Inhoudsopgave

1 Inleiding 5
2 Itemontwikkeling 6
2.1 Itemconstructie 6
2.2 Itemtypes 6
3 Uitvoering Pretest 8
3.1 Productie- en afnameomgeving, ... 8
3.1.1 Pretest 2015 8
3.1.2 Ontwikkeling 8
3.2 Design 8
3.2.1 Randvoorwaarden en schattingen 8
3.2.2 Globale pretestdesigns 9
3.3 Dataverwerking en analyses 10
4 Resultaten 12
4.1 Schoolrapportages pretest 12
4.1.1 Verschillen tussen tweetalig onderwijs en niet tweetalig onderwijs in de DTT pretest 2015 13
4.2 Responstijden 14
4.3 Resultaten itemanalyses 15
4.3.1 Schrijfvaardigheid Engels 16
4.3.2 Resultaten schrijfvaardigheid Nederlands 19
4.3.3 Resultaten Wiskunde 24
4.4 Standaardbepaling door experts na de eerste pretest van de DTT 29
4.4.1 Standaardbepaling met experts 29
4.4.2 Rapportage gezette standaarden 32
4.5 Indeling in blokken en simulatie 33
4.5.1 Adaptieve structuur voor eerste adaptieve afname 33
4.5.2 Methode 34
4.5.3 Resultaten 36
4.5.4 Conclusie 39
5 Referenties en eerdere verslagen 40
6 Bijlagen 42
6.1 Bijlagen design 42
6.1.1 Initiële schattingen per vak voor het toetsdesign 42
6.1.2 Globale designs per vak en leerweg 43
6.2 Bijlagen pretest schoolrapportages 46
6.3 Bijlage standaardbepaling 49
6.4 Bijlagen expert meeting 51
6.4.1 Programma Expert Meeting 51

Opening session 51
- Plenary 2: Brigitte Grugeon (Paris): Diagnostic digital assessment using « Pepite » software: test, cognitive profiles and learning courses 51
Dutch Design: 51
- Plenary 5: Chris Sangwin (Loughborough): Objective automatic assessment of mathematics: establishing mathematical properties, partial credit and feedback 51
Round table: Technical aspects of automated assignment of partial credit 51
Round table: Technical aspects of automated scoring of graphs & geometry items 51
Wrap-up with plenary speakers and organizers 51
Plenary closing session 51
6.4.2 Overzicht relevante software 54

1 Inleiding

Het ministerie van OCW heeft een herziene opdracht verstrekt voor de ontwikkeling en implementatie van de *diagnostische tussentijdse toets* (DTT). Volgens de opdracht wordt de DTT doorontwikkeld in een driejarige pilot met scholen. Er zijn diagnostische meetinstrumenten voorzien voor de vakken Nederlands, Engels en wiskunde aan het einde van de onderbouw in vmbo (leerjaar 2) en havo/vwo (leerjaar 3). Deze meetinstrumenten worden voor vijf verschillende beheersingsniveaus ontwikkeld: vmbo-bb, vmbo-kb, vmbo-gl/tl, havo en vwo.

De eerste daadwerkelijke adaptieve afname van de diagnostische toetsen op de scholen is gepland in het schooljaar 2015 – 2016 (gelijktijdig met de tweede pretest). De toetsontwikkeling vindt gefaseerd per vaardigheid/domein plaats en doorloopt vier opeenvolgende fases. Voorafgaand aan de invoering vindt een voorstudie, try-out(-s), grootschalige opgave-ontwikkeling en pretesten van de opgaven op scholen plaats.

In 2015 heeft de eerste grootschalige pretest plaatsgevonden voor wiskunde en schrijfvaardigheid Nederlands en Engels met ongeveer de helft van alle items. Het doel van de pretestafname was om empirische informatie te krijgen over de kwaliteit en (statistische) kenmerken van de items, dusdanig dat de items gekalibreerd kunnen worden. Na het afnemen van de pretest zijn de gegevens geanalyseerd en heeft de standaardbepaling plaatsgevonden. In dit verslag worden de uitkomsten van die eerste pretestafname gepresenteerd en besproken.

Naast de pretestafname hebben ook onderzoeksactiviteiten en met name ICT-ontwikkelingen plaatsgevonden voor verschillende ketens. Vooral ten bate van de eerste adaptieve afname in 2016 hebben veel ontwikkelingen plaatsgevonden. Ook is een mockup van de rapportage opgeleverd die is gebruikt om pilotscholen te demonstreren hoe de diagnoses gerapporteerd gaan worden en heeft gefungeerd als specificatie voor de rapportagemodule bij DUO.

In het psychometrisch onderzoek ging de aandacht uit naar het kunnen realiseren van de analyses besproken in dit verslag. De nadruk lag op het schrijven en testen van de code die noodzakelijk was voor de adhoc standaardbepaling en kalibratie (gerapporteerd in 4.1) en de definitieve kalibratie. Ook is voorafgaand aan de simulaties voor het samenstellen van de blokken (beschreven in 4.5) , de simulatieprocedure getest met fictieve data en indien nodig verbeterd. Verder is een bijdrage geleverd aan de ontwikkeling van de adaptieve module.

In dit verslag wordt in hoofdstuk 2 beschreven hoe de itemontwikkeling voor de pretest is uitgevoerd. Hoofdstuk 3 gaat in op het proces van afname en verwerking. In hoofdstuk 4 worden de resultaten van de pretest besproken en in hoofdstuk 5 voorstellen voor onderzoeksactiviteiten op het gebied van (automatische) beoordeling.

2 Itemontwikkeling

2.1 Itemconstructie

De grootschalige itemontwikkeling voor de pretest van februari 2015 is gestart in 2013/2014. De eerste voorvaststelling van de items vond plaats voor de zomer van 2014. In het najaar zijn alle items definitief vastgesteld. Tabel 2-1 geeft een overzicht van de aantallen item(responses) die zijn opgeleverd in relatie tot het vereiste volgens de constructieopdrachten. Op basis van deze items is een design voor afname gemaakt.

Tabel 2-1. Overzicht van het aantal items in de constructieopdracht pretest

	Engels	Nederlands	Wiskunde
vmbo	**Item-responses**	**Item-responses**	**Item-responses**
Opdracht totaal	432	576	576
Opdracht 1e pretest	216	288	288
Havo/Vwo			
Opdracht totaal	324	468	540
Opdracht 1e pretest	162	234	216

2.2 Itemtypes

Voor DTT is gebruik gemaakt van een verscheidenheid aan itemtypes binnen Questify en Facet. Hieronder een korte toelichting van de itemtypes en het soort vragen waarvoor ze gebruikt worden. Zie ook het Verslag try-out (Cito, 2014) voor voorbeelden en een beschrijving van de itemtypes.

Combinatie-item

Een itemtype waarbij meerdere soorten items (bijvoorbeeld een meerkeuzevraag en een sleepvraag) op één scherm worden gecombineerd.

Correctie-item

De leerling verbetert in de tekst. Geschikt om spelling te toetsen.

Dropdown-item

Een itemtype waarbij de leerling een antwoord kiest uit een keuzelijst. De leerling ziet daarna direct het effect van zijn keuze in de tekst.

Kort open-item

De leerling typt zelf een antwoord in. Dit itemtype wordt gebruikt om spelling te toetsen en om leerlingen zelf input te laten geven in de vorm van getallen en formules bij wiskundige opgaven, via een speciale formule-editor.

Markeer-item

De leerling klikt aan in de tekst. Geschikt om antwoorden te zoeken in de tekst, bijvoorbeeld een citaat of fout gespelde woorden.

Matrix-item

De leerling kiest uit (meestal) twee mogelijkheden, zoals ‘wel’ of ‘niet’. Geschikt voor het bevragen van details. De matrixvraag levert onafhankelijke itemresponses op.

Meerkeuze-item

De leerling kiest het juiste antwoord uit een lijst van alternatieven.

Multiple respons-item
De leerling kiest één of meer antwoorden uit een lijst van alternatieven.

Sleep-item
De leerling sleept een antwoord in de vorm van tekst of een afbeelding. Geschikt om te matchen of te sorteren.

Sleep-item afbeelding
De leerling sleept een antwoord. Geschikt om plaatjes aan woorden te koppelen.

Volgorde-item
De leerling sleept een antwoord. Geschikt om tekstdelen op volgorde te zetten.

3 Uitvoering Pretest

3.1 Productie- en afnameomgeving,

3.1.1 Pretest 2015

Bij de pretest is voor de productie van de items gebruik gemaakt van QuestifyBuilder (Cito), en voor de afname van Facet 3.1 (CvTE/DUO). Beide systemen zijn aangepast op basis van requirements die door Cito zijn opgesteld op basis van de opzet van de DTT.

Gerealiseerde aanpassingen betroffen met name een extra laag van metadatering om:

- alle onafhankelijke responses binnen een opgave als een apart item te kunnen verwerken.
- te kunnen 'scoren' op een diagnostisch profiel in plaats van op één vaardigheid ten behoeve van een zak/slaag beslissing.

De invoermogelijkheid in QuestifyBuilder van de diagnostische metadatering was nog niet volledig gerealiseerd voor de afname van de pretest in februari. Bij een vijftal itemtypes was de mogelijkheid om te scoren op (sub)attributen met behulp van zogenaamde conceptcoderingen nog niet beschikbaar.

De leerling-antwoorden op de formule-items voor wiskunde werden door Facet in een ander format geretourneerd dan het format waarin de sleutels in Questify Builder waren aangemaakt. Hierdoor moesten de antwoorden na afname eerst worden omgezet, waarna ze pas konden worden gescoord.

Beide bovengenoemde aspecten betekenden voor Cito een extra verwerkingsslag bij de dataverwerking om alle responses op de voor de DTT gewenste wijze te verwerken.

3.1.2 Ontwikkeling

Voor volgende afnames (pretesten en adaptieve afnames) is gedurende deze pretestperiode het volgende ontwikkeld:

- De adaptieve module voor de DTT is doorontwikkeld, getest (door Cito, Citrus en DUO) en waar nodig aangepast, met als doel dat adaptieve DTT-afnames in 2016 mogelijk zijn.
- Voor de aansturing van de CAS (computeralgebrasysteem) in Facet, zijn door Cito requirements opgesteld en templates gemaakt waarin verschillende innovatieve scoringswijzen zijn geïmplementeerd. Door middel van de CAS kunnen wiskunde-opgaven automatisch gescoord worden. Bijvoorbeeld een formule-response van een leerling kan niet alleen op gelijkheid maar ook op equivalentie (gelijkwaardigheid) beoordeeld worden. Dit maakt een meer interactieve toetsing van wiskunde mogelijk, samen met een efficiënt verwerkingsproces.
- Om bij de DTT-wiskunde ook op het domein van grafieken en tabellen voldoende opgaven te kunnen maken, heeft Cito de integratie in Facet van zogenaamde DWO-items gerealiseerd. Het Freudenthal Instituut heeft in opdracht van Cito een specifieke exportfunctie gerealiseerd van de DWO (digitale wiskunde omgeving). Cito zelf heeft de scoring via de CAS en de conceptscoring van DWO-items in nieuwe templates toegevoegd.

3.2 Design

3.2.1 Randvoorwaarden en schattingen

Om alle items te kunnen pretesten is met een toetsdesign gewerkt. We zijn begonnen met een globaal indicatief design, waarin beschreven was hoe de toets eruit zou kunnen zien gegeven de randvoorwaarden (beschikbare afnametijd en benodigd aantal observaties) en schattingen van het aantal items en responsen, de toetstijd en het aantal leerlingen.

Het streven is om binnen elk schooltype voor elk deelattribuut uiteindelijk 24 responsen te hebben voor de adaptieve afname. In verband met de koerswijziging dit jaar van OCW is besloten om in plaats van één grote pretest af te nemen met alle items, in twee opeenvolgende jaren twee kleinere pretesten met de helft

van de items af te nemen. Dit betekent voor de talen (schrijfvaardigheid Engels en schrijfvaardigheid Nederlands) dat bij de eerste pretest voor elk schooltype en *elk deelattribuut minstens 12 responsen* beschikbaar moesten zijn. Bij wiskunde betekent dit dat bij de eerste pretest voor elk schooltype minstens *24 responsen beschikbaar moesten zijn voor elk van de drie wiskunde aspecten* binnen twee domeinen.

Voor de schaling van de items en op termijn eventueel af- en opstroomindicaties te kunnen geven, gaat de constructieopdracht uit van *50% overlap* tussen de itembanken voor aangrenzende schooltypes (zie de Constructieopdracht itembanken DTT wiskunde en schrijfvaardigheid Nederlands en Engels van CvTE, 2013). De gerealiseerde overlap tijdens de afname kan groter of kleiner zijn afhankelijk van het aantal beschikbare items. Bijvoorbeeld, bij de bb-leerlingen worden dus de bb-items afgenomen en een deel van de kb-items.

Een belangrijke randvoorwaarde was de beschikbare afnametijd. Per leerling waren *drie klokuren* beschikbaar. Met name voor vmbo-leerlingen is dit erg lang en moeten we rekenen op een ruime pauze. Om te voorkomen dat we te weinig observaties krijgen bij sommige items moest het design zo zijn dat *90% van de leerlingen voldoende tijd* zou hebben voor de toets, dus ook de trage leerling (twee standaarddeviaties boven het gemiddelde).

- Vmbo: geschatte toetstijd + 2 standaarddeviaties = maximaal 2,5 uur
- Havo/vwo: geschatte toetstijd + 2 standaarddeviaties = maximaal 3 uur

Op grond van de waargenomen toetstijd in de pretest (zie bijlage 7.1) is een initiële schatting gemaakt van de benodigde toetstijd wanneer elk item twee responsen op zou leveren en alle items afgenomen zouden worden.

Een tweede cruciale randvoorwaarde was het aantal benodigde observaties. Het minimale aantal observaties dat nodig is per respons per schooltype om te kunnen kalibreren is 500. Om *500 observaties te realiseren* moeten 625 leerlingen gepland worden per item (uitgaande van 20% uitval). Op grond van een schatting van het aantal items en responsen, de gegeven toetstijd, en de geschatte toetstijd is een wervingsdoel geformuleerd.

Het wervingsdoel voor Engels en wiskunde was minstens 3750 leerlingen en voor Nederlands 5000 leerlingen[1]. Vanwege verschillen in de geschatte antwoordtijd gingen we bij wiskunde uit van 625 leerlingen per vmbo-type (1875 vmbo-leerlingen) en 940 leerlingen voor havo en voor vwo (1880 havo/vwo leerlingen).

Tabel 3-1. Overzicht randvoorwaarden

- Voor elk deelattribuut uit het leerlingmodel minstens 12 responsen (of 6 items) bij de talen en 24 responsen per wiskundige aspect binnen een domein;
- Beschikbare afnametijd is 3 uur;
- 90% v.d. leerlingen moet voldoende tijd hebben om de gehele toets te maken;
- Het minimaal benodigd aantal observaties per item is 500;
- Varianten met andere volgorde om te voorkomen dat dezelfde items achteraan in de toets zitten;
- Meerdere hoofdattributen per toets om de samenhang tussen attributen te kunnen analyseren;
- Overlap tussen de toetsen van aangrenzende schooltypes. Indien er veel items waren minder dan 50% overlap, indien er relatief weinig items waren voor een deelattribuut meer overlap.

3.2.2 Globale pretestdesigns

Op grond van de ontwikkelde items is een globaal design gemaakt. In het globale design staat aangegeven uit welke onderdelen een pretesttoets bestaat, hoeveel observaties verwacht worden en in welke volgorde de onderdelen worden afgenomen. In Figuur 3-1 staat een voorbeeld van een dergelijk globaal design. In bijlage 7.1.2 staan alle globale toetsdesigns voor de vakken.

[1] Het aantal benodigde leerlingen bij Nederlands is groter, omdat het leerlingmodel omvangrijker is.

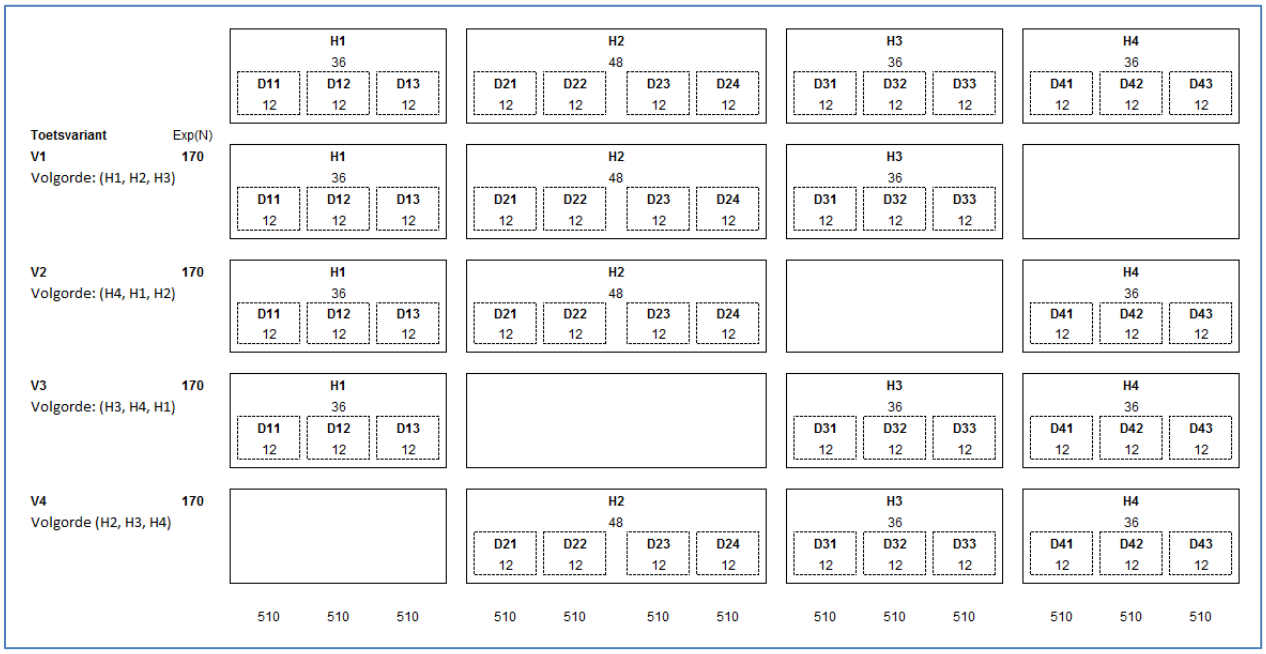
Naar schatting konden bij schrijfvaardigheid Engels zowel op vmbo als op havo/vwo alle items afgenomen worden binnen de beschikbare afnametijd. Bij Engels zijn vier varianten gemaakt waarin de items voor de vier hoofdattributen in verschillende volgorde werden gezet. Ook bij schrijfvaardigheid Nederlands konden op vmbo-niveau alle items afgenomen worden en waren er vier varianten met verschillende volgorde. Op havo/vwo niveau konden de items van drie van de vier hoofdattributen afgenomen worden. In het design is gezorgd dat elke combinatie van hoofdattributen even vaak voorkwam (in twee van de vier toetsvarianten) en dat de volgorde van de hoofdattributen varieerde.

Bij wiskunde was op het vmbo naar schatting voldoende tijd om de items van vijf van de zes wiskundige aspecten binnen een domein af te nemen en op havo/vwo vier van de zes. In het design is gezorgd dat een domein altijd in zijn geheel werd gemeten en daarnaast respectievelijk 2 (vmbo) of 1 (havo/vwo) wiskundige aspect(en) van het tweede domein. De volgorde van de domeinen werd in de varianten systematisch afgewisseld en alle combinaties van wiskundige aspecten binnen een domein en tussen domeinen kwamen even vaak voor.

Afhankelijk van de items die vastgesteld waren door de vaststellingscommissie en het aantal vastgestelde items konden in de definitieve designs de exacte aantallen per deelattribuut (12 responsen) hoger zijn en kon de overlap variëren (hoger of lager dan 50%).

Per leerweg en deelattribuut zijn de items in blokjes ingedeeld qua moeilijkheid (makkelijk tot medium of medium tot moeilijk) en overlap (wel of geen overlap-item). Omdat het aantal blokjes erg groot was zijn de blokjes vervolgens samengevoegd tot secties (per attribuut – zie de onderdelen in de designs). Uiteindelijk zijn met de secties de verschillende toetsvarianten gemaakt, die samen zijn opgenomen in een package. Tijdens de afname kregen de leerlingen automatisch een van de varianten toegewezen[2].

Figuur 3-1. Design schrijfvaardigheid Nederlands voor één schooltype havo/vwo



3.3 Dataverwerking en analyses

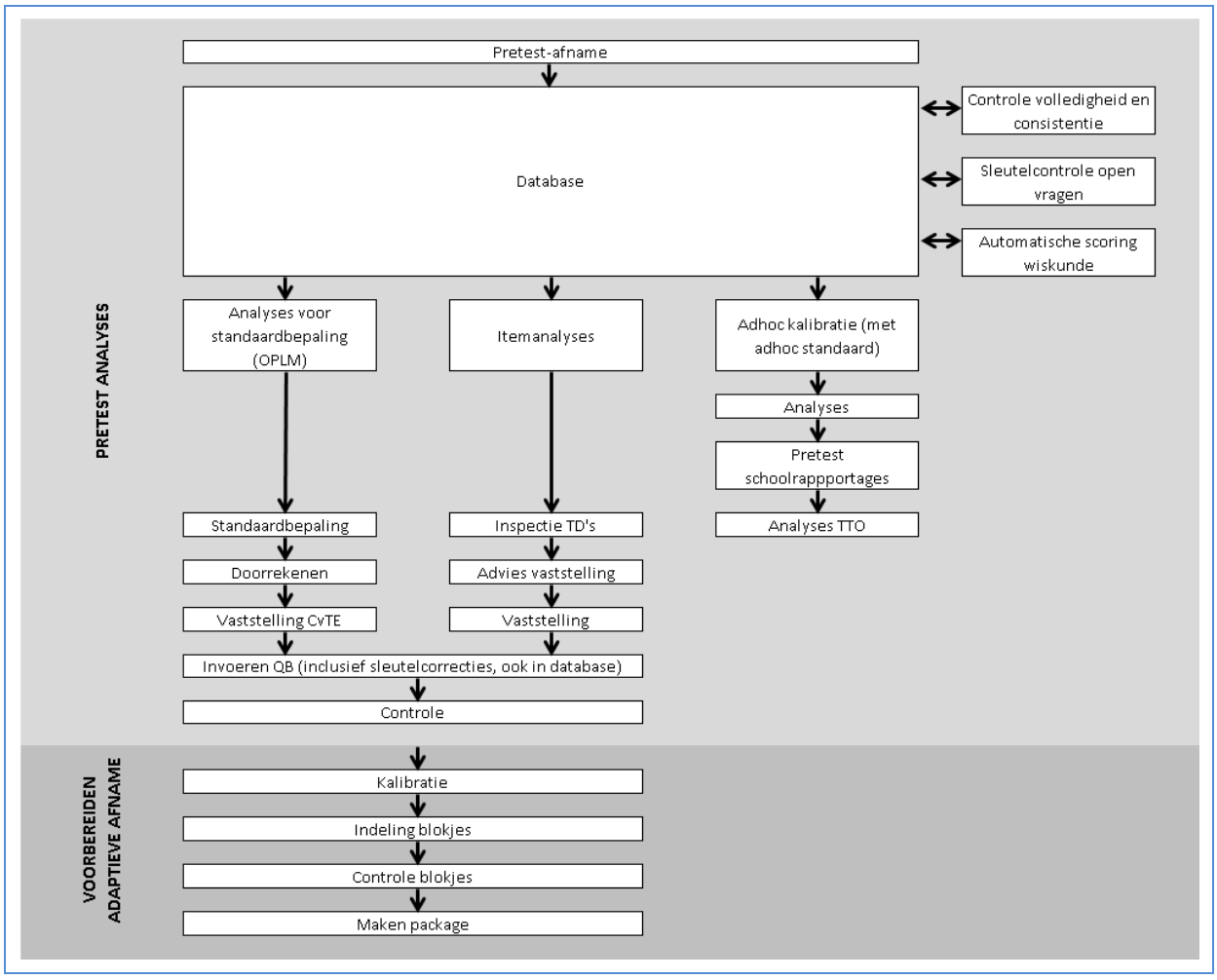
Na de pretestafname zijn de xml-bestanden ingelezen in een database en gecontroleerd op volledigheid en consistentie. De kwesties die hierbij aan het licht kwamen, onder andere ontbrekende gegevens en negatieve responstijden, zijn gerapporteerd aan DUO en indien nodig verbeterd. Het inlezen van de pretest-data in de database ging goed. Er is nog gewerkt met een separate database omdat voor het eerst met

[2] Bij de eerste kandidaat van een groep wordt aselect een variant gekozen. De rest van de varianten wordt dan doorgenummerd. De eerste kandidaat krijgt bijvoorbeeld aselect variant 3 toegewezen. De tweede kandidaat krijgt dan variant 4 en de derde kandidaat krijgt variant 5, etc.

coderingen op sub-item niveau werd gewerkt en nog niet alle open vragen automatisch beoordeeld konden worden (met name bij wiskunde) ten tijde van de pretest-afname. Daarnaast moest de data in verschillende formats aangeleverd kunnen worden voor analyses. De antwoorden die leerlingen hebben gegeven op de open vragen zijn na afloop gecontroleerd en indien nodig zijn de sleutels aangepast en de open vragen opnieuw automatisch beoordeeld.

De formule-items van wiskunde konden nog niet automatisch beoordeeld worden via Facet. Deze wiskunde opgaven zijn achteraf via een eigen ontworpen automatische beoordelingstool (gebaseerd op het sympy pakket) beoordeeld en de resulterende beoordelingen zijn aan de database toegevoegd.

Figuur 3-2. Proces van dataverwerking en analyses



Na de dataverwerking (zie Figuur 3-2) zijn globaal voor drie verschillende doeleinden analyses gedaan. Allereerst zijn direct na de dataverwerking de data geanalyseerd ten bate van de pretestrapportages aan scholen. Hierbij zijn aanvullende analyses uitgevoerd voor scholen die tweetalig onderwijs aanbieden.

Om de kwaliteit van de items te kunnen evalueren zijn daarnaast de individuele items geanalyseerd en zijn voor de standaardbepaling schaalanalyses uitgevoerd. Vlak voor de zomervakantieperiode heeft het CvTE de cesuren vastgesteld die gebruikt gaan worden bij de eerste DTT-afname in februari 2016. Op grond van die cesuren zullen de analyses uitgevoerd worden voor definitieve itemkalibratie en adaptieve toetssamenstelling. In dit verslag worden de uitkomsten van de diverse analyses beschreven.

4 Resultaten

4.1 Schoolrapportages pretest

In een pretest kan om verschillende redenen nog niet op leerlingniveau gerapporteerd worden. Allereerst kunnen individuele leerlingen gezien het aantal pretestitems en de beschikbare toetstijd slechts een subset van alle items maken[3], waardoor een individueel leerlingrapport niet mogelijk is. Bovendien moet de kwaliteit van de items na de pretest nog onderzocht worden en is er ten tijde van de pretest nog geen standaard gezet om de diagnoses te kunnen stellen. Om scholen toch feedback te kunnen geven is besloten om de data te analyseren met een adhoc standaard.

In de adhoc standaard zijn per attribuut de 20% best presterende leerlingen als boven-niveau beschouwd, de 20% minst presterende leerlingen als onder-niveau en de overige 60% als op-niveau. Nadat de adhoc standaard was bepaald zijn alle responsen gekalibreerd. Met behulp van deze adhoc kalibratie zijn vervolgens voor elk vak (wiskunde, schrijfvaardigheid Engels en schrijfvaardigheid Nederlands) de data geanalyseerd om per school en leerweg een rapportage te kunnen maken.

Voor schrijfvaardigheid Engels zijn 121 schooluitkomsten opgeleverd betreffende 5336 leerlingen, voor schrijfvaardigheid Nederlands zijn 135 schooluitkomsten opgeleverd betreffende 6420 leerlingen en voor wiskunde zijn 145 schooluitkomsten opgeleverd betreffende 7002 leerlingen. Leerlingen die meer dan 99% ontbrekende antwoorden hadden zijn niet meegenomen in de rapportages. De aantallen leerlingen liggen iets hoger dan vermeld in de notitie Evaluatie pilot DTT jaar 1 (CvTE, dd. 6 mei 2015) omdat later gegevens nog toegevoegd zijn (zie § 3.3 Dataverwerking en analyses) en alle gegevens zijn gebruikt, ook wanneer de toets niet (correct) was afgesloten

Tabel 4-1. Aantal berekende schooluitkomsten per leerweg en per vak.

	Schrijfvaardigheid Engels		Schrijfvaardigheid Nederlands		Wiskunde	
Leerweg	**Aantal scholen**	**Aantal leerlingen**	**Aantal scholen**	**Aantal leerlingen**	**Aantal scholen**	**Aantal leerlingen**
vmbo-bb	22	505	27	747	23	618
vmbo-kb	23	793	26	787	32	1117
vmbo-TL	29	1321	38	1791	33	1651
havo	23	1269	21	1321	25	1777
vwo	24	1448	23	1774	32	1839
Eindtotaal	121	5336	135	6420	145	7002

Opmerking. Op sommige scholen zijn voor meerdere leerwegen en/of meerdere vakken toetsuitkomsten opgeleverd.

De resultaten zijn vervolgens aan DUO opgeleverd (zie Tabel 4-2 voor een voorbeeld). DUO heeft vervolgens de rapportages vervaardigd (zie de voorbeelden in de bijlage 7.2) overeenkomstig het voorbeeld dat Cito heeft verstrekt. Het CvTE heeft een toelichting geschreven bij de rapportage. In de rapportage wordt per attribuut het aantal leerlingen dat een diagnose boven niveau heeft, het aantal leerlingen dat een diagnose op niveau heeft en het aantal leerlingen dat een diagnose onder niveau heeft gevisualiseerd (zie bijlage 7.2).

De adhoc kalibratie en de analyses met de nieuwe programma's (prototypes) werkten goed en de adhoc kalibratie liet geen bijzonderheden zien die negatieve consequenties zouden kunnen hebben voor de eerste adaptieve afname (zie ook de bespreking in hoofdstuk 4.3 Resultaten itemanalyses).

[3] Bij Nederlands en wiskunde hebben de leerlingen een subset van alle items gemaakt. Bij Engels hebben de leerlingen wel alle items gemaakt.

Tabel 4-2. Voorbeeld van de wijze waarop de gegevens aan DUO zijn aangeleverd.

				Schrijfvaardigheid Engels											
				1. Afstemming op publiek en doel			2. Coherentie			3. Woordenschat en woordgebruik			4 Grammatica, spelling en interpunctie		
Brin	Brinvn	Vak	Leerweg	Onder niveau	Op niveau	Boven niveau	Onder niveau	Op niveau	Boven niveau	Onder niveau	Op niveau	Boven niveau	Onder niveau	Op niveau	Boven niveau
XXXX	0	Engels schrijfvaardigheid	HAVO	5	19	35	13	36	10	13	35	11	11	38	10
XXXX	0	Engels schrijfvaardigheid	HAVO	9	30	13	12	31	9	8	31	13	4	33	15
XXXX	1	Engels schrijfvaardigheid	HAVO	3	5	0	2	4	2	2	5	1	4	2	2
XXXX	2	Engels schrijfvaardigheid	HAVO	5	15	6	4	18	4	5	16	5	7	18	1
XXXX	5	Engels schrijfvaardigheid	HAVO	5	6	1	5	6	1	7	3	2	7	4	1
XXXX	0	Engels schrijfvaardigheid	HAVO	22	68	36	26	66	34	24	55	47	23	66	37
XXXX	0	Engels schrijfvaardigheid	HAVO	10	40	4	16	28	10	12	28	14	12	32	10
XXXX	0	Engels schrijfvaardigheid	HAVO	20	51	17	25	50	13	19	45	24	25	48	15
XXXX	5	Engels schrijfvaardigheid	HAVO	6	21	2	3	22	4	5	18	6	8	18	3

4.1.1 Verschillen tussen tweetalig onderwijs en niet tweetalig onderwijs in de DTT pretest 2015

In de pilot hebben een paar scholen meegedaan die tweetalig onderwijs bieden en die geïnteresseerd waren in de verschillen tussen TTO- en niet-TTO leerlingen. Er zijn verschillende redenen waarom een verschil zou kunnen zijn tussen de diagnoses van TTO-leerlingen en niet-TTO leerlingen. Allereerst is bij tweetalig onderwijs vaak sprake van een (zelf-)selectie. Het is denkbaar dat de betere leerlingen of de meer gemotiveerde leerlingen zich aanmelden of worden toegelaten tot TTO. In dit geval zou je verwachten dat TTO leerlingen het beter doen. Het is ook mogelijk dat juist leerlingen met een aanleg voor talen zich aanmelden of worden toegelaten tot TTO en dat echte beta-leerlingen minder vaak voor TTO kiezen. In dit geval zou je verwachten dat TTO leerlingen het minder goed doen bij wiskunde en juist beter bij Engels.

Leerlingen in het TTO gebruiken daarnaast vaak andere boeken (met een andere methode) dan niet-TTO leerlingen. Het is mogelijk dat de gebruikte methodes minder (of juist meer) aansluiten bij de einddoelen en tussendoelen. In dit geval kan men verwachten dat TTO-leerlingen op enkele aspecten het minder goed doen wanneer deze nog niet aan bod zijn geweest en wellicht op andere aspecten het beter doen.

Tot slot zou een instructietaal (Engels) die leerlingen nog niet zeer goed beheersen het wiskunde leerproces kunnen hinderen. In dit geval zou men verwachten dat TTO-leerlingen het minder goed doen bij wiskunde dan niet TTO-leerlingen. De leerlingen zouden ook hinder kunnen ondervinden van het feit dat de DTT wiskunde afgenomen wordt in het Nederlands, terwijl de instructietaal Engels is. De DTT wiskunde zou voor deze leerlingen moeilijker zijn doordat zij onbekend zijn met de Nederlandse terminologie.

Van vier scholen hebben wij informatie gekregen over TTO-afnames en niet TTO-afnames. Voor deze scholen is een vergelijking gemaakt binnen de school tussen TTO en niet TTO-afnames en een vergelijking gemaakt met alle overige pilotscholen. Bij drie van de vier scholen is de DTT wiskunde afgenomen en bij 1 school de DTT schrijfvaardigheid Engels - vwo. Bij de school waar schrijfvaardigheid Engels - vwo is afgenomen, is een vergelijking gemaakt binnen de school tussen TTO en niet TTO-afnames en een vergelijking gemaakt met alle overige pilotscholen (niet TTO-afnames).

De drie TTO scholen waarvoor informatie over wiskunde beschikbaar was hadden TTO-leerlingen in VMBO-tl, havo en vwo. Aangezien van slechts drie scholen informatie beschikbaar is over drie verschillende leerwegen zijn de vergelijkingen die gemaakt konden worden zeer beperkt. Bij zowel vmbo-tl als bij havo was slechts één TTO school waar wiskunde was afgenomen. Voor deze twee scholen is een vergelijking gemaakt binnen de school tussen TTO en niet TTO-afnames en een vergelijking gemaakt met alle overige pilotscholen.

Het aantal TTO-scholen en leerlingen is zo gering dat geen algemene uitspraken over de verschillen tussen TTO en niet-TTO gedaan kunnen worden. De komende afname en pretest kunnen de verschillen wederom bekeken worden. Alleen verder en grootschaliger onderzoek kan meer duidelijkheid verschaffen over de aanwezigheid van verschillen en de oorzaak.

4.2 Responstijden

Gemiddeld hadden de leerlingen (ruim) voldoende tijd voor de pretestafname. Dit was cruciaal om voldoende observaties te verkrijgen voor elk item (minstens 500, zie 3.2 Design over de randvoorwaarden voor kalibratie).

Aan de responstijden was te zien dat op sommige scholen gebruik was gemaakt van de mogelijkheid om leerlingen een pauze te geven (zie Figuur 4-1). Deze pauzes zijn lastig uit de data te filteren, omdat de pauzes per leerling bij een ander item kan zijn genomen. Naast extreem hoge responstijden, kwamen ook extreem lage responstijden voor. Door de lokale computersystemen kon het voorkomen dat er negatieve responstijden werden geregistreerd. Ook waren er enkele leerlingen die niet serieus de items hadden beantwoord (met een gemiddelde responstijd van 2 seconde). Door deze kwesties is het lastig om tot een nauwkeurige inschatting te komen van de benodigde toetstijd.

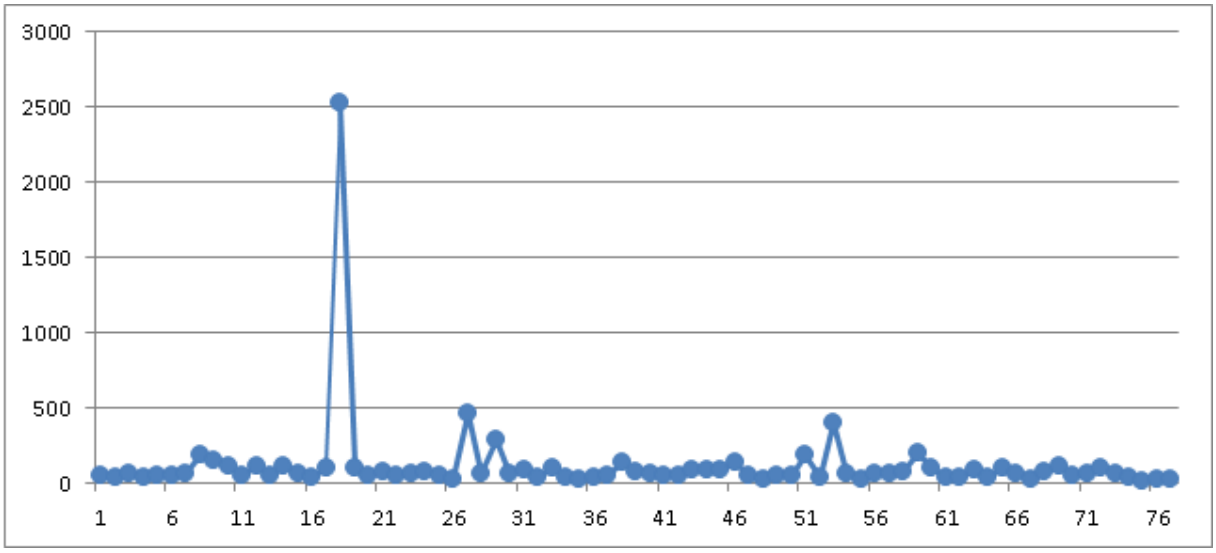
We zien als de extreme waardes zijn verwijderd dat 2 klokuren voldoende is voor 95% van de leerlingen voor het aantal items dat in de pretest was opgenomen. In de tweede pretest kunnen dus indien nodig meer items afgenomen worden of de reële toetstijd (zonder pauze) teruggebracht worden. Bij het bepalen van het design en het samenstellen van de toetsen zal hierover met het CvTE afgestemd worden.

Tabel 4-3. Totale antwoordtijd zonder uitbijters[4] van leerlingen die 75% of meer hebben beantwoord.

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	F	df	Sig,	Eta² Squared
Schrijfvaardigheid Engels									
N	499	787	1312	1263	1443	436,31	4,5299	0,000	0,25
Gemiddeld	30,07	34,87	32,78	31,21	25,71				
St.deviatie	6,00	6,18	6,19	5,67	4,58				
95 Percentiel	39,82	44,85	43,00	40,47	32,84				
Schrijfvaardigheid Nederlands									
N	650	645	1645	1307	1757	287,66	4,5999	0,000	0,16
Gemiddeld	88,17	97,84	102,62	80,61	82,93				
St.deviatie	27,42	26,09	22,69	16,71	16,99				
95 Percentiel	130,85	136,87	135,94	109,54	110,54				
Wiskunde									
N	601	1057	1575	1741	1801	221,06	4,677	0,000	0,12
Gemiddeld	55,89	67,75	63,50	70,12	78,46				
St.deviatie	15,53	15,37	16,13	20,76	21,89				
95 Percentiel	81,23	92,20	88,74	102,09	112,41				

[4] Waarnemingen die extreem afwijken (3*standaarddeviaties hoger of lager dan het gemiddelde)

Figuur 4-1. Voorbeeld van pauze-effect



4.3 Resultaten itemanalyses

Om de kwaliteit van de items te kunnen evalueren zijn net zoals voorgaande jaren beschrijvende itemanalyses uitgevoerd die informatie geven over de (statistische) kenmerken van de ontwikkelde items. Voor elk vak waren binnen elke leerweg verschillende toetsversies zodat alle items gepretest zouden worden. Voor alle drie de vakken zijn per leerweg de statistische eigenschappen van de (sub-)items bepaald in elke toetsversie met OPLM en R (R Development Core Team, 2008). De responsen van leerlingen met zeer veel ontbrekende antwoorden (25% of meer) zijn niet meegenomen in de itemanalyses om vertekening te voorkomen.

Voor schrijfvaardigheid Nederlands zijn 1626 itemresponsen uit 419 items (gespreid over de 4 attributen in 5 leerwegen) geëvalueerd. In die eerste itemanalyses was de gemiddelde moeilijkheid goed en ook de betrouwbaarheid was goed.

Voor schrijfvaardigheid Engels zijn 687 itemresponsen uit 383 items (gespreid over de 4 attributen in 5 leerwegen) geëvalueerd. In die eerste itemanalyses was de gemiddelde moeilijkheid goed, maar wel iets eenvoudiger dan bij Nederlands, en ook de betrouwbaarheid was goed.

Bij wiskunde zijn 645 itemresponsen uit 444 items (gespreid over de 3 wiskundige aspecten van 2 domeinen in 5 leerwegen) geëvalueerd. In die eerste itemanalyses bleek dat de items gemiddeld vrij moeilijk waren vergeleken met de items voor schrijfvaardigheid. Ook de betrouwbaarheid was iets minder, maar nog steeds voldoende.

Mede op grond van deze itemanalyses en de uitkomsten van de adhoc kalibratie hebben de toetsdeskundigen de kwaliteit van de items geëvalueerd en advies gegeven aan de vaststellings-commissies over de items (goedkeuren voor de adaptieve afname, aanpassen en opnemen in de volgende pretest of afkeuren). In de volgende secties worden de resultaten van de evaluatie per vak beschreven.

Tabel 4-4. Globaal overzicht itemanalyses pretest 2015

	Aantal items	Aantal onafhankelijke responsen	Gem. P-waarde	Gemiddelde Cronbach's alpha	Gemiddelde alpha40
Schrijfvaardigheid Nederlands	419	1626	0,56	0,95	0,72
Schrijfvaardigheid Engels	383	687	0,68	0,93	0,80
Wiskunde	444	645	0,37	0,85	0,64

4.3.1 Schrijfvaardigheid Engels

Operationalisatie van het leerlingmodel schrijfvaardigheid Engels

Het leerlingmodel, op basis waarvan de pretest is afgenomen, is weergegeven in Tabel 4-5. Dit is het leerlingmodel zoals geaccordeerd door de Toetswijzercommissie.

Er zijn totaal 687 itemresponses afgenomen, verdeeld over 383 items. Sommige items zijn in twee of drie leerwegen afgenomen. Het totaal aantal unieke itemresponses was 542, verdeeld over 305 items.

Tabel 4-5. Leerlingmodel schrijfvaardigheid Engels met aantallen items.

	Totaal afgenomen		Totaal uniek	
	Items	Responsen	Items	Responsen
1. Afstemming op publiek en doel	77	176	61	140
1.1 toonzetting en register afstemmen op publiek en schrijfdoel	35	52	26	39
1.2 conventies behorend bij een tekstsoort gebruiken	42	124	35	101
2. Coherentie	98	141	65	92
2.1 tekststructuur en verbanden aanbrengen	48	69	21	30
2.2 passende structuurwoorden gebruiken	50	72	44	62
3. Woordenschat en woordgebruik	104	205	90	165
3.1 passende woorden en woordcombinaties gebruiken	53	130	47	104
3.2 het woordgebruik functioneel variëren	51	75	43	61
4. Spelling, interpunctie en grammatica	104	165	91	145
4.1 woordvolgorde en zinsconstructie functioneel hanteren	53	94	46	80
4.2 passende spelling en interpunctie hanteren	51	71	45	65
Totaal	383	687	307	542

Tabel 4-6. Gemiddelde p-waarden per hoofd-/deelattribuut en leerweg bij schrijfvaardigheid Engels

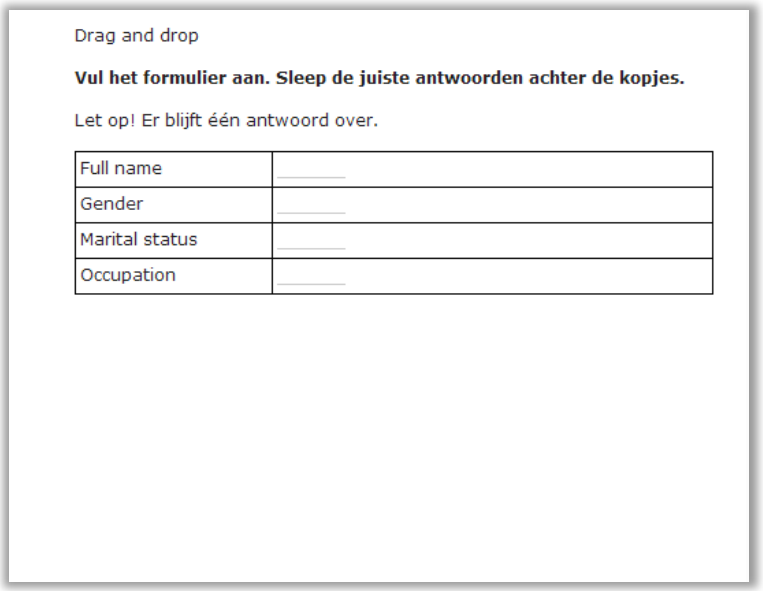
	bb		kb		gt		havo		vwo		Totaal	
(Deel-) attribuut	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit
1	0,68	0,31	0,66	0,34	0,65	0,26	0,75	0,24	0,88	0,25	0,71	0,29
11	0,50	0,22	0,57	0,24	0,67	0,29	0,82	0,27	0,86	0,21	0,66	0,25
12	0,75	0,34	0,69	0,38	0,64	0,24	0,73	0,23	0,88	0,26	0,72	0,30
2	0,50	0,36	0,54	0,39	0,61	0,37	0,72	0,31	0,79	0,31	0,62	0,35
21	0,39	0,34	0,46	0,35	0,56	0,36	0,59	0,30	0,70	0,25	0,51	0,34
22	0,66	0,39	0,63	0,43	0,66	0,38	0,85	0,32	0,81	0,32	0,73	0,37
3	0,69	0,38	0,75	0,48	0,76	0,32	0,78	0,34	0,76	0,25	0,75	0,35
31	0,71	0,35	0,82	0,51	0,82	0,21	0,82	0,37	0,82	0,23	0,80	0,33
32	0,66	0,43	0,60	0,43	0,66	0,52	0,69	0,26	0,69	0,27	0,67	0,36
4	0,49	0,33	0,56	0,31	0,56	0,31	0,66	0,29	0,77	0,26	0,62	0,30
41	0,41	0,34	0,59	0,33	0,67	0,31	0,73	0,31	0,83	0,27	0,67	0,31
42	0,56	0,32	0,53	0,30	0,42	0,32	0,52	0,27	0,69	0,24	0,54	0,29
Totaal	0,61	0,34	0,63	0,38	0,65	0,32	0,73	0,29	0,78	0,26	0,68	0,32

Globale bespreking van de items per leerweg - vmbo

vmbo bb-items

De bb-items zijn over het algemeen uitstekend op niveau. De p-waarden liggen gemiddeld rond de 0,61 en de gemiddelde Rit is 0,34. De hoogste p-waarden komen voor onder deelattribuut 1.2 (conventies behorend bij een tekstsoort). Deze p-waarden maken echter in alle gevallen deel uit van een testlet en betreffen slechts één of twee responsen van een item.

Afbeelding 4-1. Voorbeelditem



De kopjes die naar een naam, datum of telefoonnummer vragen gaan gepaard met de hoge p-waarden. Toch geven deze simpele itemresponsen in sommige gevallen ook waardevolle informatie, zoals bij *Full name* in Afbeelding 4-1. Hoewel het respons een hoge p-waarde heeft, hangt het item goed samen met de andere items in de toets (Rit 0,35) en zouden dit soort responsen iets kunnen bijdragen aan het onderscheid tussen de diagnoses *op niveau* en *onder niveau*.

Tabel 4-7. P-waardes en Rit van de voorbeeldopgave uit Afbeelding 4-1.

QTI-item label	Conceptrespons	Flag	HA	DA	Popula-tion	N	P	Average rit(u)
dtt-en-vm-0424-SC	CONCEPTRESPONSE_ENG-1-2	!	1	12	BB	498	0.97	0.18
	CONCEPTRESPONSE2_ENG-1-2		1	12	BB	498	0.84	0.28
	CONCEPTRESPONSE3_ENG-1-2	!	1	12	BB	498	0.93	0.35
	CONCEPTRESPONSE4_ENG-1-2		1	12	BB	498	0.33	0.27

De laagste p-waarden kwamen voor onder deelattribuut 2.1 (tekststructuur en verbanden aanbrengen). De items die een lage p-waarden hebben, betreffen vooral volgorde vragen; vragen waarin de leerling een instructie of een verhaal in de juiste volgorde moet zetten. Belangrijk om hierbij te realiseren, is dat deze items als één respons worden gecodeerd: alleen als een leerling alle alternatieven op de juiste volgorde heeft gezet, is de diagnose voor dat item positief. Dit maakt de items betrekkelijk moeilijk en het kost de leerling veel tijd om ze te maken.

vmbo kb-items

De p-waarden van de kb-items liggen gemiddeld rond de 0,63 en de gemiddelde Rit is 0,37. Net als bij vmbo-bb, zien we bij kb dezelfde trend van p-waarden onder deelattribuut 1.2 (conventies behorend bij een tekstsoort).

Deelattribuut 3.1 heeft een gemiddelde p-waarde van 0,81. De items die deze hoge waarde veroorzaken, zijn voornamelijk sleepitems, waarbij (een mate van) afhankelijkheid een rol speelt. Ook betreffen deze sleepitems basisonderwerpen zoals groenten of vertrekken in een huis.

Bij deelattribuut 4.2 (passende spelling en interpunctie hanteren) gaan vooral de korte open vragen gepaard met een lage p-waarde, maar onderscheiden deze open schrijf items wel goed. Het beste onderscheid ligt hierbij vaak tussen de *op niveau* en *boven niveau* laag.

vmbo gt-items

De p-waarden van de gt-items liggen gemiddeld rond de 0,65 en de gemiddelde Rit is 0,32. Bij de gt-items viel deelattribuut 3.1 (passende woorden en woordcombinaties gebruiken) op: deze items hadden een gemiddelde p-waarde van 0,82 en een Rit van 0,21. De items die voor een te hoge p-waarde en het minste onderscheid zorgden, waren vooral de items waarin een woord naar het juiste plaatje gesleept moest worden. Hierbij is het belangrijk om op te merken dat er sprake was van afhankelijkheid: er waren evenveel responsmogelijkheden als plekken voor het antwoord.

Globale bespreking van de items per leerweg - havo/vwo

Havo items

De havo-items hadden een gemiddelde p-waarde van 0,73 en een gemiddeld Rit van 0,29. Over het algemeen onderscheiden de items goed, de testlets items behorend bij 4.2 onderscheidden het minst goed.

Vwo items

De vwo-items hadden een gemiddelde p-waarde van 0,78 en een gemiddeld Rit van 0,26. Opvallend was dat er bij dit niveau een aantal items op deelattribuut 4.1 (woordvolgorde en zinsconstructie functioneel hanteren) een erg hoge p-waarde hadden. Deze items maakten overigens wel goed onderscheid. Het betrof vooral dropdown items waarin drie-keuze mogelijkheden zaten. Wellicht is een ander itemtype beter geschikt voor het bevragen van deze onderwerpen op het gebied van woordvolgorde en zinsconstructie.

Globale bespreking van de 'afkeur' en de redenen daarvoor, signalering van trends en mogelijke oplossingen.

Er zijn totaal twee items afgekeurd. Deze items waren te makkelijk voor het bb-niveau. Daarnaast worden er 21 items opnieuw getest op een lager of hoger niveau met eventuele aanpassingen. Deze items liggen verdeeld onder verschillende deelattributen, waarvan de meeste onder 3.1 (passende woorden en woordcombinaties gebruiken) vallen. Dit deelattribuut was gemiddeld genomen op een te makkelijk niveau geconstrueerd.

In de pretest zat één correctie-item, waarin leerlingen twee woorden in een zin moesten herkennen als fout, en deze vervolgens verbeteren. Ongeveer 4% van de leerlingen laat het fout geschreven woord staan en typt het juiste geschreven woord erachter. Het is goed om hier kritisch naar de instructie te kijken. Er zaten ook een aantal kort open vragen in de pretest. De kort open vragen die naar aanhef vroegen, leverden teveel verschillende responsen op. Dit soort items is niet eenduidig genoeg, de vraag moet op een andere manier gesteld gaan worden. De items die afhankelijkheid hebben, door onder andere de technische eigenschappen van een sleepitem, lijken erg gemakkelijk te zijn voor leerlingen. Om meer variatie in moeilijkheid en meer onderscheid te krijgen in deze items, kan het een verbetering zijn om meer alternatieven toe te voegen of het item als één conceptrespons te coderen.

Over het algemeen zouden we een gelijke gemiddelde p-waarde over de niveaus verwachten. Uit de analyses blijkt dat de p-waarde per niveau iets omhoog gaat. Er zou dus meer spreiding moeten zitten tussen de moeilijkheid van de verschillende niveaus: met andere woorden, de havo en vwo moeten ten opzichte van elkaar en ten opzichte van de niveaus op vmbo moeilijker geconstrueerd worden.

4.3.2 Resultaten schrijfvaardigheid Nederlands

Operationalisatie van het leerlingmodel schrijfvaardigheid Nederlands

Het leerlingmodel op basis waarvan de pretest is uitgevoerd, is weergegeven in Tabel 4-8**Fout! Verwijzingsbron niet gevonden.** Dit is het leerlingmodel zoals geaccordeerd door de Toetswijzercommissie.

Tabel 4-8. Leerlingmodel schrijfvaardigheid Nederlands met aantallen items.

	Totaal afgenomen		Totaal uniek	
	Items	Responsen	Items	Responsen
I retorische vaardigheden (doel en publiek)	98	409	52	212
1.1 voorkennis en informatiebehoefte bij de lezers inschatten	30	185	15	100
1.2 toonzetting afstemmen op de lezer (o.a. formeel-informeel)	34	74	19	41
1.3 schrijfdoel bepalen (bv. informeren, overtuigen, uitleggen, uitnodigen)	34	150	18	71
II tekststructurele vaardigheden (structuur)	134	295	92	190
2.1 tekstelementen kiezen, rekening houdend met het genre[5]	43	87	25	55
2.2 passende volgorde van tekstelementen bepalen en correcte indeling en lay-out aanbrengen[6]	41	46	35	39
2.3 samenhang tussen tekstelementen aanbrengen (coherentie)[7]	32	127	18	70
2.4 een standpunt weergeven en van passende argumenten voorzien (alleen h/v)	18	35	14	26
III linguïstische vaardigheden (woord- en zinsniveau)	102	366	58	215
3.1 een correcte zinsbouw hanteren	37	131	18	64
3.2 een passende en cohesieve[8] schrijfstijl hanteren	32	113	20	75
3.3 passend en gevarieerd woordgebruik laten zien[9]	33	122	20	76
IV orthografische vaardigheden (spelling en interpunctie)	92	556	42	268
4.1 correct spellen van werkwoorden	31	158	13	77
4.2 correct toepassen van overige regelgeleide spelling	30	155	13	69
4.3 leestekens en hoofdletters correct hanteren	31	243	16	122
Totaal	426	1626	244	885

[5] Te denken valt aan: briefconventies en titels.
[6] Dit betreft de volgorde van de inhoudelijke elementen en de bijbehorende lay-outtechnische aspecten en (voor havo/vwo) tussenkopjes.
[7] Coherentie betreft de inhoudelijke samenhang in tekst (bijv. via inhoudswoorden en woordvelden).
[8] Cohesie betreft de syntactische samenhang in tekst; deze kan worden versterkt door gebruikmaking van verwijswoorden, lexicale samenhang door middel van synoniemen en hyponiemen en door voegwoorden.
[9] Hiertoe behoort ook figuurlijk taalgebruik (voor havo-vwo)

Tabel 4-9. Gemiddelde p-waarden per hoofd-/deelattribuut en leerweg bij schrijfvaardigheid Nederlands

	bb		kb		gt		havo		vwo		Totaal	
(Deel-) attribuut	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit	Gem. P	Gem. Rit
1	0,57	0,21	0,63	0,20	0,70	0,19	0,80	0,14	0,81	0,15	0,70	0,18
11	0,58	0,18	0,71	0,18	0,77	0,18	0,80	0,12	0,83	0,13	0,74	0,16
12	0,55	0,26	0,47	0,25	0,55	0,22	0,80	0,17	0,73	0,19	0,60	0,22
13	0,58	0,22	0,64	0,20	0,71	0,20	0,79	0,18	0,83	0,19	0,69	0,20
2	0,47	0,32	0,54	0,27	0,58	0,24	0,64	0,21	0,68	0,22	0,59	0,25
21	0,53	0,27	0,64	0,25	0,61	0,20	0,66	0,19	0,69	0,17	0,62	0,22
22	0,23	0,29	0,34	0,26	0,23	0,17	0,49	0,23	0,59	0,26	0,40	0,24
23	0,49	0,37	0,53	0,30	0,65	0,27	0,54	0,23	0,61	0,25	0,57	0,29
24							0,81	0,19	0,84	0,21	0,82	0,20
3	0,53	0,32	0,57	0,29	0,64	0,25	0,48	0,30	0,49	0,31	0,54	0,29
31	0,40	0,27	0,46	0,24	0,53	0,25	0,28	0,67	0,28	0,65	0,40	0,39
32	0,57	0,33	0,56	0,31	0,60	0,23	0,33	0,12	0,51	0,18	0,52	0,23
33	0,64	0,38	0,73	0,33	0,81	0,27	0,71	0,11	0,66	0,15	0,72	0,23
4	0,37	0,36	0,46	0,37	0,48	0,32	0,39	0,31	0,53	0,31	0,45	0,33
41	0,53	0,33	0,61	0,24	0,54	0,22	0,44	0,28	0,61	0,28	0,54	0,27
42	0,32	0,35	0,42	0,49	0,48	0,24	0,48	0,35	0,59	0,35	0,46	0,36
43	0,31	0,38	0,40	0,39	0,46	0,40	0,28	0,30	0,40	0,30	0,37	0,36
Totaal	0,48	0,30	0,55	0,29	0,59	0,26	0,55	0,25	0,62	0,25	0,56	0,27

Algemeen

In Tabel 4-9 staan links de onderscheiden deelattributen voor Schrijfvaardigheid opgesomd, daarna volgen voor de onderscheiden onderwijstypen de gemiddelde p-waarden en de Rit-waarden. Het aantal leerlingen aan wie de opgaven zijn voorgelegd, verschilt per schooltype en soms per opgave binnen schooltype: de *minimum*-aantallen reponses per item zijn:

- voor bb: 655 leerlingen
- voor kb: 648 leerlingen
- voor gt: 1640 leerlingen
- voor havo: 975 leerlingen
- voor vwo: 1309 leerlingen

Het eerste wat opvalt is, dat bij de hoofdattributen 1 en 2 de gemiddelde p-waarde hoger is naarmate het een hoger schooltype betreft: de gemiddelde p is voor kb hoger dan voor bb, voor gt hoger dan bij kb en voor havo en vwo weer hoger dan voor gt. Daarbij past nog de opmerking dat in de pretest andere items zijn gebruikt voor vmbo en voor h/v en dat in het constructieproces geprobeerd is de items voor h/v een hogere moeilijkheidsgraad mee te geven. Waar het de attributen 3 en 4 betreft, is er een ander beeld: daar is de gemiddelde p-waarde voor havo en vwo niet systematisch hoger dan voor de andere schooltypen.

Verder past bij de getallen voor deelattribuut 4.3 *leestekens en hoofdletters correct hanteren* de opmerking, dat er gerede twijfel is bij de waarde van de voorgelegde items. Het gebruikte itemformat (open bevraging, waarbij leerlingen meer en andersoortige antwoorden konden invullen dan beoogd en gevraagd) bleek minder geschikt voor het bevragen van interpunctie en vereiste bovendien een handmatige correctieslag. In een volgende pretest worden andere itemtypes beproefd.

Vmbo-items

De moeilijkheid van de items die bij de vmbo-leerlingen zijn afgenomen, is gemiddeld goed (tussen de .20 en .80). Er is een viertal deelattributen waarvan de gemiddelde p-waarde aan de lage kant is, namelijk 2.2, 3.1, 4.2 en 4.3. Deelattribuut 2.2 *passende volgorde van tekstelementen bepalen en correcte indeling en lay-out aanbrengen* heeft de laagste p-waarde met 0,23. Deze lage p-waarde wordt deels bepaald door een item met een p-waarde van 0,03, maar ook zonder dit item blijft het met een gemiddelde p-waarde van 0,26 een moeilijk deelattribuut voor bb-leerlingen. Bij deelattribuut 2.2 moeten leerlingen bijvoorbeeld alinea's van een tekst in de goede volgorde zetten of aangeven op welke plek in de tekst een nieuwe alinea begint. Deze taak blijkt voor alle vmbo-leerlingen complex, bij kb heeft dit attribuut een gemiddelde p-waarde van 0,34 (waarbij ook een (ander) item een bijzonder lage p-waarde heeft van 0,06) en bij gt een gemiddelde p-waarde van 0,23. Naast een inhoudelijke verklaring waarbij de complexiteit van de tekst een factor is, kan bij de 'volgorde'-items de keuze van het itemtemplate wellicht een rol spelen. Bij het slepen van de alinea's in de juiste volgorde is het denkbaar dat de leerling het overzicht kwijtraakt. Wellicht is een template waarbij de leerling alinea's naar lege vakken moet slepen overzichtelijker; een dergelijk itemtype kan beproefd worden in een volgende pretest. Verder speelt nog een rol dat bij bepaalde itemtypen de leerling alle itemresponsen goed moet hebben, om een positieve diagnose op dit item te krijgen; dit geldt bij de items waarbij de alinea's in de goede volgorde geplaatst moeten worden en bij de items waarbij de tekst in alinea's verdeeld moet worden.

Opvallend is ook het feit dat de p-waarden bij bb en gt een stuk lager zijn dan die bij kb. Een mogelijke verklaring is dat de opgaven voor gt voor dit deelattribuut relatief moeilijker waren. Inhoudelijk ligt het niet voor de hand om uit te gaan van een mindere vaardigheid op dit deelattribuut bij gt-leerlingen.

Deelattribuut 4.3 *leestekens en hoofdletters correct hanteren* valt ook op door zijn hoge moeilijkheidsgraad. De gemiddelde p-waarde bij dit deelattribuut is 0,31. Zoals al opgemerkt is deze lage p-waarde wellicht deels te wijten aan het gebruikte itemtype (de *hottext correction*). Bij dit itemtype moet een leerling in een tekst de plek aanklikken waar de hoofdletter of het leesteken moet komen. De leerling moet vervolgens het juiste leesteken plaatsen. Het antwoordgedrag van de leerling bij dit itemtype is echter zo onvoorspelbaar dat het correctievoorschrift niet dicht te timmeren is, waardoor handmatige correctie nodig was. Zo gebruiken leerlingen verschillende tekens als leestekens (bijvoorbeeld komma's als aanhalingstekens) en zetten ze vreemde tekens ("*characters")* voor of na hun antwoord. In een volgende pretest zullen andere itemtypes uitgeprobeerd worden, waaronder meerkeuzevragen en vragen waarbij de leerling leestekens naar de juiste plek moet slepen.

Ook deelattribuut 4.2 *correct toepassen van overige regelgeleide spelling* lijkt moeilijk voor leerlingen in het vmbo. De gemiddelde p-waarde bij bb is 0,32, bij kb 0,40 en bij gt 0,46. Deze gegevens lijken aan te sluiten bij het beeld dat docenten schetsen van de zwakke spelvaardigheid van vmbo-leerlingen. Daarnaast is ook hier de technische verklaring van toepassing dat er bij een aantal items gebruik is gemaakt van het itemtype *hottext correction*. Dit itemtype heeft een veelheid aan (soms juiste) responses uitgelokt die omwille van de vorm fout gerekend zijn in de geautomatiseerde beoordeling, en daarna handmatig aangepast moesten worden. Leerlingen blijken bijvoorbeeld vaak het fout gespelde woord dat in de opgave stond, te laten staan en daarachter hun juiste antwoord te typen waardoor het geheel een fout antwoord oplevert. Andere itemtypes waarbij de automatische scoring beter verliep, waren de *hottext* waarbij de leerling moet aanklikken welke woorden aan elkaar moeten worden geschreven en de *gaps inline* waarbij de leerling zelf een woord moet invullen.

Een ander deelattribuut dat aan de moeilijke kant is, is het deelattribuut 3.1 *een correcte zinsbouw hanteren.* Bij bb is gemiddelde p-waarde 0,40 en deze loopt op bij kb naar 0,46 en bij gt naar 0,53. De leerling moet bij dit deelattribuut voornamelijk aangeven welke zinnen in een stuk tekst een fout in de zinsbouw bevatten. De fouten bestaan bijvoorbeeld uit incongruentie in tijd, incongruentie in getal, onjuiste verwijzingen en onduidelijke verwijzingen. Dit levert regelmatig problemen op voor de leerlingen. Daarnaast speelt hier het feit dat een aantal items op alle niveaus is ingezet een rol. Bij dit deelattribuut is te zien dat een aantal van de items die op meer niveaus zijn ingezet echt te moeilijk zijn voor bb.

Het feit dat een aantal items op alle niveaus is afgenomen, zou wellicht deels kunnen verklaren dat de moeilijkheidsgraad van de items bij de meeste deelattributen afneemt naarmate het niveau van de leerweg

toeneemt. Het is niet verwonderlijk dat een leerling op bb-niveau een bepaald item minder goed maakt dan een leerling op gt-niveau. Daarnaast lijkt het erop dat de items voor de hogere niveaus makkelijker gevonden worden door die doelgroep.

Naast een aantal items met een relatief hoge moeilijkheidsgraad zijn er bij gt ook twee deelattributen met een relatief lage moeilijkheidsgraad, namelijk 1.1 en 3.3. Deelattribuut 1.1 *voorkennis en informatiebehoefte bij de lezers inschatten* heeft een p-waarde van 0,77. Deze hoge p-waarde wordt voornamelijk veroorzaakt doordat er binnen een itemscherm een aantal itemresponses voorkomt met een p-waarde van 0,90 of hoger. Bij de volgende pretest moeten we overwegen om de itemresponses waarvan we verwachten dat ze een erg hoge p-waarde zullen hebben, al als goed antwoord 'weg' te geven of ze niet mee te nemen in de codering, zodat ze in de resultaten geen rol spelen.

Deelattribuut 3.3 *passend en gevarieerd woordgebruik laten zien* heeft bij gt ook een hoge p-waarde van 0,81. Ook hierbij is er sprake van een aantal extreem hoge p-waarden (boven de 0,90) bij een aantal itemresponses. Daarnaast heeft de grote meerderheid van itemresponsen een p-waarde van boven de 0,70. Wellicht komt dit doordat dit deelattribuut met name receptief wordt getoetst en de gt-leerling een grotere receptieve woordenschat heeft dan de bb- en de kb-leerling. We kunnen echter niet uitsluiten dat de items voor gt eenvoudiger waren dan die voor bb en kb.

Items havo/vwo

Items van attribuut 1*, Retorische vaardigheden*, worden door leerlingen over het algemeen goed gemaakt: anders dan op voorhand door de vaksectie ingeschat werd, zijn leerlingen van klas 3 op h/v-niveau kennelijk al behoorlijk bedreven in het maken van retorische overwegingen die voorafgaan aan het schrijven. Dit geldt voor alle drie de deelattributen. Overigens kunnen de relatief hoge p-waarden ook anders worden verklaard: het blijkt telkens lastig om voor dit attribuut goede opgaven te maken waarbij volledige consensus bestaat over de goede antwoorden. Daardoor is in het constructieproces gekozen voor zogenaamde clear cases: items waarbij het antwoord onbetwist en onomstotelijk is en daardoor eenvoudiger.

Bij het tweede hoofdattribuut, *Tekststructurele vaardigheden*, scoren de h/v-leerlingen wat minder hoog dan op deelattribuut 1. De vwo-leerlingen scoren structureel iets hoger dan de havo-leerlingen. Ook hierbij geldt de opmerking dat dit een vertekend beeld kan zijn, omdat we de relatieve moeilijkheidsgraad van de afzonderlijke items nog niet kennen. Opvallend is de hoge score op deelattribuut *2.4 een standpunt weergeven en van passende argumenten voorzien*. De verwachting was dat leerlingen met argumentatieve vaardigheden nog veel moeite zouden hebben, temeer omdat deze deelvaardigheid niet structureel in het onderwijs aan de orde komt. Een aantal items in de pretest is, achteraf bezien, wat te eenvoudig voor de beoogde populatie: met name de gebruikte afleiders lijken aan de eenvoudige kant, waardoor het item als geheel eenvoudiger wordt.

Overigens waren er binnen dit deelattribuut ook enkele items met een erg lage p-waarde (<0,10). Op advies van de vaststellingscommissie zijn deze erg moeilijke items:

- ofwel gehandhaafd: de leerling moet dit wel beheersen volgens de commissie, ook al blijkt de praktijk momenteel anders;
- ofwel minimaal veranderd om opnieuw te worden gepretest: in betreffende items wordt gevraagd naar de mening van leerlingen, waarmee de suggestie wordt gedaan dat elk antwoord goed is.

Bij attribuut 3*, linguïstische vaardigheden op woord- en zinsniveau*, en attribuut 4, *spelling en interpunctie*, is de gemiddelde p-waarde lager dan bij de eerste twee attributen. Binnen attribuut 3 valt op dat de gemiddelde p-waarden nogal fluctueren: bij deelattribuut 3.1, *correcte zinsbouw hanteren*, is de gemiddelde p-waarde 0,28 en bij deelattribuut 3.3, *passend en gevarieerd woordgebruik*, is de gemiddelde p-waarde 0,71. Bij laatstgenoemd deelattribuut wordt leerlingen bijvoorbeeld gevraagd welke woorden wel respectievelijk niet passend zijn in een bepaalde context. Deze uiteenlopende moeilijkheidsgraad doet zich ook voor bij de populaties van vmbo. Hiermee is niet gezegd dat deelattribuut 3.3 eenvoudiger zou zijn voor de leerlingen: de relatief hoge scores lijken, net als bij attribuut 1, deels veroorzaakt doordat de foute antwoorden te duidelijk fout zijn, waardoor de opgaven te gemakkelijk worden. Verder zou de vorm van de matrixvraag een rol kunnen spelen; om die reden is het goed om in komende pretests een deel van de

opgaven op deelattribuut 3.3 een andere vorm te geven, bijvoorbeeld de vorm van een meerkeuzevraag waarbij één woord moet worden aangewezen dat passend is in de gegeven context, of juist niet passend.

Op attribuut 4, *spelling en interpunctie*, is de gemiddelde p-waarde weer lager dan bij attribuut 3. Dit wordt mede veroorzaakt door de lage p-waarden op deelattribuut 4.3, *leestekens en hoofdletters correct hanteren*. De lage p-waarden op dit deelattribuut kunnen, zoals ook al aangeduid bij de bespreking van de vmbo-items, een artefact zijn van de gehanteerde vraagvorm: doordat dit deelattribuut in open vorm is bevraagd hadden leerlingen de mogelijkheid om een bonte variëteit van antwoorden te geven waarin niet was voorzien in het gehanteerde antwoordmodel. Te denken valt aan door de toetsconstructeurs onvoorziene woordherhaling, meerdere leestekens op één invulplek en veelvuldig spatiegebruik. Elk van deze afwijkingen leidt in het geautomatiseerde proces van beoordelen tot afkeuring van het antwoord, ook als de respons in essentie wel laat zien dat de leerling het gevraagde antwoord wel weet. Het is tijdens de scoringprocedure niet goed gelukt de hier beschreven datavervuiling handmatig helemaal weg te filteren. Net als bij vmbo is de conclusie getrokken dat in komende pretests dit deelattribuut in meer gesloten vorm moet worden bevraagd, bijvoorbeeld in meerkeuzevorm of met gebruikmaking van een leesplankje waar de leerling het beoogde antwoord van af kan slepen.

Afgekeurde items

Na afname van de items is gekeken of er items moeten worden afgekeurd omdat ze extreme psychometrische waarden hebben. Bij de analyse is uitgegaan van de volgende richtlijnen:

- Extreem hoge p-waardes; itemresponses met extreem hoge p-waardes zijn in principe niet zo bruikbaar. Soms maakt de itemresponse echter deel uit van een item met meerdere responses die betere waardes geven. Als minimaal de helft informatief is, hebben we het item behouden;
- Extreem lage p-waardes; ook hiervoor geldt dat deze itemresponses in principe niet zo bruikbaar zijn, maar soms deel uitmaken van een item met meerdere responses die betere waardes hebben.
- Rit-waardes onder 0,10 (of negatief) zijn niet informatief, items met dergelijke Rit-waardes hebben we in principe laten vervallen, tenzij de vaststellingscommissie heeft benadrukt dat het betreffende item gehandhaafd zou moeten worden (dit was het geval bij enkele items voor subattribuut 2.4: een standpunt weergeven en van passende argumenten voorzien).

Bij vmbo vallen in totaal elf items af. Twee items vallen om technische redenen af. Dit zijn items van het hottext correctionsoort (correctie-item). Twee items vallen om inhoudelijke redenen af. Bij beide items kan achteraf bezien discussie ontstaan over de goede antwoorden. Zeven items zijn niet meer bruikbaar, omdat ze in de voorbeeldtoetsen zijn opgenomen. Naast de items die afvallen zijn er drie items die opnieuw moeten worden meegenomen in de pretest, omdat één of meer responsen van dat item minder gelukkig zijn en het item om die reden aangepast worden. Daarnaast zijn er drie items waarbij de codering inhoudelijk niet klopt. Deze items worden ook opnieuw aangeboden in de volgende pretest.

Bij havo zijn in totaal 14 items afgekeurd, waarvan vijf voor deelattribuut 1.1 vanwege te hoge p-waarden. De overige afgekeurde items zijn gespreid over de deelattributen; de afkeuring heeft te maken met:

- ofwel een te lage p-waarde, veroorzaakt door een te lastige taakstelling (bijvoorbeeld Klik de verkeerd gespelde woorden aan en verbeter ze);
- ofwel te lage Rit-waarden hetgeen duidt op een te lage of zelfs negatieve bijdrage aan de betrouwbaarheid van de adaptieve toets
- ofwel een codering die ervoor zorgde dat responses op het item anders werden gewaardeerd dan bedoeld.

Bij vwo zijn er in totaal 10 (deels met havo overlappende) items afgekeurd omdat:

- ofwel de p-waarde te hoog was: in deze gevallen wordt het item nog eens op haar merites bekeken en vervolgens gepretest voor een lager onderwijstype
- ofwel de p-waarde te laag was: dit was het geval bij enkele klik- en verbeter-opgaven.
- ofwel de Rit te laag was, zodat het item weinig of niets bij zal dragen aan de betrouwbaarheid van de adaptieve toets.

4.3.3 Resultaten Wiskunde

Algemeen

De pretest van DTT wiskunde in 2015 bevatte items uit de domeinen Meten en meetkunde (D) en Verbanden en formules (E). Elk domein bevatte items van alle drie de kenmerken uit het leerlingmodel. De andere domeinen, B en C voor vmbo en B, C en F voor havo/vwo zullen in de pretest van het voorjaar van 2016 aan de orde komen.

De minimum aantallen leerlingen per schooltype zijn:

- vmbo-bb: 495
- vmbo-kb: 187
- vmbo-tl: 1304
- havo: 1152
- vwo: 1192

Het leerlingmodel van wiskunde bestaat uit een matrix. In deze matrix worden de rijen gevormd door de vakinhoudelijke domeinen uit de tussendoelen. De kolommen worden gevormd door de vakdidactische kenmerken, zie Tabel 4-10.

Tabel 4-10. Leerlingmodel wiskunde

Domein	Structuur (1)	Meerduidigheid (2)	Samenhang (3)
B: Getallen (havo/vwo en variabelen)			
C: Verhoudingen			
D: Meten en meetkunde			
E: Verbanden en formules			
(havo/vwo F: Informatieverwerking en onzekerheid)			

Bij de constructie van de items bij vmbo bleek het vooral voor het domein D Samenhang en het domein E Meerduidigheid lastig om items te construeren. In de uiteindelijke pretest zijn de geconstrueerde items aan leerlingen van alle niveaus van de populatie voorgelegd. In het komende constructiejaar zal extra aandacht gegeven worden aan deze onderdelen bij de constructie. Wellicht dat nieuwe technische ontwikkelingen ervoor kunnen zorgen dat deze onderdelen beter aan bod komen.

Vmbo-items

In Tabel 4-11 staat een overzicht van de aantallen itemresponses per deelattribuut die zijn afgenomen per vmbo-niveau (bb, kb en tl). Er zijn meer erg moeilijke itemresponses geweest dan erg makkelijke. Van de itemresponses die bij vmbo-bb zijn afgenomen, hebben 17 van de 125 een p-waarde die lager is dan 0,10. Deze items zijn daarmee te moeilijk. Voor vmbo-kb is dit bij 17 van de 140 itemresponses het geval. Bij vmbo-tl hebben 13 van de 137 responses een p-waarde lager dan 0,10.

Het aantal itemresponses dat aan de gemakkelijke kant is, is veel kleiner. Er zijn geen itemresponses met een p-waarde boven 0,90 bij vmbo-bb vastgesteld. Voor vmbo-kb waren er 3 itemresponses met een p-waarde boven 0,90 en bij vmbo-tl slechts 1.

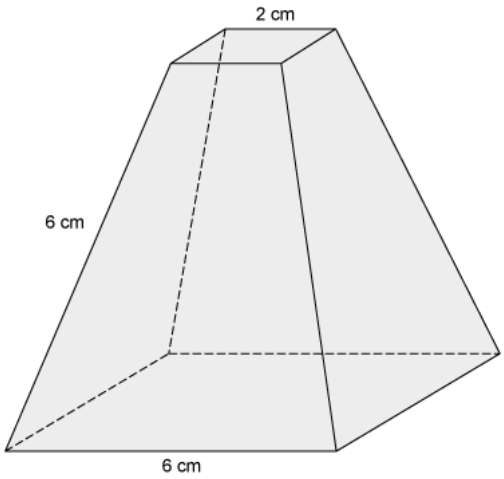
Tabel 4-11. Aantallen itemresponses vmbo

	totaal aantal itemresponses			aantal itemresponses met p-waarde <0,10			aantal itemresponses met p-waarde >0,90			aantal itemresponses met Rit <0,10		
Domein en kenmerk	bb	kb	tl	bb	kb	tl	bb	kb	tl	bb	kb	tl
D structuur	26	26	22	4	3	2	-	-	-	1	3	4
D meerduidigheid	25	29	28	1	1	3	-	-	-	1	2	10
D samenhang	19	19	19	4	3	1	-	-	1	3	3	3
E structuur	22	25	28	1	2	3	-	3	-	1	0	2
E meerduidigheid	13	13	13	3	2	2	-	-	-	2	2	2
E samenhang	20	28	27	4	6	2	-	-	-	4	2	5
totaal aantal itemresponses	125	140	137	17	17	13	-	3	1	12	12	26

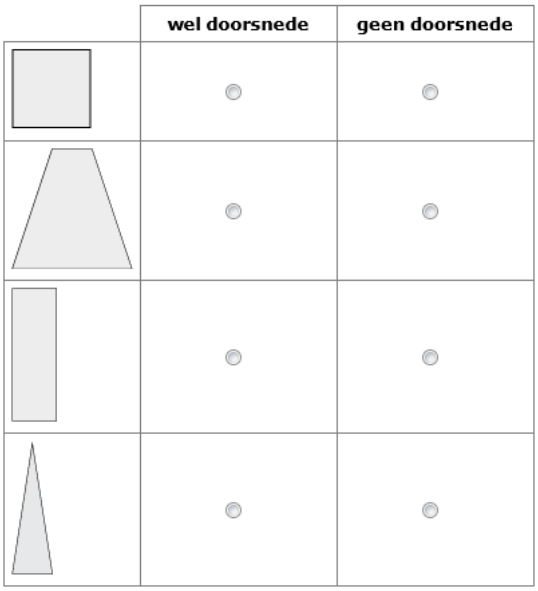
Opvallend is het hoge aantal responses bij vmbo-gt met een Rit-waarde kleiner dan 0,10. Dit wordt vooral veroorzaakt door het grote aantal items met een lage Rit bij het domein D Meerduidigheid. Als bijzonderheid in dit kader is wellicht op te merken dat er hier twee items zijn met respectievelijk 3 en 4 itemresponses waarvan er respectievelijk 2 en 3 een Rit-waarde lager dan 0,10 hebben. De items hebben geen lage p-waarde, omdat het eigenlijk steeds twee-keuze vragen zijn. Als dergelijke vragen lastig zijn, bestaat de mogelijkheid dat dat niet in de analyses aan de p-waarden afgelezen kan worden, omdat er een grote gokkans is om een dergelijk item goed te beantwoorden. Wel is, als waarschijnlijk gevolg van dat gokaspect, de itemrestcorrelatie in zo'n situatie laag, zie Voorbeeld 4-1.

Voorbeeld 4-1. (dtt-wi-vm-0260-D2): Item met een lage Rit-waarde

Je ziet een afgeknotte piramide. Het grondvlak en het bovenvlak zijn vierkant.



Geef bij elke figuur aan of het een doorsnede van de piramide kan zijn.



De items waarvan de itemresponses een zeer hoge of lage p-waarde hadden of een zeer lage rit-waarde zijn inmiddels besproken met de vaststellingscommissie. In totaal waren dit 57 items. Op grond van de resultaten in de pretest heeft de vaststellingscommissie besloten 13 van deze 'omstreden' items te laten vervallen voor het vervolg van de DTT-bankontwikkeling. Bij 22 van deze items is besloten om ze aan te passen en ze dus vervolgens in de toekomst opnieuw te pretesten. 12 items bleken te lastig, maar zijn wellicht bruikbaar op havo of vwo. De overige items waren wel onderscheidend op een bepaald niveau en worden om die reden toch behouden.

Tijdens een observatie door toetsdeskundigen wiskunde tijdens de afname van de pretests bleek dat de toetsduur van 3 uur voor veel vmbo-leerlingen, met name voor leerlingen van vmbo-bb, te lang was.Leerlingen konden zich na zo'n anderhalf uur niet meer goed concentreren waardoor de laatste items van de toets niet meer serieus werden gemaakt[10]. Aan de afnametijden is te zien dat leerlingen gemiddeld veel minder tijd nodig hadden en veel leerlingen een pauze kregen of namen, met name bij bb - dat was ook toegestaan (zie hoofdstuk 4.2 Responstijden). Andere bevindingen tijdens deze observatie zijn beschreven in een door de toetsdeskundigen ingevulde checklist en zijn zeker zinvol om mee te nemen in het verdere constructie-, ontwikkel- en organisatieproces. Meer informatie over de observaties kan gevonden worden in het pretestverslag van CvTE (mei 2015).

Havo/vwo-items

In Tabel 4-12 staat een overzicht van de aantallen itemresponses die per niveau (havo en vwo) zijn afgenomen per deelattribuut voor havo en voor vwo. Opvallend is dat de itemresponses in de pretest aan de moeilijke kant zijn geweest. Van de itemresponses die op havo zijn afgenomen hebben er 27 van de 129 een p-waarde die lager is dan 0,10 en zijn daarmee te moeilijk. Op het vwo hebben 21 van de 114 itemresponses een p-waarde lager dan 0,10. Het aantal itemresponses dat aan de gemakkelijke kant is, is veel kleiner. Op de havo was er 1 itemrespons met een p-waarde boven 0,90; op het vwo waren dat er 3. Voor de toekomstige constructie dienen toetsdeskundigen zich in ieder geval bewust te zijn dat het oppassen geblazen is dat items niet te moeilijk worden en herkenbaar blijven voor leerlingen.

Tabel 4-12 Aantallen itemresponses havo/vwo

	totaal aantal itemresponses		aantal itemresponses met p-waarde <0,10		aantal itemresponses met p-waarde >0,90		aantal itemresponses met Rit <0,10	
Domein en kenmerk	havo	vwo	havo	vwo	havo	vwo	havo	vwo
D structuur	17	16	6	3	-	2	1	3
D meerduidigheid	23	23	2	1	-	1	1	1
D samenhang	17	17	4	1	-	-	2	1
E structuur	25	23	3	1	-	-	5	1
E meerduidigheid	20	15	3	6	-	-	2	1
E samenhang	27	20	9	9	1	-	1	3
Totaal aantal itemresponses	129	114	27	21	1	3	13	10

Verder valt op dat er ook hier diverse itemresponses zijn met een lage itemrestcorrelatie. Op havo zijn dat er in totaal 13 en op vwo zijn dat er 10. De items die een lage p-waarde of een lage itemrestcorrelatie hebben, zijn besproken met de vaststellingscommissie.

In totaal zijn er op basis van deze pretestresultaten 54 items besproken die problematisch zijn met de vaststellingscommissie. Van deze items is besloten om er 7 te laten vervallen omdat ze niet goed hebben gefunctioneerd in de pretest. Van de overige items is besloten om er 17 te handhaven omdat ze bijvoorbeeld toch onderscheidend bleken te zijn. Van de overige 31 items is besloten om ze aan te passen en ze dus opnieuw te pretesten.

Een reden om items te laten vervallen, kan zijn dat ze buiten de toetswijzer vallen. Dit betreft dan bijvoorbeeld onderwerpen die traditioneel aan het einde van het 3e jaar aan bod komen en buiten de DTT dienen te worden gehouden omdat de toets niet aan het einde van het 3e jaar wordt afgenomen.

Een andere reden om items te laten vervallen is een lage Rit-waarde. Een voorbeeld van zo'n item is Voorbeeld 4-3. In deze vraag heeft de leerling de keuze uit drie verschillende antwoorden. De lage Rit-waarde van dit item (0,04) zou veroorzaakt kunnen zijn door gokgedrag van de leerlingen.

[10] Om vertekening door dergelijke antwoordtendenties te voorkomen stonden in de verschillende toetsvarianten steeds andere opgaven achteraan, zie § 3.2 over het design.

Voorbeeld 4-2: Item met een lage Rit-waarde, vermoedelijk veroorzaakt door gokgedrag van de leerlingen

Mark heeft 0,2 liter bitter lemon in zijn bekertje. Er kan maximaal 275 cm^3 in. Het bitter lemonflesje heeft een diameter van 5 cm. Er staat nog 4 cm drank in het flesje.

Kan het restant uit het flesje nog helemaal in het bekertje bijgeschonken worden?

- Ja, de drank past er nog helemaal in.
- Nee, de drank past er niet helemaal meer in.
- Er zijn niet genoeg gegevens om dit te berekenen.

De digitale omgeving waarin de items worden afgenomen is nog volop in ontwikkeling. In de pretest is een aantal problemen met de digitale omgeving aan het licht gekomen. Zo was er een opgave waarin leerlingen de gemengde breuk $5\frac{2}{3}$ moesten invoeren. In de beoordeling werd dit gelezen als $5\times\frac{2}{3}$, waardoor goede anwoorden van leerlingen als fout beoordeeld werden. Dit werd tijdens de dataverwerking geconstateerd. Voor de toekomst is extra alertheid op dergelijke aspecten geboden.

In een ander item werd van leerlingen gevraagd een vergelijking op te stellen van een lijn die loodrecht staat op een gegeven lijn. De software kon, in tegenstelling tot de verwachtingen van de toetsdeskundigen wiskunde ten tijde van het ontwikkelen van deze items, nog niet overweg met de veelheid aan verschillende juiste antwoorden die hier gegeven zijn.

De itemresponses met hoge p-waarden (>0,90) zijn zonder uitzondering onderdeel van een item met meerdere itemresponses. De eerste itemrespons kan dan inleidend zijn, bedoeld als een opstapje. In die gevallen is de itemrespons gehandhaafd, ook al is de p-waarde erg hoog. Een andere reden kan zijn dat juist de combinatie van meerdere vragen interessant is in de context.

In de analyse van de items is gekeken naar de mate waarin ze onderscheidend zijn op ‘onder’, ‘op’ of ‘boven niveau’. Voor de adaptieve afname is het van belang om items te hebben die sterk zijn in het onderscheiden van leerlingen in deze drie groepen. Dat onderscheidend vermogen is bij items een reden geweest om dat item te behouden, ook al was de p-waarde laag of juist hoog.

Bij het bestuderen van de leerlingenantwoorden valt niet aan de conclusie te ontkomen dat leerlingen de items van deze pretests in grote mate serieus beantwoord hebben. Er zijn uiteraard leerlingen die ‘onzin’-antwoorden geven, maar de overgrote meerderheid van de gegeven antwoorden betreft serieuze pogingen om tot het juiste antwoord te komen.

Dit past in het beeld dat toetsdeskundigen hebben die een afname van de DTT wiskunde op een school hebben bijgewoond. Ook op de bezochte school waren leerlingen serieus met de opgaven aan het werk. De bezoekende toetsdeskundigen hebben in het bijzonder gelet op de technische zaken, zoals het gebruik van de rekenmachine in FACET en de formule-editor. De leerlingen op de bezochte school maakten geen gebruik van een muis maar van een touchpad. Dat betekent dat het selecteren van knoppen in het scherm door te klikken wat meer moeite kost dan wanneer leerlingen een muis hebben.

De formule-editor leek goed te werken: leerlingen vinden hun weg daar wel in. Opvallend is dat leerlingen de variabele die klaarstond boven de opgave, toch vaak opnieuw intikten. Dit kon ook teruggezien worden in de leerlingantwoorden in de database. In de nieuwere versie van de formule-editor is dit probleem gelukkig verholpen en staat de variabele al in het venster waarin de leerling zijn of haar antwoord intypt.

Leerlingen zijn geneigd om in de formule-editor tekst te typen om zo de overgang van de ene formule naar de andere formule te verduidelijken. Dit geeft problemen met de automatische beoordeling van items.

Resultaat en conclusies

- Na afloop van de pretest zijn op grond van de analyses op basis van voorstellen van de toetsdeskundigen de volgende beslissingen genomen door de vaststellingscommissie:
    - een groot aantal items is geschikt gebleken en kan ter kalibratie in de verdere toekomst worden ingezet;
    - een aantal items is aangepast en daarmee klaar voor een volgende pretest;
    - een aantal items is verwijderd uit de DTT-itemset;
    - enkele items zijn van een andere niveau-indicatie voorzien en kunnen daar alsnog in een volgende pretest ondergebracht worden.
- Verder hebben vaststellingscommissie en toetsdeskundigen als gevolg van de pretest en de analyses een veel beter en realistischer beeld gekregen van na te streven moeilijkheidsgraad van DTT-items.
- Ook de beperkingen van de huidige toetsomgeving en de gevolgen daarvan voor het uitvoerbaar zijn van scorebepaling zijn helder naar voren gekomen.

De analyse van de pretest heeft zich nog niet gericht op de empirische onderbouwing van het leerlingmodel. Deze analyse wordt binnenkort uitgevoerd en is van eminent belang voor de verdere ontwikkeling van DTT wiskunde. Mocht het model empirisch niet onderbouwd kunnen worden, zal op domein gerapporteerd worden of een alternatief ontwikkeld moeten worden (wat een ingrijpend en tijdsintensief traject is).

4.4 Standaardbepaling door experts na de eerste pretest van de DTT

Bij de eerste adaptieve afname van de Diagnostische Tussentijdse Toets (DTT) in februari 2016 zullen voor elke leerling diagnoses gesteld worden voor schrijfvaardigheid Nederlands, schrijfvaardigheid Engels en wiskunde. Daartoe moet voor vijf leerniveaus (vmbo-bb, kb, gt, havo en vwo) per vakaspect (deel-attribuut) bepaald worden in welke beheersingscategorie de leerling valt: onder, op of boven niveau. Om de grenzen tussen de beheersingscategorieën goed te kunnen plaatsen, hebben in 2015 de eerste toetsgeoriënteerde standaardbepalingen met docenten plaatsgevonden. Er moest per vak voor elke leerweg een standaardbepaling met experts plaatsvinden, in totaal dus vijftien.

Om de pretest in omvang te beperken (en het benodigd aantal leerlingen) zijn niet alle items in de eerste pretest afgenomen, maar ongeveer de helft. Ook het diagnostisch profiel wordt in twee pretests opgebouwd. Deze stapsgewijze opbouw betekent dat ook de standaardbepaling in twee stappen opgebouwd moet worden. Bij de talen zijn na de eerste pretest grenzen gezet voor de vier hoofdaspecten van schrijfvaardigheid. Bij wiskunde zijn grenzen gezet voor de vakdidactische aspecten (structuur, meerduidigheid en samenhang) binnen twee domeinen (domein D: Meten en meetkunde en domein E: Verbanden en formules).

Na de standaardbepalingen zijn de standaarden doorgerekend en gerapporteerd aan het CvTE. Vervolgens heeft een besluitvormingsproces bij het CvTE plaatsgevonden en zijn de cesuren voor de eerste adaptieve afname vastgesteld (CvTE, juni 2015).

4.4.1 Standaardbepaling met experts

Experts

Voor elke standaardbepaling zijn 10 a 12 experts geworven door het CvTE. Het was wenselijk om de standaardbepalingen van een vak (voor de vijf verschillende leerwegen) door, in ieder geval deels, dezelfde groep experts te laten bepalen. Voor elk leerweg is er itemoverlap met de aangrenzende niveaus. Beoordeling door dezelfde experts kan voorkomen dat er voor de lagere niveaus items zijn die goed gemaakt moeten worden die bij een hoger niveau niet goed gemaakt hoeven worden.

Er is vooral onder de pilotscholen geworven omdat het erg belangrijk is dat de inhoudelijk experts mensen zijn die recente ervaring hebben met het onderwijzen van de doelgroep. Op deze wijze wordt zoveel mogelijk kennis en informatie gebruikt bij het zetten van de standaarden: aan de ene kant hebben we de tussendoelen en de toetswijzer, daarnaast hebben we de empirie in de vorm van opgaven en pretestdata en deze worden dan aangevuld met de ervaring en kennis van de docent. Ruim honderd docenten hadden zich opgegeven voor de standaardbepalingen. Veertig procent van de docenten deed aan twee of meer standaardbepalingen mee (zie Tabel 4-13).

De standaardbepaling voor een leerweg binnen een vak heeft plaatsgevonden op één dag. Er waren dus 3 vakken * 5 leerwegen = 15 sessies van één dag nodig om alle standaarden te kunnen bepalen. De standaardbepaling heeft plaatsgevondenvan 18 mei tot en met 4 juni 2015. In deze periode vonden maximaal twee standaardbepalingen van twee verschillende vakken op één dag plaats. Op deze wijze konden de experts aan meerdere standaardbepalingen deelnemen en was er in geval van nood een inval mogelijk voor de betrokken toetsdeskundigen en psychometrici. Bij elke standaardbepaling waren tussen de 7 en 12 docenten aanwezig (zie Tabel 4-14).

Tabel 4-13. Aantal standaardbepalingen per docent

	Aantal standaardbepalingen per docent			Totaal aantal docenten	
	1	2	3	Deelname	Geworven
Schrijfvaardigheid Engels	16	14	2	32	34
Schrijfvaardigheid Nederlands	18	10	1	29	36
Wiskunde	23	9	2	34	39

Tabel 4-14. Aantal docenten per standaardbepaling

	Aantal docenten per standaardbepaling					
	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Totaal
Schrijfvaardigheid Engels	8	10	9	12	11	50
Schrijfvaardigheid Nederlands	7	9	7	8	10	41
Wiskunde	8	10	11	8	10	47

Standaardbepalingsmethode

De standaardbepalingsmethode die hiervoor gebruikt is, is een voor de DTT aangepaste vorm van de 3DC-methode, ofwel de *Data-Driven Direct Consensus* methode (zie hoofdstuk 5 van het onderzoeksverslag uit 2014, Cito). Deze methode is eerder succesvol toegepast bij de standaardbepaling van de toetsen MBO-COE Nederlands en MBO-COE Rekenen en voor het bepalen van de prestatiestandaarden voor het Europees Referentie Kader (ERK) in 2013 (Feskens, Keuning, Van Til & Verheyen, 2014). Het voornaamste verschil is dat we bij de DTT geïnteresseerd zijn in de grenzen die per attribuut (cluster) gezet worden en niet, zoals bij de COE, in de grenzen per vak en/of latente trek.

Bij de standaardbepaling worden zoveel mogelijk verschillende informatiebronnen gebruikt om te komen tot gefundeerde standaarden: inhoudelijke descriptoren (de tussendoelen en toetswijzer), toetsmateriaal, ervaring van de experts en empirie (de pretest gegevens).

Analyse en selectie items

De moeilijkheid van alle items die meegaan in de standaardbepaling is bepaald door alle vmbo-items op één vaardigheidsschaal te leggen en alle havo/vwo-items op één vaardigheidsschaal te leggen met behulp van een model uit de itemresponstheorie. Hiervoor zijn de leerlingantwoorden ten bate van de standaardbepaling geanalyseerd met het One Parameter Logistic Model (Verhelst, Glas & Verstralen, 1995).

Het aantal items oversteeg wat de experts op 1 dag zouden kunnen bekijken. Op grond van de ervaring met de standaardbepaling bij de COE is een aantal van 12 items per aspect als richtlijn genomen:

(1) Talen: maximaal 12 items die zoveel mogelijk evenredig verdeeld zijn over de deelattributen.
(2) Wiskunde: maximaal 12 items per wiskundig aspect binnen een domein.

Items met responsen die bij verschillende aspecten horen zijn niet gebruikt. Ook items waarbij de meerderheid van de responsen problematisch waren (extreme p-waardes, lage Rit of negatieve discriminatie-index) zijn niet gebruikt tijdens de standaardbepalingen. Indien er na uitsluiting van de complexe en problematische items nog teveel items waren is een random selectie gebruikt.

Programma

Voorafgaand aan de standaardbepaling kregen de docenten een brief met een korte uitleg van het doel en het programma van de standaardbepaling. Daarnaast kregen zij de publieksversie van de toetswijzer (CvTE, 2014) en de tussendoelen van SLO (2012). De experts werden verzocht deze voorafgaand aan de standaardbepaling te bestuderen.

Tabel 4-15. Programma standaardbepaling

- Inloop + regelen formele zaken
- Welkom en toelichting
  - DTT in vogelvlucht
  - Leerlingmodel
  - Uitleg standaardbepaling
- Aanvang beoordelen
- Lunch
- Vervolg beoordelen

Voorafgaand aan de feitelijke standaardbepaling gaf een toetsdeskundige een presentatie over de diagnostische tussentijdse toets en het onderliggende leerlingmodel aan de experts. Tijdens de presentatie werd uitgelegd wat we willen diagnosticeren, hoe de items ontwikkeld worden en waarom een standaardbepaling nodig is. Aansluitend werd de werkwijze tijdens een standaardbepaling en het gebruik van de tools uitgelegd. Via de computer konden de experts de items en de sleutel zien, maar er was ook een geprinte versie van alle items. Hun oordelen konden de experts via een speciale tool ingeven.

Beoordelingen

Per aspect moesten twee standaarden gezet worden: een voor de onder/op grens en een voor de op/boven grens. Bij de talen werd in de ochtend de standaard voor de ondergrens gezet voor de vier hoofdaspecten en in de middag de standaard voor de bovengrens (dus 8 standaarden per standaardbepalingsdag). Bij wiskunde werden in de ochtend voor het eerste domein beide standaarden gezet bij de drie wiskunde aspecten en in de middag voor het tweede domein(dus 12 standaarden per standaardbepalingsdag, zie Tabel 4-16. Aantal standaarden).

Tabel 4-16. Aantal standaarden

Schrijfvaardigheid Engels		
2 grenzen **4 hoofdaspecten**	8 standaarden per leerweg 5 leerwegen	40 standaarden
Schrijfvaardigheid Nederlands		
2 grenzen **4 hoofdaspecten**	8 standaarden per leerweg 5 leerwegen	40 standaarden
Wiskunde		
2 grenzen **2 domeinen** **3 wiskundige aspecten binnen elk domein**	12 standaarden per leerweg 5 leerwegen	60 standaarden voor wiskunde

Per te zetten standaard waren er twee beoordelingsrondes. De experts kregen via de previewer van Facet per aspect de items (en sleutels) te zien. Per aspect dienden ze te bepalen hoeveel itemresponsen de leerling naar verwachting correct zou moeten beantwoorden op dat aspect als zijn/haar vaardigheid zich precies op de grens onder/op niveau bevindt. Hun evaluatie van de items konden ze aangeven op een kladpapier (Voorbeeld 4-6**Fout! Verwijzingsbron niet gevonden.**). Als ze alle items van een aspect hadden beoordeeld konden ze het standaardbepalingsformulier invullen (Voorbeeld 4-67**Fout! Verwijzingsbron niet gevonden.**). Op dit formulier konden de experts ook zien hoe de aspecten van een vak (de clusters) zich empirisch tot elkaar verhouden.

Als alle beoordelaars hun beoordeling hadden ingevuld, werden de beoordelingen samengevoegd en gepresenteerd aan de groep door de moderator. De moderator nodigde daarna de mildste en strengste experts uit om hun oordeel toe te lichten en argumenten te delen met de groep. Vervolgens volgde er een discussie over de beoordelingen. Als alle argumenten gedeeld waren volgde ronde 2, waarin de experts hun definitieve oordeel konden geven. De definitieve beoordelingen werden dan weer samengevoegd tot een grenswaarde, die weer aan de groep werd gepresenteerd door de moderator. Hierna volgde eventueel nog een korte discussie over de definitieve grenswaarde en herziening.

Deze hele procedure werd tenslotte herhaald voor de op/bovengrens. Alle beoordelingen (zowel van de eerste als tweede ronde) en standaarden zijn aan de pretestdatabase toegevoegd.

Voorbeeld 4-3. Kladpapier

Naam beoordelaar:

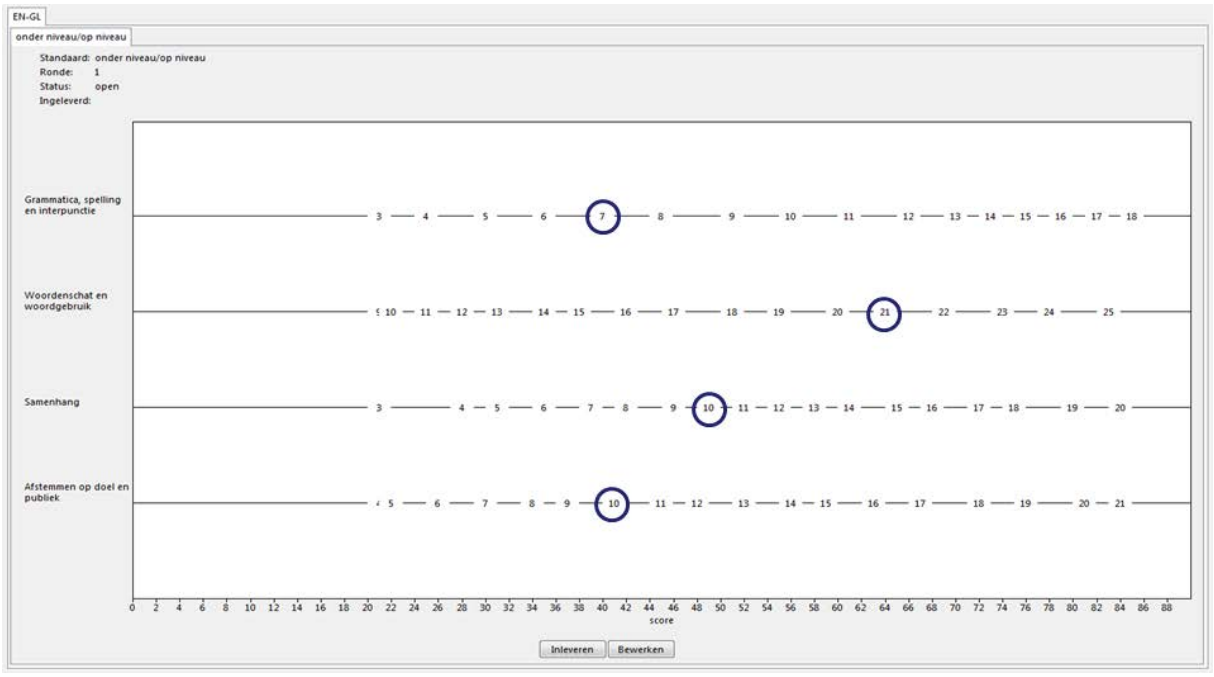
Hoofdattribuut 1: afstemming op publiek en doel

Vraag	Titel vraag	Aantal responsen	Max	Net op-niveau leerling behaalt Oordeel 1	Net op-niveau leerling behaalt Oordeel 2	Net boven-niveau leerling behaalt Oordeel 1	Net boven-niveau leerling behaalt Oordeel 2	Opmerkingen
dtt-en-hv-0184-SC	Hand in later	1	1					
dtt-en-hv-0291-SC	Dollars	2	3					
dtt-en-hv-0289-SC	currency	1	1					
dtt-en-hv-0185-SC	Pity	1	1					
dtt-en-hv-0294-SC	Oxford Hotel	7	7					
dtt-en-hv-0175-SC	Congratulations	3	3					

Hoofdattribuut 2: coherentie

Vraag	Titel vraag	Aantal responsen	Max	Net op-niveau leerling behaalt Oordeel 1	Net op-niveau leerling behaalt Oordeel 2	Net boven-niveau leerling behaalt Oordeel 1	Net boven-niveau leerling behaalt Oordeel 2	Opmerkingen
dtt-en-hv-0124-SC	Computers	1	1					
dtt-en-hv-0002-SC	Wordfile	1	1					
dtt-en-hv-0141-SC	Artikel	5	5					
dtt-en-hv-0169-SC	Review	1	1					
dtt-en-hv-0170-SC	Story	1	1					
dtt-en-hv-0198-SC	Salary	1	1					

Voorbeeld 4-4. Standaardbepalingsformulier



4.4.2 Rapportage gezette standaarden

In totaal zijn er 140 standaarden gezet: 40 voor Engels, 40 voor Nederlands en 60 voor wiskunde (zie Tabel 4-16). Per vak zijn verschillende informatieve maten berekend per leerweg, te diagnosticeren aspect en grens (zie 7.3 Bijlage standaardbepaling):

- Aantal voorgelegde items (en responsen)
- Gemiddeld oordeel (aantal correcte responsen)
- Spreiding in oordelen
- % responsen correct van maximum

Daarnaast is berekend hoeveel leerlingen in de categorieën onder, op en boven niveau zouden vallen onder de gezette standaard. Het gemiddelde oordeel en de range van alle oordelen is ook in het standaardbepalingsformulier weergegeven, zodat te zien is hoe de aspecten van een vak en de standaarden zich empirisch tot elkaar verhouden.

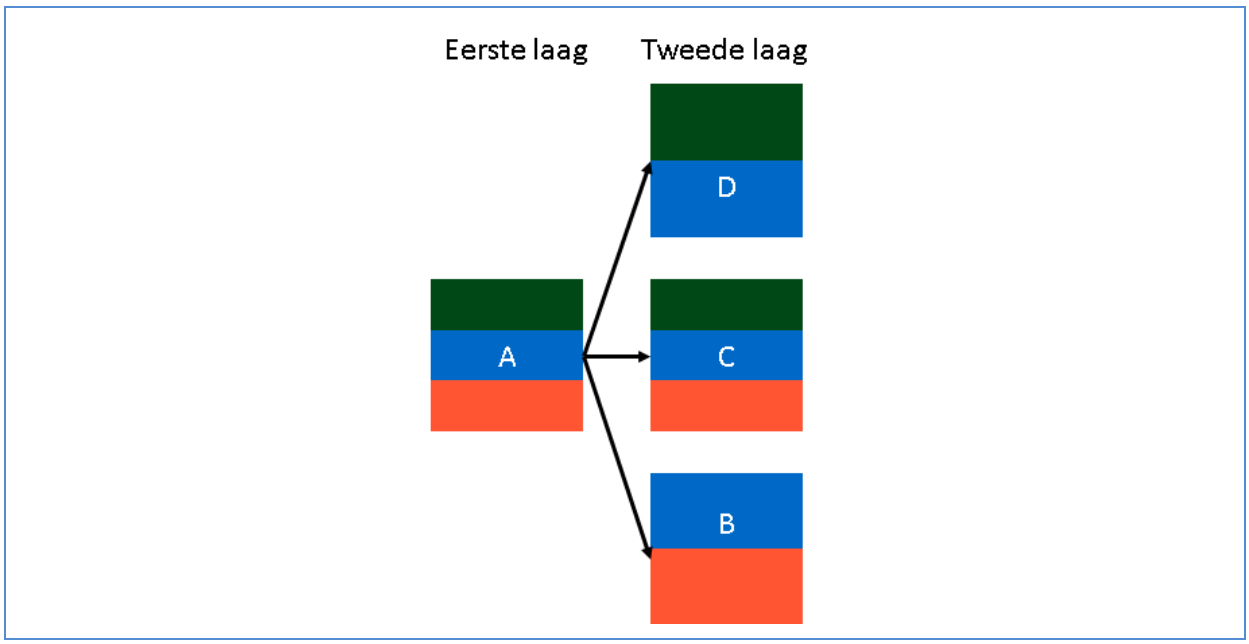
Per vak werden de tabellen met de maten, de grafische weergave van de verdeling en de standaardbepalingsformulieren met de gezette standaarden aan het CvTE opgeleverd. Op deze wijze kon een goed overzicht verkregen worden voor het besluitvormingsproces bij CvTE. Op grond van de cesuren die door het CvTE zijn vastgesteld is per vakaspect bepaald in welke beheersingscategorie elke pretestleerling valt (onder, op of boven niveau) en zijn de definitieve kalibraties uitgevoerd.

4.5 Indeling in blokken en simulatie

4.5.1 Adaptieve structuur voor eerste adaptieve afname

Aangezien we de adaptieve toets stapsgewijs opbouwen met in het eerste jaar de helft van de items en diagnoses, wordt de adaptiviteit ook stapsgewijs opgebouwd. Bij de eerste afname wordt adaptiviteit op blokniveau toegepast (zie ook hfst 3, onderzoeksverslag 2014), waarbij twee lagen van blokken worden gebruikt (zie Figuur 4-2). Bij de talen is per hoofdattribuut een blokstructuur en bij wiskunde per wiskundig aspect binnen een domein.

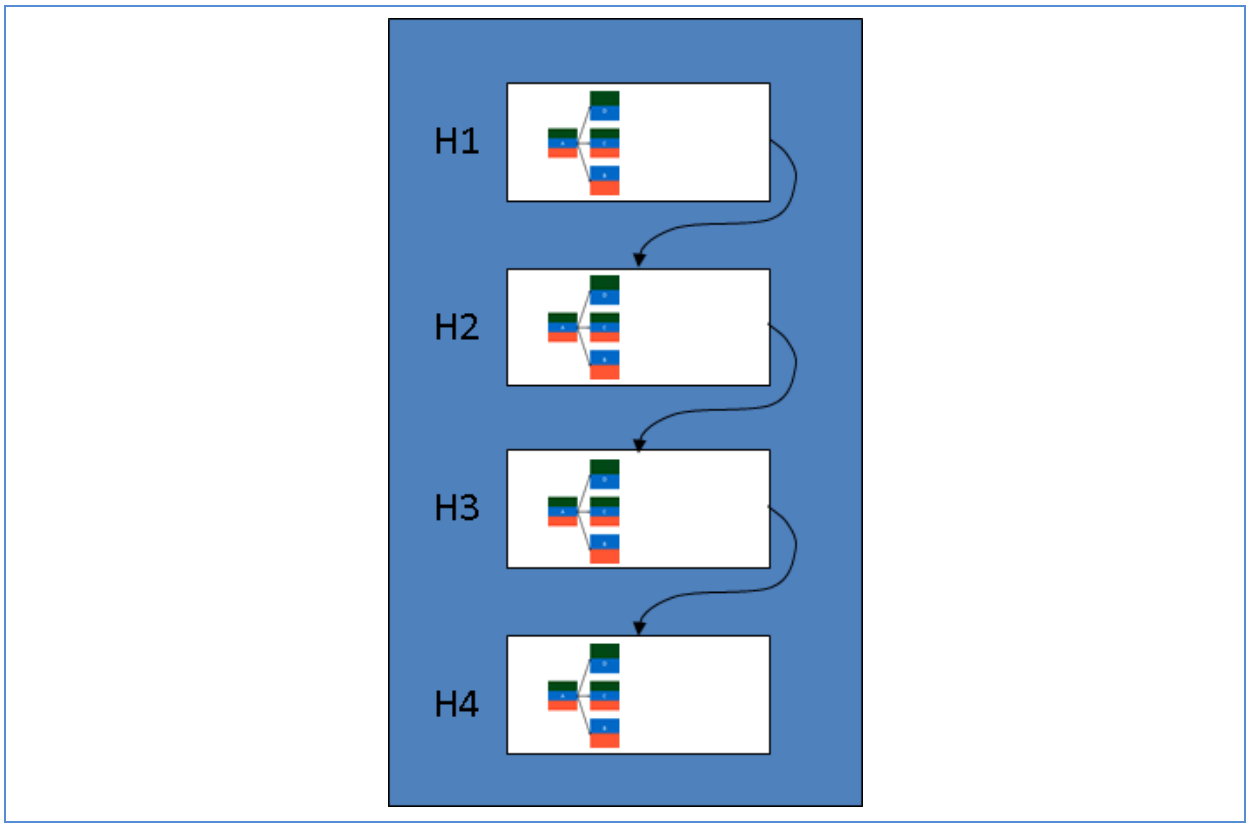
Figuur 4-2. Blokkenstructuur per aspect dat gediagnosticeerd wordt in de eerste adaptieve afname



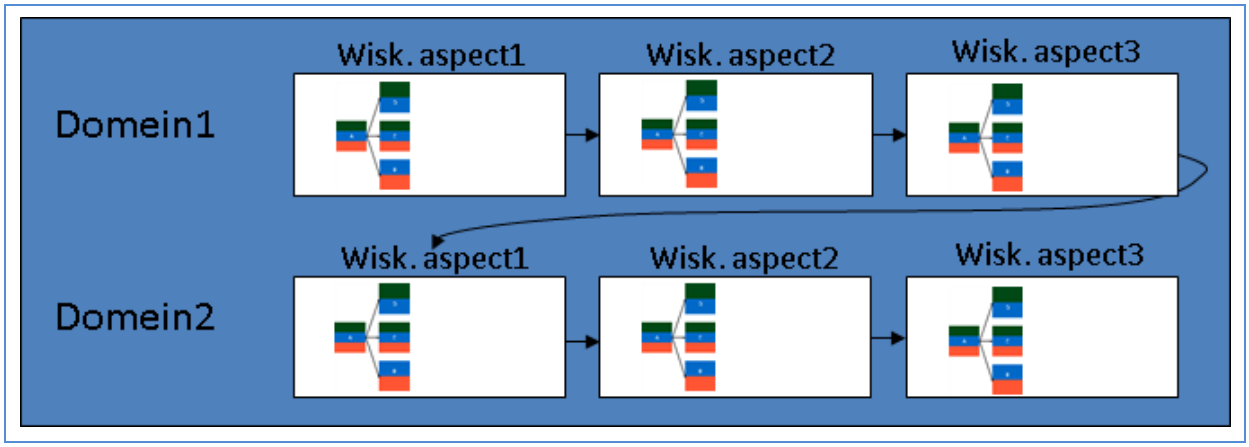
Blokje A wordt aan alle leerlingen aangeboden. Hierna wordt blokje B aangeboden aan de leerlingen die op grond van blokje A de grootste kans op “onder niveau” hadden, blokje C aan de leerlingen die op grond van blokje A de grootste kans op “op niveau” hadden en blokje D aan de leerlingen die op grond van blokje A de grootste kans op “boven niveau” hadden.

Het gehele toetsverloop – voor alle aspecten die gediagnosticeerd worden bij de eerste adaptieve afname– ziet er bij de talen uit zoals weergegeven in Figuur 4-3 en bij wiskunde zoals weergegeven in Figuur 4-5. De volgorde waarin de verschillende aspecten aan bod komen is bepaald door de vaststellingscommissie.

Figuur 4-4 Adaptieve toetsverloop in 2016 bij de talen



Figuur 4-5 Adaptieve toetsverloop in 2016 bij wiskunde



4.5.2 Methode

Na de kalibratie zijn de items behorend bij één attribuut aan vier toetsblokjes toegewezen. Met simulatieonderzoek wordt nagegaan wat de beste verdeling over blokjes is en wordt nagegaan hoe accuraat de diagnoses zijn.

Het streven is om de beschikbare QTI-items (schermen) zo over de vier blokjes te verdelen dat de accuratesse van de diagnoses zo hoog mogelijk is. Het basisprincipe is dat per hoofdattribuut bij de talen of wiskundig aspect binnen een domein de beschikbare QTI-items **1000 keer** over drie blokjes (A, B en D in Figuur 4-2) worden verdeeld. Hierbij wordt gezorgd dat 1/3^{e} van het aantal responsen ± 2 in blokje A en B komen en de rest van de conceptresponsen in blokje D. Het vierde blokje (C) bestaat uit de helft van de QTI-items van de twee andere blokjes in de tweede laag (B en D). Voor de eerste adaptieve afname

hebben we default gewerkt met een verschil 2 (aangezien de meeste QTI-items meer dan 1 conceptrespons opleveren). Indien de analyse “stukloopt” met deze grenswaarde dan wordt door degene die analyseert de grenswaarde opgehoogd.

Per oplossing wordt vervolgens gecontroleerd of aan de randvoorwaarden wordt voldaan:

- Bij de talen moeten de QTI-items per deelattribuut gelijk verdeeld zijn over de blokjes
- De som van de responstijden van de schermen in een route (deze tijden kunnen worden gevonden op de 20e regel van descrip.txt) moet kleiner zijn dan de maximaal beschikbare tijd voor een diagnose.

Indien de oplossing niet aan de randvoorwaarden voldoet, wordt de oplossing weggegooid.

Vervolgens wordt voor elk van die 1000 oplossingen data (antwoorden) gesimuleerd voor 10.000 leerlingen onder niveau, 10.000 leerlingen op niveau en 10.000 leerlingen boven niveau en wordt uitgerekend welke diagnose die leerlingen zouden krijgen. De oplossing met het hoogste percentage correcte classificaties worden aan de toetsdeskundigen aangeboden om inhoudelijk (QTI-item niveau) te controleren.

4.5.3 Resultaten

Schrijfvaardigheid Engels

De simulaties laten zien dat de beste blokindeling bij schrijfvaardigheid Engels tot accurate classificaties leiden. Hoe goed de oplossingen zijn kan gezien worden in Tabel 4-17. Bijvoorbeeld bij de diagnose "Afstemming op publiek en doel" in vmbo-bb zien we dat 90,2% van de onder-niveau leerlingen correct gediagnosticeerd worden, 89,2% van de op-niveau leerlingen en 89,2% van de boven-niveau leerlingen.

In doorsnee ligt bij schrijfvaardigheid Engels de proportie correcte classificaties rond de 0,85. Slechts in twee gevallen (vmbo-gt "Woordenschat en woordgebruik" en havo "Coherentie") ligt de proportie tussen de 0,75 en 0,80, bij alle andere diagnoses ligt deze boven de 0,80 (zie Tabel 4-17).

Tabel 4-17 Proporties correcte classificaties schrijfvaardigheid Engels, aantal QTI-items en responsen

Leerweg	Hoofdattribuut	Onder niveau	Op niveau	Boven niveau	Aantal QTI-items	Aantal itemres-ponsen
Vmbo-bb	1. Afstemming op publiek en doel	0,90	0,89	0,89	18	41
	2. Coherentie	0,88	0,88	0,88	23	29
	3. Woordenschat en woordgebruik	0,90	0,90	0,90	20	31
	4. Spelling, interpunctie en grammatica	0,85	0,84	0,85	17	18
Vmbo-kb	1. Afstemming op publiek en doel	0,84	0,85	0,85	19	36
	2. Coherentie	0,91	0,92	0,92	22	32
	3. Woordenschat en woordgebruik	0,87	0,88	0,88	21	32
	4. Spelling, interpunctie en grammatica	0,93	0,92	0,93	21	44
Vmbo-gt	1. Afstemming op publiek en doel	0,84	0,84	0,84	13	25
	2. Coherentie	0,87	0,87	0,87	24	33
	3. Woordenschat en woordgebruik	0,76	0,77	0,75	15	16
	4. Spelling, interpunctie en grammatica	0,86	0,86	0,87	22	29
Havo	1. Afstemming op publiek en doel	0,91	0,91	0,90	16	43
	2. Coherentie	0,80	0,80	0,80	12	18
	3. Woordenschat en woordgebruik	0,83	0,83	0,84	19	34
	4. Spelling, interpunctie en grammatica	0,86	0,86	0,87	23	34
Vwo	1. Afstemming op publiek en doel	0,85	0,82	0,85	6	16
	2. Coherentie	0,83	0,82	0,83	11	21
	3. Woordenschat en woordgebruik	0,85	0,85	0,85	19	47
	4. Spelling, interpunctie en grammatica	0,83	0,82	0,82	13	21

Schrijfvaardigheid Nederlands

De simulaties laten zien dat de beste blokindelingen bij schrijfvaardigheid Nederlands ook tot accurate classificaties leiden in havo/vwo en tot zeer accurate classificaties in vmbo (zie Tabel 4-18). De simulaties laten zien dat bij havo/vwo in doorsnee de proportie correcte classificaties rond de 0,85 ligt, net zoals bij Engels. In twee gevallen (havo “1. Doel en publiek” en “3. Woord- en zinsniveau”) ligt de proportie tussen de 0,75 en 0,80. Bij vmbo ligt de proportie correcte classificaties rond de 0,93. Slechts in één geval ligt de proportie net onder de 0,90, namelijk bij vmbo-gt “1. Doel en publiek”.

Tabel 4-18 Proporties correcte classificaties schrijfvaardigheid Nederlands, aantal QTI-items en responsen

Leerweg	Hoofdattribuut	Onder niveau	Op niveau	Boven niveau	Aantal QTI-items	Aantal itemresponsen
Vmbo-bb	1. Doel en publiek	0,92	0,91	0,92	15	74
	2. Structuur	0,95	0,94	0,95	16	49
	3. Woord- en zinsniveau	0,96	0,90	0,92	14	49
	4. Spelling en interpunctie	0,98	0,94	0,94	15	84
Vmbo-kb	1. Doel en publiek	0,92	0,91	0,92	15	69
	2. Structuur	0,91	0,89	0,91	19	51
	3. Woord- en zinsniveau	0,96	0,95	0,95	18	76
	4. Spelling en interpunctie	0,96	0,94	0,95	14	79
Vmbo-gt	1. Doel en publiek	0,93	0,85	0,88	13	60
	2. Structuur	0,92	0,88	0,92	17	46
	3. Woord- en zinsniveau	0,97	0,91	0,93	15	64
	4. Spelling en interpunctie	0,97	0,96	0,97	13	70
Havo	1. Doel en publiek	0,78	0,78	0,77	10	29
	2. Structuur	0,85	0,85	0,85	27	59
	3. Woord- en zinsniveau	0,77	0,77	0,79	16	65
	4. Spelling en interpunctie	0,93	0,92	0,92	14	89
Vwo	1. Doel en publiek	0,80	0,81	0,80	10	28
	2. Structuur	0,88	0,87	0,88	27	61
	3. Woord- en zinsniveau	0,85	0,85	0,85	17	59
	4. Spelling en interpunctie	0,97	0,96	0,97	15	92

Wiskunde

De beste blokindelingen bij wiskunde leiden tot redelijk accurate classificaties (zie Tabel 4-19). De accuratesse ligt lager dan bij de talen, vermoedelijk omdat ook het aantal responsen in veel gevallen aanzienlijk lager is. In doorsnee ligt de proportie correcte classificaties rond de 0,75. Bij havo ligt de accuratesse lager, omdat in één geval de accuratesse op de ondergrens van 0,60 zit (bij D-Meten en meetkunde: meerduidigheid). Ook bij vmbo-bb ligt de accuratesse bij één wiskundig aspect in domein D onder de 0,70, namelijk bij samenhang.

Tabel 4-19 Proporties correcte classificaties wiskunde, aantal QTI-items en responsen

Leerweg	Hoofdattribuut	Onder niveau	Op niveau	Boven niveau	Aantal QTI-items	Aantal itemresponsen
Vmbo-bb	D-Meten en meetkunde: structuur	0,79	0,78	0,80	16	20
	D-Meten en meetkunde: meerduidigheid	0,80	0,80	0,79	10	23
	D-Meten en meetkunde: samenhang	0,66	0,64	0,62	8	9
	E-Verbanden en formules: structuur	0,77	0,69	0,70	10	12
	E-Verbanden en formules: meerduidigheid	0,74	0,64	0,64	7	7
	E-Verbanden en formules: samenhang	0,73	0,73	0,73	9	12
Vmbo-kb	D-Meten en meetkunde: structuur	0,75	0,73	0,74	18	19
	D-Meten en meetkunde: meerduidigheid	0,74	0,73	0,75	11	22
	D-Meten en meetkunde: samenhang	0,79	0,79	0,79	10	14
	E-Verbanden en formules: structuur	0,83	0,84	0,83	16	23
	E-Verbanden en formules: meerduidigheid	0,76	0,73	0,73	8	9
	E-Verbanden en formules: samenhang	0,81	0,79	0,80	8	14
Vmbo-gt	D-Meten en meetkunde: structuur	0,78	0,79	0,80	13	13
	D-Meten en meetkunde: meerduidigheid	0,77	0,77	0,78	9	15
	D-Meten en meetkunde: samenhang	0,77	0,78	0,77	10	11
	E-Verbanden en formules: structuur	0,83	0,83	0,83	15	20
	E-Verbanden en formules: meerduidigheid	0,75	0,75	0,78	8	9
	E-Verbanden en formules: samenhang	0,80	0,80	0,80	9	16
Havo	D-Meten en meetkunde: structuur	0,77	0,76	0,77	8	12
	D-Meten en meetkunde: meerduidigheid	0,58	0,59	0,63	8	14
	D-Meten en meetkunde: samenhang	0,76	0,76	0,76	9	11
	E-Verbanden en formules: structuur	0,82	0,81	0,81	12	16
	E-Verbanden en formules: meerduidigheid	0,70	0,71	0,71	11	18
	E-Verbanden en formules: samenhang	0,71	0,69	0,72	5	11
Vwo	D-Meten en meetkunde: structuur	0,79	0,77	0,76	8	12
	D-Meten en meetkunde: meerduidigheid	0,73	0,72	0,72	8	14
	D-Meten en meetkunde: samenhang	0,84	0,81	0,82	12	14
	E-Verbanden en formules: structuur	0,80	0,81	0,81	14	17
	E-Verbanden en formules: meerduidigheid	0,76	0,73	0,74	9	9
	E-Verbanden en formules: samenhang	0,70	0,71	0,71	5	11

4.5.4 Conclusie

De accuratesse bij de talen is goed tot zeer goed en bij wiskunde redelijk. Bij een tweetal aspecten bij wiskunde zou de accuratesse bij voorkeur hoger liggen. Door in 2017 extra items te zaaien voor deze aspecten zou de accuratesse verhoogd kunnen worden.

Na de eerste adaptieve afname in 2016 zullen de items geherkalibreerd worden (Cito, 2015) en opnieuw in blokjes ingedeeld worden, ditmaal in drie lagen indien hiervoor voldoende items zijn. Herkalibratie is van belang, omdat de wijze van afnemen verandert (adaptief in plaats van lineair) en we op grond van een relatief kleine steekproef hebben gekalibreerd.

5 Referenties en eerdere verslagen

Bokhove, C., & Drijvers, P. (2010). Digital tools for algebra education: Criteria and evaluation. *International Journal of Computers for Mathematical Learning*, *15(1)*, 45-62.

Cito (2012). *Diagnostische tussentijdse toets: verslag van een voorstudie*.

Cito (2013*). Diagnostische tussentijdse toets: Onderzoek 2013*.

Cito (2013). *Diagnostische tussentijdse toets: Verslag try-out 2013*.

Cito (2013). *Diagnostische tussentijdse toets: Voorstudies en evaluatie 2013*.

Cito (2013). *Notitie verwachte toetsduur en afnamescenario's*.

Cito (2013). *De diagnostische tussentijdse toets: Tussenstand in ontwikkeling*. Arnhem: Cito.

Cito (2014). *Diagnostische tussentijdse toets: Onderzoek 2014*.

Cito (2014). *Diagnostische tussentijdse toets: Verslag try-out 2014*.

Cito (2015), *Notitie Herkalibratie DTT 2016*.

CvTE (2013). *Constructieopdracht itembanken DTT wiskunde en schrijfvaardigheid Nederlands en Engels (pretest 2015).*

CvTE (2014). *Toetswijzers diagnostische tussentijdse toets: Nederlands, Engels en Wiskunde (versie 1 april).*

CvTE (juni 2015). *Getekend besluit standaarden DTT*.

CvTE (mei 2015). *Evaluatie pilot DTT jaar 1 (2014-2015)*

Feskens, R., Keuning, J., Van Til, A & Verheyen, R. (2014). *Prestatiestandaarden voor ERK in het examenjaar: Een internationaal ijkingsonderzoek.* Arnhem: Cito. Dit rapport is NIET openbaar.

Fife, J. H. (2011). *Automated scoring of CBAL mathematics tasks with m-rater. Research Memorandum*. Princeton, NJ: ETS. Retrieved June, 11th, 2013, from http://www.ets.org/Media/Research/pdf/RM-11-12.pdf.

Grugeon-Allys, B., Pilet, J., Chenevotot-Quentin, F., & Delozanne, E. (2012). *Diagnostic et parcours différentiés d'enseignement en algèbre élémentaire*. Recherches en Didactique des Mathématiques.

Heeren, B., & Jeuring, J. (2014). *Feedback services for stepwise exercises*. Technical report UU-CS-2014-005. Utrecht: Department of Information and Computer Sciences, Utrecht University. Retrieved February, 24th, 2015, from http://www.cs.uu.nl/research/techreps/repo/CS-2014/2014-005.pdf.

Linthorst, R. & Keune, K. (2014). Hfst 7. Beoordeling open vragen schrijfvaardigheid Nederlands. In Schouwstra, S. (red.). *Diagnostische tussentijdse toets: Onderzoek 2014*, pp. 71-82. Arnhem: Cito

Sangwin, C. (2013). Computer aided assessment of mathematics. Oxford, U.K.: Oxford University Press.

SLO (april 2012). *Concept-tussendoelen kernvakken onderbouw vo.*

Stacey, K. & Wiliam, D. (2013). Technology and assessment in mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 721-751). New York/Berlin: Springer.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995*). OPLM: One-Parameter Logistic Model. Computer program and manual*. Arnhem: Cito.

6 Bijlagen

6.1 Bijlagen design

6.1.1 Initiële schattingen per vak voor het toetsdesign

Aantal responsen

- Bij schrijfvaardigheid Engels zijn er voor zowel vmbo als havo/vwo 4 hoofd- en 8 deelattributen[11]. Dit betekent dat voor leerlingen van één schooltype items nodig waren die minstens 8*12 = 96 responsen zouden opleveren (inclusief overlap).
- Bij schrijfvaardigheid Nederlands zijn er bij vmbo 4 hoofd- en 12 deelattributen en bij havo/vwo 4 hoofd- en 13 deelattributen. Dit betekent dat voor leerlingen van één vmbo schooltype items nodig waren die 12*12 = 144 responsen (inclusief overlap) zouden opleveren en voor leerlingen van één havo/vwo schooltype items die 13*12 = 156 responsen (inclusief overlap) zouden opleveren.
- Bij wiskunde zijn er 2 domeinen met ieder 3 wiskundige aspecten. Dit betekent dat voor leerlingen van één schooltype items nodig waren die 2 domeinen*3 aspecten*24 = 144 responsen (inclusief overlap) zouden opleveren.

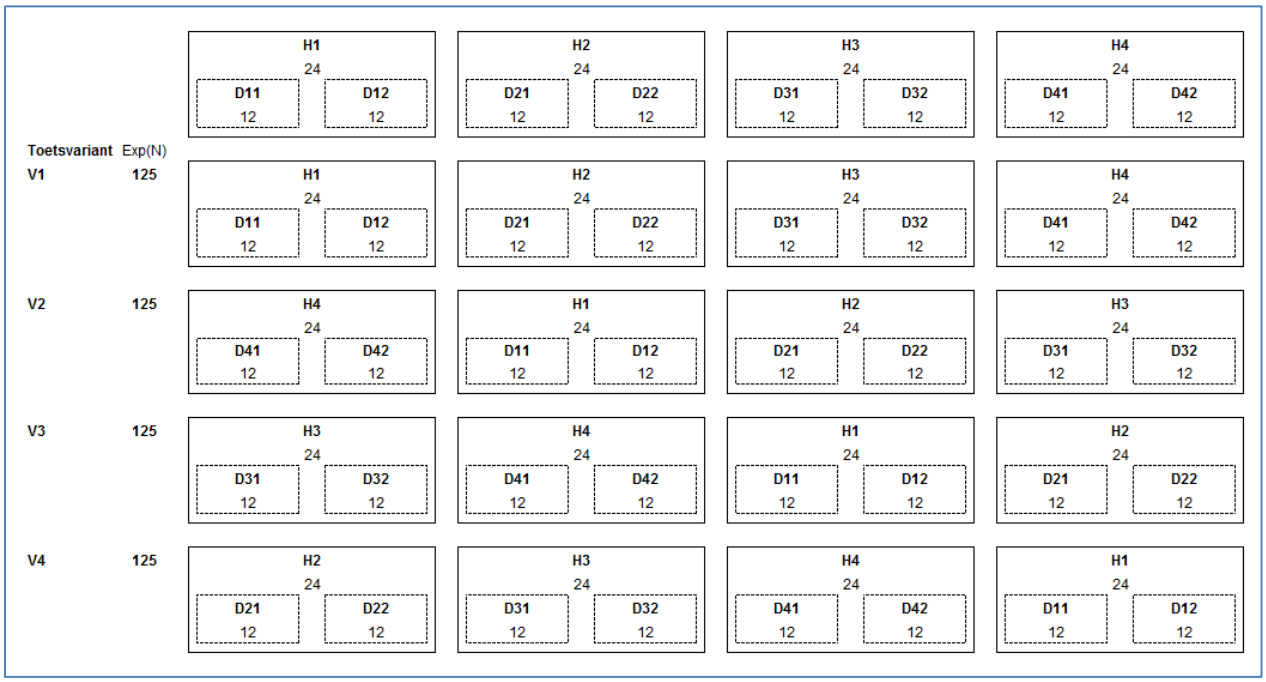
Benodigde antwoordtijd

- In de laatste try-out kostte bij schrijfvaardigheid Engels 12 schermen op vmbo gemiddeld 13,5 minuten (stdev = 4,9 min.) en op havo/vwo gemiddeld 17,8 minuten (stdev = 4,9 min.)
    - Indien minstens twee items op een scherm staan is voor het beantwoorden van de 96 items (48 schermen) op vmbo gemiddeld 54 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 93 minuten.
    - Indien minstens twee items op een scherm staan is voor het beantwoorden van de 96 items (48 schermen) op havo/vwo gemiddeld 71 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 110 minuten.
- In de laatste try-out kostte bij schrijfvaardigheid Nederlands 12 schermen op vmbo gemiddeld 14,5 minuten (stdev = 4,2 min.) en op havo/vwo gemiddeld 14,5 minuten (stdev = 6,7 min.)
    - Indien minstens twee items op een scherm staan is voor het beantwoorden van de 144 items (72 schermen) op vmbo gemiddeld 87 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 137 minuten (2 uur en een kwartier).
    - Indien minstens twee items op een scherm staan is voor het beantwoorden van de 156 items (78 schermen) op havo/vwo gemiddeld 94 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 181 minuten (3 klokuren).
- In de laatste try-out kostte bij wiskunde 12 schermen op vmbo gemiddeld 15,9 minuten (stdev = 5,3 min.) en op havo/vwo gemiddeld 28,5 minuten (stdev = 6,0 min.)
    - Indien minstens twee items op een scherm staan is gemiddeld voor het beantwoorden van 144 vmbo-items (72 schermen) 95 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 159 minuten (2 uur en 40 minuten).
    - Indien minstens twee items op een scherm staan is gemiddeld voor het beantwoorden van 144 havo/vwo-items (72 schermen) 171 minuten nodig. Voor iemand die twee standaarddeviaties boven het gemiddelde zit is de verwachte antwoordtijd 243 minuten (4 uur en 5 minuten).

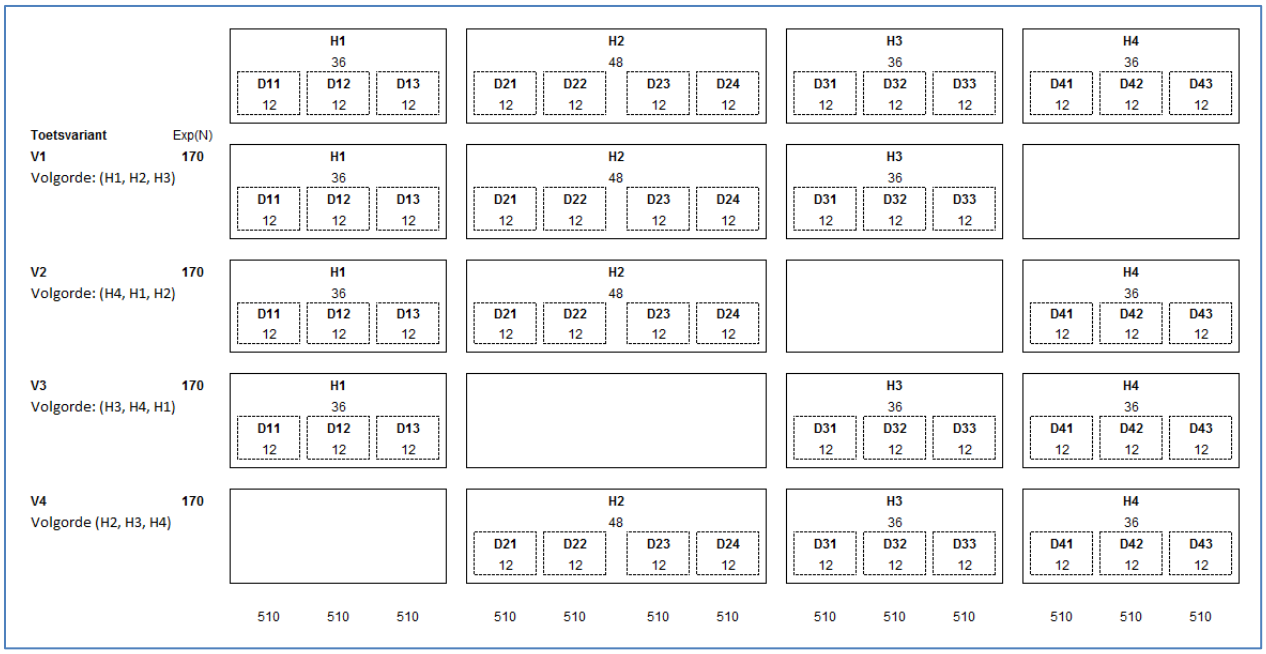
[11] Er is een deelattribuut dat uit twee subattributen bestaat. Hiervoor is ook geconstrueerd.

6.1.2 Globale designs per vak en leerweg

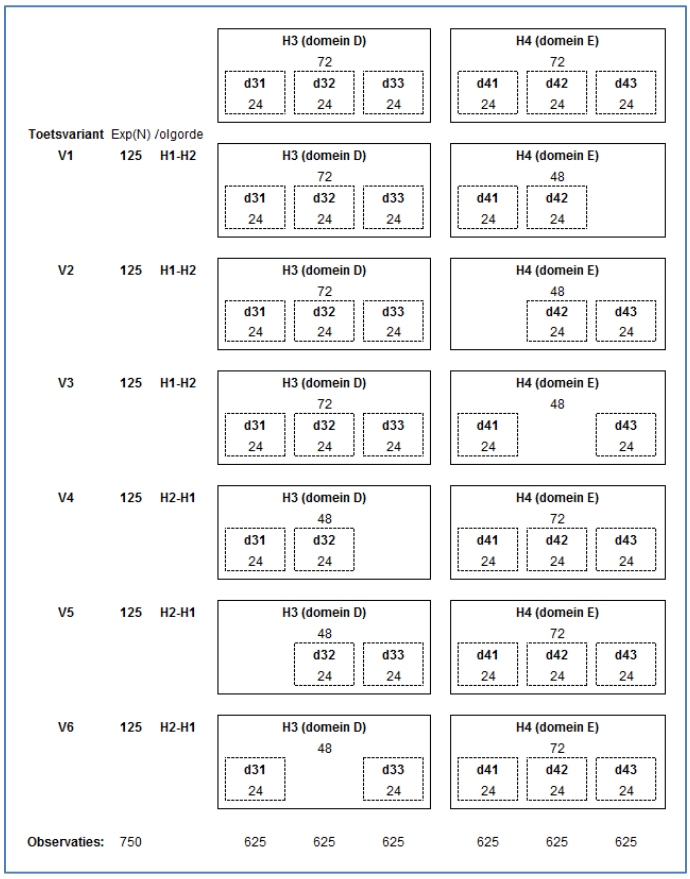
Figuur 7-1. Design schrijfvaardigheid Engels voor één schooltype (vmbo-bb, vmbo-kb, vmbo-gt, havo, vwo) en schrijfvaardigheid Nederlands vmbo (vmbo-bb, vmbo-kb, vmbo-gt)



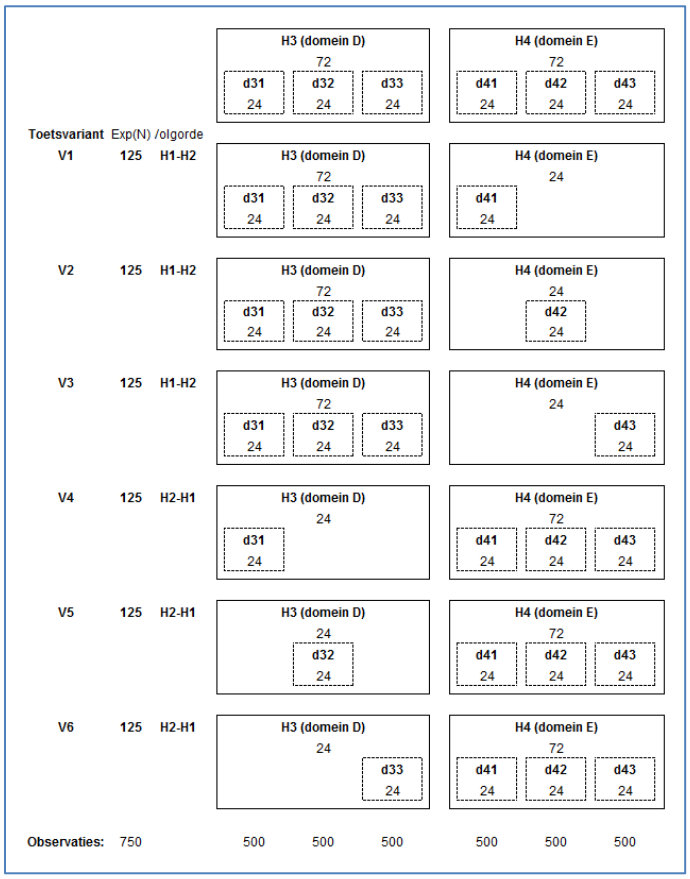
Figuur 7-2. Design schrijfvaardigheid Nederlands voor één schooltype havo/vwo



Figuur 7-3. Design wiskunde voor vmbo

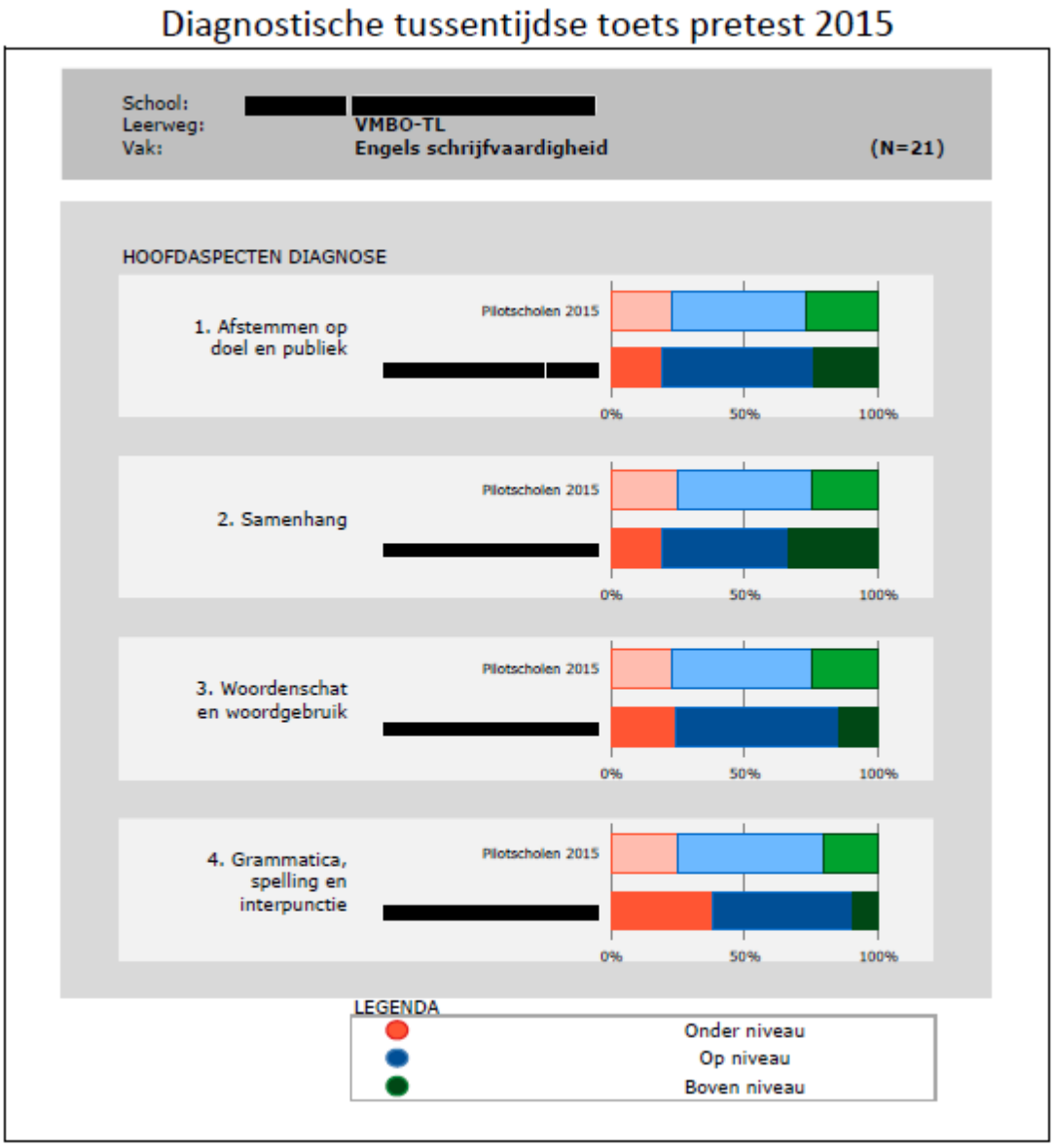


Figuur 7-4. Design wiskunde voor havo/vwo



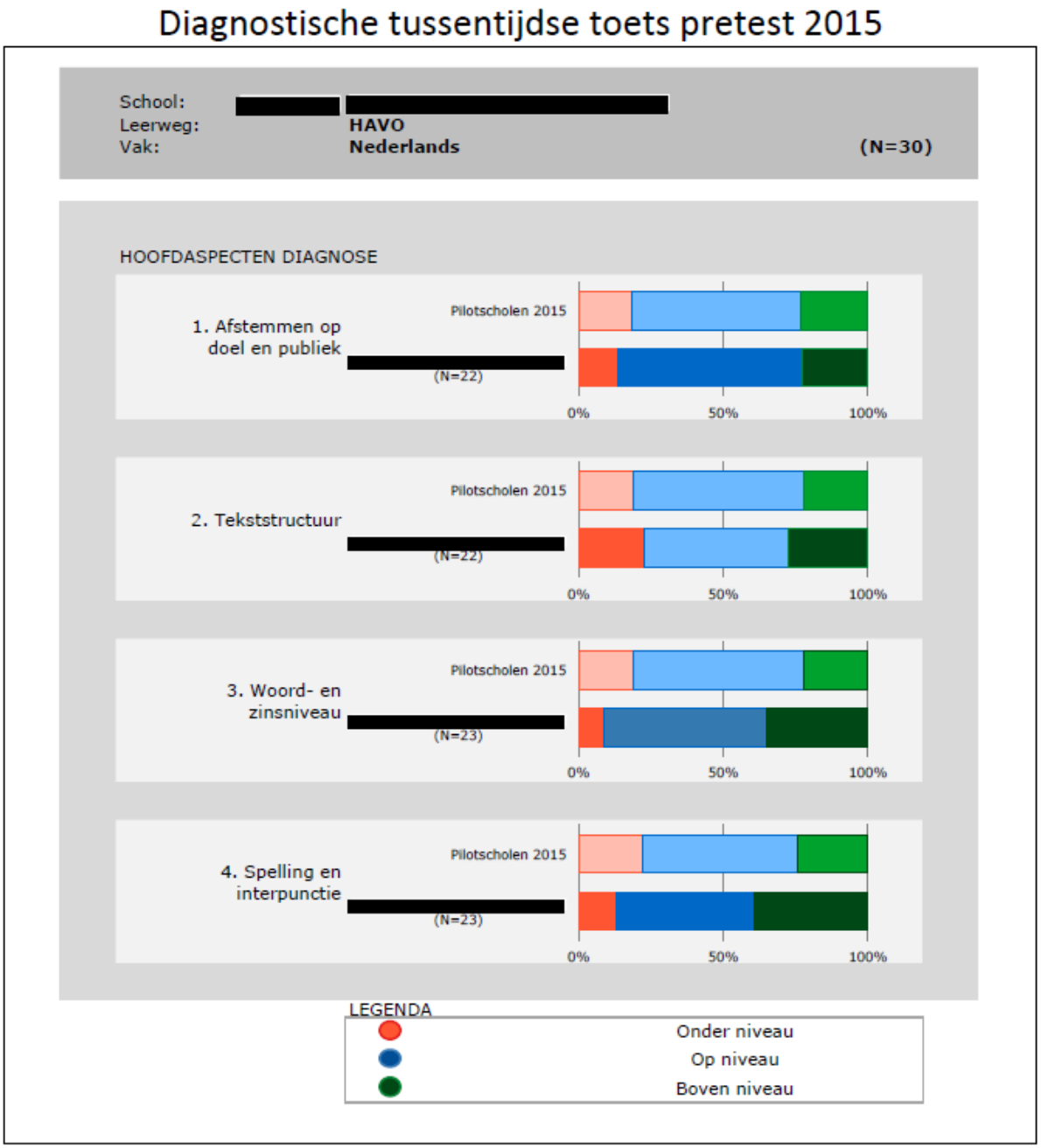
6.2 Bijlagen pretest schoolrapportages

Voorbeeld 7-1 Schoolrapportage schrijfvaardigheid Engels pretest 2015

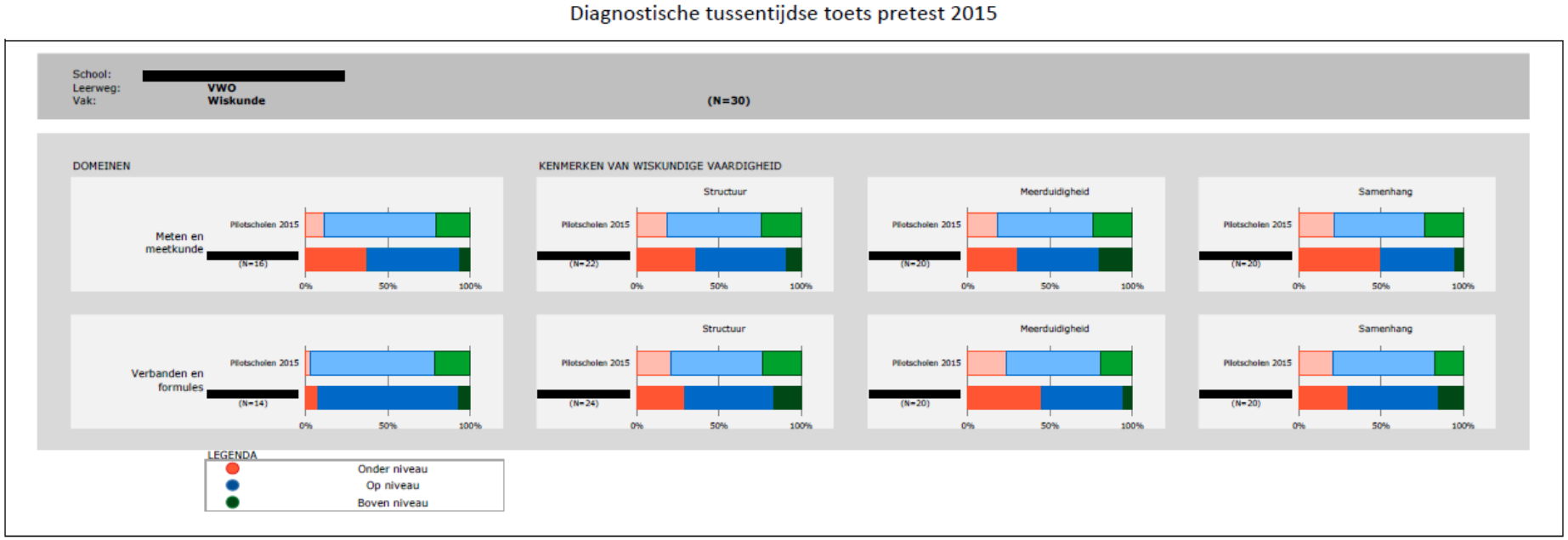


Vertrouwelijk Concept

Voorbeeld 7-2 Schoolrapportage schrijfvaardigheid Nederlands pretest 2015



Voorbeeld 7-3 Schoolrapportage wiskunde pretest 2015



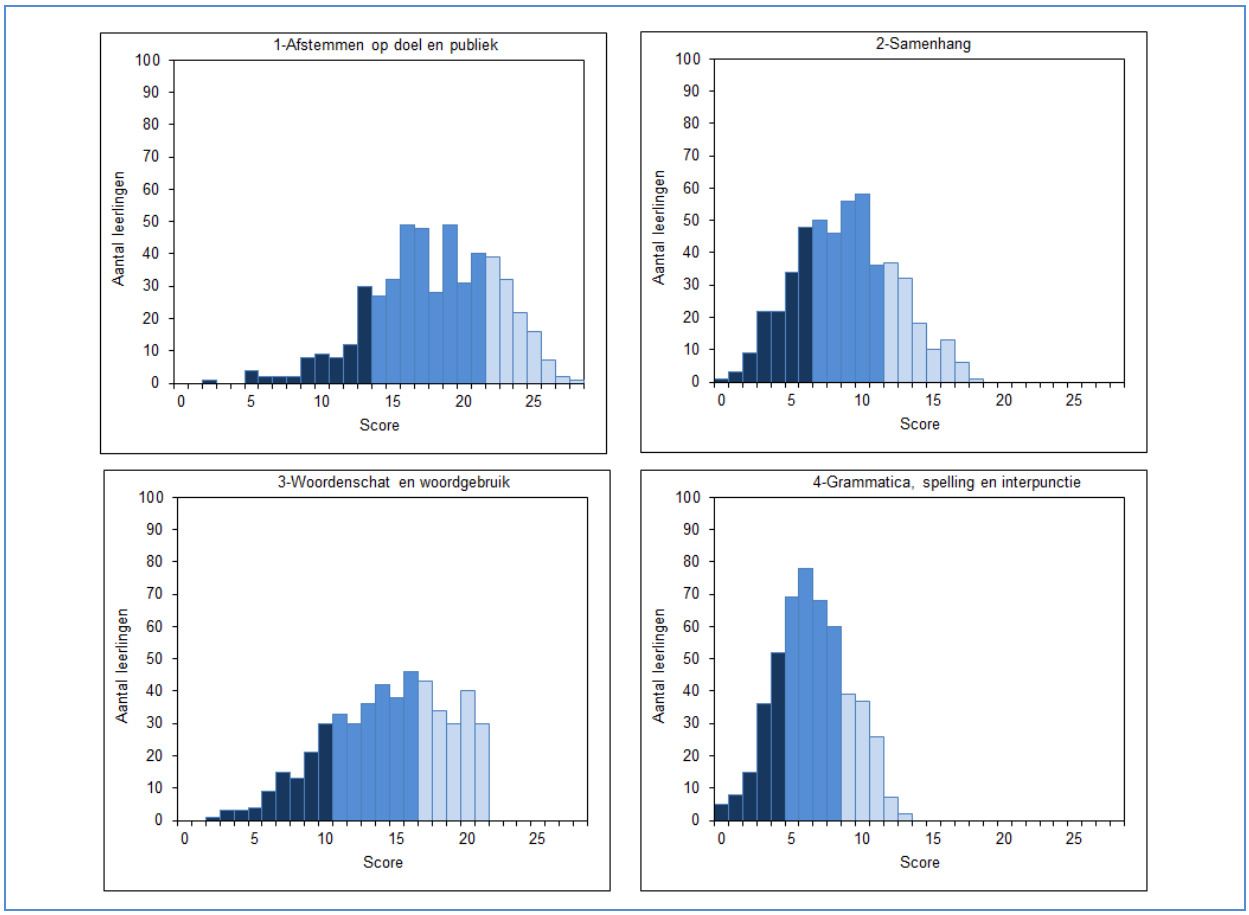
Vertrouwelijk concept

6.3 Bijlage standaardbepaling

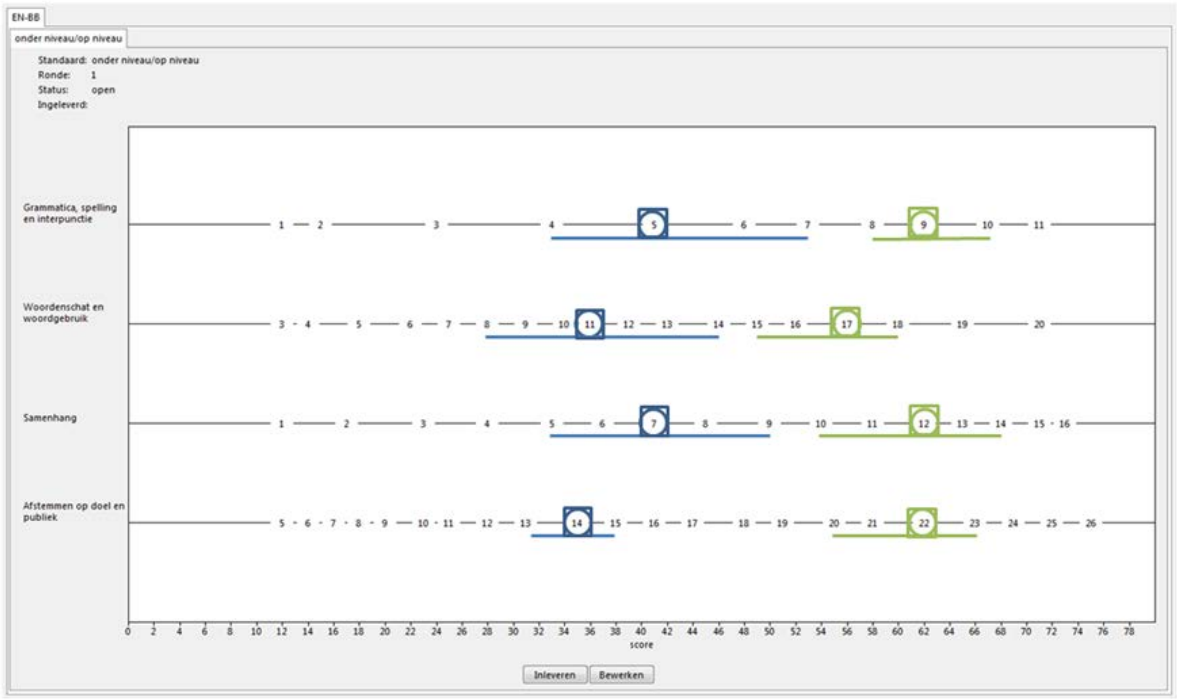
Voorbeeld 7-4. Rapportage maten

Uit standaardsetting (DTT variant van Data driven direct consensus-methode)				
Datum	11 juni 2015			
Vak	**Schrijfvaardigheid Engels**			
Leerweg	**vmbo-BB leerjaar 2**			
Aantal experts	8			
	1-Afstemmen op doel en publiek	2-Samenhang	3-Woordenschat en woordgebruik	4-Grammatica, spelling en interpunctie
Aantal responsen	28	18	21	13
Verdeeld over	12 items	12 items	12 items	12 items
	Cesuur onder-op	**Cesuur onder-op**	**Cesuur onder-op**	**Cesuur onder-op**
Gem. aantal correct responsen	14	7	11	5
% van maximum	0.50	0.39	0.52	0.38
Range	13-15	5-9	8-14	4-7
SD	0.64	1.49	1.91	1.25
	Cesuur op-boven	**Cesuur op-boven**	**Cesuur op-boven**	**Cesuur op-boven**
Gem. aantal correct responsen	22	12	17	9
% van maximum	0.79	0.67	0.81	0.69
Range	20-23	10-14	15-18	8-10
SD	1.19	1.51	1.41	0.64
Aantal pretestleerlingen onder-niveau	78	139	99	116
	16%	28%	20%	23%
Aantal pretestleerlingen op-niveau	304	246	225	275
	61%	49%	45%	55%
Aantal pretestleerlingen boven-niveau	119	117	177	111
	24%	23%	35%	22%

Voorbeeld 7-5.Grafische weergave verdeling



Voorbeeld 7-6. Formulier met gemiddeld oordeel en range



6.4 Bijlagen expert meeting

6.4.1 Programma Expert Meeting

Time slot	Activity	Location
Monday, February 16th: What's the problem?		
13:00 - 14:00	Welcome with lunch	Hotel Haarhuis, Arnhem
14:00 - 16:00	Opening session Chair: Martin van Reeuwijk (CvTE) - Opening by Jacob Raap (CvTE) - Introduction by Paul Drijvers (Cito / UU): Setting the scene - Plenary 1: James Fife (ETS): Automated Scoring of Mathematics Items - Interactive session	Hotel Haarhuis, Arnhem
16:30 - 18:00	- Plenary 2: Brigitte Grugeon (Paris): Diagnostic digital assessment using « Pepite » software: test, cognitive profiles and learning courses - Interactive session	Hotel Haarhuis, Arnhem
Tuesday, February 17th: In-depth solutions		
09:00 - 11:00	Dutch Design: - Plenary 3: Peter Boon (UU): Automated scoring within the DME-project, challenges and choices - Plenary 4: Johan Jeuring and Bastiaan Heeren (UU / OU): Assessing derivations of mathematical exercises	Cito, Meeting rooms 1-2, Arnhem
11:00 - 12:30	- Plenary 5: Chris Sangwin (Loughborough): Objective automatic assessment of mathematics: establishing mathematical properties, partial credit and feedback	Cito, Meeting rooms 1-2, Arnhem
12:30 - 13:30	Lunch	Watermuseum
13:30 - 15:30	Round table: Technical aspects of automated assignment of partial credit	Cito, Meeting rooms 1-2 Arnhem
16:00 - 18:00	Round table: Technical aspects of automated scoring of graphs & geometry items	Cito, Arnhem
Wednesday, February 18th: Synthesis and actions		
09:00 - 10:30	Wrap-up with plenary speakers and organizers	Hotel Haarhuis, Arnhem
11:00 - 12:30	Plenary closing session Chair: Martin van Reeuwijk - state of the art? - next steps? - what, who and how?	Hotel Haarhuis, Arnhem
12:30 - 14:00	Closing lunch	Hotel Haarhuis, Arnhem

Setting the Scene

Dr. Paul Drijvers, Cito

In this short introductory lecture I will set the scene for this expert meeting in two respects. First, I will sketch the state-of-the-art of the work done in the Netherlands by CvTE and Cito on the automated scoring of mathematics. Second, I will briefly highlight the main challenges and dilemmas that we encounter in this field to set this meeting's agenda.

Paul Drijvers is senior assessment expert at CITO and professor in mathematics education at Utrecht University´s Freudenthal Institute.

Plenary 1: Automated Scoring of Mathematics Items

Dr. James Fife, ETS, USA

In this talk I will discuss the automated scoring of mathematics items for a variety of response types. When the response is an equation or mathematical expression, a computer algebra system can determine if the student's response is mathematically equivalent to the correct response. Such systems rarely make errors, but sometimes they assign a score that is different from what a human reader would assign. For items whose response is a graph or geometric figure, the student response can be captured in a graph editor that facilitates automated scoring. For both equation and graph responses, the student's response can be scored based on the student's responses to previous items. Additionally, partial credit can be assigned based on features of the student's response. Alternatively, there is some evidence to indicate that machine-learning techniques may be used to score freely-drawn graphs. In addition, machine-learning techniques have been used successfully to score mathematics items with short text responses. Current research involves the scoring of longer text responses (responses to such prompts as "Explain your answer.") and the scoring of text responses with embedded equations.

Dr. James Fife is a researcher at ETS, USA, and one of the developers of CBAL and m-rater.

Plenary 2: Diagnostic digital assessment using « Pepite » software: test, cognitive profiles and learning courses

Prof. dr. Brigitte Grugeon, Université Paris Est Créteil, Paris, France.

What automatic analysis of the student responses to a diagnostic digital assessment is possible in order to organize learning courses adapted to the learning needs of students? For this purpose, we will present the automatic diagnostic assessment model on a mathematical domain - elementary algebra – for students aged 15-16 (Grugeon,1997). The aim is to build the student profile on a specific mathematical domain. We have based the diagnosis on an epistemological and didactical study of algebraic field: aspects of algebraic skills constitute a reference to locate the knowledge and skills built by students in algebra.

We will point out three points which characterize "Pépite" diagnostic digital assessment:

- The diagnostic test covers the characteristic types of problems of the algebraic field, open-ended questions or QCM.
- Automatic analysis of responses describes the success or failure to every exercise, but also, the use of letters, the type of calculation rules, the type of translation, the type of justification.
- Cross-sectional analysis of students' responses aims to automatically characterize the profile of the students on three dimensions (mobilizing algebraic tool, computing capacity and flexibility through the semiotic registers).

Prof. dr. Brigitte Grugeon is a researcher at the Université Paris Est Créteil and one of the designers of the Pépite diagnostic test environment.

Plenary 3: Automated scoring within the DME-project, challenges and choices.

Peter Boon, Freudenthal Institute, Utrecht University

For automated scoring of mathematical tasks, especially more complex tasks, usually a price has to be paid in two ways. First, the technology often is advanced and complex and therefore expensive. Second, mathematical tasks need to be transformed in order to make automated scoring doable, so we need to do concessions and to agree on a suboptimal version of the task. In our DME (Digital Mathematics Environment) we use a hybrid approach. On the one hand, we firmly take up the challenge to make automated scoring possible within a broad range of mathematical activities. On the other hand, we want to prevent paying the second type of price described above. We don't accept the need for automated scoring changing the task in undesirable ways.

In this presentation I will present a versatile approach of automated scoring within a range of subjects, I will share the problems that we are facing in the design of these features and I hope to discuss with you the choices that we are making in this process.

Peter Boon is a researcher at the Freudenthal Institute, Utrecht University and the principal designer of the Digital Mathematics Environment.

Plenary 4: Assessing derivations of mathematical exercises

Dr. Bastiaan Heeren, Open University, Heerlen, and/or prof. dr. Johan Jeuring Utrecht University / Open University

We show how we can automatically assess not only the final answer to an algebra exercise, such as solving a quadratic equation, or solving a linear inequation, but also the process leading to the answer. To assess an exercise of a particular class, we use a strategy for solving that class of exercises. Furthermore, we show how we can use the same strategy to build up a student model that contains information about the student's competencies in the domain of the exercise.

Dr. Bastiaan Heeren and prof. dr. Johan Jeuring are researchers and co-designers of the IDEAS feedback engine.

Plenary 5: Objective automatic assessment of mathematics: establishing mathematical properties, partial credit and feedback.

Dr. Chris Sangwin, Loughborough University, U.K.

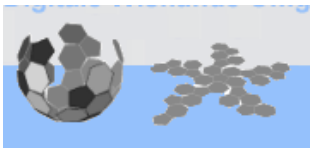
In this talk I will discuss computer aided assessment (CAA) of mathematics, with a focus on automatically establishing the properties of answers which have mathematical content. Based on these objective properties partial credit, feedback and cohort statistics can be generated. In particular, I will illustrate this talk using the system STACK with examples from high-school algebra, and introductory calculus. I will briefly compare this with other available technology. We shall demonstrate how such questions can be automatically assessed, and report use with existing students.

Dr. Chris Sangwin is a Senior Lecturer in Mathematics Education in the Mathematics Education Centre at Loughborough University in the United Kingdom and the designer of the STACK system.

6.4.2 Overzicht relevante software

Behalve Facet en Questify Builder is tijdens de expert meeting de volgende software aan de orde gekomen.

Digitale Wiskunde Omgeving
http://www.fisme.science.uu.nl/dwo/



De Digitale Wiskunde Omgeving (DWO) is een omgeving waarin activiteiten niet alleen digitaal worden aangeboden, maar waarin de verschillende activiteiten ook digitaal kunnen worden uitgevoerd. Tevens kunnen de resultaten worden vastgelegd. Er zijn standaardmodules (oefeningen, lessen), maar de docent kan ook zelf materiaal aanmaken binnen de DWO.

De DWO richt zich vooral op het voortgezet onderwijs (havo/vwo en vmbo), maar er zijn zeker ook mogelijkheden voor basisonderwijs en mbo.

ETS Graphing Editor
http://www.ets.org/research/topics/as_nlp/math/

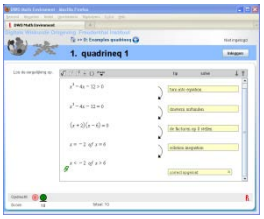
Automated Scoring of Math Responses

ETS's *m-rater* scoring engine is used for scoring open-ended mathematical responses, such as those which take the form of mathematical expressions, equations or graphs. Dating from the late 1990s, the *m-rater* scoring engine is one of the first ETS capabilities for automated scoring to be developed. The scores generated by the *m-rater* engine demonstrate very strong agreement with human ratings.

Geogebra
http://www.geogebra.org

Geogebra is a multi-platform mathematics software that gives everyone the chance to experience the extraordinary insights that math makes possible

IDEAS: INTERACTIVE DOMAIN REASONERS
http://ideas.cs.uu.nl/www/



Domain reasoners help students solving interactive exercises in learning environments, such as ActiveMath, MathDox, and the Digital Mathematical Environment from the Freudenthal Institute.

Ideas is a framework for developing domain reasoners that give intelligent feedback. Using Ideas, we have developed domain reasoners for solving linear, quadratic and higher-degree equations, gaussian elimination, and many other domains.

Maxima, a Computer Algebra System
http://maxima.sourceforge.net

Maxima is a system for the manipulation of symbolic and numerical expressions, including differentiation, integration, Taylor series, Laplace transforms, ordinary differential equations, systems of linear equations, polynomials, sets, lists, vectors, matrices and tensors. Maxima yields high precision numerical results by using exact fractions, arbitrary-precision integers and variable-precision floating-point numbers. Maxima can plot functions and data in two and three dimensions.

The Maxima source code can be compiled on many systems, including Windows, Linux, and MacOS X. The source code for all systems and precompiled binaries for Windows and Linux are available at the SourceForge file manager.

Maxima is a descendant of Macsyma, the legendary computer algebra system developed in the late 1960s at the Massachusetts Institute of Technology. It is the only system based on that effort still publicly available and with an active user community, thanks to its open source nature. Macsyma was revolutionary in its day, and many later systems, such as Maple and Mathematica, were inspired by it.

STACK
http://stack.bham.ac.uk/moodle/

STACK provides a question type for the Moodle quiz which is specifically designed to enable sophisticated computer-aided assessment in Mathematics and related disciplines, with emphasis on formative assessment.