric mean of a's = 2.970

Nr.	label	a	beta

1	Cl_T1_01	2	-.668
2	Cl_T1_02	5	-.058
3	Cl_T1_03	2	.097
4	Cl_T1_04	3	.143
5	Cl_T1_05	2	-.042
6	Cl_T1_06	3	-.456
7	Cl_T1_07	4	.112
8	Cl_T1_08	4	-.400
9	Cl_T1_09	3	-.109
10	Cl_T1_10	2	-.429
11	Cl_T1_11	4	-.116
12	Cl_T1_12	3	.263
13	Cl_T2_01	7	.341
14	Cl_T2_02	3	.547
15	Cl_T2_03	3	.747
16	Cl_T2_04	3	-.325
17	Cl_T2_05	4	-.265
18	Cl_T2_06	6	.642
19	Cl_T2_07	3	-.229
20	Cl_T2_08	4	.088
21	Cl_T2_09	3	-1.487
22	Cl_T2_10	4	.168
23	Cl_T2_11	3	.937
24	Cl_T2_12	4	.011
25	Cl_T3_01	3	-.236
26	Cl_T3_02	5	.040
27	Cl_T3_03	2	.304
28	Cl_T3_04	4	-.202
29	Cl_T3_05	5	.292
30	Cl_T3_06	5	.380
31	Cl_T3_07	3	.134
32	Cl_T3_08	3	.809
33	Cl_T3_09	3	.139
34	Cl_T3_10	5	.050
35	Cl_T3_11	4	.423
36	Cl_T3_12	5	-.503
37	Cl_T3_13	4	.161
38	Cl_T3_14	4	.382
39	Cl_T3_15	4	.348
40	Cl_T4_01	3	.646
41	Cl_T4_02	4	-.106
42	Cl_T4_03	4	.079
43	Cl_T4_04	4	-.099
44	Cl_T4_05	4	-.563
45	Cl_T4_06	1	1.099
47	Cl_T4_08	3	-.048
48	Cl_T4_09	4	-.209
49	Cl_T4_10	2	.922
50	Cl_T4_11	5	.087
53	Op_T1_02	2	.003
54	Op_T1_3c	1	-.358
			.067
55	Op_T1_5c	3	-.623
			-.601
			-.321
56	Op_T1_08	3	.578
57	Op_T1_09	3	-.061
58	Op_T1_10	4	-.509

Nr.	label	a	beta

59	Op_T1_11	3	-.220
60	Op_T1_12	1	-.117
61	Op_T1_13	3	.498
63	Op_T2_01	2	.550
64	Op_T2_02	3	-.528
65	Op_T2_03	3	-.392
66	Op_T2_04	4	.556
67	Op_T2_05	2	.148
68	Op_T2_06	3	.184
69	Op_T2_07	4	-.205
70	Op_T2_08	3	.025
71	Op_T2_09	3	.296
72	Op_T3_01	2	.302
73	Op_T3_02	3	-.408
74	Op_T3_03	5	-.306
75	Op_T3_04	3	-.462
76	Op_T3_05	2	.353
77	Op_T3_06	4	.062
78	Op_T3_07	2	.207
79	Op_T3_8c	3	-.227
			-.013
80	Op_T3_10	3	.011
82	Op_T4_02	2	-.037
83	Op_T4_03	4	-.114
84	Op_T4_04	2	.165
85	Op_T4_05	2	.284
86	Op_T4_06	4	-.152
87	Op_T4_7c	3	-.203
			-.395
			.216
88	Op_T4_10	1	1.325
89	Op_T4_11	4	.556
90	Op_T4_12	4	.459
91	1162	3	.142
92	1235	2	-.092
93	1181	3	-.418
94	1394	3	-.348
95	1166	2	-.170
96	1230	2	-.036
97	1168	2	-.418
98	1167	3	.610
99	1378	3	-.129
100	1176	5	-.189
101	1252	4	-.139
102	1158	4	-.022
103	1365	3	-.243
104	1066	3	-.298
105	1091	4	-.205
106	1208	2	-.168
107	998	3	-.155
108	1411	2	-.365
109	1188	3	-.167
110	1223	2	-.119
111	984	1	-.695
112	1172	1	.103
113	1276	4	-.453
114	1186	3	-.350
115	1358	3	-.058
116	1150	2	.382
117	1362	3	-.107
118	1170	4	-.328

Bijlage A2;
Resultaten beoordelaarsovereenstemmingsonderzoek Nederlands

item	OV(1)*	(90% betr. int.)
1	0.98	(0.98 - 0.99)
2	perfecte overeenstemming	
3	0.61	(0.50 - 0.73)
4	0.52	(0.39 - 0.65)
5	0.17	(0.09 - 0.30)
6	0.62	(0.51 - 0.74)
7	0.23	(0.14 - 0.37)
8	0.41	(0.29 - 0.55)
9	perfecte overeenstemming	
10	0.91	(0.87 - 0.95)
11	0.97	(0.95 - 0.98)
12	0.87	(0.82 - 0.92)
13	perfecte overeenstemming	
14	0.87	(0.82 - 0.92)
15	0.95	(0.93 - 0.97)
16	0.77	(0.68 - 0.85)
17	0.91	(0.86 - 0.94)
18	0.42	(0.30 - 0.56)
19	0.76	(0.67 - 0.84)
20	0.86	(0.80 - 0.91)
21	0.86	(0.80 - 0.91)
22	0.47	(0.34 - 0.62)
23	0.43	(0.31 - 0.58)
24	0.97	(0.96 - 0.98)
25	0.96	(0.93 - 0.97)
26	0.94	(0.91 - 0.96)
27	0.95	(0.93 - 0.97)
28	0.82	(0.75 - 0.89)
29	0.93	(0.90 - 0.96)
30	0.96	(0.94 - 0.98)
31	0.51	(0.39 - 0.64)
32	0.51	(0.39 - 0.64)
33	0.97	(0.95 - 0.98)
34	0.92	(0.89 - 0.95)
35	0.95	(0.93 - 0.97)
36	0.47	(0.35 - 0.61)
37	0.98	(0.97 - 0.99)
38	0.47	(0.35 - 0.61)
39	0.95	(0.92 - 0.97)
40	0.40	(0.28 - 0.54)
41	0.87	(0.81 - 0.92)
42	0.71	(0.61 - 0.81)
43	0.95	(0.93 - 0.97)
44	0.96	(0.94 - 0.97)
45	0.32	(0.21 - 0.46)

* Beoordelaarsovereenstemmingscoëfficiënt geschat voor één willekeurige beoordelaar (zie formule (1.5) p. 20)

Bijlage B1; Parameters voor de items Engels

geometric mean of a's = 1.965

Nr.	label	a	beta

1	Item_1	2	-.278
3	Item_3	1	-.430
4	Item_4	3	-.833
5	Item_5	2	-.457
6	Item_6	2	-.205
7	Item_7	3	-.470
8	Item_8	3	-.458
10	Item_10	2	.487
11	Item_11	5	-.299
12	Item_12	3	-.268
13	Item_13	3	-.315
14	Item_14	2	-.481
15	Item_15	1	.617
16	Item_16	2	-.307
17	Item_17	3	-.365
18	Item_18	3	-.931
19	Item_19	2	.600
20	Item_20	1	.156
21	Item_21	2	.141
22	Item_22	1	-.368
24	Item_24	2	-.220
25	Item_25	2	.221
26	Item_26	3	-.165
27	Item_27	2	.240
28	Item_28	3	.259
29	Item_29	2	-.876
31	Item_31	2	-.299
32	Item_32	1	.895
33	Item_33	4	-.342
34	Item_34	3	-.251
35	Item_35	3	.060
36	Item_36	2	.390
37	Item_37	2	.353
39	Item_39	2	.730
41	Item_41	1	.420
42	Item_42	2	1.077
43	Item_43	2	-.064
44	Item_44	2	-.006
45	Item_45	2	.274
46	Item_46	1	-.038
47	Item_47	2	-.139
48	Item_48	1	-.038
49	Item_49	1	-.260
51	Item_51	3	-.439
52	Item_52	2	-.546
53	Item_53	3	-.439
54	Item_54	1	.024
55	Item_55	1	.081
56	Item_56	2	.028
57	Item_57	1	.562
58	Item_58	3	-.386
59	Item_59	2	-.107
60	Item_60	2	-.007
61	Item_61	1	.584
62	Item_62	1	-1.119
63	Item_63	2	.230
64	Item_64	2	-.971
65	Item_65	2	-.898

Nr.	label	a	beta
66	Item_66	1	-.079
67	Item_67	1	.904
68	Item_68	2	-.246
69	Item_69	2	-.058
70	Item_70	2	.293
71	Item_71	1	-.606
72	Item_72	1	.535
73	Item_73	2	-.228
74	Item_74	2	-.509
75	Item_75	3	-.425
76	Item_76	2	-.199
77	Item_77	1	-.598
78	Item_78	2	-.291
79	Item_79	3	-.296
80	Item_80	2	.300
81	Item_81	1	-1.184
82	Item_82	2	-.765
83	Item_83	2	-1.529
84	Item_84	2	-.884
87	Item_87	2	-.907
88	Item_88	1	1.234
89	Item_89	1	-1.162
90	Item_90	1	-.002
91	Item_91	2	-.867
92	Item_92	2	-1.492
95	Item_95	1	-.034
97	Item_97	2	-1.415
99	Item_99	3	-.718
100	Item_100	1	-.124
101	Item_101	2	.047
102	Item_102	1	.104
103	Item_103	2	1.112
104	Item_104	3	.289
105	Item_105	2	.465
106	Item_106	2	.079
107	Item_107	2	.339
108	Item_108	2	.869
109	Item_109	2	.328
110	Item_110	2	.878
111	Item_111	1	.595
112	Item_112	4	.438
113	Item_113	3	.423
114	Item_114	2	.618
115	Item_115	3	.640
116	Item_116	2	.001
117	Item_117	3	.311
118	Item_118	3	-.029
119	Item_119	2	.748
120	Item_120	2	.791
121	Item_121	3	-.505
122	Item_122	4	-.163
123	Item_123	3	.090
124	Item_124	3	.367
125	Item_125	3	.903
126	Item_126	2	-.717
127	Item_127	3	.187
128	Item_128	3	.642
129	Item_129	3	-.023
130	Item_130	4	.351
131	Item_131	2	.612
132	Item_132	2	.848
133	Item_133	2	1.212
134	Item_134	2	.271
135	Item_135	2	.148

Nr.	label	a	beta
136	Item_136	2	.779
137	Item_137	2	.053
138	Item_138	3	.268
139	Item_139	3	.727
140	Item_140	3	.768
141	Item_141	2	.417
142	Item_142	2	-.301
143	Item_143	3	.773
144	Item_144	2	-.405
145	Item_145	1	-.072
146	Item_146	2	.722

Bijlage B2;
Resultaten beoordelaarsovereenstemmingsonderzoek Engels

item	OV(1) *	(90% betr. int.)
1	0.85	(0.79 - 0.91)
2	0.94	(0.90 - 0.96)
3	0.71	(0.61 - 0.81)
4	0.92	(0.88 - 0.95)
5	0.85	(0.78 - 0.90)
6	0.95	(0.93 - 0.97)
7	0.97	(0.95 - 0.98)
8	0.93	(0.90 - 0.96)
9	0.62	(0.51 - 0.74)
10	0.86	(0.81 - 0.92)
11	0.89	(0.85 - 0.93)
12	0.97	(0.95 - 0.98)
13	0.86	(0.79 - 0.91)
14	0.78	(0.70 - 0.86)
15	0.87	(0.82 - 0.92)
16	0.89	(0.84 - 0.93)
17	0.89	(0.83 - 0.93)
18	0.93	(0.90 - 0.96)
19	0.71	(0.61 - 0.80)
20	0.71	(0.61 - 0.81)
21	0.78	(0.70 - 0.86)
22	0.96	(0.95 - 0.98)
23	perfecte overeenstemming	
24	0.95	(0.92 - 0.97)
25	0.40	(0.28 - 0.55)
26	0.81	(0.73 - 0.88)
27	0.81	(0.73 - 0.88)
28	0.73	(0.63 - 0.82)
29	0.92	(0.88 - 0.95)
30	0.87	(0.81 - 0.92)
31	0.90	(0.86 - 0.94)
32	0.91	(0.87 - 0.95)
33	0.63	(0.52 - 0.75)
34	0.61	(0.49 - 0.73)
35	0.78	(0.70 - 0.86)
36	0.43	(0.31 - 0.58)
37	0.62	(0.50 - 0.74)
38	0.72	(0.62 - 0.81)
39	0.46	(0.34 - 0.61)
40	0.66	(0.56 - 0.77)
41	0.50	(0.38 - 0.64)
42	0.82	(0.75 - 0.89)
43	0.30	(0.17 - 0.46)
44	0.88	(0.83 - 0.93)
45	0.90	(0.86 - 0.94)
46	0.72	(0.61 - 0.81)

* Beoordelaarsovereenstemmingscoëfficiënt geschat voor één willekeurige beoordelaar (zie formule (1.5) p.20)

Bijlage C1; Parameters voor de items wiskunde

geometric mean of a's = 3.054			
Nr.	label	a	beta

1	Item_1	2	-.337
2	Item_2	2	-.430
3	Item_3	2	-.166
4	Item_4	3	.151
5	Item_5	4	-.426
6	Item_6	3	.288
			.207
7	Item_7	4	-.365
8	Item_8	3	.305
			-.081
9	Item_9	4	.235
10	Item_10	3	.472
11	Item_11	3	.770
			-.029
13	Item_13	2	-.358
			.499
14	Item_14	2	.247
			-.617
15	Item_15	5	-.030
16	Item_16	4	-.078
17	Item_17	6	.295
18	Item_18	4	.275
21	Item_21	3	-.537
22	Item_22	5	-.354
23	Item_23	3	-.469
			-.605
24	Item_24	4	-.193
25	Item_25	2	-.818
26	Item_26	3	-.251
			.306
27	Item_27	2	-.452
28	Item_28	3	.091
29	Item_29	4	-.518
30	Item_30	3	-.078
31	Item_31	3	-.115
			-.019
32	Item_32	2	.257
			-.712
34	Item_34	3	.527
			-.063
35	Item_35	5	.199
36	Item_36	4	.461
37	Item_37	4	-.197
38	Item_38	3	-.381
39	Item_39	2	-.695
40	Item_40	3	-.118
			.059
41	Item_41	3	-.525
42	Item_42	2	-.310
			-.260
43	Item_43	4	-.190
46	Item_46	4	-.384
47	Item_47	3	.155
			-.078
48	Item_48	2	-.095
49	Item_49	3	.563
50	Item_50	2	1.015
51	Item_51	5	.118

Nr.	label	a	beta
52	Item_52	6	.271
53	Item_53	4	.524
54	Item_54	2	.439
55	Item_55	3	.374
56	Item_56	2	.602
			-.420
57	Item_57	4	-.043
58	Item_58	3	-1.204
59	Item_59	4	-.655
60	Item_60	5	-.155
61	Item_61	2	.029
			-.590
63	Item_63	1	1.422
64	Item_64	3	.008
			.000
65	Item_65	6	-.243
66	Item_66	4	.343
67	Item_67	7	-.176
68	Item_68	4	.291
69	Item_69	5	-.045
70	Item_70	4	.269
71	Item_71	4	.238
72	Item_72	4	.201
74	Item_74	3	.504
75	Item_75	4	.141
76	Item_76	2	.316
			.451
77	Item_77	2	.060
			-.056
79	Item_79	2	-.443
80	Item_80	2	.159
81	Item_81	1	-.296
82	Item_82	2	-.620
83	Item_83	3	.108
84	Item_84	3	.062
85	Item_85	4	.404
86	Item_86	4	.279
87	Item_87	3	.116
88	Item_88	2	.201
89	Item_89	3	.328
			-.290
90	Item_90	4	-.240
91	Item_91	3	.425
			.050
92	Item_92	2	.701

Bijlage C2;
Resultaten beoordelaarsovereenstemmingsonderzoek wiskunde

item	OV(1)*	(90% betr. int.)
1	0.79	(0.70 - 0.87)
2	0.74	(0.64 - 0.83)
3	0.88	(0.82 - 0.93)
4	0.81	(0.73 - 0.88)
5	0.87	(0.81 - 0.92)
6	0.93	(0.90 - 0.96)
7	0.91	(0.87 - 0.95)
8	0.59	(0.46 - 0.72)
9	0.83	(0.75 - 0.89)
10	0.35	(0.20 - 0.51)
11	0.91	(0.86 - 0.95)
12	0.61	(0.48 - 0.73)
13	0.73	(0.60 - 0.83)
14	0.83	(0.75 - 0.89)
15	0.96	(0.94 - 0.98)
16	0.77	(0.68 - 0.85)
17	0.92	(0.88 - 0.95)
18	0.92	(0.88 - 0.95)
19	0.92	(0.88 - 0.95)
20	0.91	(0.87 - 0.95)
21	0.83	(0.76 - 0.90)
22	0.93	(0.90 - 0.96)
23	-0.01	(-0.07 - 0.09)
24	0.89	(0.83 - 0.93)
25	0.36	(0.23 - 0.52)
26	0.61	(0.48 - 0.73)
27	0.62	(0.50 - 0.75)
28	0.91	(0.87 - 0.95)
29	0.92	(0.88 - 0.95)
30	0.92	(0.88 - 0.95)
31	0.78	(0.69 - 0.86)
32	0.85	(0.79 - 0.91)
33	0.35	(0.22 - 0.50)
34	0.75	(0.62 - 0.85)
35	0.31	(0.17 - 0.48)
36	0.92	(0.88 - 0.95)
37	0.90	(0.84 - 0.94)
38	0.74	(0.63 - 0.84)
39	0.82	(0.75 - 0.89)
40	0.82	(0.74 - 0.89)
41	0.68	(0.54 - 0.79)
42	0.84	(0.76 - 0.90)
43	0.83	(0.75 - 0.89)
44	0.45	(0.32 - 0.60)
45	0.45	(0.32 - 0.60)
46	0.83	(0.76 - 0.90)
47	0.83	(0.76 - 0.89)
48	0.80	(0.71 - 0.87)
49	0.71	(0.60 - 0.81)
50	0.42	(0.29 - 0.57)

item	OV(1)*	(90% betr. int.)
51	0.93	(0.90 - 0.96)
52	0.96	(0.93 - 0.97)
53	0.78	(0.69 - 0.86)
54	0.33	(0.20 - 0.49)
55	0.85	(0.78 - 0.90)
56	0.90	(0.85 - 0.94)
57	0.89	(0.84 - 0.93)
58	Perfekte overeenstemming	
59	Perfekte overeenstemming	
60	0.97	(0.95 - 0.98)
61	0.91	(0.86 - 0.94)
62	0.93	(0.90 - 0.96)
63	0.94	(0.90 - 0.96)
64	0.96	(0.93 - 0.97)
65	0.81	(0.73 - 0.88)
66	0.95	(0.93 - 0.97)
67	0.49	(0.36 - 0.63)
68	0.87	(0.81 - 0.92)
69	0.92	(0.88 - 0.95)
70	0.74	(0.64 - 0.83)
71	0.79	(0.70 - 0.87)
72	0.86	(0.79 - 0.91)
73	0.79	(0.70 - 0.87)
74	0.69	(0.58 - 0.80)
75	0.94	(0.90 - 0.96)
76	0.79	(0.68 - 0.87)
77	0.85	(0.78 - 0.90)
78	0.34	(0.22 - 0.50)
79	0.86	(0.80 - 0.92)
80	0.92	(0.87 - 0.95)
81	0.83	(0.73 - 0.90)
82	0.85	(0.75 - 0.92)
83	0.93	(0.89 - 0.96)
84	0.89	(0.83 - 0.93)
85	0.71	(0.60 - 0.81)
86	0.84	(0.75 - 0.91)
87	0.92	(0.87 - 0.95)
88	0.88	(0.82 - 0.93)
89	0.69	(0.57 - 0.80)
90	0.78	(0.69 - 0.86)
91	0.85	(0.78 - 0.90)
92	0.85	(0.78 - 0.91)
93	0.83	(0.76 - 0.89)

* Beoordelaarsovereenstemmingscoëfficiënt geschat voor één willekeurige beoordelaar (zie formule (1.5) p. 20)

Bijlage D; Normtabellen bij de plaatsingstoetsen

Tabel D.1
Normtabellen bij de toetsen voor Nederlands

Decielscore	Scores toets 3 havo	Scores toets 3 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 D.2
Normtabellen bij de toetsen voor Engels

Decielscore	Scores toets 3 havo	Scores toets 3 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 D.3
Normtabellen bij de toetsen voor wiskunde

Decielscore	Scores toets 3 havo	Scores toets 3 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 E; Resultaten SAUL-analyses voor rapportcijfer en docentoordeel

Tabel E1.1
Resultaten SAUL-analyse rapportcijfer Nederlands voor drie havo

Rapportcijfer	effect	SE	n	z	size
1-5	0		54		
6	.152	.085	274	1.788	.341
7	.282	.086	240	3.277	.632
8-10	.534	.112	49	4.781	1.198
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.130	.050		2.583	.291
8-10 - 6	.382	.087		4.381	.857
8-10 - 7	.252	.088		2.860	.566

Tabel E1.2
Resultaten SAUL-analyse rapportcijfer Nederlands voor drie vwo

Rapportcijfer	effect	SE	n	z	size
1-5	0		53		
6	.100	.083	293	1.208	.232
7	.255	.080	484	3.179	.593
8-10	.545	.084	258	6.516	1.267
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.155	.041		3.831	.361
8-10 - 6	.445	.047		9.494	1.035
8-10 - 7	.290	.042		6.847	.674

Tabel E1.3
Resultaten SAUL-analyse docentoordeel Nederlands voor drie havo

Docentoordeel	effect	SE	n	z	size
slecht	0		101		
middelmatig	.227	.064	294	3.574	.541
goed	.344	.079	92	4.330	.820
Pairwise contrasts					
	effect	SE		z	size
goed - middelmatig	.117	.066		1.779	.279

Tabel E1.4
Resultaten SAUL-analyse docentoordeel Nederlands voor drie vwo

Docentoordeel	effect	SE	n	z	size
slecht	0		104		
middelmatig	.251	.059	540	4.289	.580
goed	.531	.063	277	8.433	1.228
Pairwise contrasts					
	effect	SE		z	size
goed - middelmatig	.280	.041		6.871	.647

Tabel E2.1
Resultaten SAUL-analyse rapportcijfer Engels voor drie havo

Rapportcijfer	effect	SE	n	z	size
1-5	0		206		
6	.377	.062	371	6.112	.603
7	.713	.064	327	11.218	1.139
8-10	1.020	.078	145	13.100	1.630
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.336	.054		6.174	.536
8-10 - 6	.643	.071		9.110	1.027
8-10 - 7	.307	.072		4.258	.491

Tabel E2.2
Resultaten SAUL-analyse rapportcijfer Engels voor drie vwo

Rapportcijfer	effect	SE	n	z	size
1-5	0		107		
6	.478	.081	305	5.921	.758
7	.777	.077	481	10.103	1.232
8-10	1.251	.079	385	15.858	1.986
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.299	.053		5.643	.474
8-10 - 6	.773	.056		13.849	1.227
8-10 - 7	.475	.050		9.474	.753

Tabel E2.3
Resultaten SAUL-analyse docentoordeel Engels voor drie havo

	effect	SE	n	z	size
slecht	0		208		
middelmatig	.468	.058	569	8.140	.745
goed	.953	.068	234	13.955	1.517
Pairwise contrasts					
goed - middelmatig	.485	.056		8.620	.772

Tabel E2.4
Resultaten SAUL-analyse docentoordeel Engels voor drie vwo

	effect	SE	n	z	size
slecht	0		174		
middelmatig	.556	.063	552	8.797	.875
goed	1.051	.069	329	15.316	1.654
Pairwise contrasts					
goed - middelmatig	.495	.052		9.596	.779

Tabel E3.1
Resultaten SAUL-analyse rapportcijfer wiskunde voor drie havo

Rapportcijfer	effect	SE	n	z	size
1-5	0		341		
6	.258	.052	311	5.000	.563
7	.533	.056	223	9.580	1.163
8-10	.897	.068	112	13.219	1.956
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.275	.056		4.894	.600
8-10 - 6	.639	.068		9.357	1.393
8-10 - 7	.364	.071		5.100	.793

Tabel E3.2
Resultaten SAUL-analyse rapportcijfer wiskunde voor drie vwo

Rapportcijfer	effect	SE	n	z	size
1-5	0		278		
6	.242	.048	339	5.081	.573
7	.403	.046	372	8.673	.954
8-10	.644	.046	367	13.897	1.524
Pairwise contrasts					
	effect	SE		z	size
7 - 6	.161	.043		3.721	.381
8-10 - 6	.402	.043		9.319	.951
8-10 - 7	.241	.042		5.759	.570

Tabel E3.3
Resultaten SAUL-analyse docentoordeel wiskunde voor drie havo

Docentoordeel	effect	SE	n	z	size
slecht	0		180		
middelmatig	.336	.059	483	5.697	.711
goed	.685	.069	189	9.970	1.449
Pairwise contrasts					
	effect	SE		z	size
goed - middelmatig	.349	.055		6.371	.737

Tabel E3.4
Resultaten SAUL-analyse docentoordeel wiskunde voor drie vwo

Docentoordeel	effect	SE	n	z	size
slecht	0		228		
middelmatig	.245	.044	624	5.558	.610
goed	.656	.048	341	13.606	1.631
Pairwise contrasts					
	effect	SE		z	size
goed - middelmatig	.411	.037		11.053	1.021

Bijlage F;

Resultaten SAUL-analyses voor het succescriterium

Tabel F1.1

Resultaten SAUL-analyse Nederlands voor succescriterium drie havo

Succes	effect	SE	n	z	size
geen succes	0		118		
succes	.245	.061	388	3.998	.530

Tabel F1.2

Resultaten SAUL-analyse Nederlands voor succescriterium drie vwo

Succes	effect	SE	n	z	size
geen succes	0		95		
succes	.214	.062	793	3.482	.471

Tabel F2.1

Resultaten SAUL-analyse Engels voor succescriterium drie havo

Succes	effect	SE	n	z	size
geen succes	0		215		
succes	.131	.060	673	2.198	.194

Tabel F2.2

Resultaten SAUL-analyse Engels voor succescriterium drie vwo

Succes	effect	SE	n	z	size
geen succes	0		140		
succes	.225	.072	1040	3.107	.313

Tabel F3.1

Resultaten SAUL-analyse wiskunde voor succescriterium drie havo

Succes	effect	SE	n	z	size
geen succes	0		175		
succes	.393	.071	578	5.556	.570

Tabel F3.2

Resultaten SAUL-analyse wiskunde voor succescriterium drie vwo

Succes	effect	SE	n	z	size
geen succes	0		130		
succes	.381	.074	983	5.146	.549

Reeds eerder in 1998 verschenen OPD Memoranda

- 98-1 A.P.J.M. Heuvelmans. Constructie en verwerking van vragenlijsten.
- 98-2 A.P.J.M. Heuvelmans en J. Kuijt. Grafische afbeeldingen voor Computer Based Toetsen.
- 98-3 H.H.F.M. Verstralen. Het groeperen van frequentieverdelingen

10