Diagnostische tussentijdse toets

Onderzoek (2014)

zeker weten

Diagnostische tussentijdse toets

Onderzoek (2014)

Cito | Arnhem

Auteurs: Hfst. 2: Herbert Hoijtink en Aniek Sies
Hfst. 3: Herbert Hoijtink en Aniek Sies
Hfst. 4: Herbert Hoijtink en Sanneke Schouwstra
Hfst. 5: Karen Keunen en Sanneke Schouwstra
Hfst. 6: Paul Drijvers, Saskia Visser, Hendrik Straat, Sanneke Schouwstra
Hfst. 7: Roelien Linthorst, Karen Keune
Hfst. 8: Rick Godschalk, Wilma Vrijs, Sanneke Schouwstra

Eindredactie: Sanneke Schouwstra

Managementsamenvatting

De ontwikkeling van de diagnostisch tussentijdse toets (DTT) is nauw omgeven door onderzoeks-activiteiten, omdat de DTT een innovatief product is. In 2014 zijn tegelijk met de try-outs en grootschalige itemontwikkeling de onderzoeksactiviteiten voor de uitwerking van het diagnostische toetsontwerp voortgezet door Cito. In dit onderzoeksverslag worden de uitkomsten gepresenteerd van het vervolg-onderzoek dat gedurende de tweede try-outperiode uitgevoerd is. Het psychometrische onderzoek heeft zich gericht op de verdere onderbouwing van de adaptieve samenstelling van de diagnostische tussentijdse toets (DTT). Daarnaast is ook onderzoek verricht voor de inzet van automatische beoordeling binnen de DTT zowel op het terrein van wiskunde als op het terrein van schrijfvaardigheid.

Psychometrisch onderzoek

Een reconstructie van de succeskansen in de eerste try-out

Vorig jaar (2012-2013) is in simulatieonderzoek gebruik gemaakt van hypothetische itemeigenschappen (succeskansen). In het vervolgonderzoek dit jaar zijn reële data gebruikt, namelijk de data uit de eerste try-out. De try-outdata zijn gebruikt om items te detecteren die zeer gunstige eigenschappen hebben en items die minder gunstige eigenschappen hebben voor het diagnosticeren. Het onderzoek wijst uit dat het voor schrijfvaardigheid Engels, wiskunde en schrijfvaardigheid Nederlands mogelijk is items te construeren die zeer goed onderscheid maken tussen leerlingen die onder, op en boven niveau zitten. Aan de andere kant zijn er voor elk van de vakken ook items te vinden die geen goed onderscheid maken tussen leerlingen die onder, op en boven niveau zitten. Deze items kunnen als voorbeeld gebruikt worden door de itemconstructeurs bij de verdere itemontwikkeling.

De try-outdata zijn ook gebruikt in nieuw simulatieonderzoek. Het simulatieonderzoek diende een indruk te geven hoe goed de diagnoses zijn (onder, op of boven niveau) wanneer gebruik wordt gemaakt van empirisch bepaalde succeskansen in plaats van hypothetische succeskansen. Het simulatie-onderzoek laat zien dat wanneer we alleen zeer goed onderscheidende items zouden hebben, het aantal items dat nodig is om leerlingen met een kans van .80 correct te diagnosticeren, beduidend lager ligt dan verwacht op grond van de eerste geheel hypothetische simulatiestudie.

Adaptieve procedure met itemblokken

Er is ook nader onderzoek verricht naar de mate van adaptiviteit. Alhoewel bij maximale adaptiviteit de minste items nodig zijn voor een accurate diagnose, kleven er wel een aantal nadelen aan maximale adaptiviteit. Maximale adaptiviteit is vaak niet mogelijk omdat sommige opgaven in blokken georganiseerd zijn (zoals meerdere opgaven over dezelfde tekst). Belangrijke nadelen van maximale adaptiviteit zijn verder dat herkalibratie na de pretest onmogelijk is en dat het lastig is (zeker real-time) om met alle inhoudelijke randvoorwaarden die aan de samenstelling gesteld kunnen worden, rekening te houden.
In simulatieonderzoek is nagegaan wat de gevolgen zijn voor de accuraatheid van de diagnoses als de adaptieve module gericht is op kleine itemblokken (bijvoorbeeld 6 items) in plaats van losse items.

Uit het onderzoek blijkt dat de adaptieve procedure met blokken goed werkt wanneer de spreiding van de succeskansen als leidraad wordt genomen voor de samenstelling van de blokken en niet de moeilijkheid van de items. Met deze procedure worden, vergeleken met de niet-adaptieve procedure uit de voorstudie, meer leerlingen correct geclassificeerd. Het informatieverlies ten opzichte van de adaptieve procedure uit de voorstudie is minimaal. De adaptieve blokkenprocedure is dus een goed alternatief voor de adaptieve procedure uit de voorstudie

Interpretatie van het initiële profiel op deelattribuutniveau

Voor de ontwikkeling van de adaptieve module dient ook nader bepaald en onderzocht te worden hoe op grond van het onderkennende diagnostische profiel een indicatie verkregen kan worden van nader te diagnosticeren deelattributen. Op grond van het onderzoek is een algoritme ontwikkeld worden voor het doortoetsen van geïndiceerde deelattributen. In het algoritme is nadere toetsing van een deelattribuut

geïndiceerd als uit het initiële profiel blijkt dat er een aanzienlijke kans is dat de beheersing van de leerling relatief verbetering behoeft (relatieve verbeterpunten).

In het algortime ligt dus de nadruk op de “onder niveau” deelattributen, omdat verwacht wordt dat de aanpak van deze verbeterpunten tot hogere prestaties zal leiden. Het kan echter wenselijk zijn om een gelijke focus te hebben op verbeterpunten en sterkte punten in het doortoetsen. Verder is in het algoritme is geen rekening gehouden met het feit dat het profiel al dusdanig zeker kan zijn dat doortoetsen niet nodig is. In het vervolgonderzoek zal daarom gewerkt worden aan een alternatief algoritme. Vervolgens zullen de algoritmes met gesimuleerde data geëvalueerd worden, zodat een overwogen keuze gemaakt kan worden.

Standaardbepaling door experts na de eerste pretest van de DTT

Om de items na de pretest te kunnen kalibreren zijn standaarden nodig. Om de grenzen tussen de beheersingscategorieën (onder, op, of boven niveau) goed te kunnen plaatsen, zal er een toets-georiënteerde standaardbepaling door inhoudelijk experts plaatsvinden. Deze grenzen zullen voor alle vakken (Nederlands, Engels en Wiskunde), vaardigheden en alle leerwegen (vmbo-bb, vmbo-kb, vmbo-gt, havo, vwo) apart bepaald worden. Dit betekent dat er voor 15 toetsen een standaardbepaling zal plaatsvinden waarin voor elk van de hoofdattributen van deze toets twee standaarden worden bepaald, namelijk de grens tussen onder en op niveau en de grens tussen op en boven niveau. Om de pretest in omvang te beperken (en het benodigd aantal leerlingen) worden de items niet in één grote pretest afgenomen, maar in twee kleinere pretesten. Deze stapsgewijze opbouw betekent dat ook de standaardbepaling in twee stappen opgebouwd moet worden.

De standaardbepalingsmethode die hiervoor gebruikt zal worden is de 3DC-methode, ofwel de *Data-Driven Direct Consensus* methode. 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. Om de validiteit en betrouwbaarheid van de standaardbepaling te verhogen, hoeven de experts bij de 3DC-methode hun inschatting van de grens tussen onder en op niveau en de grens tussen op en boven niveau niet alleen op inhoudelijke descriptoren te baseren, maar kunnen zij ook zien hoe de clusters zich empirisch tot elkaar verhouden. Op grond van de 3DC-methode is de werkwijze, aanpak en globale planning van de standaardbepaling voor de eerste pilotafname van de DTT uitgewerkt.

Onderzoek voor automatische beoordeling

Digitale diagnostische toetsing van wiskunde: onderzoek van een nieuwe omgeving

Wiskunde stelt een aantal vakspecifieke eisen van hoge complexiteit aan een digitale omgeving voor het construeren, afnemen en beoordelen van adaptieve toetsen. Vorig jaar is literatuuronderzoek verricht om de complexiteit van digitale toetsen voor rekenen en wiskunde in kaart te brengen en op literatuurstudie gefundeerde uitgangspunten te formuleren. De geformuleerde uitgangspunten en keuzes leidden tot een aantal vereisten (requirements) voor de verdere ontwikkeling van de afname-omgeving Facet en de auteursomgeving QuestifyBuilder.

De mogelijkheden binnen Facet om formules, grafieken, meet kundige figuren en tussenstappen te registreren en automatisch te coderen zijn momenteel nog gering. Daarom zijn de mogelijkheden onderzocht van de Digitale Wiskunde Omgeving (DWO, Freudenthal Instituut) die minder wiskundige beperkingen oplegt dan Facet en QuestifyBuilder. Het doel was om (1) na te gaan uit welk type items een diagnostische toets kan bestaan als er meer digitale mogelijkheden zijn en (2) te onderzoeken of afname van deze items leidt tot een adequate en efficiënte toetsing van het leerlingmodel.

Daartoe zijn in de DWO 42 items ontworpen die in “boekjes” zijn voorgelegd aan 578 leerlingen uit 3 havo/vwo. Uit de data-analyse blijkt dat de toets een lage betrouwbaarheid kende en te lang was voor de beschikbare toetstijd. Op basis van de resultaten kan desondanks worden geconcludeerd dat een omgeving zoals de Digitale Wiskunde Omgeving mogelijkheden biedt voor gevarieerde, rijke en productieve items die in principe een meer genuanceerde diagnose mogelijk maken. Hierbij spelen de mogelijkheden van automatische beoordeling een grote rol.

Uit het onderzoek vloeit als belangrijkste aanbeveling voort de – al dan niet native – integratie van wiskundige mogelijkheden in de toetsontwikkel- en toetsafname-omgeving, waarbij met name het invoeren en beoordelen van formules, grafieken en meetkundige constructies prioriteit verdienen.

Beoordeling open vragen Nederlands: onderzoek naar de betrouwbaarheid en bruikbaarheid van drie beoordelingsmodellen

Op dit moment zijn er binnen Facet ook nog geen mogelijkheden om langere tekst automatisch te beoordelen. Eventueel verder inhoudelijk onderzoek naar de bruikbaarheid en kwaliteit (validiteit, betrouwbaarheid en transparantie) van automatische beoordeling voor de DTT kan pas verricht worden als een component gekozen of ontwikkeld is. In 2014 zijn derhalve wederom op kleine schaal enkele open schrijfopdrachten Nederlands en Engels voorgelegd aan leerlingen. Het doel van deze open schrijf-opdrachten was tweeledig. Allereerst levert de afname extra teksten en docentoordelen op die gebruikt kunnen worden bij het ontwikkelen van een component voor automatische beoordeling. Daarnaast is de verwachting dat menselijke beoordelaars nodig zullen blijven. Om scholen tegemoet te komen bij de beoordeling van de leerlingteksten in het open deel van de toets, zijn de leerlingteksten die nu verzameld zijn, gebruikt voor een onderzoek naar drie verschillende beoordelingsmodellen die door docenten gehanteerd zouden kunnen worden.

De interbeoordelaarsovereenstemming lijkt voor alle modellen redelijk goed te zijn. Bij het globale model en het analytische model zijn geen attributen die opvallend lager of hoger dan de andere attributen beoordeeld werden, bij het ankermodel echter wel. Het ankermodel heeft ook zeker niet de eerste voorkeur van de beoordelaars. Het globale model was volgens de beoordelaars het efficiëntst, maar het analytische model sprak de beoordelaars het meest aan. Volgens de beoordelaars kunnen ze door de vele beoordelings-punten in het analytische model het meest recht doen aan de leerlingteksten. Bij alle modellen gaven de beoordelaars overigens verbeterpunten aan. Voordat de modellen uitgezet zouden worden in het veld is het raadzaam te onderzoeken wat de effecten van deze aanpassingen zullen zijn.

Beoordeling open vragen schrijfvaardigheid Engels

Er is ook onderzoek verricht naar de beoordeling van de open schrijftaken van Engels. Bij Engels lag de focus op twee beoordelingsschema's die kunnen worden gebruikt om de globale kwaliteit en de communicatieve effectiviteit te beoordelen. De toetswijzercommissie heeft geadviseerd om communicatieve effectiviteit op te nemen binnen het hoofdattribuut 'afstemming op doel en publiek'. Communicatieve effectiviteit is echter niet opgenomen in het leerlingmodel, omdat dit lastig toetsbaar is. Aangezien dit wel een belangrijk aspect is, is communicatieve effectiviteit de basis geworden van het holistisch, globaal beoordelingsmodel voor open schrijftaken dat onderzocht is.

De resultaten van het onderzoek wijzen uit dat het model nog niet in alle opzichten erg geschikt is voor het beoordelen van open schrijfopdrachten. Alhoewel de niveau-inschatting van de schrijfproducten erg betrouwbaar was, bleken de beoordelaars niet in staat een helder onderscheid te maken tussen de communicatieve effectiviteit en globale kwaliteit van de schrijfproducten. Het verdient de aanbeveling om verder onderzoek te doen naar de beoordelingsmodellen. Bij vervolgonderzoek zouden ervaren docenten ingezet moeten worden om de bruikbaarheid en validiteit van de modellen te onderzoeken.

Inhoudsopgave

Managementsamenvatting 3
Psychometrisch onderzoek 3
Onderzoek voor automatische beoordeling 4
1 Inleiding 9
1.1 Referenties 10
2 Een reconstructie van de succeskansen in de eerste try-out 11
2.1 Inleiding 11
2.2 Procedure try-out succeskansen 12
2.3 Resultaten try-out succeskansen 14
2.4 Eén simulatie met op de try-out gebaseerde succeskansen 20
2.5 Opzet simulatie 20
2.6 Resultaten simulatie 20
2.7 Conclusie 21
2.8 Referenties 22
2.9 Bijlagen 23
3 Adaptieve procedure met itemblokken 33
3.1 Inleiding 33
3.2 Adaptieve procedure met blokken items samengesteld op grond van itemmoeilijkheid 33
3.3 Adaptieve procedure met blokken items samengesteld op grond van spreiding in succeskansen . 36
3.4 Conclusie 38
3.5 Referenties 38
3.6 Bijlagen 39
4 Interpretatie van het initiële profiel op deelattribuutniveau 41
4.1 Inleiding 41
4.2 Methode voor het afnemen van de diagnostische tussentijdse toets 41
4.3 Algoritme voor de selectie van deelattributen ter verdere evaluatie 44
4.4 Conclusie 45
5 Standaardbepaling door experts na de eerste pretest van de DTT 47
5.1 Standaardbepalingsmethode 47
5.2 Werkwijze voor de eerste DTT standaardbepaling 48
5.3 Aanpak tijdens standaardbepalingssessie 48
5.4 Globale planning 52
5.5 Referenties 52
6 Digitale diagnostische toetsing van wiskunde: onderzoek van een nieuwe omgeving 53
6.1 Probleemstelling 53
6.2 Methode 53
6.3 Resultaten 55

6.4 Conclusies en aanbevelingen 60
6.5 Referenties 62
6.6 Bijlagen 62
7 Beoordeling open vragen schrijfvaardigheid Nederlands 71
7.1 Inleiding 71
7.2 De beoordelingsmodellen 72
7.3 Methode 73
7.4 Resultaten 74
7.5 Conclusie en discussie 80
7.6 Referenties 81
7.7 Bijlagen 82
8 Beoordeling open vragen schrijfvaardigheid Engels 83
8.1 Inleiding 83
8.2 Methode en procedure 83
8.3 Resultaten 85
8.4 Conclusie 86
8.5 Referenties 87
8.6 Bijlagen 87

1 Inleiding

De ontwikkeling van de DTT is nauw omgeven door onderzoeksactiviteiten, omdat de DTT een innovatief product is. Tegelijk met de try-out en grootschalige itemontwikkeling zijn de psychometrische, vakinhoudelijke en onderwijskundige onderzoeksactiviteiten voor de uitwerking van het diagnostische toetsontwerp voorgezet. In dit onderzoeksverslag worden de uitkomsten gepresenteerd van het vervolg-onderzoek dat gedurende de tweede try-outperiode (2013-2014) uitgevoerd is.

Het psychometrische onderzoek heeft zich gericht op de verdere onderbouwing van een adaptieve samenstelling van de diagnostische tussentijdse toets (DTT). Vorig jaar is in het verrichtte simulatie-onderzoek gebruik gemaakt van hypothetische itemeigenschappen (succeskansen).

In het vervolgonderzoek beschreven in hoofdstuk 2 worden reële data gebruikt, namelijk de data uit de eerste try-out (Cito, 2013). De try-outdata zijn gebruikt in verder psychometrisch onderzoek om items te detecteren die zeer gunstige eigenschappen hebben en items die minder gunstige eigenschappen hebben. Deze items kunnen als voorbeeld gebruikt worden door de itemconstructeurs bij de verdere item-ontwikkeling. De try-outdata zijn ook gebruikt in nieuw simulatieonderzoek. Het simulatieonderzoek diende een indruk te geven hoe goed de diagnoses zijn (onder, op of boven niveau) wanneer gebruik wordt gemaakt van empirisch bepaalde succeskansen in plaats van hypothetische succeskansen.

Er is ook nader onderzoek verricht naar de mate van adaptiviteit, beschreven in hoofdstuk 0. Alhoewel bij maximale adaptiviteit de minste items nodig zijn voor een accurate diagnose, kleven er wel een aantal nadelen aan maximale adaptiviteit. Maximale adaptiviteit is vaak niet mogelijk omdat sommige opgaven in blokken georganiseerd zijn (zoals meerdere opgaven over dezelfde tekst). Belangrijke nadelen van maximale adaptiviteit zijn verder dat herkalibratie na de pretest onmogelijk is en dat het lastig is (zeker real-time) om met alle inhoudelijke randvoorwaarden die aan de samenstelling gesteld kunnen worden rekening te houden. In simulatieonderzoek is nagegaan wat de gevolgen zijn voor de accuraatheid van de diagnoses als de adaptieve module gericht is op kleine itemblokken (bijvoorbeeld 6 items) in plaats van losse items.

Voor de ontwikkeling van de adaptieve module dient ook nader bepaald en onderzocht te worden hoe op grond van het onderkennende diagnostische profiel een indicatie verkregen kan worden van nader te diagnosticeren deelattributen. Op grond van het onderzoek is een algoritme ontwikkeld worden voor het doortoetsen van geïndiceerde deelattributen. Het algoritme wordt beschreven in hoofdstuk 0.

Om de items na de pretest te kunnen kalibreren zijn standaarden nodig. In hoofdstuk 5 is de standaardbepalingsmethode voor de eerste afname van de diagnostisch tussentijdse toets uitgewerkt. Om de grenzen tussen de beheersingscategorieën (onder, op, of boven niveau) goed te kunnen plaatsen, zal er een toetsgeoriënteerde standaardbepaling door inhoudelijk experts plaatsvinden.

Naast het psychometrische onderzoek is ook verder onderzoek verricht voor de inzet van automatische beoordeling binnen de DTT zowel op het terrein van wiskunde als op het terrein van schrijfvaardigheid. Het is belangrijk, voor onder meer de (indruks-) validiteit, het draagvlak in het veld voor de DTT en de diagnostische waarde, dat de DTT niet uitsluitend bestaat uit gesloten vragen, maar ook opdrachten omvat die langere open antwoorden vereisen. De mogelijkheid om deze langere open antwoorden automatisch te laten beoordelen is van grote toegevoegde waarde. Geautomatiseerde beoordeling kan de docenten ontlasten, de adaptiviteit van de toets ondersteunen en een snelle rapportage van de toetsresultaten mogelijk maken.

De mogelijkheden binnen Facet om formules, grafieken, meet kundige figuren en tussenstappen te registreren en automatisch te coderen zijn momenteel nog gering. Derhalve kan met behulp van Questify/Facet in 2014 nog geen verder inhoudelijk onderzoek naar automatische beoordeling verricht worden. Aan de hand van een alternatieve digitale omgeving voor het ontwerpen en afspelen van digitale wiskundetoetsen is nagegaan welk type items bruikbaar zijn voor een diagnostische toets als er meer digitale mogelijkheden binnen Facet zijn en is de bruikbaarheid van deze items nagegaan voor diagnoses zoals beoogd in de toetswijzer (het leerlingmodel).

Op dit moment zijn er binnen Facet ook nog geen mogelijkheden om langere tekst automatisch te beoordelen. Eventueel verder inhoudelijk onderzoek naar de bruikbaarheid en kwaliteit (validiteit, betrouwbaarheid en transparantie) van automatische beoordeling voor de DTT kan pas verricht worden als een component gekozen of ontwikkeld is. In 2014 zijn derhalve wederom op kleine schaal enkele open schrijfopdrachten Nederlands en Engels voorgelegd aan leerlingen. Het doel van deze open schrijf-opdrachten was tweeledig. Allereerst levert de afname extra teksten en docentoordelen op die gebruikt kunnen worden bij het ontwikkelen van een component voor automatische beoordeling. Daarnaast is de verwachting dat menselijke beoordelaars nodig zullen blijven. Om scholen tegemoet te komen bij de beoordeling van de leerlingteksten in het open deel van de toets, zijn de leerlingteksten die nu verzameld zijn gebruikt voor een onderzoek, beschreven in hoofdstuk 7, naar verschillende beoordelingsmodellen die door docenten gehanteerd zouden kunnen worden.

Er is ook onderzoek verricht naar de beoordeling van de open schrijftaken van Engels, beschreven in hoofdstuk 8. Bij Engels lag de focus op twee beoordelingsschema's die kunnen worden gebruikt om de globale kwaliteit en de communicatieve effectiviteit te beoordelen. De toetswijzercommissie Engels heeft geadviseerd om communicatieve effectiviteit op te nemen binnen het hoofdattribuut 'Afstemming op doel en publiek'. Communicatieve effectiviteit is echter niet opgenomen in het leerlingmodel, omdat dit lastig toetsbaar is. Aangezien dit wel een belangrijk aspect is, is communicatieve effectiviteit de basis geworden van het holistisch, globaal beoordelingsmodel voor open schrijftaken dat onderzocht is.

1.1 Referenties

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

2 Een reconstructie van de succeskansen in de eerste try-out

2.1 Inleiding

In 2013 is onderzocht hoeveel (sub-)items[1] nodig zouden zijn om met betrekking tot een hoofdattribuut 80% van de leerlingen correct aan de beheersingscategorieën onder, op en boven niveau toe te kunnen wijzen (Cito, 2013a, hfst. 2). In de simulatiestudies is uitgegaan van de hypothetische succeskansen (zoals weergegeven in Tabellen 1-1, 1-2 en 1-3). In de simulatiestudie is nagegaan hoeveel items nodig zouden zijn wanneer het verschil (de spreiding) in succeskans tussen de drie beheersingscategorieën klein (.10), medium (.15) of groot (.20) is.

Bijvoorbeeld, in de simulatie met een grote spreiding in succeskansen (zie Tabel 2-1) had een “onder niveau” leerling hypothetisch .5 kans om een makkelijk item correct te beantwoorden, een “op niveau” leerling .7 en een leerling die boven niveau zat .9 kans. In de simulatie met een kleine spreiding in succeskansen (zie Tabel 2-3) daarentegen had een leerling die onder niveau zat hypothetisch .75 kans om een makkelijk item correct te beantwoorden, een leerling die op niveau zat .85 en een leerling die boven niveau zat .95.

De conclusies waren ruwweg dat als een toets lineair afgenomen wordt en bestaat uit items met een grote spreiding er 36 items nodig zijn om 80% van de leerlingen correct te diagnosticeren. De corresponderende aantallen voor items met een medium spreiding van .15 dan wel een kleine spreiding van .10, waren 48 en “meer dan 48”. In een adaptieve afname kon worden volstaan met 18 items met een grote spreiding en ongeveer 42 items met een kleine spreiding in de succeskansen.

Hoe reëel de hypothetische succeskansen en dus de conclusies met betrekking tot het aantal benodigde items waren, kon destijds nog niet beantwoord worden omdat de items van de diagnostische tussentijdse toets nog niet afgenomen waren. Echter, met behulp van de gegevens resulterend uit de eerste try-out kan nu een eerste indruk worden verkregen (Cito, 2013b).

Allereerst zal onderzocht worden hoe de hypothetische succeskansen zich verhouden tot de succeskansen bepaald op grond van de data resulterend uit de try-out. In de volgende paragraaf zal de gebruikte procedure uiteen worden gezet. Daarna zullen de resultaten uitgesplitst naar schrijfvaardigheid Engels, wiskunde en schrijfvaardigheid Nederlands worden gepresenteerd. Vervolgens worden de items met een grote en kleine spreiding in de succeskansen gepresenteerd. Dit zou de itemschrijvers een eerste indruk moeten geven over wat wel en wat niet werkt.

Daarna wordt er één simulatie met op de try-out gebaseerde succeskansen gepresenteerd om een, wederom eerste, indruk te krijgen hoe goed de diagnose van leerlingen in de categorieën onder, op en boven niveau werkt met empirisch bepaalde in plaats van hypothetische succeskansen. Dit hoofdstuk wordt afgesloten met een paragraaf waarin de conclusies worden samengevat.

Tabel 2-1 Hypothetische succeskansen met grote spreiding

	Beheersingscategorie		
Itemmoeilijkheid	**< onder niveau**	**= op niveau**	**> boven niveau**
Makkelijk	.500	.700	.900
Medium	.400	.600	.800
Moeilijk	.300	.500	.700

1 (sub)-items: vragen dan wel onderdelen van vragen waarop een leerling een score kan behalen

Tabel 2-2 Hypothetische succeskansen met medium spreiding

Itemmoeilijkheid	Beheersingscategorie < onder niveau	= op niveau	> boven niveau
Makkelijk	.600	.750	.900
Medium	.525	.675	.825
Moeilijk	.450	.600	.750

Tabel 2-3 Hypothetische succeskansen met kleine spreiding

Itemmoeilijkheid	Beheersingscategorie < onder niveau	= op niveau	> boven niveau
Makkelijk	.750	.850	.950
Medium	.650	.750	.850
Moeilijk	.550	.650	.750

2.2 Procedure try-out succeskansen

2.2.1 Datapreparatie

In de try-out hebben leerlingen groepen items behorend bij elk van de hoofdattributen van schrijfvaardigheid Engels, wiskunde en schrijfvaardigheid Nederlands beantwoord. In de try-out zijn niet alle items voldoende vaak beantwoord om betrouwbaar de succeskansen te kunnen berekenen. Ook niet alle leerlingen hebben voldoende items beantwoord om tot een betrouwbare evaluatie van de leerling te kunnen komen. In dit onderzoek beperken we ons derhalve tot die attributen waarvoor er voldoende items zijn die minimaal 100 keer beantwoord zijn (stap 1) én er voldoende personen zijn die minimaal 8 items beantwoord hebben (stap 2). Het resultaat is een datamatrix met J kolommen corresponderend met *J* items en *N* rijen corresponderend met *N* personen zoals geschetst in Figuur 2-1. Een 0 betekent dat het antwoord op een item fout is, een 1 betekent dat het antwoord goed is en een *X* betekent dat het antwoord ontbreekt.

Voor de kalibratieprocedure is het noodzakelijk dat we een complete set antwoorden van *N* leerlingen op *J* items hebben. In de derde stap worden daarom de eventueel ontbrekende waarnemingen vervangen door een schatting van het antwoord dat de leerling zou hebben gegeven als hij het item wel had beantwoord (stap 3). Tweewegimputatie (Van Ginkel, Van der Ark, Sijtsma en Vermunt, 2007) is een eenvoudige manier waarmee dit kan worden gerealiseerd. Zoals in Figuur 2-1 kan worden gezien beantwoord leerling *i* 5 van de 8 items correct ($p_{.i}$ = 5/8). Dit is geen heel goede maar ook geen heel slechte uitkomst. Item *j* wordt door 3 van de 15 leerlingen correct beantwoord ($p_{j.}$ = 3/15) dat betekent dat het een moeilijk item is. De kans dat een willekeurige leerling een willekeurig item goed beantwoord is .6 ($p_{..}$ = .6), dat wil zeggen dat de toets overall een medium moeilijkheid heeft. Deze drie kansen worden gecombineerd tot *P*+, de kans dat de leerling het betreffende item goed beantwoord zou hebben:

$P+ = p_{j.} + p_{.i} - p_{..} = 3/15 + 5/8 - 6/10 = .225$.

Als *P*+ groter of gelijk is aan .5 wordt de ontbrekende waarneming vervangen door een goed antwoord dat wil zeggen de itemscore 1. Als *P*+ kleiner is dan .5 wordt de itemscore 0 toegekend. In het voorbeeld wordt de ontbrekende waarneming dan ook vervangen door de itemscore 0.

Nadat de ontbrekende waarnemingen zijn geïmputeerd, worden de items in twee groepen ingedeeld: de items met een *p*-waarde tussen .10 en .90; en items met een *p*-waarde kleiner dan .10 of groter dan .90 (stap 4). Alleen antwoorden gegeven op de items met een *p*-waarde tussen de .10 en .90 zullen worden gebruikt om de leerlingen in de groepen onder, op en boven niveau in te delen.

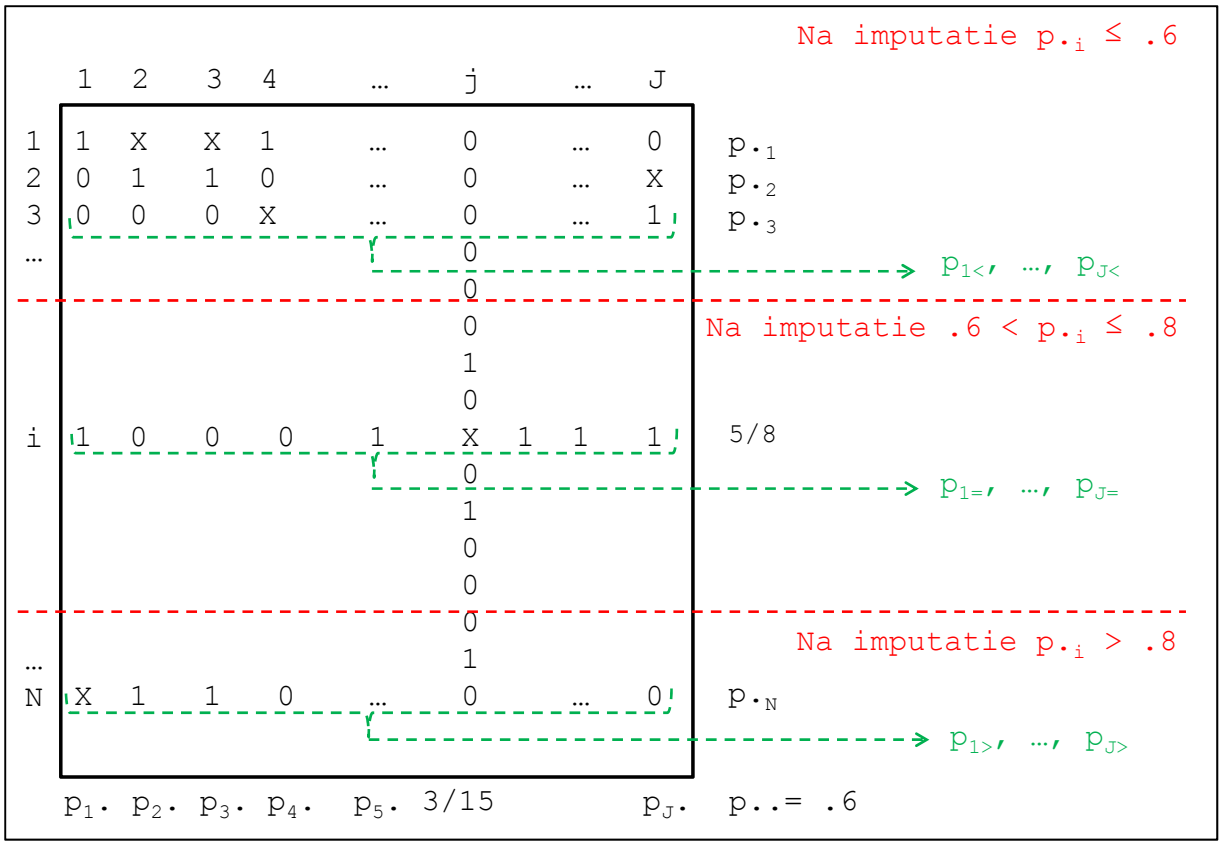
De eerste twee stappen (stap 1 en stap 2) zorgen ervoor dat er voor elke overblijvende leerling voldoende informatie beschikbaar is om relatief betrouwbaar de ontbrekende antwoorden te kunnen imputeren en de leerling in te kunnen delen in de categorieën onder, op en boven niveau; en dat er voor elk item voldoende informatie beschikbaar is om de succeskansen voor leerlingen, die in de onder, op, dan wel boven niveau categorie zitten, te kunnen schatten. In stap 3 worden de ontbrekende data weggewerkt. De gebruikte

benadering neemt aan dat de geobserveerde antwoorden voldoende informatie over de leerlingen en items bevatten om een goede inschatting van de vaardigheid van de leerlingen en de moeilijkheid van de items te kunnen maken. Het is noodzakelijk om de ontbrekende waarnemingen weg te werken, omdat de kalibratie-procedure die uiteen gezet wordt in de volgende paragraaf gebaseerd is op een complete set antwoorden van *N* leerlingen op *J* items. De vierde stap in de datavoorbereiding (stap 4) zorgt ervoor dat alleen de items die passen bij het vaardigheidsniveau van de leerlingen van de betreffende leerweg gebruikt worden in de kalibratie. Items die te makkelijk (*p*-waarde kleiner dan .10) of te moeilijk (*p*-waarde groter dan .90) zijn krijgen geen rol in de kalibratie, maar er zal wel over gerapporteerd worden.

2.2.2 Kalibratie

Voor de analyse van de try-outdata zal de kalibratie fase uit twee stappen bestaan. Ten eerste wordt er een ad-hoc standaardbepaling uitgevoerd. In Figuur 2-1 is met behulp van de rode lijnen aangegeven dat leerlingen die 60% of minder van de items met een *p*-waarde tussen de .10 en de .90 goed beantwoorden aan de groep “onder niveau” worden toegewezen. De motivatie voor deze keuze is dat deze leerlingen een onvoldoende halen op een toets bestaande uit items die zijn toegesneden op hun vaardigheidsniveau. Leerlingen die vanaf 60% tot en met 80% van deze items goed beantwoorden worden toegewezen aan de groep “op niveau” en leerlingen die meer dan 80% van deze items goed beantwoorden worden toegewezen aan de groep “boven niveau”. Omdat deze standaardbepaling ad hoc is, zullen ook resultaten met toedeling van leerlingen aan de groepen onder, op en boven niveau op basis van grenswaarden van 40% en 60% en 50% en 70% gerapporteerd worden. Echter, in de bespreking hiervan zal de nadruk liggen op de resultaten verkregen met de grenswaarden 60% en 80%.

Vervolgens worden voor alle items, ook de items waarvoor de *p*-waarde kleiner is dan .10 of groter dan .90, de kansen berekend dat het item goed beantwoord wordt door leerlingen die het attribuut onder niveau ($p_{i<}$), beheersen, leerlingen die het attribuut op niveau ($p_{i=}$) beheersen en door leerlingen die het attribuut boven niveau ($p_{i>}$) beheersen. Dit is in Figuur 2-1 in de kleur groen weergegeven. In het navolgende zullen de resulterende succeskansen separaat worden weergeven voor de items met een *p*-waarde tussen .10 en .90 en de items met een *p*-waarde buiten deze range.



Figuur 2-1. Een illustratie van tweeweg imputatie en de daaropvolgende kalibratie

2.3 Resultaten try-out succeskansen

Voor elk van de onderdelen schrijfvaardigheid Engels, Wiskunde en schrijfvaardigheid Nederlands zal de combinatie van leerweg en hoofdattribuut waarvoor het grootste aantal itemantwoorden verkregen is in dit verslag besproken worden. Evaluaties van andere combinaties lijken niet zinvol omdat het aantal itemantwoorden, dan wel het aantal personen, dan wel het aantal items zo klein wordt dat de analyses geen betrouwbare resultaten op zouden leveren.

2.3.1 Engels – vmbo gt – hoofdattribuut 3: Woordenschat en Woordgebruik

Bij vmbo-gt leerlingen zijn 51 items afgenomen waarvan er 22 door minimaal 100 leerlingen zijn beantwoord (zie Tabel 2-4, vierde kolom). Van deze 22 items hebben 15 items een *p*-waarde tussen .1 en .9 en 7 een *p*-waarde buiten dit interval. Na de standaardbepaling met grenswaarden 60% en 80% vielen respectievelijk 90, 105 en 45 leerlingen in de groepen onder, op en boven niveau.

In de vierde kolom van Tabel 2-5 wordt een overzicht gegeven van de geobserveerde spreiding in de succeskansen. Wat opvalt is dat het verschil in succeskansen tussen leerlingen onder niveau en leerlingen op niveau ($p_{i<}$ - $p_{i=}$) voor 5 van de 18 items met .1 < *p* < .9 (veel) groter is dan de grootste spreiding (.20, zie Tabel 2-1) gebruikt in de simulatiestudie gerapporteerd in het Cito rapport (Cito, 2013b). Hetzelfde geldt voor het verschil in succeskansen tussen de leerlingen op niveau en de leerlingen boven niveau ($p_{i=}$ - $p_{i>}$).

Gegeven de gebruikte methode van standaardbepaling (de wijze waarop leerlingen aan de groepen onder, op en boven niveau worden toegedeeld), kan dan ook worden geconcludeerd dat het mogelijk is items te construeren die goed tot zeer goed onderscheid kunnen maken tussen leerlingen die onder niveau zitten en leerlingen die op niveau zitten en om items te construeren die goed tot zeer goed onderscheid kunnen maken tussen leerlingen die op niveau zitten en leerlingen die boven niveau zitten. Echter, men zal ook moeten proberen te vermijden dat er items worden geschreven met slechts een kleine spreiding in de succeskansen. In de vierde kolom van Tabel 2-5 kan worden gezien dat dit eenmaal voorkomt voor met

betrekking tot het onderscheid tussen onder en op niveau en viermaal met betrekking tot het onderscheid tussen op en boven niveau. Deze conclusies blijven in essentie onveranderd als er naar de tweede en derde kolom van Tabel 2-4 en Tabel 2-5 wordt gekeken waar de standaardbepaling met andere grenzen is uitgevoerd.

Aangezien zeven van de items die door minimaal 80 leerlingen zijn beantwoord een *p*-waarde groter dan .9 hebben, kan ook worden geconcludeerd dat ongeveer een derde van de items te makkelijk is voor de groep vmbo-gt leerlingen. Er dient bij het construeren van de items dan ook een betere afstemming op de doelgroep plaats te vinden. Zoals in Tabel 2-5 kan worden gezien, zat er bij deze 7 items slechts 1 met een spreiding van .20 of groter voor $p_{i=}$ - $p_{i>}$.

Voor alle 22 items die door 80 of meer leerlingen zijn beantwoord, zijn de succeskansen bepaald (zie Bijlage 2-1). Door deze te bestuderen kan een goede indruk worden gekregen welke items wel en welke items niet een goed onderscheid maken tussen de drie beheersingscategorieën (onder, op en boven niveau). De resulterende indrukken kunnen meegenomen worden bij het construeren van nieuwe items. Merk op dat deze resultaten deels afhangen van de in de standaardbepaling gebruikte grenzen van 60% en 80%. Om die reden zijn ook de succeskansen bepaald nadat de standaard is gezet met andere grenzen. Deze paragraaf wordt afgesloten met enkele voorbeelden van goed en slecht onderscheidende items.

Voorbeelden van items die zeer goed onderscheiden tussen onder en op niveau
- De items t13.054_6 en t13.054_7 zijn onderdeel van een tekst waarin met behulp van drop-down menus woorden moeten worden ingevuld;
- Item t13.028_2: plaats het lichaamsdeel “chest” correct in een figuur van een lichaam.

Voorbeelden van items die goed tot zeer goed onderscheiden tussen op en boven niveau
- Item t13.019_1: “Are these jobs or not” eerste van vijf items “a bargain”;
- Item t13.055_5 is onderdeel van een tekst waarin met behulp van dropdown menus woorden moeten worden ingevuld;
- Items t13.028_2 en t13.028_3 waarin de lichaamsdelen “chest” en “thigh” die samen met “ankle” en “waist” in een tekening van een lichaam op de juiste plaats moeten worden gezet.

Voorbeelden van items die geen onderscheid maken tussen de drie beheersingsniveaus:
- Item t13.015_1: “Where can I go and watch a movie”, a) a cinema, b) a library, c) a stadium, d) a theatre;
- Item t13.025_1: kies de juiste omschrijving voor het huishoudelijke apparaat “dishwasher” uit een drop-down menu;
- Item t13.055_3 is deel van een brief met invulvelden via een drop-down menu, I’m Robin Schreurs and …

2.3.2 Wiskunde – vwo – hoofdattribuut 1: domein getallen en variabelen

Bij vwo leerlingen zijn 54 items afgenomen waarvan er 43 door minimaal 80 leerlingen zijn beantwoord (zie Tabel 2-4, zevende kolom). Van deze 43 items hebben 18 items een *p*-waarde tussen .1 en .9 en 25 een *p*-waarde buiten dit interval. Na de standaarsetting vielen respectievelijk 52, 58 en 60 leerlingen in de groepen onder, op en boven niveau.

In de zevende kolom van Tabel 2-5 wordt een overzicht gegeven van de geobserveerde spreiding in de succeskansen. Wat opvalt is dat van de 18 items er 14 een grote (.20) tot zeer grote (> .20) spreiding hebben in de succeskansen voor de onder niveau leerlingen en de op niveau leerlingen. Eén conclusie is dat ook dat het mogelijk is items te construeren die goed tot zeer goed onderscheid maken tussen leerlingen de onder niveau zitten en leerling die op niveau zitten. Zoals in de vijfde en zesde kolom van Tabel 2-5 kan worden gezien, hangt deze conclusie niet af van de standaardbepaling met grenswaarden van 60% en 80%, maar gelden ze ook voor de standaardbepalingen met grenswaarden van 40% en 60% en 50% en 70%. Dezelfde conclusie kan worden getrokken met betrekking tot het construeren van items die onderscheidt maken tussen leerlingen die op dan wel boven niveau zitten.

Aangezien 22 van de 43 items die door minimaal 80 leerlingen zijn beantwoord een *p*-waarde groter dan .9 hebben, kan er ook worden geconcludeerd dat *iets meer dan de helft van de items te makkelijk is voor de groep vwo leerlingen. Er dient bij het construeren van de items dan ook een betere afstemming op de doelgroep plaats te vinden.* Zoals in Tabel 2-5 kan worden gezien, zaten er bij deze 25 items met een te grote dan wel te kleine *p*-waarde slechts 2 met een spreiding van ongeveer .20 voor $p_{i<} - p_{i=}$ en 1 met een spreiding van ongeveer .20 voor $p_{i=} - p_{i>}$.

Voor alle 43 items die door 80 of meer leerlingen zijn beantwoord, zijn de succeskansen bepaald (zie Bijlage 2-2). Door deze te bestuderen kan een goede indruk worden gekregen welke wiskunde-items wel en welke items niet een goed onderscheid maken tussen de drie beheersingscategorieën (onder, op en boven niveau). De resulterende indrukken kunnen meegenomen worden bij het construeren van nieuwe items. Merk op dat deze resultaten deels afhangen van de in de standaardsetting gebruikte grenzen van 60% en 80%. Om die reden zijn ook de succeskansen bepaald nadat de standaard is gezet met andere grenzen. Deze paragraaf wordt afgesloten met enkele voorbeelden van goed en slecht onderscheidende items en voorbeelden van te makkelijke en te moeilijke items.

Voorbeelden van items die zeer goed onderscheiden tussen de onder en op niveau leerlingen

- Item hv.t13.036_4: “Als je van een geheel getal x weet dat het even, een vijfvoud en een kwadraat is, wat kun je dan zeggen over het getal 2x?” “Is het een priemgetal?” met de antwoorden juist, niet juist en je weet niet of dit juist of niet juist is.
- Item hv.t13.056_5: “Maak in onderstaande Tabel steeds een keuze uit eens of oneens”. “$\sqrt{4} \times \sqrt{7} + \sqrt{7} = 3\sqrt{7}$, met de antwoorden eens/oneens.
- Het item hv.t13.084a_1_bd_2 toont een figuur van een driehoek met de vraag de omtrek van de driehoek te berekenen. Dit item is handmatig gescoord.

Voorbeelden van items die zeer goed onderscheiden tussen de op en boven niveau leerlingen

- Het eerste item hv.t13.036_5 is hierboven beschreven. Hier betreft het het alternatief “Is het een kwadraat?”
- Het tweede item hv.t13.060_2 vraagt om een vergelijking met “waar” dan wel “niet waar” te evalueren “a + a = (a - 3) + (a + 3)”.

Voorbeelden van items die te makkelijk zijn (beide tegen de 100% correcte antwoorden):

- Het item hv.t13.060_3: “Kies bij ieder van de volgende regels voor waar of niet waar: -a –a –a –a = -4a” met de antwoorden waar/niet waar.
- Het item ov.t13.004_1: “Je ziet zes plaatjes. Kies voor elk plaatje de wolk waar dat plaatje bij hoort.” Het plaatje is een 3x3 tegelpatroon waarin de diagonale tegels zwart zijn en de andere tegels wit. Antwoorden: wolk met daarin het getal 1/2 en wolk met daarin het getal 1/3.

Voorbeelden van items die te moeilijk zijn(beide onder de 5% correcte antwoorden):

- Het item hv.t13.046_1: Joris heeft van vier gewichten (.5 kg, 1.25 kg, 2.5 kg, 5 kg) ieder twee schijven om aan een halter te hangen. Hoeveel gewichten kan Joris met deze schijven maken als aan elke kant van de halter hetzelfde gewicht moet hangen.

2.3.3 Nederlands – vmbo k – hoofdattribuut 3: Tekststructuur

Bij vmbo-k leerlingen zijn 39 items afgenomen waarvan er 34 door minimaal 80 leerlingen zijn beantwoord (zie Tabel 2-4, tiende kolom). Van deze 34 items hebben 32 items een *p*-waarde tussen .1 en .9 en 7 een *p*-waarde buiten dit interval. Na de standaarsetting vielen respectievelijk 99, 88 en 21 leerlingen in de groepen onder, op en boven niveau. Het kleine aantal leerlingen in de categorie boven niveau betekend dat de opmerkingen hierover met een grote slag om de arm gemaakt worden. Zoals kan worden gezien in de negende kolom van Tabel 2-4 is de verdeling van de leerlingen over de drie beheersingsniveaus beter voor de standaardbepaling met de grenzen 50% en 70% (respectievelijk 60, 83 en 65 leerlingen in de categorieën onder, op en boven niveau). Daarom zullen in het navolgende de resultaten voor deze standaardbepaling besproken worden.

In de negende kolom van Tabel 2-5 wordt een overzicht gegeven van de geobserveerde spreiding in de succeskansen. Wat opvalt is dat het verschil in succeskansen tussen leerlingen onder niveau en leerlingen op niveau ($p_{i<} - p_{i=}$) voor 7 van de 32 items met $.1 < p < .9$ (veel) groter is dan de grootste spreiding (.20, zie

Tabel 2-1) gebruikt in de simulatiestudie gerapporteerd in het CiTo (Cito, 2013a). Hetzelfde geldt voor het verschil in succeskansen tussen de leerlingen op niveau en de leerlingen boven niveau ($p_{i=}$ - $p_{i>}$): 5 van de 32 items hebben een grotere spreiding dan .20.

Gegeven de gebruikte methode van standaardbepaling, dat wil zeggen, de wijze waarop leerlingen aan de categorieën onder, op en boven niveau worden toegedeeld, kan dan ook worden geconcludeerd dat het mogelijk is items te construeren die goed tot zeer goed onderscheid kunnen maken tussen leerlingen die onder, op en boven niveau zitten. Hoe meer van deze items er geschreven kunnen worden, hoe beter de leerlingen uiteindelijk onderscheiden zullen kunnen worden. Maar merk op dat er ook veel items zijn met een matige (.10) tot redelijke (.20) spreiding. Dit soort items dient zoveel mogelijk vermeden te worden.

Voor alle 34 items die door 80 of meer leerlingen zijn beantwoord, zijn de succeskansen bepaald (zie Bijlage 2-3). Door deze te bestuderen kan een goede indruk worden gekregen welke items wel en welke items niet een goed onderscheid maken tussen de drie beheersingscategorieën (onder, op en boven niveau). De resulterende indrukken kunnen meegenomen worden bij het construeren van nieuwe items. Merk op dat deze resultaten deels afhangen van de in de standaardbepaling gebruikte grenzen van 60% en 80%. Om die reden zijn ook de succeskansen bepaald nadat de standaard is gezet met andere grenzen. Deze paragraaf wordt afgesloten met enkele voorbeelden van goed en slecht onderscheidende items.

Voorbeelden van items die zeer goed onderscheiden tussen onder en op niveau

- De items t13.014_x: deze opdracht betreft een brief waarvan de onderdelen gegeven zijn en op de juiste volgorde gezet moeten worden. De meeste onderdelen x van deze opdracht maken prima onderscheid tussen de leerlingen;
- Item ov.t13.037_1.bd_13: “Schrijf een brief van 300 woorden aan de directie van het zwembad. De gebruikte scoring maakt goed onderscheid tussen leerlingen die onder en op niveau zijn.

Voorbeelden van items die goed onderscheiden tussen op en boven niveau

- Item t13.14_x. Ook besproken in de vorige alinea. Deze opdracht maakt ook goed onderscheid tussen leerlingen die op en boven niveau zitten.

Items die minder tot slecht onderscheid maken tussen de leerlingen

- De items t13.016_x: de onderdelen van deze opdracht (geef aan waar in een brief een nieuwe alinea begint) maken minder tot slecht onderscheid tussen de leerlingen;
- De items t13.018_x: ook de onderdelen van deze opdracht (voeg op een aantal plaatsen in een tekst een woord uit een drop-down menu in) maken niet een heel goed onderscheid tussen de leerlingen.

Tabel 2-4 Aantallen items en aantallen personen

leerweg x attribuut	vmbo-gt engels-h3, grenzen .4 en .6	vmbo-gt engels-h3, grenzen .5 en .7	vmbo-gt engels-h3, grenzen .6 en .8	vwo wiskunde-h1, grenzen .4 en .6	vwo wiskunde-h1, grenzen .5 en .7	vwo wiskunde-h1, grenzen .6 en .8	vmbo-k Nederlands-h3, grenzen .4 en .6	vmbo-k Nederlands-h3, grenzen .5 en .7	vmbo-k Nederlands-h3, grenzen .6 en .8
item totaal	51	51	51	54	54	54	39	39	39
N > 80	22	22	22	43	43	43	34	34	34
.1 < p < .9	15	15	15	18	18	18	32	32	32
p < .1	0	0	0	3	3	3	0	0	0
p > .9	7	7	7	22	22	22	2	2	2
N									
onder	27	40	90	9	42	52	16	60	99
op	63	89	105	43	42	58	83	83	88
boven	150	111	45	118	86	60	109	65	21

Tabel 2-5 Spreiding in de succeskansen

leerweg x attribuut	vmbo-gt engels-h3, grenzen .4 en .6	vmbo-gt engels-h3, grenzen .5 en .7	vmbo-gt engels-h3, grenzen .6 en .8	vwo wiskunde-h1, grenzen .4 en .6	vwo wiskunde-h1, grenzen .5 en .7	vwo wiskunde-h1, grenzen .6 en .8	vmbo-k Nederlands-h3, grenzen .4 en .6	vmbo-k Nederlands-h3, grenzen .5 en .7	vmbo-k Nederlands-h3, grenzen .6 en .8
.1 < p < .9									
$p_{i<} - p_{i=}$	< .10: 4 ≈ .20: 3 ≈ .30: 5 ≈ .40: 2 > .50: 1	< .10: 4 ≈ .20: 4 ≈ .30: 2 ≈ .40: 4 > .50: 1	< .10: 1 ≈ .20: 9 ≈ .30: 3 ≈ .40: 1 > .50: 1	< .10: 6 ≈ .20: 4 ≈ .30: 3 ≈ .40: 3 > .50: 2	< .10: 5 ≈ .20: 10 ≈ .30: 2 ≈ .40: 1	< .10: 4 ≈ .20: 6 ≈ .30: 6 ≈ .40: 2	< .10: 10 ≈ .20: 9 ≈ .30: 7 ≈ .40: 5 > .50: 1	< .10: 7 ≈ .20: 18 ≈ .30: 5 ≈ .40: 1 > .50: 1	< .10: 8 ≈ .20: 16 ≈ .30: 2 ≈ .40: 2 > .50: 4
$p_{i=} - p_{i>}$	< .10: 1 ≈ .20: 8 ≈ .30: 4 ≈ .40: 2	< .10: 4 ≈ .20: 4 ≈ .30: 5 ≈ .40: 1 > .50: 1	< .10: 4 ≈ .20: 6 ≈ .30: 2 ≈ .40: 0 > .50: 3	< .10: 1 ≈ .20: 6 ≈ .30: 5 ≈ .40: 4 > .50: 2	< .10: 5 ≈ .20: 7 ≈ .30: 2 ≈ .40: 1 > .50: 3	< .10: 7 ≈ .20: 4 ≈ .30: 3 ≈ .40: 2 > .50: 2	< .10: 8 ≈ .20: 16 ≈ .30: 2 ≈ .40: 2 > .50: 4	< .10: 17 ≈ .20: 9 ≈ .30: ≈ .40: 1 > .50: 4	< .10: 16 ≈ .20: 10 ≈ .30: 5 ≈ .40: 1
p < .1 & p > .9									
$p_{i<} - p_{i=}$	< .10: 4 ≈ .20: 3	< .10: 4 ≈ .20: 3	<.10: 6 ≈.20: 1	< .10: 18 ≈ .20: 3 ≈ .30: 2 ≈ .40: 2	< .10: 21 ≈ .20: 4	< .10: 23 ≈ .20: 2	< .10: 2	< .10: 2	< .10: 2
$p_{i=} - p_{i>}$	< .10: 7	< .10: 7	< .10: 7	<.10: 23 ≈.20: 2	<.10: 23 ≈.20: 2	< .10: 24 ≈ .20: 1	<.10: 2	<.10: 2	< .10: 2

2.4 Eén simulatie met op de try-out gebaseerde succeskansen

In 2013 zijn simulatiestudies uitgevoerd om te bepalen hoeveel items er moeten worden afgenomen om 80% van de leerlingen correct te diagnosticeren in de niveaus onder, op en boven niveau (Cito, 2013a, hfst. 2). Deze simulaties zijn uitgevoerd met hypothetische succeskansen. In de vorige paragraaf is uiteengezet hoe deze hypothetische succeskansen zich verhouden tot de succeskansen bepaald op grond van de data resulterend uit de try-out. De op de try-out gebaseerde succeskansen zijn weergegeven in Bijlagen 1-1, 1-2 en 1-3. De vraag is hoe hoe goed de diagnose van leerlingen in de categorieën onder, op en boven niveau werkt met deze empirisch bepaalde succeskansen in plaats van hypothetische succeskansen. Om dit na te gaan is een simulatieonderzoek verricht met de op de try-out gebaseerde succeskansen.

2.5 Opzet simulatie

Wat vorig jaar nog niet kon, was het adequaat representeren van testlets in het gebruikte psychometrische model. Het onderzoek toonde wel aan dat het negeren van de testletstructuur een negatief effect had op de kans om een leerling correct te diagnosticeren. Ondertussen is het psychometrisch model zo uitgebreid dat wel op een adequate manier met testlets rekening gehouden kan worden. Een beschrijving en evaluatie van dit model zal worden gegeven in Sies (2014). Hier zal nu kort een eerste evaluatie van het uitgebreide psychometrische model worden gepresenteerd gebaseerd op de succeskansen resulterend uit de try-outdata voor Engels – vmbo gt – hoofdattribuut 3: Woordenschat en Woordgebruik.

In Bijlage 2-4 staan de succeskansen weergegeven voor afzonderlijke items, dat wil zeggen, items die niet tot een testlet behoren (dit zijn er twee) en voor testlets, dat wil zeggen, groepjes bij elkaar horende (sub-) items. Er zijn vijf testlets, die respectievelijk, 5, 3, 3, 4 en 4, (sub-)items bevatten. Zoals kan worden gezien, “hangen er aan” een afzonderlijk item drie succeskansen: de kansen dat een leerling die onder, op en boven niveau zit het item correct beantwoord. Als er in een testlet n items zitten die goed=1 dan wel fout=0 kunnen worden beantwoord, zijn er 2^n mogelijke antwoordpatronen mogelijk. Zoals in Bijlage 2-4 gezien kan worden, zijn er op een testlet met 5 (sub-)items 2^5=32 mogelijke antwoordpatronen. Voor elk van de antwoord-patronen zijn er drie “succeskansen”, dat wil zeggen, de kans dat een leerling die onder, op en boven niveau zit het betreffende antwoordpatroon oplevert.

Er zijn in totaal 21 items waarvan er 5+3+3+4+4 in testlets zijn georganiseerd. Met behulp van de succeskansen resulterend uit de try-outdata (één set voor elke van de drie standaardbepalingen) zijn data voor 5000 leerlingen gesimuleerd voor elk van de categorieën onder, op en boven niveau. Vervolgens is het met testlets uitgebreide psychometrische model gebruikt om elk van de leerlingen toe te kennen aan de categorieën onder, op en boven niveau. Dit is gedaan met prior model kansen van .33, .33, .33 en zonder de items adaptief af te nemen.

2.6 Resultaten simulatie

Het resultaat kan worden gevonden in Tabellen 1-6, 1-7 en 1-8. Zoals kan worden gezien wordt voor elk van de standaardbepalingen gemiddeld ruwweg 80% van de leerlingen correct geclassificeerd. Merk op dat dit nu gebeurd met 21 in plaats van de 36 (of meer) items die uit de simulatiestudie gepresenteerd in het Cito rapport (Cito, 2013a, hfst. 2) resulteerden. Hiervoor zijn twee redenen: de succeskansen zijn anders dan in de simulatiestudie; en er wordt verdisconteerd voor het feit dat items in testlets zijn georganiseerd. Kortom, dit suggereert dat als de itemschrijvers erin slagen items op te leveren van dezelfde kwaliteit als gebruikt in deze sectie en als we verdisconteren voor testlets in het psychometrische model, dat de aantallen items nodig om leerlingen met een kans van .80 correct te diagnosticeren beduidend lager liggen dan verwacht op grond van de simulatiestudie.

Tabel 2-6 Diagnoses Engels vmbo-gt attribuut 3 volgend op standaardbepaling met grenzen .4 en .6

		Gediagnosticeerd beheersingsniveau		
		<	=	>
Echt beheersings-niveau	<	.858	.129	.013
	=	.118	.747	.135
	>	.014	.120	.866

De rijen geven het echte beheersingsniveau aan, de kolommen welke proportie van de leerlingen in elk beheersingsniveau gediagnosticeerd is. Bijvoorbeeld: van de leerlingen onder niveau wordt ongeveer 86% ook daadwerkelijk als onder niveau gediagnosticeerd.

Tabel 2-7 Diagnoses Engels vmbo-gt attribuut 3 volgend op standaardbepaling met grenzen .5 en .7

		Gediagnosticeerd beheersingsniveau		
		<	=	>
Echt	<	.874	.112	.014
Beheersings-	=	.115	.743	.142
niveau	>	.020	.161	.819

De rijen geven het echte beheersingsniveau aan, de kolommen welke proportie van de leerlingen in elk beheersingsniveau gediagnosticeerd is.

Tabel 2-8 Diagnose Engels vmbo-gt attribuut 3 volgend op standaardbepaling met grenzen .6 en .8

		Gediagnosticeerd beheersingsniveau		
		<	=	>
Echt	<	.852	.120	.028
Beheersings-	=	.142	.664	.194
niveau	>	.032	.126	.842

De rijen geven het echte beheersingsniveau aan, de kolommen welke proportie van de leerlingen in elk beheersingsniveau gediagnosticeerd is.

2.7 Conclusie

Na aanleiding van de analyse van de try-outdata kan het volgende geconcludeerd worden:

- Het is voor schrijfvaardigheid Engels, wiskunde en schrijfvaardigheid Nederlands mogelijk items te construeren die een goed onderscheid maken tussen leerlingen die onder, op en boven niveau zitten.
- Aan de andere kant zijn er voor elk van de vakken ook items te vinden die geen goed onderscheid maken tussen leerlingen die onder, op en boven niveau zitten.

Advies: kijk grondig naar de items die wel en niet goed onderscheid maken en probeer daaruit te leren wat wel en wat niet goed werkt bij de constructie van een item

- Er zijn vooral bij schrijfvaardigheid Engels en wiskunde 30%-50% items die te makkelijk zijn. Dit zou kunnen komen doordat items voor havo en vwo lijken te zijn geschreven. Deze zijn voor vwo misschien te makkelijk.

Advies: kijk grondig naar de items die een goede moeilijkheidsgraad hebben én de items die te makkelijk zijn en probeer daaruit te leren wat wel en wat niet goed werkt bij de constructie van een item.

- Op grond van een eerste simulatie gebaseerd op de succeskansen resulterend uit de try-out voor schrijfvaardigheid Engels vmbo gt hoofdattribuut 3 kan worden geconcludeerd dat het aantal items dat nodig is om 80% van de leerlingen correct te diagnosticeren waarschijnlijk kleiner is dan geconcludeerd op basis van de simulatiestudie gerapporteerd in het Cito rapport (Cito, 2013a, hfst. 2). Het lijkt er sterk op dat zonder adaptief af te nemen we toe kunnen met 21 in plaats van 36 items.

Advies: Schrijf items met een goed onderscheidend vermogen en een goede moeilijkheidsgraad en baseer het vervolg van de ontwikkeling van de DTT op 24 in plaats van 36 niet adaptief afgenomen items. Merk op

dat het waarschijnlijk is dat een adaptieve afname tot een verdere reductie van het aantal benodigde items zal leiden. Complicatie bij zo'n adaptieve afname is echter wel dat er veel testlets zijn die in hun geheel aangeboden dienen te worden wat niet ten goede komt aan de adaptieve flexibiliteit. Hoe dat het beste kan worden georganiseerd is een punt van verdere studie.

2.8 Referenties

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

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

Sies, A. (2014). *Cognitive Diagnostic Testing Using Calibrated Hypotheses*. Master Thesis, Methods and Statistics for the Behavioral, Biomedical, and Social Sciences, University Utrecht.

Van Ginkel, J.R., Van der Ark, L. A., Sijtsma, K. en Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis*, *51* (8), 4013-4027. http://www.sciencedirect.com/science/article/pii/S0167947306004993

2.9 Bijlagen

Bijlage 2-1 Succeskansen Engels – vmbo gt – hoofdattribuut 3 “Woordenschat en Woordgebruik”

	Met grenzen ,4 en ,6			Met grenzen ,5 en ,7			Met grenzen ,6 en ,8		
Item	Groep "Onder niveau" 0,0-0,4	Groep "Op niveau" 0,4-0,6	Groep "Boven niveau" 0,6-1,0	Groep "Onder niveau" 0,0-0,5	Groep "Op niveau" 0,5-0,7	Groep "Boven niveau" 0,7-1,0	Groep "Onder niveau" 0,0-0,6	Groep "Op niveau" 0,6-0,8	Groep "Boven niveau" 0,8-1,0
dtt.en.vmbo.t13.014_1	0,56	0,71	0,91	0,60	0,81	0,91	0,67	0,90	0,93
dtt.en.vmbo.t13.015_1	0,96	0,98	0,99	0,98	0,99	0,99	0,98	1,00	0,98
dtt.en.vmbo.t13.017_1	0,63	0,78	0,91	0,57	0,89	0,91	0,73	0,90	0,96
dtt.en.vmbo.t13.019_1	0,15	0,40	0,64	0,15	0,43	0,73	0,32	0,50	0,96
dtt.en.vmbo.t13.019_2	0,19	0,46	0,67	0,28	0,47	0,73	0,38	0,55	0,93
dtt.en.vmbo.t13.019_3	0,93	0,97	0,97	0,95	0,96	0,98	0,96	0,97	0,98
dtt.en.vmbo.t13.019_4	0,59	0,59	0,85	0,60	0,62	0,92	0,59	0,79	1,00
dtt.en.vmbo.t13.019_5	0,33	0,65	0,85	0,48	0,67	0,89	0,56	0,80	0,98
dtt.en.vmbo.t13.025_1	0,48	0,78	0,94	0,50	0,88	0,95	0,69	0,92	0,98
dtt.en.vmbo.t13.025_2	0,93	0,98	0,99	0,92	1,00	0,99	0,97	1,00	0,98
dtt.en.vmbo.t13.025_3	0,81	0,97	0,97	0,85	0,99	0,96	0,92	0,97	0,98
dtt.en.vmbo.t13.025_4	0,63	0,87	0,96	0,68	0,94	0,95	0,80	0,95	0,98
dtt.en.vmbo.t13.028_1	0,30	0,76	0,94	0,42	0,82	0,96	0,62	0,91	1,00
dtt.en.vmbo.t13.028_2	0,22	0,24	0,63	0,22	0,26	0,75	0,23	0,49	0,96
dtt.en.vmbo.t13.028_3	0,19	0,29	0,61	0,20	0,30	0,71	0,26	0,46	0,96
dtt.en.vmbo.t13.028_4	0,81	0,95	0,97	0,85	0,98	0,96	0,91	0,96	1,00
dtt.en.vmbo.t13.054_6	0,22	0,52	0,93	0,20	0,73	0,95	0,43	0,91	0,98
dtt.en.vmbo.t13.054_7	0,22	0,57	0,89	0,30	0,67	0,93	0,47	0,86	0,96
dtt.en.vmbo.t13.054_13	0,41	0,65	0,93	0,48	0,75	0,95	0,58	0,93	0,91
dtt.en.vmbo.t13.055_3	1,00	1,00	0,97	1,00	0,99	0,96	1,00	0,97	0,96
dtt.en.vmbo.t13.055_5	0,04	0,02	0,25	0,02	0,03	0,32	0,02	0,10	0,58
dtt.en.vmbo.t13.055_11	0,33	0,92	0,89	0,52	0,92	0,87	0,74	0,90	0,87

Van de oorspronkelijke 51 items zijn 15 bruikbare items overgebleven (p-waardes tussen .1 en .9 en door minimaal 100 leerlingen beantwoord). De 7 items onder de streep zijn de items met een p-waarde kleiner dan .1 of groter dan .9. Eerder zijn 29 items weggevallen omdat ze door minder dan 100 leerlingen zijn beantwoord.

Bijlage 2-2 Succeskansen Wiskunde – vwo – hoofdattribuut 1 “domein getallen en variabelen”

	Met grenzen ,4 en ,6			Met grenzen ,5 en ,7			Met grenzen ,6 en ,8		
Item	Groep "Onder niveau" 0,0-0,4	Groep "Op niveau" 0,4-0,6	Groep "Boven niveau" 0,6-1,0	Groep "Onder niveau" 0,0-0,5	Groep "Op niveau" 0,5-0,7	Groep "Boven niveau" 0,7-1,0	Groep "Onder niveau" 0,0-0,6	Groep "Op niveau" 0,6-0,8	Groep "Boven niveau" 0,8-1,0
dtt.wi.hv.t13.031a_1	1,00	0,98	1,00	0,98	1,00	1,00	0,98	1,00	1,00
dtt.wi.hv.t13.031b_1	1,00	0,98	1,00	0,98	1,00	1,00	0,98	1,00	1,00
dtt.wi.hv.t13.036_1	0,78	0,91	0,98	0,86	0,95	1,00	0,88	0,97	1,00
dtt.wi.hv.t13.036_2	0,78	0,91	0,97	0,86	0,95	0,99	0,88	0,95	1,00
dtt.wi.hv.t13.036_3	0,33	0,70	0,98	0,67	0,86	0,99	0,63	0,97	1,00
dtt.wi.hv.t13.036_4	0,11	0,56	0,92	0,45	0,74	0,97	0,48	0,84	0,98
dtt.wi.hv.t13.036_5	0,22	0,40	0,66	0,33	0,36	0,79	0,37	0,40	0,92
dtt.wi.hv.t13.036_6	0,56	0,23	0,38	0,26	0,33	0,41	0,29	0,31	0,45
dtt.wi.hv.t13.037_1_bd_2	0,56	0,74	0,98	0,69	0,90	1,00	0,71	0,97	1,00
dtt.wi.hv.t13.046_1	0,00	0,07	0,06	0,05	0,02	0,08	0,06	0,07	0,05
dtt.wi.hv.t13.056_1	0,67	0,84	0,95	0,79	0,93	0,95	0,81	0,90	1,00
dtt.wi.hv.t13.056_2	0,56	0,93	0,98	0,83	0,98	0,99	0,87	0,98	0,98
dtt.wi.hv.t13.056_3	0,56	0,60	0,86	0,60	0,69	0,92	0,60	0,72	1,00
dtt.wi.hv.t13.056_4	0,89	0,95	0,97	0,95	0,90	0,99	0,94	0,93	1,00
dtt.wi.hv.t13.056_5	0,00	0,28	0,75	0,24	0,33	0,90	0,23	0,53	0,97
dtt.wi.hv.t13.057_1	1,00	0,93	0,97	0,95	0,93	0,98	0,94	0,93	1,00
dtt.wi.hv.t13.057_2	1,00	0,98	0,99	0,98	0,98	1,00	0,98	0,98	1,00
dtt.wi.hv.t13.057_3	0,11	0,23	0,40	0,17	0,31	0,44	0,21	0,29	0,50
dtt.wi.hv.t13.057_4	0,89	0,84	0,94	0,86	0,83	0,98	0,85	0,88	1,00
dtt.wi.hv.t13.057_5	0,89	0,88	0,97	0,93	0,88	0,99	0,88	0,95	1,00
dtt.wi.hv.t13.057_6	0,67	0,44	0,83	0,43	0,67	0,90	0,48	0,72	0,93
dtt.wi.hv.t13.060_1	0,22	0,49	0,94	0,50	0,74	0,95	0,44	0,90	0,98
dtt.wi.hv.t13.060_2	0,33	0,23	0,70	0,21	0,36	0,84	0,25	0,47	0,93
dtt.wi.hv.t13.060_3	1,00	0,98	0,98	0,98	1,00	0,98	0,98	0,97	1,00
dtt.wi.hv.t13.060_4	1,00	0,91	0,99	0,90	0,98	1,00	0,92	0,98	1,00
dtt.wi.hv.t13.084a_1_bd_2	0,11	0,23	0,74	0,17	0,33	0,90	0,21	0,50	0,97
dtt.wi.hv.t13.084a_1_bd_3	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
dtt.wi.hv.t13.084a_1_bd_4	0,22	0,16	0,01	0,17	0,07	0,00	0,17	0,02	0,00
dtt.wi.hv.t13.084b_1_bd_2	0,00	0,37	0,82	0,21	0,62	0,91	0,31	0,66	0,98
dtt.wi.hv.t13.087_1_bd_2	0,44	0,56	0,85	0,55	0,64	0,91	0,54	0,71	0,98
dtt.wi.hv.t13.095_1	0,33	0,84	0,97	0,69	0,93	0,99	0,75	0,95	0,98
dtt.wi.hv.t13.095_2	0,56	0,86	0,97	0,76	0,98	0,98	0,81	0,97	0,98
dtt.wi.hv.t13.095_3	0,44	0,72	0,97	0,64	0,90	0,98	0,67	0,95	0,98
dtt.wi.hv.t13.095_4	0,67	0,93	0,99	0,86	0,98	1,00	0,88	0,98	1,00
dtt.wi.hv.t13.095_5	0,56	0,98	0,97	0,88	0,98	0,98	0,90	0,97	0,98
dtt.wi.hv.t13.095_6	0,22	0,79	0,95	0,67	0,88	0,97	0,69	0,91	0,98

dtt.wi.ov.t13.004_1	1,00	1,00	0,99	1,00	1,00	0,99	1,00	0,98	1,00
dtt.wi.ov.t13.004_2	0,89	0,93	0,99	0,90	1,00	0,99	0,92	1,00	0,98
dtt.wi.ov.t13.004_3	1,00	0,98	1,00	0,98	1,00	1,00	0,98	1,00	1,00
dtt.wi.ov.t13.004_4	1,00	0,95	0,97	0,95	0,98	0,98	0,96	0,98	0,97
dtt.wi.ov.t13.004_5	1,00	0,98	1,00	0,98	1,00	1,00	0,98	1,00	1,00
dtt.wi.ov.t13.004_6	1,00	0,93	0,97	0,93	0,93	0,99	0,94	0,93	1,00
dtt.wi.ov.t13.055_1	0,56	0,47	0,54	0,48	0,48	0,57	0,48	0,48	0,60

Van de oorspronkelijke 54 items zijn 18 bruikbare items overgebleven (p-waardes tussen .1 en .9 en door minimaal 100 leerlingen beantwoord). De 25 items onder de streep zijn de items met een p-waarde kleiner dan .1 of groter dan .9. Eerder zijn 11 items weggevallen omdat ze door minder dan 100 leerlingen zijn beantwoord.

Bijlage 2-3 Succeskansen Nederlands – vmbo k – hoofdattribuut 3 “Tekststructuur”

	Met grenzen ,4 en ,6			Met grenzen ,5 en ,7			Met grenzen ,6 en ,8		
Item	Groep "Onder niveau" 0,0-0,4	Groep "Op niveau" 0,4-0,6	Groep "Boven niveau" 0,6-1,0	Groep "Onder niveau" 0,0-0,5	Groep "Op niveau" 0,5-0,7	Groep "Boven niveau" 0,7-1,0	Groep "Onder niveau" 0,0-0,6	Groep "Op niveau" 0,6-0,8	Groep "Boven niveau" 0,8-1,0
dtt.nl.ov.t13.037_1_bd_12	0,38	0,83	0,88	0,68	0,87	0,89	0,76	0,86	0,95
dtt.nl.ov.t13.037_1_bd_13	0,06	0,41	0,84	0,17	0,76	0,83	0,35	0,84	0,86
dtt.nl.vmbo.t13.014_1	0,19	0,23	0,39	0,23	0,30	0,38	0,22	0,40	0,33
dtt.nl.vmbo.t13.014_2	0,06	0,07	0,28	0,07	0,12	0,35	0,07	0,26	0,33
dtt.nl.vmbo.t13.014_3	0,00	0,22	0,77	0,10	0,43	0,92	0,18	0,72	1,00
dtt.nl.vmbo.t13.014_4	0,06	0,24	0,78	0,13	0,43	0,95	0,21	0,73	1,00
dtt.nl.vmbo.t13.014_5	0,00	0,20	0,70	0,08	0,36	0,89	0,17	0,64	0,95
dtt.nl.vmbo.t13.014_6	0,00	0,16	0,71	0,07	0,34	0,89	0,13	0,65	0,95
dtt.nl.vmbo.t13.014_7	0,19	0,46	0,84	0,32	0,60	0,98	0,41	0,81	1,00
dtt.nl.vmbo.t13.015_1	0,31	0,71	0,87	0,60	0,76	0,92	0,65	0,84	1,00
dtt.nl.vmbo.t13.015_2	0,12	0,64	0,83	0,50	0,70	0,89	0,56	0,80	1,00
dtt.nl.vmbo.t13.015_3	0,44	0,82	0,93	0,68	0,89	0,94	0,76	0,91	1,00
dtt.nl.vmbo.t13.015_4	1,00	0,98	0,99	0,97	1,00	0,98	0,98	0,99	1,00
dtt.nl.vmbo.t13.016_1	0,56	0,61	0,83	0,55	0,75	0,85	0,61	0,80	0,95
dtt.nl.vmbo.t13.016_2	0,69	0,88	0,93	0,78	0,92	0,95	0,85	0,93	0,90
dtt.nl.vmbo.t13.016_3	0,50	0,37	0,37	0,45	0,33	0,38	0,39	0,35	0,43
dtt.nl.vmbo.t13.016_4	0,50	0,54	0,72	0,50	0,67	0,69	0,54	0,70	0,76
dtt.nl.vmbo.t13.016_5	0,62	0,63	0,88	0,58	0,77	0,91	0,63	0,86	0,95
dtt.nl.vmbo.t13.016_6	0,25	0,40	0,55	0,38	0,46	0,55	0,37	0,55	0,57
dtt.nl.vmbo.t13.016_7	0,38	0,41	0,50	0,33	0,52	0,49	0,40	0,49	0,57
dtt.nl.vmbo.t13.016_8	0,75	0,61	0,83	0,68	0,67	0,88	0,64	0,81	0,95
dtt.nl.vmbo.t13.016_9	0,31	0,46	0,66	0,48	0,47	0,72	0,43	0,58	1,00
dtt.nl.vmbo.t13.018_1	0,62	0,64	0,81	0,60	0,81	0,74	0,64	0,78	0,90
dtt.nl.vmbo.t13.018_10	0,12	0,41	0,43	0,23	0,45	0,49	0,36	0,42	0,48
dtt.nl.vmbo.t13.018_2	0,31	0,71	0,82	0,53	0,78	0,86	0,65	0,80	0,90
dtt.nl.vmbo.t13.018_3	0,12	0,47	0,75	0,27	0,65	0,82	0,41	0,72	0,90
dtt.nl.vmbo.t13.018_4	0,62	0,86	0,89	0,80	0,82	0,95	0,82	0,88	0,95
dtt.nl.vmbo.t13.018_5	0,81	0,96	0,99	0,95	0,96	0,98	0,94	0,99	1,00
dtt.nl.vmbo.t13.018_6	0,06	0,49	0,61	0,35	0,54	0,66	0,42	0,57	0,81
dtt.nl.vmbo.t13.018_7	0,31	0,61	0,74	0,48	0,71	0,75	0,57	0,72	0,86

dtt.nl.vmbo.t13.018_8	0,56	0,83	0,92	0,68	0,93	0,92	0,79	0,92	0,90
dtt.nl.vmbo.t13.018_9	0,69	0,76	0,89	0,72	0,83	0,91	0,75	0,86	1,00
dtt.nl.vmbo.t13.033C_1	0,25	0,47	0,52	0,35	0,52	0,55	0,43	0,50	0,62
dtt.nl.vmbo.t13.034_2_bd_1	0,12	0,43	0,73	0,27	0,63	0,77	0,38	0,68	0,95

Van de oorspronkelijke 39 items zijn 32 bruikbare items overgebleven (p-waardes tussen .1 en .9 en door minimaal 100 leerlingen beantwoord). De 2 items onder de streep zijn de items met een p-waarde kleiner dan .1 of groter dan .9. Eerder zijn 5 items weggevallen omdat ze door minder dan 100 leerlingen zijn beantwoord.

Bijlage 2-4 Succeskansen voor afzonderlijke items en testlets uit de toets Engels vmbo-gt attribuut 3 gebaseerd op standaardbepaling met grenzen ,4 en ,6; met grenzen ,5 en ,7 en met grenzen ,6 en ,8

	Afzonderlijke items	Met grenzen ,4 en ,6			Met grenzen ,5 en ,7			Met grenzen ,6 en ,8		
		Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	dtt.en.vmbo .t13.014_1	0,550	0,710	0,910	0,600	0,800	0,900	0,660	0,900	0,910
	dtt.en.vmbo .t13.017_1	0,620	0,770	0,910	0,570	0,880	0,900	0,730	0,890	0,940

Elk item heeft 3 succeskansen, een voor elk beheersingscategorie (op, onder of boven niveau).

Testlets	Antwoordpatroon					Met grenzen ,4 en ,6			Met grenzen ,5 en ,7			Met grenzen ,6 en ,8		
dtt.en.vmbo .t13.019	dtt.en.vmbo .t13.019_1	dtt.en.vmbo .t13.019_2	dtt.en.vmbo .t13.019_3	dtt.en.vmbo .t13.019_4	dtt.en.vmbo .t13.019_5	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	0	0	0	0	0	0,034	0,011	0,005	0,028	0,008	0,007	0,018	0,007	0,013
	0	0	0	0	1	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	0	0	0	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	0	1	0	0	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	0	0	1	0	0	0,169	0,042	0,022	0,153	0,050	0,007	0,116	0,029	0,013
	0	0	0	1	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	1	1	1	1	0,017	0,053	0,346	0,042	0,050	0,420	0,045	0,175	0,519
	1	1	1	1	0	0,034	0,053	0,044	0,028	0,050	0,049	0,054	0,051	0,026
	0	1	1	1	1	0,017	0,074	0,071	0,042	0,091	0,049	0,062	0,088	0,026
	1	0	1	1	1	0,051	0,063	0,060	0,042	0,066	0,063	0,071	0,066	0,039
	1	1	0	1	1	0,017	0,011	0,011	0,014	0,008	0,014	0,009	0,007	0,026

	1	1	1	0	1	0,017	0,021	0,027	0,014	0,041	0,014	0,018	0,036	0,013
	1	1	0	0	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	0	1	0	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	0	0	1	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	0	0	0	1	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	0	1	1	0	0	0,017	0,021	0,011	0,014	0,017	0,014	0,018	0,015	0,013
	0	1	0	1	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	0	1	0	0	1	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	0	0	1	0	1	0,017	0,105	0,011	0,056	0,066	0,007	0,089	0,015	0,013
	0	0	1	1	0	0,085	0,042	0,016	0,083	0,041	0,007	0,071	0,022	0,013
	0	0	0	1	1	0,017	0,021	0,005	0,014	0,017	0,007	0,018	0,007	0,013
	1	1	1	0	0	0,017	0,063	0,027	0,014	0,058	0,028	0,054	0,036	0,013
	1	0	1	1	0	0,017	0,032	0,005	0,014	0,025	0,007	0,027	0,007	0,013
	1	0	0	1	1	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	0	1	0	1	0,034	0,053	0,044	0,028	0,074	0,028	0,054	0,058	0,013
	1	1	0	1	0	0,017	0,011	0,005	0,014	0,008	0,007	0,009	0,007	0,013
	1	1	0	0	1	0,017	0,011	0,011	0,014	0,017	0,007	0,009	0,015	0,013
	0	1	1	1	0	0,068	0,053	0,033	0,069	0,058	0,021	0,071	0,044	0,013
	0	1	0	1	1	0,034	0,021	0,016	0,028	0,025	0,014	0,027	0,022	0,013
	0	1	1	0	1	0,017	0,042	0,011	0,028	0,025	0,014	0,036	0,015	0,013
	0	0	1	1	1	0,102	0,084	0,154	0,111	0,116	0,140	0,116	0,197	0,026
dtt.en.vmbo .t13.055	**dtt.en.vmbo .t13.055_3**	**dtt.en.vmbo .t13.055_5**	**dtt.en.vmbo .t13.055_11**			Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	0	0	0			0,029	0,014	0,019	0,021	0,010	0,025	0,010	0,018	0,038
	1	1	1			0,029	0,014	0,196	0,021	0,031	0,244	0,010	0,080	0,434
	0	0	1			0,029	0,014	0,019	0,021	0,021	0,017	0,010	0,027	0,019
	1	0	0			0,514	0,070	0,063	0,396	0,072	0,059	0,224	0,071	0,057
	0	1	0			0,029	0,014	0,006	0,021	0,010	0,008	0,010	0,009	0,019

	1	1	0	0,057	0,028	0,044	0,042	0,021	0,059	0,031	0,035	0,075
	0	1	1	0,029	0,014	0,013	0,021	0,010	0,017	0,010	0,009	0,038
	1	0	1	0,286	0,831	0,639	0,458	0,825	0,571	0,694	0,752	0,321
dtt.en.vmbo .t13.054	**dtt.en.vmbo .t13.054_6**	**dtt.en.vmbo .t13.054_7**	**dtt.en.vmbo .t13.054_13**	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	0	0	0	0,257	0,099	0,006	0,250	0,041	0,008	0,153	0,009	0,019
	1	1	1	0,029	0,254	0,747	0,042	0,412	0,798	0,184	0,708	0,736
	0	0	1	0,257	0,169	0,019	0,250	0,103	0,017	0,204	0,027	0,019
	1	0	0	0,143	0,042	0,019	0,104	0,041	0,017	0,071	0,027	0,019
	0	1	0	0,114	0,127	0,013	0,125	0,082	0,008	0,122	0,018	0,019
	1	1	0	0,057	0,099	0,057	0,042	0,103	0,050	0,082	0,044	0,094
	0	1	1	0,086	0,085	0,051	0,125	0,062	0,042	0,082	0,062	0,038
	1	0	1	0,057	0,127	0,089	0,062	0,155	0,059	0,102	0,106	0,057
dtt.en.vmbo .t13.025	**dtt.en.vmbo .t13.025_1**	**dtt.en.vmbo .t13.025_2**	**dtt.en.vmbo .t13.025_3**	**dtt.en.vmbo .t13.025_4**	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	0	0	0	0	0,047	0,013	0,006	0,036	0,010	0,008	0,019	0,008	0,016
	1	1	1	1	0,256	0,582	0,825	0,286	0,733	0,795	0,528	0,777	0,721
	0	1	0	0	0,047	0,013	0,006	0,036	0,010	0,008	0,019	0,008	0,016
	1	0	0	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	0	1	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	0	0	1	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	1	1	0	0	0,047	0,013	0,012	0,036	0,010	0,016	0,019	0,017	0,016
	1	0	0	1	0,023	0,013	0,012	0,018	0,010	0,016	0,009	0,008	0,033
	1	0	1	0	0,023	0,025	0,006	0,036	0,010	0,008	0,019	0,008	0,016
	0	1	0	1	0,047	0,038	0,006	0,054	0,019	0,008	0,038	0,008	0,016
	0	1	1	0	0,186	0,063	0,030	0,161	0,038	0,039	0,113	0,033	0,033
	0	0	1	1	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	1	1	1	0,116	0,114	0,036	0,161	0,076	0,024	0,123	0,050	0,016

	1	0	1	1	0,047	0,013	0,006	0,036	0,010	0,008	0,019	0,008	0,016
	1	1	0	1	0,047	0,013	0,018	0,036	0,010	0,024	0,019	0,025	0,016
	1	1	1	0	0,023	0,051	0,012	0,036	0,029	0,016	0,038	0,017	0,016
dtt.en.vmbo .t13.028	**dtt.en.vmbo .t13.028_1**	**dtt.en.vmbo .t13.028_2**	**dtt.en.vmbo .t13.028_3**	**dtt.en.vmbo .t13.028_4**	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"	Groep "Onder niveau"	Groep "Op niveau"	Groep "Boven niveau"
	0	0	0	0	0,116	0,013	0,006	0,089	0,010	0,008	0,047	0,008	0,016
	1	1	1	1	0,070	0,152	0,536	0,089	0,190	0,622	0,132	0,380	0,721
	0	1	0	0	0,047	0,013	0,006	0,036	0,010	0,008	0,019	0,008	0,016
	1	0	0	0	0,023	0,013	0,012	0,018	0,010	0,016	0,009	0,017	0,016
	0	0	1	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	0	0	1	0,209	0,101	0,024	0,214	0,067	0,016	0,151	0,033	0,016
	1	1	0	0	0,023	0,051	0,024	0,036	0,029	0,031	0,038	0,033	0,016
	1	0	0	1	0,163	0,443	0,301	0,232	0,505	0,205	0,387	0,397	0,049
	1	0	1	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	1	0	1	0,093	0,025	0,024	0,071	0,029	0,024	0,047	0,033	0,016
	0	1	1	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	0	0	1	1	0,093	0,101	0,024	0,089	0,086	0,016	0,104	0,033	0,016
	0	1	1	1	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	1	0	1	1	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	1	1	0	1	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016
	1	1	1	0	0,023	0,013	0,006	0,018	0,010	0,008	0,009	0,008	0,016

Deze tabel geeft voor elk mogelijk antwoordpatroon de kans dat dit voorkomt in elk beheersingscategorie (op, onder of boven niveau).

3 Adaptieve procedure met itemblokken

3.1 Inleiding

Vorig jaar is onderzoek gedaan naar de proportie correct gediagnosticeerde leerlingen in elke beheersings-categorie (op, onder of boven niveau) met verschillende aantallen items. Om de proporties correct gediagnosticeerde leerlingen zo hoog mogelijk te krijgen met zo min mogelijk items werd ook een adaptieve procedure geëvalueerd. In dat onderzoek zijn we uitgegaan van adaptiviteit op itemniveau. Adaptiviteit op itemniveau is zeer gunstig om zo efficiënt en nauwkeurig mogelijk te meten. Er kleven echter ook psychometrische, technische, toetsinhoudelijke en praktische nadelen aan toetsen op item-niveau (zie onder andere Feskens, Zwitser, Verschoor, 2013; Hendrickson, 2007;Van der Linden & Glas, 2000; Wainer, 2000). Zo is onder andere de noodzaak tot controle over itemgebruik (“item exposure control”) lastig, is de controle over niet-psychometrische eigenschappen van de toets moeilijk te realiseren (zoals inhouds-dekking), is inspectie van de itemvolgorde door toetsdeskundigen om bijvoorbeeld context-effecten te voorkomen niet mogelijk en stelt adaptiviteit op itemniveau zware technische eisen aan het afname systeem.

Een groot nadeel is verder dat bij adaptiviteit op itemniveau geen rekening wordt gehouden met de aanwezigheid van testlets , zoals bij de DTT het geval is. Een testlet is een set van opgaven of opgave-onderdelen die niet onafhankelijk van elkaar (kunnen) worden beantwoord en beoordeeld. De meest bekende vorm van testlets zijn opgaven die een context gemeenschappelijk hebben, bijvoorbeeld een tekst waarover meerdere vragen gesteld worden. Testlets zijn lastig bij het samenstellen van een adaptieve toets en bij het schatten van de vaardigheid, maar hebben inhoudelijk grote voordelen (leestijd, authentieke handeling/materiaal).

Een ander groot nadeel van adaptiviteit op itemniveau is het feit dat men over een omvangrijke gekalibreerde itembank moet beschikken voordat de adaptieve toets afgenomen kan worden. Ook bij de DTT is dit lastig te realiseren, omdat veel items gekalibreerd moeten worden in een omvangrijke pretest (Schouwstra, 2013). De omvang van de pretest kan gereduceerd worden wanneer bij elke afname de items geherkalibreerd kunnen worden.

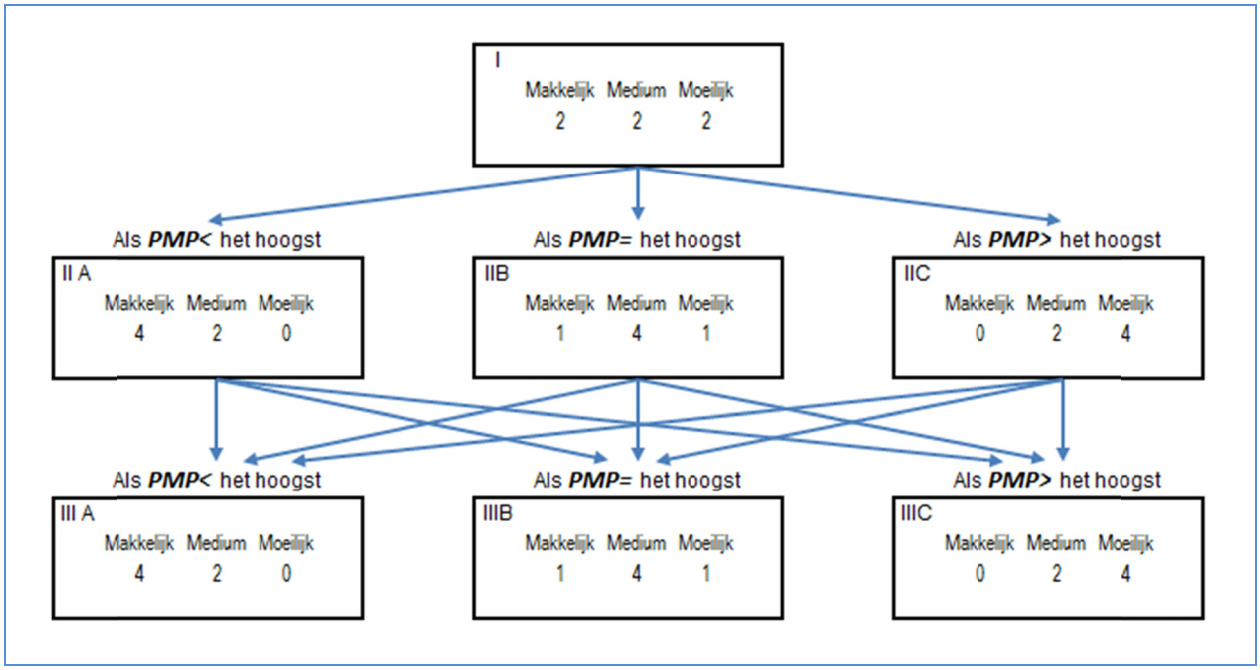
Vanwege deze laatste twee nadelen is gekeken naar een nieuwe adaptieve procedure waarbij gebruik wordt gemaakt van een blokkenstructuur. De vraag is hoe de proporties correct geclassificeerde leerlingen aan de hand van deze nieuwe procedure zich verhouden tot de proporties aan de hand van de niet-adaptieve en adaptieve procedure uit de voorstudie. In de volgende sectie zal eerst de blokkenstructuur worden uitgelegd. Vervolgens zullen de resultaten van de nieuwe procedure worden gegeven en vergeleken met de resultaten van de procedures uit de voorstudies.

3.2 Adaptieve procedure met blokken items samengesteld op grond van itemmoeilijkheid

3.2.1 Procedure eerste simulatie

In het eerste simulatieonderzoek naar de adaptieve procedure met blokken, zijn de blokken items samengesteld op grond van de moeilijkheid van de items, zoals weergegeven in Figuur 3-1. In deze adaptieve procedure krijgt elke leerling eerst het bovenste blok aangeboden dat bestaat uit zes items: 2 makkelijke, 2 medium en 2 moeilijke items. Na het beantwoorden van de items in het bovenste blok, wordt de posterior model probabilities (*PMPs*) voor elk beheersingscategorie berekend. Als de *PMP* voor “onder niveau” ($PMP_<$) het hoogste is, krijgt deze leerling in de tweede ronde blok 2A (zie Figuur 3-1). Als, echter, de *PMP* voor “op niveau” ($PMP_=$) het hoogste is, krijgt de leerling in de tweede ronde blok 2B. Als de *PMP* voor “boven niveau” ($PMP_>$) daarentegen het hoogste is, krijgt de leerling blok 2C in de tweede ronde. In blok A krijgt een leerling 4 makkelijke en 2 medium items aangeboden. In blok B krijgt een leerling 1 makkelijk, 4 medium en 1 moeilijk item aangeboden. In blok C krijgt een leerling 2 medium en 4 moeilijke items aangeboden.

Na de tweede ronde worden de *PMP*s geüpdatet en worden leerlingen volgens hetzelfde principe aan een blok in ronde 3 toegewezen. Na de derde ronde, als de leerling 18 items heeft beantwoord, wordt de diagnose voor de leerling gesteld. De diagnose is gelijk aan de beheersingscategorie met de hoogste *PMP*. Bijvoorbeeld, als de *PMP* van de beheersingscategorie "boven niveau" het hoogste is, dan krijgt de leerling de diagnose "boven niveau".



Figuur 3-1 De adaptieve procedure met blokken samengesteld op grond van itemmoeilijkheid

De blokken worden samengesteld met behulp van de hypothetische itembank uit de voorstudie (Cito, 2013, p. 45), die bestond uit 27 makkelijke, 27 medium en 27 moeilijke items met een grote spreiding.
De succeskansen van makkelijke items zijn variaties van .05 op de kansen .50, .70 en .90 voor onder, op en boven niveau. Voor medium items zijn variaties op de kansen .40, .60 en .80 en voor moeilijke items variaties op .30, .50 en .70 onder op en boven niveau (zie Bijlage 3-1). De items zijn aselect geselecteerd uit de lijsten makkelijke, medium en moeilijke items. Bijvoorbeeld voor het eerste blok zijn aselect twee items uit de lijst met makkelijke items getrokken, twee items aselect uit de lijst met medium items en 2 items uit de lijst met moeilijke item. De succeskansen van de gebruikte items zijn weergeven in Tabel 3-1.

Aan de hand van deze procedure is een simulatie uitgevoerd. De resultaten van deze simulatie zijn vergeleken met de resultaten uit de voorstudie.

Tabel 3-1 Succeskansen van items in de blokken samengesteld op grond van itemmoeilijkheid

		Beheersings-categorie			Beheersings-categorie			Beheersings-categorie		
		Onder niveau <	Op niveau =	Boven niveau >	Onder niveau <	Op niveau =	Boven niveau >	Onder niveau <	Op niveau =	Boven niveau >
Ronde 1	Item 1				0,45	0,65	0,95			
	Item 2				0,50	0,75	0,85			
	Item 3				0,35	0,60	0,85			
	Item 4				0,35	0,55	0,80			
	Item 5				0,35	0,45	0,75			
	Item 6				0,25	0,50	0,70			
		BLOK A			**BLOK B**			**BLOK C**		
Ronde 2	Item 1	0,45	0,70	0,95	0,50	0,70	0,85	0,40	0,55	0,75
	Item 2	0,55	0,75	0,90	0,40	0,60	0,85	0,45	0,55	0,85
	Item 3	0,50	0,65	0,85	0,45	0,55	0,80	0,30	0,45	0,65
	Item 4	0,45	0,75	0,95	0,40	0,60	0,75	0,30	0,50	0,75
	Item 5	0,35	0,55	0,75	0,40	0,55	0,80	0,25	0,50	0,65
	Item 6	0,40	0,55	0,85	0,35	0,50	0,70	0,30	0,55	0,70
		BLOK A			**BLOK B**			**BLOK C**		
Ronde 3	Item 1	0,50	0,70	0,95	0,45	0,75	0,90	0,35	0,65	0,85
	Item 2	0,50	0,65	0,90	0,40	0,65	0,75	0,40	0,65	0,80
	Item 3	0,50	0,75	0,90	0,35	0,60	0,75	0,35	0,45	0,70
	Item 4	0,55	0,70	0,95	0,35	0,65	0,75	0,35	0,50	0,65
	Item 5	0,35	0,60	0,80	0,45	0,65	0,85	0,30	0,50	0,65
	Item 6	0,35	0,55	0,85	0,30	0,45	0,70	0,30	0,55	0,65

3.2.2 Resultaten eerste simulatie

Zoals weergeven in Tabel 3-2, leidde in het onderzoek vorig jaar een niet-adaptieve procedure met 18 items (met een grote spreiding en gelijke PMP's) tot een correcte diagnose bij 78% van de leerlingen onder niveau, 65% van de leerlingen op niveau, en 85% van de leerlingen boven niveau. Een adaptieve procedure met deze 18 items leidde in de voorstudie tot een correcte diagnose bij 87% van de leerlingen onder niveau, bij 80% van de leerlingen op niveau en bij 94% van de leerlingen boven niveau.

De simulatie laat zien dat de adaptieve procedure met blokken, zoals weergeven in Figuur 3-1, na 18 items ertoe leidt dat 83% van de leerlingen onder niveau, 65% van de leerlingen op niveau en 85% van de leerlingen boven niveau juist gediagnosticeerd worden. Dit betekent dat, vergeleken met de niet-adaptieve procedure uit de voorstudie, op en boven niveau ongeveer evenveel leerlingen correct gediagnosticeerd worden. Onder niveau worden ongeveer 5% meer leerlingen correct geclassificeerd. Echter, vergeleken met de adaptieve procedure uit de voorstudie leidt de adaptieve procedure met blokken tot een relatief groot informatieverlies: bij op niveau worden minder correcte diagnoses gesteld (15%).

Tabel 3-2 Percentages correct gediagnosticeerde leerlingen per procedure

	Leerlingen in de beheersingscategorie		
	Onder niveau <	Op niveau =	Boven niveau >
Niet adaptief (onderzoek 2013)	78%	65%	85%
Adaptiviteit met items (onderzoek 2013)	87%	80%	94%
Adaptiviteit met blokken, samengesteld op grond van moeilijkheid	83%	65%	85%
Adaptiviteit met blokken, samengesteld op grond van spreiding	82%	78%	87%

3.3 Adaptieve procedure met blokken items samengesteld op grond van spreiding in succeskansen

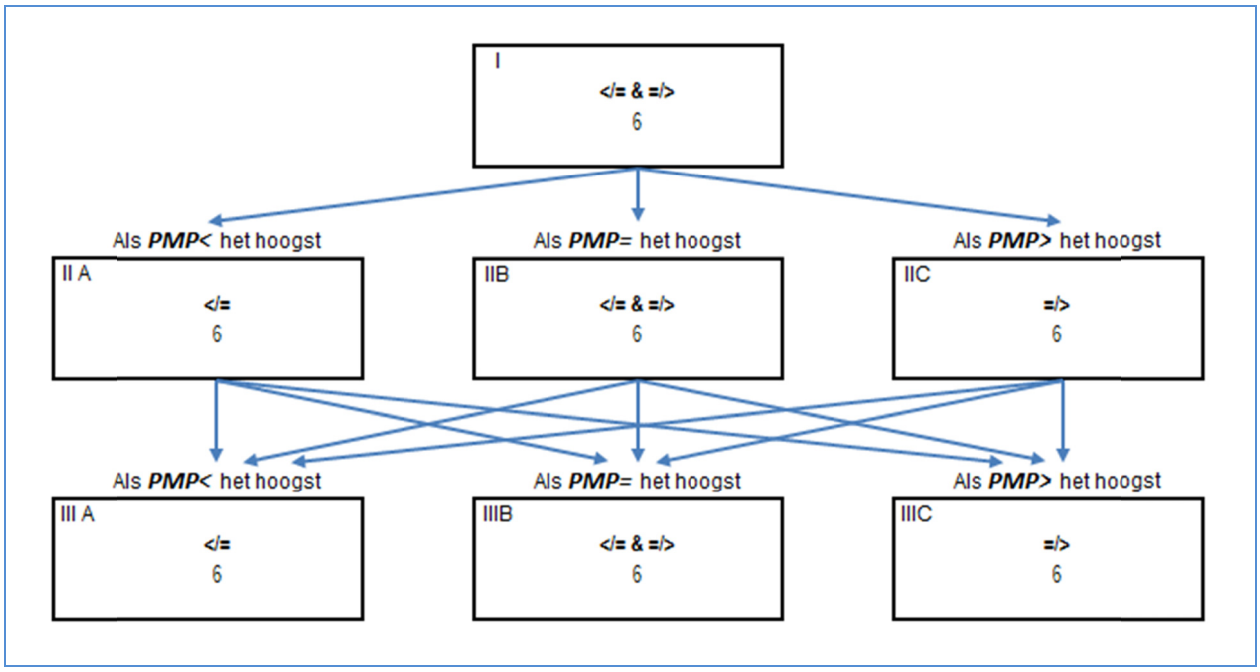
3.3.1 Procedure tweede simulatie

Omdat de adaptieve procedure met blokken niet leidde tot de verwachte resultaten (teveel informatieverlies), is er nagedacht hoe een aanpassing van deze procedure wel tot de gewenste resultaten zou leiden. In de tweede simulatie zijn de blokken samengesteld op grond van de spreiding van de succeskansen tussen de beheersingsniveaus (onderscheidend vermogen), zie Figuur 3-2.

Elke leerling krijgt het bovenste blok aangeboden dat bestaat uit 6 items met een groot verschil tussen de succeskans voor leerlingen uit de beheersingscategorie “onder niveau” en voor leerlingen uit de beheersingscategorie “op niveau”, en een groot verschil tussen de succeskans voor leerlingen uit de beheersingscategorie “op niveau” en voor leerlingen uit de beheersingscategorie “boven niveau”. Nadat de leerling deze 6 items beantwoord heeft, worden de *PMP*s berekend en worden de leerlingen aan blok IIA, IIB of IIC toegewezen:

- Als $PMP_{<}$ het hoogste is, gaat de leerling naar blok IIA en worden 6 nieuwe items aangeboden met een groot verschil tussen de succeskans voor leerlingen uit de beheersingscategorie “onder niveau” en voor leerlingen uit de beheersingscategorie “op niveau” (groot onderscheidend vermogen tussen *onder* en *op* niveau).
- Als $PMP_{=}$ het hoogste is, wordt de leerling toegewezen aan blok B en worden 6 items aangeboden met een groot onderscheidend vermogen tussen zowel *onder* en *op* niveau als tussen *op* en *boven* niveau.
- Als $PMP_{>}$ het hoogste is, komt de leerling in blok IIC terecht en worden 6 items aangeboden met een groot onderscheidend vermogen tussen *op* en *boven* niveau.

Hierna worden de *PMP*s geupdatet en worden leerlingen in de derde ronde toegewezen aan blok IIIA, IIIB of IIIC volgens hetzelfde principe. Na de derde ronde, als de leerling 18 items heeft beantwoord, wordt de diagnose voor de leerling gesteld. De diagnose is gelijk aan de beheersingscategorie met de hoogste PMP. De succeskansen van de gebruikte items zijn weergeven in Tabel 3-3.



Figuur 3-2 De adaptieve procedure met blokken samengesteld op grond van de spreiding in succeskansen van de beheersingscategorieën (onderscheidend vermogen)

Tabel 3-3 Succeskansen van items in de blokken samengesteld op grond van spreiding

		Beheersings-categorie			Beheersings-categorie			Beheersings-categorie		
		Onder niveau <	Op niveau =	Boven niveau >	Onder niveau <	Op niveau =	Boven niveau >	Onder niveau <	Op niveau =	Boven niveau >
Ronde I	Item 1				0,45	0,70	0,95			
	Item 2				0,45	0,65	0,85			
	Item 3				0,35	0,60	0,85			
	Item 4				0,35	0,55	0,75			
	Item 5				0,25	0,50	0,75			
	Item 6				0,25	0,45	0,65			
		BLOK A			**BLOK B**			**BLOK C**		
Ronde II	Item 1	0,45	0,75	0,85	0,45	0,70	0,90	0,45	0,65	0,90
	Item 2	0,35	0,65	0,75	0,35	0,60	0,80	0,35	0,55	0,80
	Item 3	0,25	0,55	0,65	0,25	0,50	0,70	0,25	0,45	0,70
	Item 4	0,50	0,75	0,95	0,50	0,65	0,85	0,55	0,65	0,95
	Item 5	0,40	0,65	0,85	0,40	0,55	0,75	0,45	0,55	0,85
	Item 6	0,30	0,55	0,75	0,30	0,45	0,65	0,35	0,45	0,75
		BLOK A			**BLOK B**			**BLOK C**		
Ronde III	Item 1	0,25	0,55	0,75	0,35	0,55	0,75	0,25	0,45	0,75
	Item 2	0,35	0,65	0,85	0,50	0,70	0,90	0,30	0,45	0,75
	Item 3	0,45	0,75	0,95	0,40	0,60	0,80	0,45	0,65	0,95
	Item 4	0,45	0,75	0,90	0,30	0,50	0,70	0,40	0,55	0,85
	Item 5	0,35	0,65	0,80	0,55	0,75	0,95	0,50	0,65	0,95
	Item 6	0,25	0,55	0,70	0,45	0,65	0,85	0,45	0,65	0,95

3.3.2 Resultaten tweede simulatie

Ook aan de hand van deze procedure is een simulatiestudie gedaan. Deze procedure leidde ertoe dat onder niveau ongeveer 82% van de leerlingen juist geclassificeerd werd, op niveau 78% en boven niveau werd 87% van de leerlingen juist geclassificeerd (zie Tabel 3-2). Dit is een grote vooruitgang ten opzichte van de niet-adaptieve procedure uit de voorstudie. Ten opzichte van de adaptieve procedure uit de voorstudie, is er slechts een kleine verslechtering in correct geclassificeerde leerlingen tussen de 3 en 7% te zien.

3.4 Conclusie

De conclusie die getrokken kan worden is dat de adaptieve procedure met blokken goed werkt wanneer de spreiding van de succeskansen als leidraad wordt genomen en niet de moeilijkheid van de items. Met deze procedure worden, vergeleken met de niet-adaptieve procedure uit de voorstudie, meer leerlingen correct geclassificeerd. Het informatieverlies ten opzichte van de adaptieve procedure uit de voorstudie is minimaal. De adaptieve blokkenprocedure is dus een goed alternatief voor de adaptieve procedure uit de voorstudie. De blokken procedure maakt het mogelijk om rekening te houden met de aanwezige testlets en maakt herkalibratie makkelijker.

3.5 Referenties

Feskens, R., Zwitser, R., & Verschoor, A. (2013). *Eerste gedachten over mogelijkheden voor een Digitale Adaptieve Eindtoets in 2018*, Versie 1.1. Interne notitie, Cito.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

Schouwstra, S. J. (2013). Notitie verwachte toetsduur en afnamescenario's. Arnhem: Cito.

Van der Linden, W. J., & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer* (2nd Edition). Mahwah, NJ: Lawrence Erlbaum Associates.

3.6 Bijlagen

Bijlage 3-1 Lijst met items met hypothetische succeskansen uit het simulatieonderzoek 2013 (Cito, 2013)

	Makkelijke items				Medium items				Moeilijke items		
Beheersings-niveau	<	=	>		<	=	>		<	=	>
1	.50	.70	.90	28	.40	.60	.80	55	.30	.50	.70
2	.55	.70	.90	29	.45	.60	.80	56	.35	.50	.70
3	.50	.75	.90	30	.40	.65	.80	57	.30	.55	.70
4	.50	.70	.95	31	.40	.60	.85	58	.30	.50	.75
5	.45	.70	.90	32	.35	.60	.80	59	.25	.50	.70
6	.50	.65	.90	33	.40	.55	.80	60	.30	.45	.70
7	.50	.70	.85	34	.40	.60	.75	61	.30	.50	.65
8	.55	.75	.90	35	.45	.65	.80	62	.35	.55	.70
9	.55	.70	.95	36	.45	.60	.85	63	.35	.50	.75
10	.50	.75	.95	37	.40	.65	.85	64	.30	.55	.75
11	.45	.65	.90	38	.35	.55	.80	65	.25	.45	.70
12	.45	.70	.85	39	.35	.60	.75	66	.25	.50	.65
13	.50	.65	.85	40	.40	.55	.75	67	.30	.45	.65
14	.55	.75	.95	41	.45	.65	.85	68	.35	.55	.75
15	.45	.65	.85	42	.35	.55	.75	69	.25	.45	.65
16	.55	.65	.90	43	.45	.55	.80	70	.35	.45	.70
17	.55	.70	.85	44	.45	.60	.75	71	.35	.50	.65
18	.45	.75	.90	45	.35	.65	.80	72	.25	.55	.70
19	.50	.75	.85	46	.40	.65	.75	73	.30	.55	.65
20	.45	.70	.95	47	.35	.60	.85	74	.25	.50	.75
21	.50	.65	.95	48	.40	.55	.85	75	.30	.45	.75
22	.55	.75	.85	49	.45	.65	.75	76	.35	.55	.65
23	.55	.65	.95	50	.45	.55	.85	77	.35	.45	.75
24	.45	.75	.95	51	.35	.65	.85	78	.25	.55	.75
25	.45	.65	.95	52	.35	.55	.85	79	.25	.45	.75
26	.45	.75	.85	53	.35	.65	.75	80	.25	.55	.65
27	.50	.65	.85	54	.40	.55	.75	81	.30	.45	.65

4 Interpretatie van het initiële profiel op deelattribuutniveau

4.1 Inleiding

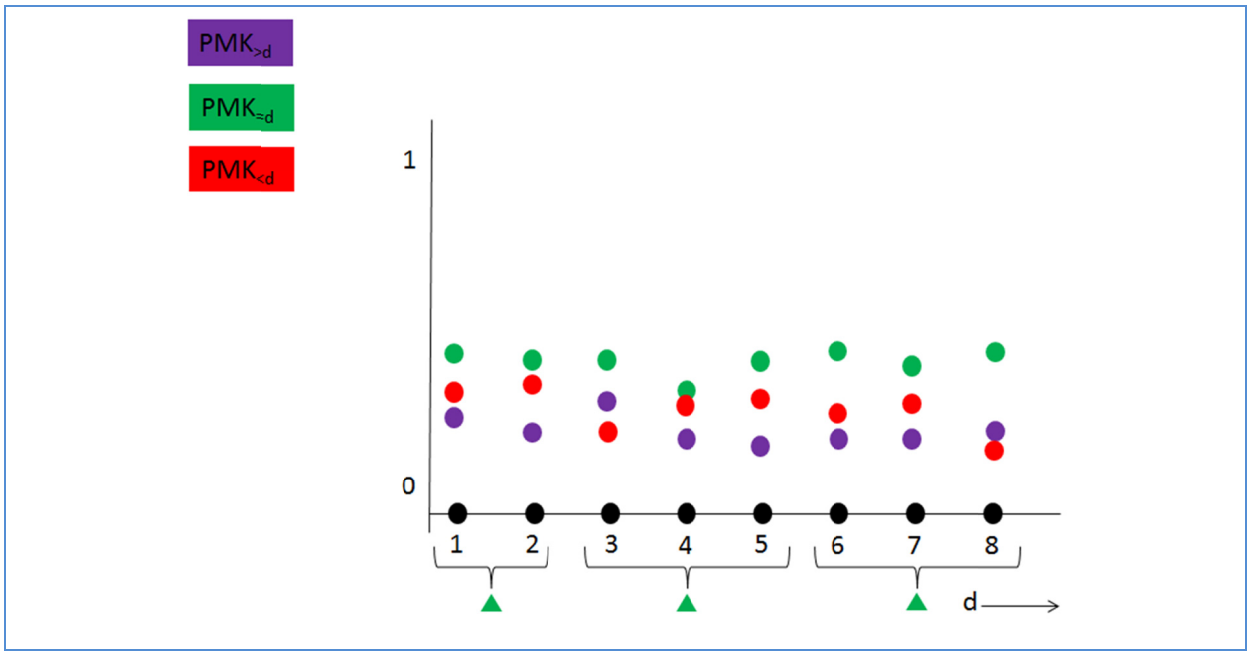
Het gebruikte psychometrische model levert voor elk attribuut drie kansen op: de kans (waarschijnlijkheid) dat de leerling, gegeven zijn antwoorden, in de categorie “onder niveau”, “op niveau”, of “boven niveau” valt. Hoe meer items er per attribuut worden afgenomen, hoe betrouwbaarder deze indeling van de leerlingen in de drie categorieën zal zijn. In de afnametijd die beschikbaar is zal het niet mogelijk zijn om voldoende items af te nemen voor een betrouwbare diagnose van elk deelattribuut. Vorig jaar (Cito, 2013) is daarom een gefaseerde afnamestrategie (in formulevorm) uitgewerkt om het aantal af te nemen items verder te reduceren.

In de eerste fase wordt bij elke leerling elk hoofdattribuut geëvalueerd met behulp van een gelijk maar zo klein mogelijk aantal items per deelattribuut. Met behulp van de afgenomen items wordt voor elk hoofd-attribuut als geheel nagegaan of een leerling “onder niveau”, “op niveau” of “boven niveau” zit. Met dezelfde informatie wordt ook een initieel profiel gemaakt van het beheersingscategorie voor elk deelattribuut. In de tweede fase worden alleen die deelattributen onderzocht wanneer dit geïndiceerd is. Nadere toetsing van een deelattribuut is geïndiceerd als uit het initiële profiel blijkt dat er een aanzienlijke kans is dat de beheersing van de leerling afwijkend is (onder of boven het verwachte niveau). Het proces van het nader toetsen van deelattributen kan doorgaan totdat de tijd op is of totdat er geen geïndiceerde deelattributen meer zijn. In dit hoofdstuk wordt beschreven hoe deze indicatiestelling automatisch in zijn werk zou kunnen gaan.

4.2 Methode voor het afnemen van de diagnostische tussentijdse toets

Op grond van de scores van een leerling op de items behorend bij elk deelattribuut kunnen $PMK_{<d}$, $PMK_{=d}$, en $PMK_{>d}$ voor elk deelattribuut worden berekend. $PMK_{<d}$ is de kans (berekent op grond van de gegeven antwoorden) dat een leerling deelattribuut d onder niveau beheerst. $PMK_{=d}$ en $PMK_{>d}$ hebben analoge interpretaties. Op deelattribuutniveau zijn de PMK_{kd}’s voor $k \in \{<,=,>\}$ de tegenhangers van de PMK_k’s op hoofdattribuut niveau. In Figuur 4-1 tot en met Figuur 4-4 staan voorbeelden van leerling profielen voor acht hypothetische deelattributen die met behulp van de PMK_{kd}’s kunnen worden gecreëerd (zie ook Cito, 2013a, sectie 2.5.2).

In Figuur 4-1 is onder de x-as met behulp van accolades weergegeven welke deelattributen bij hetzelfde hoofdattribuut horen. Door middel van een gekleurde driehoek is weergegeven of de leerling het hoofdattribuut onder, op, of boven niveau beheerst. Rood geeft aan dat $PMK_<$ het grootste is en dus dat de leerling het hoofdattribuut onder niveau beheerst, groen en paars hebben analoge interpretaties voor respectievelijk op en boven niveau. De cirkels in Figuur 4-1 geven per deelattribuut de posterior model kansen voor de drie beheersingsniveaus weer. Het profiel in Figuur 4-1past bij een prototypische op niveau leerling. De leerling beheerst elk hoofdattribuut op niveau en ook voor elke deelattribuut is op niveau de meest waarschijnlijke kwalificatie.

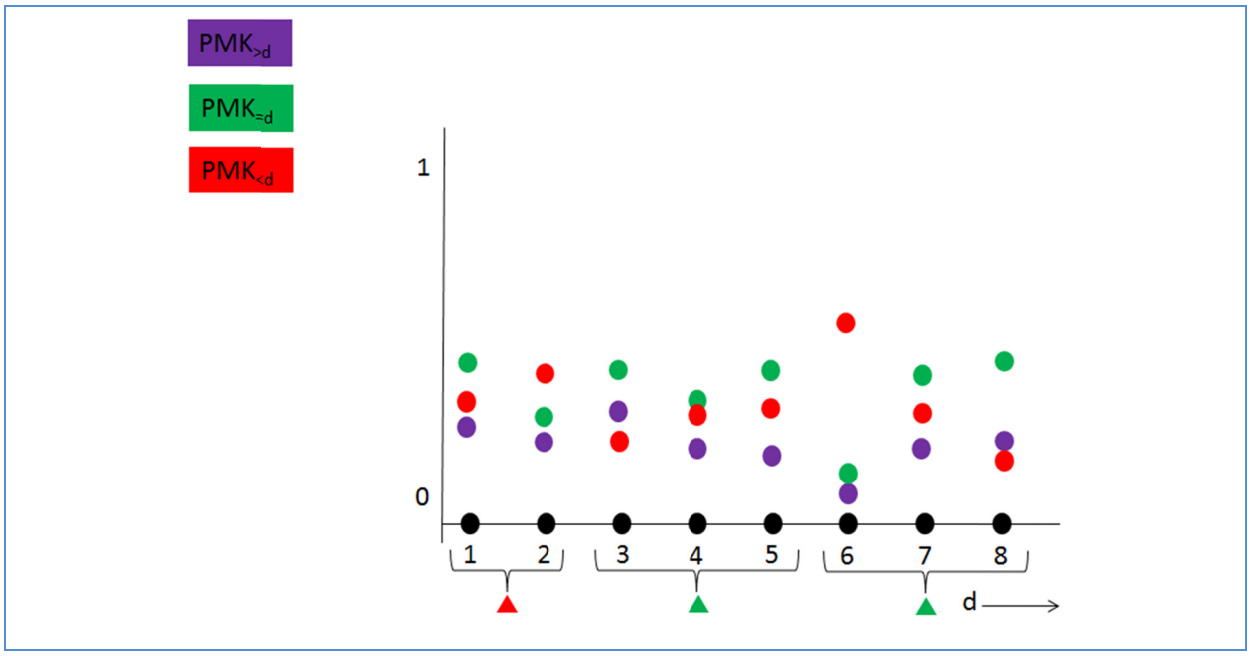


Figuur 4-1 Een prototypische "op niveau" leerling

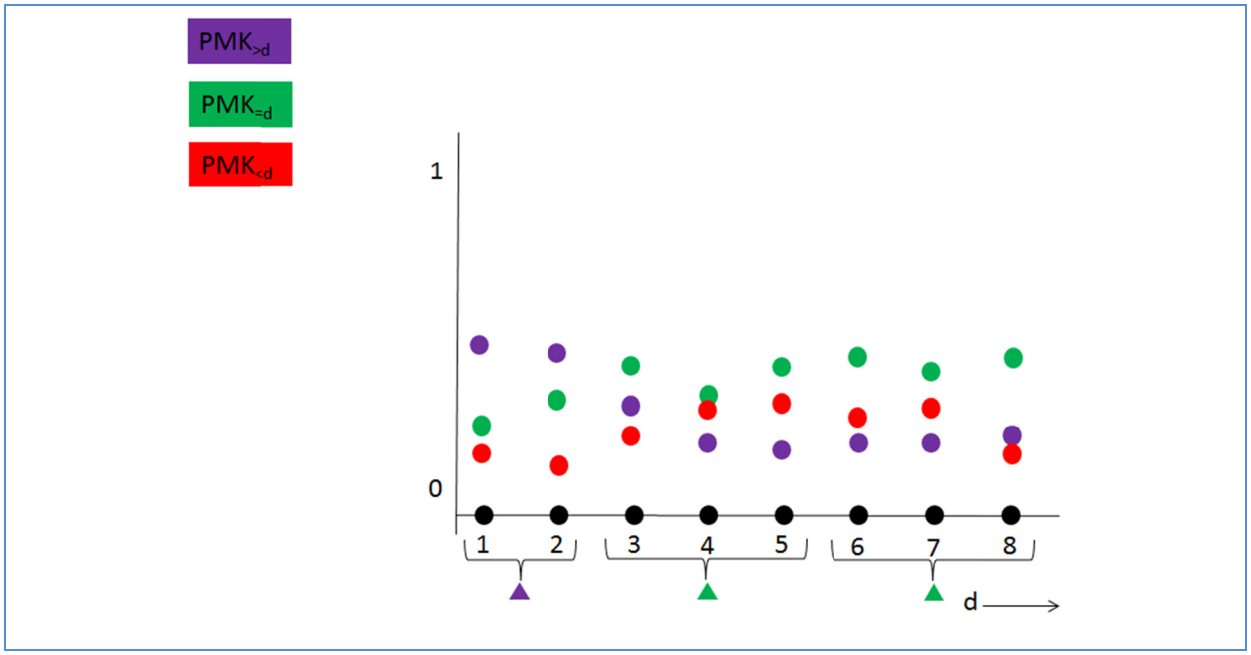
Merk op dat op hoofdattribuut niveau gestreefd wordt naar het correct diagnosticeren van 80% van de leerlingen in elk beheersingsniveau en op deelattribuutniveau naar 60% correcte diagnoses. Uit de figuur kan ook informatie worden verkregen over de zekerheid van de diagnoses op deelattribuutniveau voor de leerling die het betreft. Zo kan worden gezien dat het "groene bolletje" bij deelattribuut 8 verder boven het "paarse en rode bolletje" ligt dan bij de andere deelattributen. Dit betekend dat de zekerheid dat de leerling op niveau presteert groter is voor deelattribuut 8 dan voor de andere deelattributen. Ook kan worden gezien dat de "rode en groene bolletjes" bij deelattribuut 4 dicht bij elkaar liggen. Dit betekent dat de zekerheid dat de leerling op niveau presteert voor deelattribuut 4 kleiner is dan voor de andere deelattributen. Het zou prima zo kunnen zijn dat de leerling op deelattribuut 4 onder niveau presteert.

Het aldus geconstrueerde leerlingprofiel vormt de basis voor twee zaken. Ten eerste kan het worden gebruikt om te bepalen welke deelattributen nader geëvalueerd moeten worden om daarmee de zekerheid omtrent de diagnose van de leerling te vergroten. Ten tweede, en nadat de eerste zaak is afgehandeld, vormt het de basis voor een diagnose van de betreffende leerling gebaseerd op zijn diagnose op hoofd-attribuut niveau en zijn diagnose op de deelattributen die nader zijn onderzocht. De diagnose op de overige deelattributen blijft dermate onzeker (slechts 60% van de leerlingen wordt correct geclassificeerd) dat hier niet al te veel gewicht aan kan worden toegekend bij het maken van de diagnose. Bij de eerste zaak is het dus goed de interessante deelattributen nader te evalueren, om daarmee de zekerheid dat de leerling correct op dit deelattribuut geclassificeerd is te verhogen van 60% naar 80%. Merk op dat er qua tijd slechts ruimte zal zijn om een beperkt aantal deelattributen nader te onderzoeken.

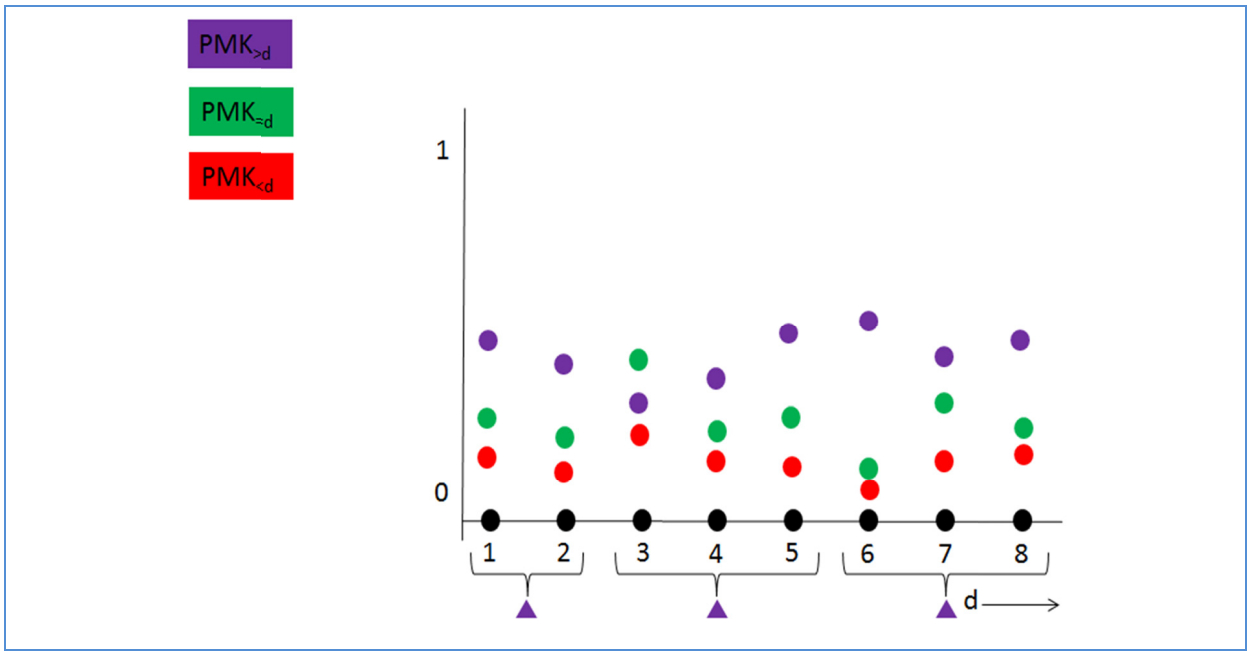
In Figuur 4-1 is deelattribuut 4 interessant omdat het best zo zou kunnen zijn dat deze op niveau leerling op dit attribuut onder niveau zit. In Figuur 4-2 zijn deelattributen 2 en 6 interessant omdat de leerling hierop onder niveau lijkt te zitten terwijl hij op alle andere attributen op niveau lijkt te zitten. Ook deelattribuut 4 is interessant omdat het groene en rode bolletje zo dicht bij elkaar liggen dat het best zo zou kunnen zijn dat de leerling onder in plaats van op niveau zit. In Figuur 4-3 zijn het vierde en vijfde deelattribuut het meest interessant omdat de kans het grootste is dat een leerling op deze deelattributen onder in plaats van op niveau zit en in Figuur 4-4 is het derde deelattribuut interessant omdat dit, alhoewel op niveau, voor deze boven niveau leerling een verbeterpunt aan lijkt te geven. Zoals uit het voorgaande blijkt, is onze strategie voor "op" en "boven" niveau leerlingen die deelattributen te selecteren voor verdere evaluatie die relatief t.o.v. het niveau van de leerling een verbeterpunt aangeven. Voor onder niveau leerlingen zullen we die deelattributen selecteren de relatief t.o.v. het niveau van de leerling aangeven dat de leerling goed presteert. Dit opdat een onder niveau leerling ook op een paar relatief sterke prestaties kan worden gewezen.



Figuur 4-2 Een leerling met verbeterpunten



Figuur 4-3 Een leerling met sterke punten



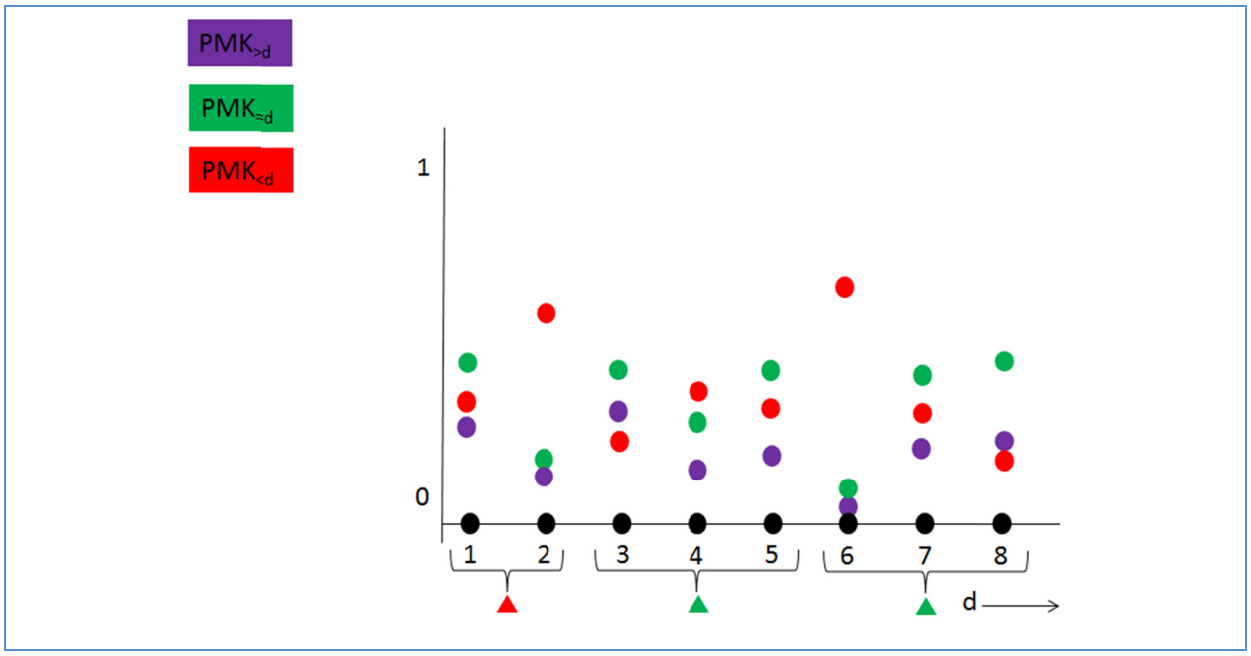
Figuur 4-4 Een grotendeels “boven niveau” leerling

4.3 Algoritme voor de selectie van deelattributen ter verdere evaluatie

Het algoritme bestaat uit twee stappen: (1) bepaal of een leerling onder (rood), op (groen), dan wel boven (paars) niveau presteert; en, (2) bepaal welke deelattributen voor deze leerling nader geëvalueerd moeten worden.

1. Bepaal hoe vaak de rode, groene, en paarse kansen de grootste zijn.
2. Als groen de dominante kleur is, orden dan de deelattributen op grond van de grootte van de rode kansen (van groot naar klein) en hanteer dat als de volgorde waarin deelattributen nader geëvalueerd gaan worden totdat de tijd op is.
3. Als paars de dominante kleur is, orden dan de deelattributen op grond van de grootte van de paarse kansen (van klein naar groot) en hanteer dat als de volgorde waarin deelattributen nader geëvalueerd gaan worden totdat de tijd op is.
4. Als rood de dominante kleur is, orden dan de deelattributen op grond van de grootte van de rode kansen (van klein naar groot) en hanteer dat als de volgorde waarin deelattributen nader geëvalueerd gaan worden totdat de tijd op is.
5. Als er geen unieke dominantie kleur is, doe hetzelfde als in stap (2).

Wanneer dit algoritme op de leerling met verbeterpunten wordt toegepast (Figuur 2-1) worden de deelattributen 2 en 6 nader getoetst. Het vermoeden dat de leerling hierop onder niveau zit word bevestigd, zie Figuur 4-5.



Figuur 4-5 Profiel van de leerling met verbeterpunten uit Figuur 4-2 na de nadere evaluatie van deelattributen 4, 2, en 6

4.4 Conclusie

In dit hoofdstuk is beschreven hoe de Interpretatie van het initiële profiel op deelattribuutniveau en het doortoetsen van deelattributen automatisch zou kunnen verlopen. In het gepresenteerde algoritme is nadere toetsing van een deelattribuut geïndiceerd als uit het initiële profiel blijkt dat er een aanzienlijke kans is dat de beheersing van de leerling relatief verbetering behoeft (relatieve verbeterpunten).

In het algortime ligt dus de nadruk op de “onder niveau” deelattributen, omdat verwacht wordt dat de aanpak van deze verbeterpunten tot hogere prestaties zal leiden. Het kan echter wenselijk zijn om een gelijke focus te hebben op verbeterpunten en sterkte punten in het doortoetsen. Verder is in het algoritme is nog geen rekening gehouden met het feit dat het profiel al dusdanig zeker kan zijn dat doortoetsen niet nodig is. In het vervolgonderzoek zal daarom gewerkt worden aan een alternatief algoritme. Vervolgens zullen de algoritmes met gesimuleerde data geëvalueerd worden, zodat een overwogen keuze gemaakt kan worden.

5 Standaardbepaling door experts na de eerste pretest van de DTT

Bij de eerste pilotafname van de DTT in 2016 zal voor elke leerling een diagnostisch profiel met onderkennende diagnoses opgeleverd worden voor schrijfvaardigheid Nederlands en Engels, waarbij per hoofdattribuut, wordt aangegeven in welke beheersingscategorie de leerling valt: onder, op of boven niveau (Cito, 2014). Voor wiskunde wordt een verdiepende diagnose voor twee domeinen opgeleverd, waarbij in beide domeinen per vakdidactisch aspect wordt aangegeven in welke beheersingscategorie de leerling valt (onder, op, of boven niveau). Om de grenzen tussen de beheersingscategorieën goed te kunnen plaatsen, zal er een toetsgeoriënteerde standaardbepaling door inhoudelijk experts plaatsvinden.

Deze grenzen zullen voor alle vakken (Nederlands, Engels en wiskunde), vaardigheden en alle leerwegen (vmbo-bb, vmbo-kb, vmbo-gt, havo, vwo) apart bepaald worden. Dit betekent dat er in 2015 voor 15 toetsen een standaardbepaling zal plaatsvinden waarin voor elk van de hoofdattributen van deze toets twee standaarden worden bepaald, namelijk de grens tussen onder en op niveau en de grens tussen op en boven niveau.

5.1 Standaardbepalingsmethode

De standaardbepalingsmethode die hiervoor gebruikt zal worden is de 3DC-methode, ofwel de *Data-Driven Direct Consensus* methode. 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).
De 3DC-methode heeft veel gelijkenissen met de *Direct Consensus*-methode (Sireci, Hambleton & Pitoniak, 2004), maar is evenals als de *Bookmark*-methode (Mitzel, Lewis, Patz & Green, 2001) datagebaseerd.

Om de 3DC-methode toe te kunnen passen, wordt verondersteld dat een toets uit meerdere opgaven bestaat en dat deze in te delen zijn in een aantal clusters met een vergelijkbaar aantal scorepunten.
Per cluster wordt dan bepaald hoeveel scorepunten behaald moeten worden om op het beoogde niveau te presteren. De som van het aantal scorepunten per cluster geeft dan het aantal scorepunten dat behaald moet worden op de gehele toets. Bij de DTT is niet zozeer sprake van scorepunten, maar van een bepaald aantal itemresponsen die correct kunnen zijn.

5.1.1 De 3DC-methode bij de DTT

De DTT bestaat uit meerdere opgaven die in te delen zijn in een aantal clusters met een vergelijkbaar aantal itemresponsen. Per cluster kan dan bepaald worden hoeveel itemresponsen correct moeten zijn om in de beoogde beheersingscategorie te vallen. De som van het aantal correcte itemresponsen per cluster geeft dan het aantal itemresponsen dat correct moet zijn om voor de gehele vaardigheid in de beoogde beheersingscategorie te vallen.

Bij DTT worden idealiter de deelattributen als clusters ingezet om zo de grenzen voor de hoofdattributen te berekenen. Om te bepalen wat de grenzen zijn tussen de beheersingscategorieën voor vaardigheid als geheel (schrijfvaardigheid bij de talen), kunnen de hoofdattributen juist als clusters ingezet worden.

Om de pretest in omvang te beperken (en het benodigd aantal leerlingen) worden de items niet in één grote pretest afgenomen, maar in twee kleinere pretesten (Schouwstra, 2014). Ook het diagnostisch profiel wordt in twee pilots opgebouwd. Bij de eerste pilot zal bij de talen een onderkennende diagnose voor alle hoofdattributen en een diagnose voor schrijfvaardigheid als geheel gegeven worden. Bij de tweede pilot zullen ook verdiepende diagnoses voor de deelattributen gegeven kunnen worden.

Bij wiskunde daarentegen zal een verdiepende diagnose voor de vakdidaktische aspecten binnen twee (van de vier a vijf) domeinen gegeven worden. Voor wiskunde als geheel kan nog geen diagnose gegeven worden. Bij de tweede pilot kunnen verdiepende diagnoses voor vier (tot vijf) domeinen gegeven worden.

Deze stapsgewijze opbouw betekent dat ook de standaardbepaling in twee stappen opgebouwd moet worden.

5.2 Werkwijze voor de eerste DTT standaardbepaling

5.2.1 Schrijfvaardigheid

Aangezien we in de eerste pilot grenzen voor schrijfvaardigheid per hoofdattribuut zullen gebruiken, ligt het voor de hand om de deelattributen als clusters in te zetten om zo een goede inschatting te kunnen maken van de juiste grenzen per hoofdattribuut. Echter, de itembank voor de eerste pilot zal voor elk deelattribuut items (computerschermen) bevatten die 6 tot 8 itemresponsen opleveren. Het maximaal aantal item-responsen per hoofdattribuut is slechts 24. Dit betekent dat de clusters eigenlijk te klein zijn om goed twee grenzen op te kunnen bepalen. Om desondanks een zo correct mogelijke grenswaarde voor elke hoofdattribuut te kunnen bepalen, zullen we toch de deelattributen als clusters voor het bepalen van de hoofdattributen inzetten. Doordat we dan informatie per deelattribuut verzamelen, kunnen we na de tweede pretest deze grenzen valideren met meer data.

Om de grenzen voor schrijfvaardigheid als geheel te bepalen zullen we de hoofdattributen als clusters inzetten. Dit betekent dat er in een cluster 24 items zitten. Aangezien we hier twee grenzen op bepalen, is dit een acceptabele grootte.

5.2.2 Wiskunde

Bij wiskunde zullen we de eerste pilotafname grenzen voor de vakdidkatische aspecten (structuur, meerduidigheid en samenhang) binnen twee domeinen (domein D: Meten en meetkunde en domein E: Verbanden en formules) gebruiken. We zullen dus de vakdidaktische aspecten als clusters in moeten zetten om een goede inschatting te krijgen per domein. Elk cluster (vakdidaktisch aspect binnen een domein) zal 24 itemresponsen bevatten. Aangezien we niet alle domeinen meten in de eerste pretest kunnen we geen grenzen voor wiskunde als geheel bepalen. Pas na de tweede pretest kunnen we de standaardbepaling uitbreiden.

5.3 Aanpak tijdens standaardbepalingssessie

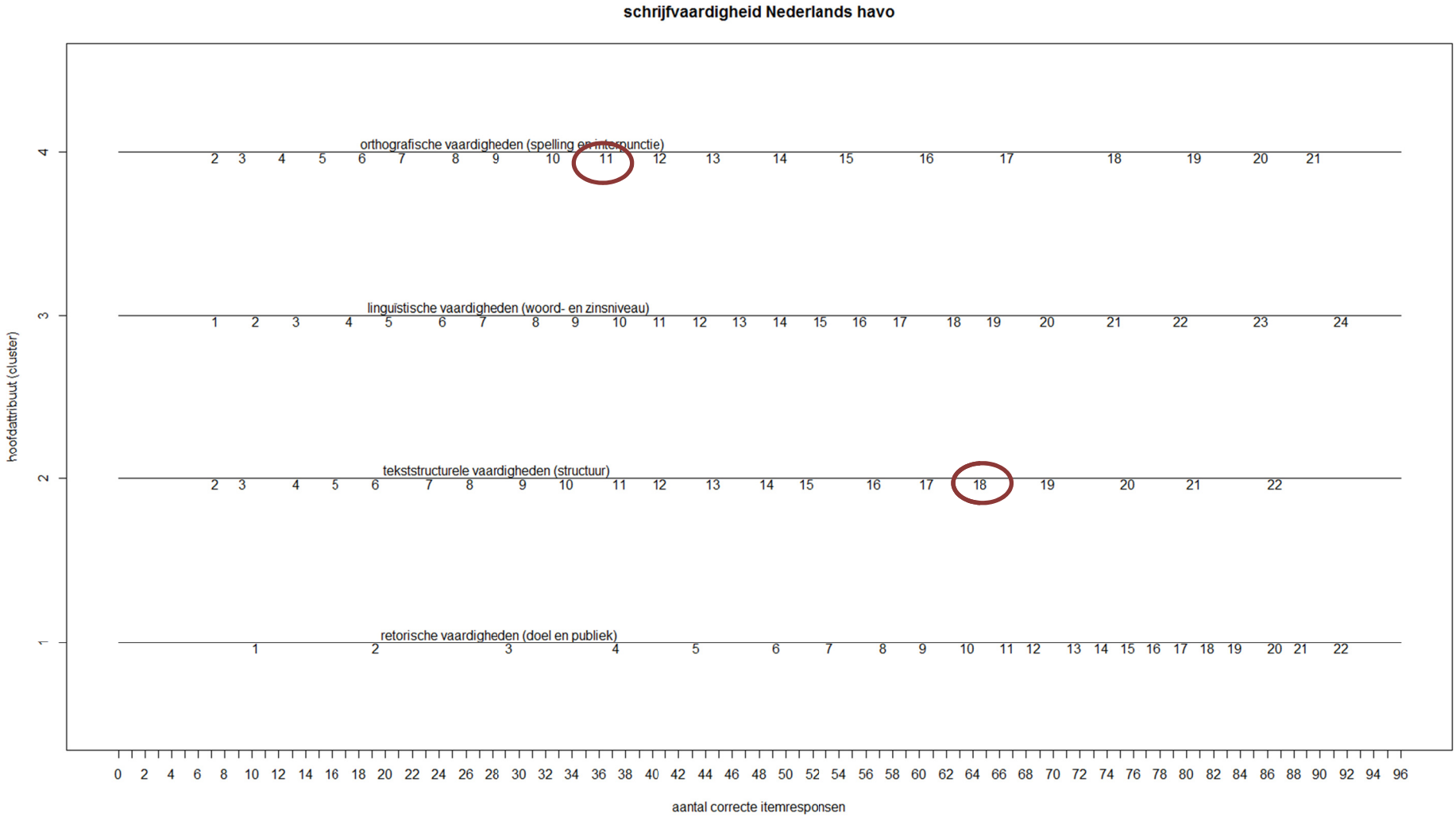
Om de validiteit en betrouwbaarheid van de standaardbepaling te verhogen, hoeven de experts bij de 3DC-methode hun inschatting van de grens tussen onder en op niveau en de grens tussen op en boven niveau per cluster niet alleen op inhoudelijke descriptoren te baseren, maar kunnen zij ook zien hoe de clusters zich empirisch tot elkaar verhouden. De aanpak tijdens de standaardbepalingssessie zal uitgelegd worden aan de hand van een mogelijk standaardbepalingsformulier.

In Figuur 5-1 staat een mogelijk standaardbepalingsformulier weergegeven voor het domein schrijf-vaardigheid bij Nederlands. Het aantal itemresponsen per hoofdattribuut is 24. De lijnen representeren de schalen die bij de clusters (hoofdattributen) horen. Op de horizontale as is de schaal voor de volledige toets weergegeven. Voor de gehele toets zijn dus 0 tot 96 (24*4) itemresponsen. Als een leerling op het hoofdattribuut “Doel en publiek” 14 itemresponsen correct beantwoord is het verwacht aantal correcte itemresponsen op de gehele toets gelijk aan 74.

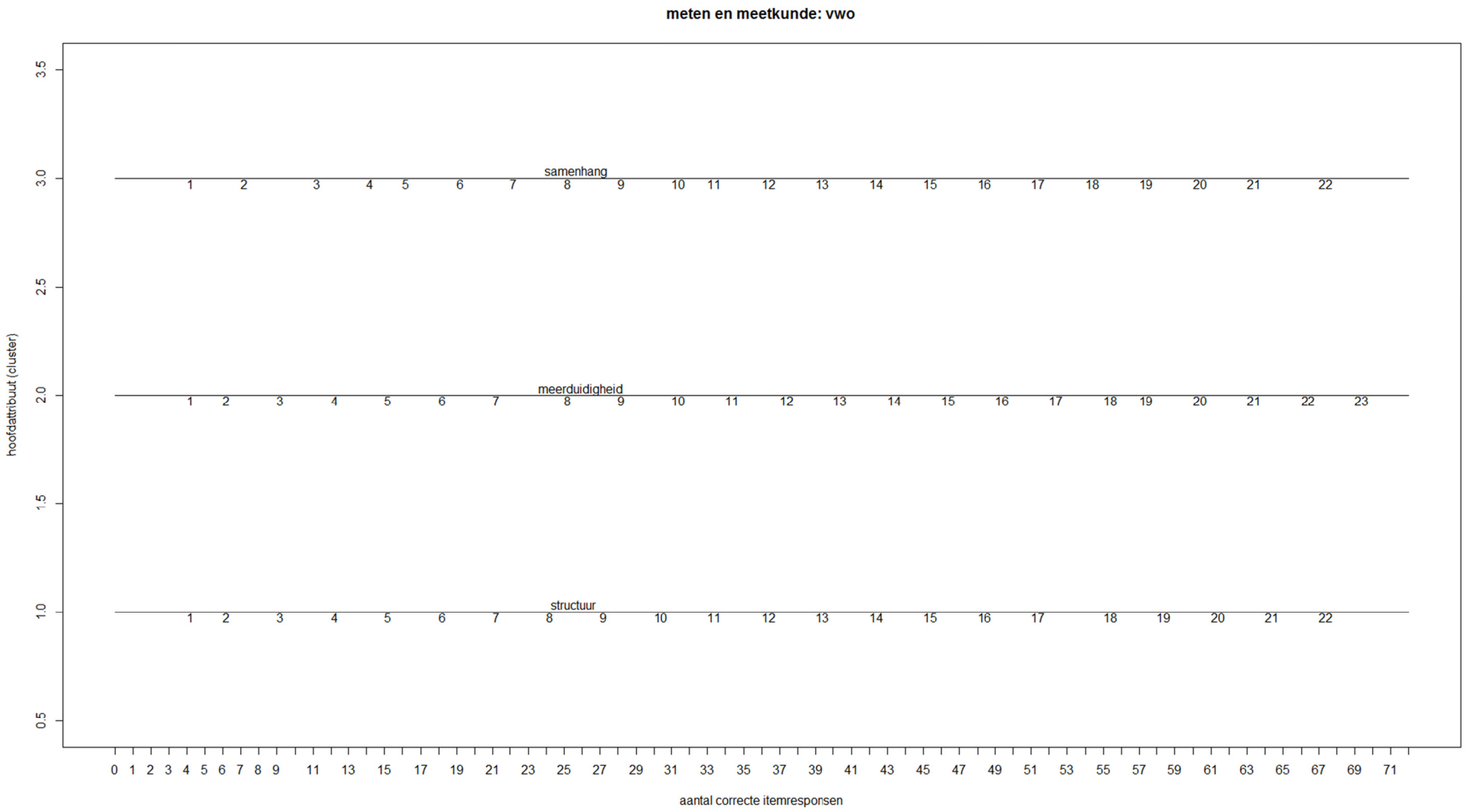
Bij geen van de hoofdattributen staat de waarde “24” gegeven. Dit betekent dat het bij de geteste leerling-populatie niet realistisch is dat een leerling alle itemresponsen correct beantwoordt. Dit geldt ook voor de waarde “1” die soms ontbreekt. Als de “1” ontbreekt, hebben alle leerlingen dus minstens 1 itemrespons correct beantwoord. Ook valt direct af te lezen dat het hoofdattribuut “Doel en publiek” relatief veel moeilijke opgaven bevat. Een leerling die slecht 8 correcte itemresponsen voor “Doel en publiek” heeft, heeft een verwacht aantal van 15 of 16 correcte itemresponsen op de andere hoofdattributen[2].

[2] Let op: dit standaardbepalingsformulier is fictief en puur ter illustratie opgenomen. Het zegt dus niets over de werkelijke moeilijkheid van de hoofdattributen.

Figuur 5-1 Standaardbepalingsformulier voor schrijfvaardigheid Nederlands waarin de verschillende hoofdattributen op basis van een IRT-model aan elkaar gerelateerd zijn.



Figuur 5-2 Standaardbepalingsformulier voor wiskunde waarin de verschillende hoofdattributen op basis van een IRT-model aan elkaar gerelateerd zijn.



De experts zien één oogopslag op welke clusters de leerlingpopulatie goed presteert en op welke minder. Op deze manier is de inschatting gebaseerd op de inhoud en de empirie. Indien bijvoorbeeld een expert op basis van de inhoudelijke descriptoren van het getoetste niveau inschat dat de grens voor het hoofdattribuut “Tekststructuur” bij 18 correcte itemresponsen moet liggen en voor het hoofdattribuut “Spelling en interpunctie” bij 11 correcte itemresponsen, dan wordt op basis van de empirie duidelijk dat deze verdeling van het verwachte aantal correcte itemresponsen niet waarschijnlijk is, omdat ze bij twee erg afwijkende vaardigheidsniveaus horen (zie Figuur 5-1).

De moeilijkheid van alle items die meegaan in de standaardbepaling wordt bepaald door alle items op één vaardigheidsschaal te leggen met behulp van een model uit de itemresponstheorie. Hiervoor zullen de leerlingantwoorden ten bate van de standaardbepaling geanalyseerd worden met het One Parameter Logistic Model (Verhelst, Glas & Verstralen, 1995).

5.3.1 Standaardbepalingssessie schrijfvaardigheid

De grenzen voor het gehele domein schrijfvaardigheid worden in twee rondes vastgesteld. Voor het bepalen van de grenzen voor de hoofdattributen, zal echter met slechts één ronde gewerkt worden: dit in verband met het lage aantal itemresponsen per cluster. In de eerste ronde wordt aan de inhoudelijk experts de volgende vraag gesteld: “Hoeveel itemresponsen beantwoordt de leerlingen naar verwachting correct op dit cluster als zijn/haar vaardigheid zich precies op de grens onder/op niveau van attribuut XX bevindt?” De experts omcirkelen deze waarde voor elk cluster in Figuur 5-1. Vervolgens beantwoorden ze dezelfde vraag maar dan voor de grens op/boven niveau.

Voor de bepaling van de grenzen voor het gehele domein schrijfvaardigheid, wordt vervolgens gediscussieerd over de uitkomsten. Alle waardes die de experts per cluster hebben genoteerd worden samengevoegd en gepresenteerd. Zo kunnen de experts zien hoe zij oordelen ten opzichte van de andere experts. Vervolgens wordt er gediscussieerd over de uitkomsten door per cluster een aantal experts met een relatief hoge grenswaardes en een aantal experts met een juist relatief lage grenswaardes aan het woord te laten om hun grenzen toe te lichten. Daarna volgt de tweede beoordelingsronde waarin de experts hun grenswaardes kunnen bijstellen.
Door vervolgens voor elke expert de grenswaardes per cluster te sommeren, wordt de grenswaarde van de expert op de gehele vaardigheid bepaald. Door grenswaardes van alle experts naast elkaar te leggen kan worden bepaald in welke range de uiteindelijke grenswaarde zal moeten liggen. Tenslotte zal de experts gevraagd worden om gezamenlijk te proberen een definitieve waarde te vinden die voor allen acceptabel is.

Ondanks dat er voor de bepaling van de grenzen voor de hoofdattributen geen aparte discussie komt per cluster, wordt elk van de hoofdattributen natuurlijk wel grondig besproken als cluster van de gehele vaardigheid. Op deze manier worden de uitkomsten voor de grenzen per hoofdattribuut toch goed onderbouwd.

5.3.2 Standaardbepalingssessie wiskunde

De grenzen voor de domeinen (Domein Meten en meetkunde, Domein Verbanden en formules) worden in twee rondes vastgesteld. In de eerste ronde wordt aan de inhoudelijk experts de volgende vraag gesteld: “Hoeveel itemresponsen beantwoordt de leerlingen naar verwachting correct op dit cluster (vakdidaktisch aspect) als zijn/haar vaardigheid zich precies op de grens onder/op niveau van domein XX bevindt?” De experts omcirkelen deze waarde voor elk cluster in een figuur als Figuur 5-2. Vervolgens beantwoorden ze dezelfde vraag, maar dan voor de grens op/boven niveau.

Voor de bepaling van de grenzen voor het gehele domein, wordt vervolgens gediscussieerd over de uitkomsten. Alle waardes die de experts per cluster hebben genoteerd worden samengevoegd en gepresenteerd. Zo kunnen de experts zien hoe zij oordelen ten opzichte van de andere experts. Vervolgens wordt er gediscussieerd over de uitkomsten door per cluster een aantal experts met een relatief hoge grenswaardes en een aantal experts met een juist relatief lage grenswaardes aan het woord te laten om hun grenzen toe te lichten. Daarna volgt de tweede beoordelingsronde waarin de experts hun grenswaardes kunnen bijstellen. Door vervolgens voor elke expert de grenswaardes per cluster te sommeren, wordt de grenswaarde van de expert op het domein bepaald.
Door grenswaardes van alle experts naast elkaar te leggen kan worden bepaald in welke range de uiteindelijke grenswaarde zal moeten liggen. Tenslotte zal de experts gevraagd worden om gezamenlijk te proberen een definitieve waarde te vinden die voor allen acceptabel is.

5.4 Globale planning

5.4.1 Experts voor de standaardbepalingssessie

Voor elke standaardbepaling zijn 10 a 12 experts nodig. Het is erg belangrijk dat de inhoudelijk experts mensen zijn die recente ervaring hebben met het onderwijzen van de doelgroep. Verder is het verstandig om de verschillende niveaus van een vak door, in ieder geval deels, dezelfde groep experts te laten bepalen. Voor elk leerweg zal er 50% itemoverlap zijn met de aanliggende niveaus. Beoordeling door dezelfde experts voorkomt hopelijk dat er voor de lagere niveaus items zijn die goed gemaakt moeten worden die bij een hoger niveau niet goed gemaakt hoeven worden. Of dit ook daadwerkelijk het geval is, zal na de standaardbepaling worden nagegaan.

5.4.2 Periode en duur van de standaardbepalingsessies

De standaardbepaling zal plaatsvinden in april/mei 2015. Aangezien de standaardbepaling datagebaseerd is, is het belangrijk dat deze datum na de afnameperiode in februari 2015 ligt.

De standaardbepaling per vak per leerweg zal plaatsvinden op één dag. Er zullen dus 3 vakken * 5 leerwegen = 15 sessies van één dag nodig zijn om alle standaarden te kunnen bepalen.

5.5 Referenties

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.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Schouwstra, S. J. (2014). Notitie door doorontwikkeling DTT, dd. 05-05-2014.Arnhem: Cito.

Sireci, S.G., Hambleton, R.K., & Pitoniak, M.J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review*, 15, 21-25.

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 Digitale diagnostische toetsing van wiskunde: onderzoek van een nieuwe omgeving

6.1 Probleemstelling

Digitale toetsing staat wereldwijd sterk in de belangstelling (Williamson, Mislevy, & Bejar, 2006). Digitale toetsing biedt aantrekkelijke praktische mogelijkheden voor flexibele afname en automatische scoring; daarnaast liggen er ook inhoudelijke kansen, bijvoorbeeld voor adaptieve toetsing en voor het gebruik van nieuwe interactieve vraagvormen die met pen en papier niet of moeilijk te realiseren zijn.

Ook voor het toetsen van wiskundige kennis en vaardigheden wordt meer en meer gebruik gemaakt van digitale middelen (Fife, 2011; Stacey & Wiliam, 2013). Het uitgangspunt van de Diagnostische Tussentijdse Toets (DTT) wiskunde is dan ook dat deze, mede vanwege het adaptieve en diagnostische karakter, digitaal zal worden afgenomen. Het probleem hierbij is echter dat een rijke en gevarieerde wiskundetoets hoge vakspecifieke eisen stelt aan de auteursomgeving waarin de items worden ontwikkeld en aan de afname-omgeving waarin de kandidaten de toets maken. Denk bijvoorbeeld aan het invoeren van formules, aan het werken met grafieken en meetkundige figuren, en aan het opslaan en automatisch coderen van tussenstappen (Drijvers, 2013).
Deze problemen zijn nog niet bevredigend opgelost, maar zijn wel cruciaal in het realiseren van toetstechnisch adequate wiskundetoetsen. Voor Cito, en de unit Centrale Toetsing en Examens in het bijzonder, is deze problematiek uitermate belangrijk, niet alleen voor het DTT-project maar ook in het licht van de rekentoetsen RVO, COE en de mogelijke verdere digitalisering van eindexamens (CBT).

De huidige situatie is dat Cito in een overgangsfase zit. Items voor digitale toetsen worden geconstrueerd in de eigen auteursomgeving Questify Builder, die nog in ontwikkeling is en waarin de wiskundige middelen vooralsnog beperkt zijn. Voor de DTT wiskunde is Facet de beoogde afname-omgeving. Facet wordt onder regie van CvE ontwikkeld en biedt momenteel onvoldoende mogelijkheden voor de afname van digitale toetsen wiskunde. Vanzelfsprekend is de intentie dat Facet op korte termijn meer wiskundespecifieke faciliteiten gaat bieden.

Het doel van het hier beschreven onderzoek is om aan de hand van een alternatieve digitale omgeving voor het ontwerpen en afspelen van digitale wiskundetoetsen (1) na te gaan uit welk type items een diagnostische toets kan bestaan als er meer digitale mogelijkheden zijn, en (2) te onderzoeken of afname van deze items leidt tot een adequate en efficiënte toetsing van het leerlingmodel. Het onderzoek kent dan ook de volgende onderzoeksvragen.

1. Welke itemtypes voor DTT wiskunde ontstaan door de beschikbaarheid van meer vakspecifieke opties in de auteursomgeving?
2. Leidt de afname van deze items tot een adequate en efficiënte toetsing van het leerlingmodel?

6.2 Methode

Binnen het onderzoek, dat een kleinschalig en exploratief karakter heeft, worden items ontworpen en voorgelegd aan een beperkt aantal leerlingen van klas 3 van havo en vwo.

6.2.1 De digitale omgeving

Als alternatief voor de combinatie Questify Builder - Facet is in dit onderzoek gekozen voor de Digitale Wiskunde Omgeving (DWO). De DWO integreert een auteursomgeving met een afname-omgeving en een leerling-volgsysteem en biedt mogelijkheden voor het werken met formules, grafieken en meetkundige figuren. Tevens worden tussenantwoorden opgeslagen en zijn er mogelijkheden voor automatische beoordeling van wiskundig equivalente antwoorden en van meetkundige constructies. In een internationaal vergelijkend onderzoek is de DWO als zeer goede omgeving aangemerkt (Bokhove & Drijvers, 2010). CvE onderzoekt momenteel in hoeverre de DWO binnen Facet kan worden ingepast voor wiskundige features op het gebied van meetkunde en grafieken. Alle grote Nederlandse educatieve uitgeverijen gebruiken de DWO en de SLO zet de DWO in voor het onderwijs in de statistiek op de zogeheten ‘doorloopscholen’ van de nieuwe curricula wiskunde per 2015. Het Freudenthal Instituut van de Universiteit Utrecht, de ontwikkelaar van de DWO, heeft de DWO voor dit onderzoek gratis ter beschikking gesteld en de uitvoering in technisch opzicht ondersteund.

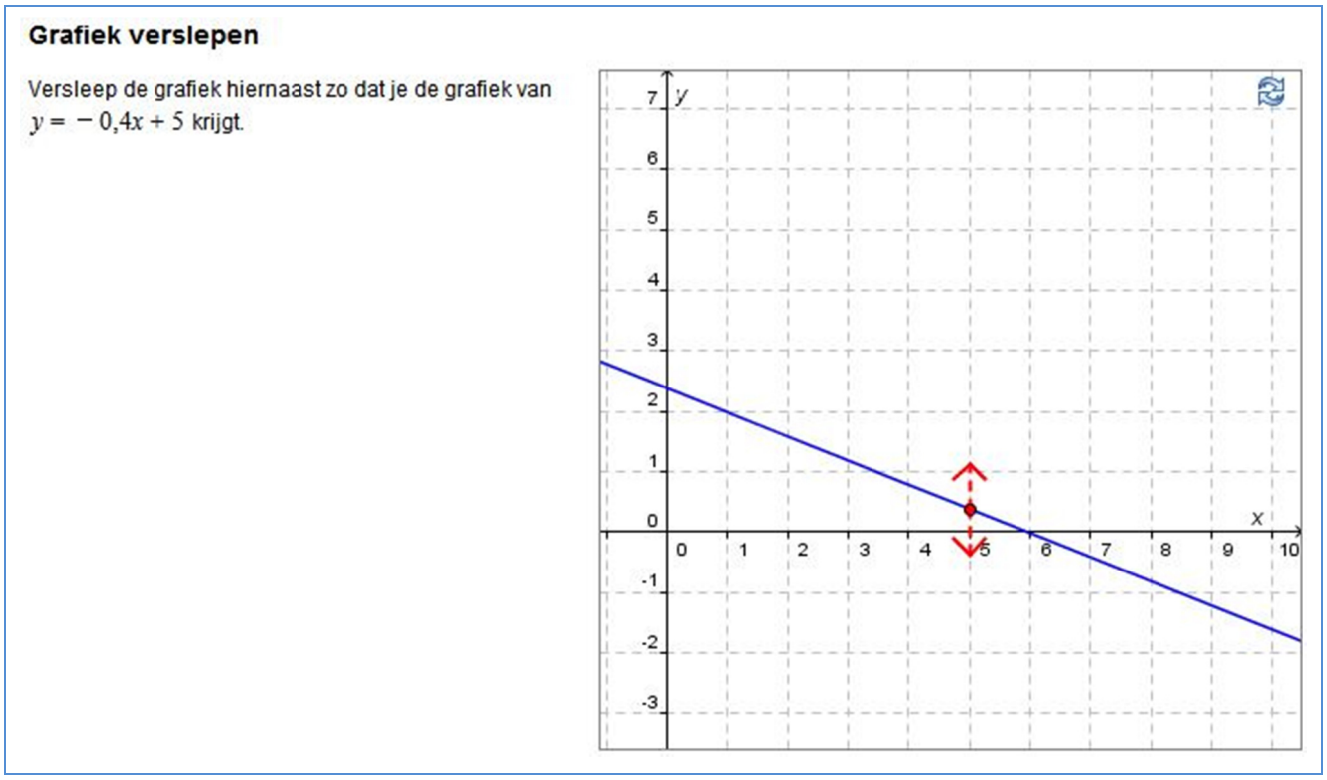
6.2.2 Itemontwerp

In de eerste fase van het onderzoek zijn 42 items in de DWO ontworpen. Het gaat om items die gebruik maken van wiskundige mogelijkheden die op dat moment niet in Questify of Facet beschikbaar waren. De items betreffen de domeinen D: Meten en meetkunde en E: Verbanden en formules, omdat dit de domeinen zijn waarin de technische beperkingen van Questify en Facet het sterkst gelden (zie de beschrijving van de tussendoelen onderbouw vmbo en havo-vwo van SLO). Aangezien de geavanceerde digitale mogelijkheden naar verwachting vooral relevant zijn voor havo en vwo, is het ontwerp op deze onderwijsniveaus gericht, dus op klas 3hv.
De ideeën voor de items zijn afkomstig van de constructiegroepen DTT, de try-out 2013 en de toetsdeskundigen van het projectteam DTT wiskunde. De 42 items zijn gelijk verdeeld over de twee domeinen en over de verschillende cellen van het SOmS vakdidactische leerlingmodel. Tabel 6-1 geeft deze opzet schematisch weer. Figuur 6-1 bevat bij wijze van voorbeeld één van de items.

Tabel 6-1 Itemverdeling over domeinen en SOmS-kenmerken

	Structuur	Objectvorming	Samenhang
Domein D: Meten en meetkunde	7	7	7
Domein E: Verbanden en formules	7	7	7

Figuur 6-1 Item B2-2: Grafiek verslepen



Voor het afnamedesign zijn deze 42 items gesplitst in twee boekjes van elk 21 items, die voor de helft domein D en voor de helft domein E betreffen, de SOmS-kenmerken op een evenredige manier overdekken en naar verwachting een vergelijkbare moeilijkheidsgraad hebben. Elk boekje was bedoeld om in een toetstijd van 50 minuten doorgewerkt te kunnen worden; in praktijk bleek de benodigde toetstijd groter te zijn. Voor een kijk op de online implementatie kunt u contact opnemen met de auteurs van dit rapport.

Behalve deze toets is een evaluatieve vragenlijst voor docenten ontwikkeld en een format voor rapportage van de resultaten aan de docenten. Deze staan in de bijlagen bij dit hoofdstuk.

6.2.2.1 Participanten

Voor de DTT-DWO try-out is geworven onder deelnemers aan een eerder project rond het gebruik van de DWO (DPICT, zie http://www.fisme.uu.nl/dpict/) en onder scholen in het netwerk van de projectuitvoerders. Op deze scholen was naar verwachting zowel de technologische infrastructuur als de animo voor deelname aan de try-out als enige ervaring met de DWO bij de leerlingen aanwezig. Randvoorwaarde was dat deze scholen niet aan de reguliere DTT try-out deelnemen. In totaal hebben zich 26 klassen aangemeld van 9 scholen. Uiteindelijk hebben 24 klassen daadwerkelijk aan de try-out deelgenomen met in totaal 578 leerlingen. In praktijk was de benodigde toetstijd te beperkt, zodat veel leerlingen het boekje niet hebben afgemaakt. Elk item is gemiddeld door 221 leerlingen beantwoord.

6.2.2.2 Toetsvoorbereiding en afname

Na aanmelding hebben de deelnemende docenten leerlinglijsten ingestuurd, waarna voor deze leerlingen specifieke logins zijn aangemaakt; specifiek in de zin dat veel leerlingen al een DWO-account hadden, maar dat het om redenen van beheersbaarheid en veiligheid beter was om hen via een nieuwe login toegang tot de Cito-DWO-omgeving te geven. De docenten zijn via email geïnformeerd over het doel van de toets, over de te treffen voorbereidingen en over de praktijk tijdens de afname (zie Bijlage 6-4). De toetsafname vond plaats in de periode 6-17 januari 2014 en de beoogde toetstijd was 50 minuten.

6.2.2.3 Data

De data uit de try-out bestaan uit online responsen van de leerlingen, gegevens over toetstijd per item en informatie over de gang van zaken bij de afname. Docenten is gevraagd een vragenlijst in te vullen met betrekking tot de relevantie van de items. Sommige docenten hebben in plaats van de vragenlijst een schriftelijke reactie gemaild. In totaal hebben 11 van de 18 docenten de enquête ingevuld en heeft een aantal anderen via de mail feedback gegeven.

6.2.2.4 Data-analyse en rapportage

De data-analyse in deze situatie met nieuwe itemtypes en een nieuwe ontwerp- en afname-omgeving is niet persé toonaangevend is voor de aanpak hiervan in de toekomst van de grootschalige afname met Facet. Ook is het rapportagemodel nog volop in ontwikkeling en dient het hier gehanteerde model dus als een eerste aanzet te worden beschouwd. De analyse van de digitale responsen is begonnen met het omschalen van de scores. Hierbij zijn deelscores van de DWO, bijvoorbeeld bij meetkundige constructies of bij algebraïsche herleidingen, omgezet. Verder zijn responsen op items die leerlingen minder dan 5 seconden in beeld hebben gehad als ontbrekend beschouwd. Vervolgens is een analyse met OPLM uitgevoerd en een bifactoranalyse. Omdat de resultaten van deze laatste analyse niet tot bevredigende resultaten leidde, is tevens een meer exploratieve principale componentanalyse uitgevoerd. Deze analyses zijn uitgevoerd door medewerkers van POK in samenspraak met de toetsdeskundigen.

Op basis van deze analyse is een rapportageformat ontwikkeld voor docenten en leerlingen (zie Bijlage 6-3). Aangezien de opbrengst met betrekking tot de vakdidactische leerlingenmodellen SOmS en TIP niet overtuigend bleken te zijn, zijn deze niet in de leerlingrapportage opgenomen. Ten slotte is een kwalitatieve analyse van de feedback van docenten uitgevoerd.

6.3 Resultaten

6.3.1 Itembeschrijving en vraagtypes

In Tabel 6-2 zijn de vraagtypes beschreven, waarbij in de naamgeving zo veel mogelijk is aangesloten bij de itemtypes die in Questify beschikbaar zijn. Zo staat GsI bijvoorbeeld voor GapsInline.

Tabel 6-2 Itemnamen en vraagtypes

Item	Naam	Vraagtype	Domein
A1-1	Wandelroute	GsI met voorwaardelijke beoordeling	E
A1-2	Vlieger maar geen ruit	Meetkundige sleepopdracht	D
A1-3	Harro en de huurauto	GsI formule met beoordeling equivalentie en tussenstappen	E
A1-4	Loodrechte lijnen	GsI formule met beoordeling equivalentie	D
A1-5	Grafiek draaien	Meetkundige sleepopdracht	E
A1-6	Gelijkvormige figuren	Meetkundige sleepopdracht	D
A1-7	Tegelpatroon	GsI geheel en formule met beoordeling equivalentie	E
A1-8	Tablet	Combi geheel en open vraag	D
A1-9	Los de volgende vergelijking op	GsI formule met beoordeling equivalentie en tussenstappen	E
A1-10	Literblikken	GsI formule met beoordeling equivalentie	D
A1-11	Parabool	GsI decimaal (3x) met breukeninvoer	E
A2-1	Kat en muis	Meetkundige sleepopdracht	D
A2-2	Schatzoeken	Meetkundige constructieopdracht	D
A2-3	Los de volgende vergelijking op	GsI formule met beoordeling equivalentie en tussenstappen	E
A2-4	Vierkant en cirkel	GsI formule met equivalentie voorw. Beoordeling en decimaal met breukeninvoer	D
A2-5	Grafieken tekenen	GsI decimaal (6x) met breukeninvoer en voorwaardelijke beoordeling; grafieken tekenen	E
A2-6	Punten even ver van A en B	Meetkundige constructieopdracht	D
A2-7	Squashen	GsI formule met beoordeling equivalentie	E
A2-8	Hoeken in een parallellogram	Meetkundige sleepopdracht	D
A2-9	Verbinden	GapMatchInline	E
A2-10	Vierkant leggen	GsI formule met beoordeling equivalentie	D
B1-1	Tabel afmaken		E
B1-2	Evenwijdige lijnen	GsI formule met beoordeling equivalentie	D
B1-3	Wasmachines en prijsverhogingen	GsI formule met beoordeling equivalentie (3x)	E
B1-4	Driehoek in rechthoek	GsI formule met beoordeling equivalentie	D
B1-5	Los de volgende vergelijking op	GsI formule met beoordeling equivalentie en tussenstappen	E
B1-6	Figuur en herleiden	GsI formule met beoordeling equivalentie (2x)	D
B1-7	Herleiden	GsI formule met beoordeling equivalentie en tussenstappen	E
B1-8	Vooraanzicht	Meetkundige sleepopdracht	D
B1-9	Lijn	GsI geheel en decimaal/breuk met intervalbeoordeling	E
B1-10	V-patronen	GsI geheel en formule met beoordeling equivalentie	E
B1-11	Evenwijdig	Meetkundige sleepopdracht	D
B2-1	Symmetrie	Meetkundige constructieopdracht	D
B2-2	Grafiek verslepen	Meetkundige sleepopdracht	E
B2-3	Hoeken in een zandloper	GsI geheel (2x)	D
B2-4	Ligging parabool	GsI geheel	E
B2-5	Driehoek in kubus	Meetkundige sleepopdracht	D
B2-6	Grafieken tekenen	GsI decimaal (6x) met breukeninvoer en voorwaardelijke beoordeling grafieken en tekenen	E
B2-7	Middelloodlijnen in een driehoek	Meetkundige sleepopdracht	D
B2-8	Squashen	GsI formule met beoordeling equivalentie en tussenstappen	E
B2-9	Vlieger knippen	GsI formule met beoordeling equivalentie	D
B2-10	Printerkosten	GsI formule met beoordeling equivalentie en tussenstappen	E

Vrijwel geen van deze items is op dit moment in Questify/Facet te realiseren. De belangrijkste oorzaken daarvan zijn de volgende.

- In totaal 10 items, overwegend maar niet uitsluitend uit domein D, maken gebruik van de mogelijkheid om meetkundige objecten te verslepen. Het hotspots vraagtype van Questify biedt hiervoor onvoldoende mogelijkheden.
- In 3 items binnen domein D worden meetkundige constructiemethoden gebruikt die in Questify/Facet niet beschikbaar zijn.
- In 19 items, zowel binnen domein D als in domein E, worden formules gevraagd. De formules zouden in Questify/Facet wel kunnen worden ingevoerd, maar kunnen daar nog niet op equivalentie worden beoordeeld. Ook is het in 7 items van domein E van belang om tussenstappen te kunnen beoordelen.
- In 3 items binnen domein E wordt de leerling gevraagd een grafiek te tekenen, wat in Questify/Facet niet mogelijk is.
- Bij 4 items is voorwaardelijke beoordeling aan de orde: De beoordeling van een respons in een invulvak is afhankelijk van de respons bij een eerder invulvak.

- Bij 5 items wordt een breuk gevraagd. Het Questify/Facet vraagtype GapsInline Decimal geeft deze mogelijkheid niet. Een formuleinvoerveld wel, maar dat heeft vanwege het grote aantal knoppen en de vele benodigde ruimte nadelen voor de opmaak van het item. Er is dus behoefte aan een kleiner formuleinvoerveld, en aan de mogelijkheid om knoppen van de formule-editor 'af te plakken'.

Samengevat zien we dat een aantal mogelijkheden van de DWO, die Questify/Facet nog niet biedt, frequent gebruikt is in de ontworpen set items. Met name gaat het om het intelligent lezen van formules en deze beoordelen op equivalentie (domeinen D en E) en om het construeren en verslepen van meetkundige objecten (domein D). Andere belangrijke mogelijkheden zijn het beoordelen van tussenstappen (bij het oplossen van vergelijkingen), het handig invoeren en beoordelen van breuken, de mogelijkheid van voorwaardelijke beoordeling en het tekenen van grafieken.

6.3.2 Itemresultaten

De leerlingen hebben boekje A of boekje B voorgelegd gekregen. Elk boekje bestond uit twee delen. Ongeveer de helft van de leerlingen is met het eerste deel begonnen en de andere helft met het tweede. Een deel van de leerlingen is nauwelijks verder gekomen dan deze eerst aangeboden helft. In Tabel 6-3 is daarom onderscheid gemaakt tussen leerlingen die overwegend één deel gemaakt hebben, of (vrijwel) de gehele toets. Voor elk geval zijn de betrouwbaarheden gegeven en de gemiddelde *p*-waarden per domein. Deze betrouwbaarheden zijn aan de lage kant. Aan het einde van deze paragraaf gaan we kort op mogelijke oorzaken daarvan in. De gemiddelde *p*-waarde voor items van domein D in boekje A zijn laag te noemen; de overige *p*-waarden zijn redelijk.

Tabel 6-3 Betrouwbaarheid (Cronbach's α) en *p*-waarden per domein

	Boek A	Boek B
Betrouwbaarheid geheel	0.631 (N=79)	0.668 (N=160)
Betrouwbaarheid deel 1	0.526 (N=167)	0.554 (N=220)
Betrouwbaarheid deel 2	0.494 (N=139)	0.450 (N=224)
Gemiddelde *p*-waarde domein D	0.29	0.47
Gemiddelde *p*-waarde domein E	0.51	0.42

Tabel 6-4 bevat voor elk item het domein, de indeling volgens het vakdidactisch leerlingmodel Structuur - Objectvorming en meervoudige betekenis - Samenhang (SOmS) en de indeling volgens het schaduwmodel Toepassen - Inzicht - Procedure (TIP). Vervolgens worden de OPLM-parameters en de p- en Rit-waarden gegeven. In deze tabel valt een aantal zaken op.

- Er is één item, B1-8, dat negatief discrimineert. Het betreft een vrij eenvoudig vooraanzicht van een ruimtelijk object, dat echter ook correcte oplossingen heeft met deelfiguren die niet als zodanig worden herkend. Daarnaast legt een aantal leerlingen figuren over elkaar, wat niet is toegestaan. Deze implementatie-moeilijkheden verklaren de negatieve discriminatie.
- Er zijn 8 items met een *p*-waarde die kleiner is dan 0,1. In vergelijking met de try-out 2013 is geprobeerd meer spreiding in moeilijkheid aan te brengen. Dit zien we hier terug in het kleinere aantal te eenvoudige items (alleen B2-7 heeft een te hoge *p*-waarde). Het aantal te moeilijke items is nu echter te groot. Een nadere blik op de uitwerkingen van leerlingen en de responsietijden leert ons dat er bij deze items geen artefacten zijn gesignaleerd; items zoals B1-4 zijn gewoon erg moeilijk en bij andere items zoals A1-10 en A1-11 lijkt tijdgebrek een rol te hebben gespeeld.

Tabel 6-4 Itemresultaten: OPLM-parameters en p- en RIt-waarden

Item	Naam	Domein	SOmS	TIP	A	B1	B2	B3	P-waarde	Rit
A1-1	Wandelroute	E	st	T	2	-0.50	-1.29	-0.56	0.83	0.36
A1-2	Vlieger maar geen ruit	D	st	I	4	-0.06			0.34	0.52
A1-3	Harro en de huurauto	E	ob	P	4	0.13			0.24	0.48
A1-4	Loodrechte lijnen	D	sa	I	3	1.10			0.05	0.19
A1-5	Grafiek draaien	E	sa	I	3	-0.55			0.63	0.33
A1-6	Gelijkvormige figuren	D	ob	I	2	0.25			0.33	0.40
A1-7	Tegelpatroon	E	st	I	3	-0.17			0.38	0.40
A1-8	Tablet	D	st	T	4	0.21			0.20	0.49
A1-9	Los de volgende vergelijking op	E	ob	P	4	0.19			0.08	0.29
A1-10	Literblikken	D	sa	I	3	1.59			0.01	0.18
A1-11	Parabool	E	sa	I	3	1.46			0.01	0.25
A2-1	Kat en muis	D	st	T	3	-0.94			0.84	0.06
A2-2	Schatzoeken	D	st	T	3	0.64			0.04	0.27
A2-3	Los de volgende vergelijking op	E	ob	P	3	-0.23			0.54	0.47
A2-4	Vierkant en cirkel	D	sa	P	3	0.09	0.56		0.14	0.37
A2-5	Grafieken tekenen	E	sa	P	3	-0.80	-0.56		0.73	0.40
A2-6	Punten even ver van A en B	D	ob	P	3	0.07			0.25	0.51
A2-7	Squashen	E	st	T	3	-0.42			0.53	0.39
A2-8	Hoeken in een parallellogram	D	ob	I	3	-0.59			0.68	0.46
A2-9	Verbinden	E	ob	P	2	-0.36			0.51	0.30
A2-10	Vierkant leggen	D	sa	T	4	0.74			0.01	0.18
B1-1	Tabel afmaken	E	sa	P	2	-0.35	-0.30		0.68	0.47
B1-2	Evenwijdige lijnen	D	sa	P	4	0.81			0.04	0.20
B1-3	Wasmachines en prijsverhogingen	E	st	T	2	-0.11	0.12		0.51	0.51
B1-4	Driehoek in rechthoek	D	st	I	3	1.27			0.03	0.22
B1-5	Los de volgende vergelijking op	E	ob	P	4	0.21			0.33	0.39
B1-6	Figuur en herleiden	D	sa	P	4	0.37	0.77		0.12	0.46
B1-7	Herleiden	E	ob	P	2	0.14			0.38	0.37
B1-8	Vooraanzicht	D	st	T	1	-1.48			0.81	0.09
B1-9	Lijn	E	ob	I	5	0.16			0.35	0.47
B1-10	V-patronen	E	st	I	4	0.11			0.37	0.50
B1-11	Evenwijdig	D	ob	I	3	-0.45			0.73	0.30
B2-1	Symmetrie	D	ob	P	2	-0.58			0.73	0.19
B2-2	Grafiek verslepen	E	sa	I	5	-0.40			0.81	0.48
B2-3	Hoeken in een zandloper	D	st	P	3	-0.49			0.74	0.42
B2-4	Ligging parabool	E	st	I	4	-0.01			0.50	0.49
B2-5	Driehoek in kubus	D	ob	I	1	0.58			0.34	0.19
B2-6	Grafieken tekenen	E	sa	P	2	-0.79	-0.09		0.69	0.37
B2-7	Middelloodlijnen in een driehoek	D	ob	I	2	-1.30			0.93	0.14
B2-8	Squashen	E	st	T	4	0.40			0.21	0.42
B2-9	Vlieger knippen	D	sa	T	4	0.52			0.14	0.41
B2-10	Printerkosten	E	sa	T	3	0.79	0.10		0.12	0.41

Een volgende vraag is of de elementen van het vakdidactische leerlingmodel in de analyse naar voren komen. De bifactoranalyse, die daartoe is uitgevoerd, heeft geen duidelijk beeld gegeven van de verschillende categorieën van het SOmS model. Evenmin zijn de drie facetten van het schaduwmodel TIP hieruit naar voren gekomen. Een meer exploratieve principale componentanalyse heeft ook niet tot de identificatie van deze of andere componenten geleid. De drie componenten van het SOmS model of het TIP model zijn in deze proefafname dus niet empirisch aantoonbaar gebleken.

Samengevat is de conclusie van deze analyse dat de toets als geheel een lage betrouwbaarheid kende en dat de vakdidactische elementen van het leerlingmodel SOmS en het schaduwmodel TIP niet meetbaar naar voren zijn gekomen. Er zijn verschillende factoren die deze resultaten kunnen verklaren.

- Er zijn vrij veel items met een te lage *p*-waarde. De responsen op deze items hebben dus weinig spreiding. Dit verklaart de lage betrouwbaarheid en heeft ook negatieve gevolgen voor de factoranalyse. Er was geen gelegenheid om de items vooraf te testen, waardoor dit verschijnsel niet was voorzien.
- De toets was te lang voor de beschikbare toetstijd. Als gevolg hiervan bevat de datamatrix veel ontbrekende waarden.
- Een aantal klassen uit het tweetalig onderwijs heeft een Engelse versie van de toets gemaakt. Deze groep leerlingen is te klein om apart te analyseren, maar wijkt mogelijk af van de reguliere klassen. Daar komt bij dat

de vertalingen van de wiskundige begrippen niet altijd spoorden met het vocabulaire dat de leerlingen hanteren.
- De items doen een beroep op de wiskundige 'knoppenvaardigheid' van leerlingen. Met name het gebruik van de formule-editor en van de meetkundige constructiemiddelen is voor de meesten van hen nieuw. Wellicht heeft dit een rol gespeeld en is de beheersing van de interface een vaardigheid die in deze toets interfereert met de wiskundige vaardigheid. Hoewel docenten hierover geen signalen hebben afgegeven, was meer oefening vooraf met de digitale omgeving wellicht ten goede gekomen aan de betrouwbaarheid van de toets.
- Vermoedelijk zijn de vakdidactische aspecten van het SOmS en het TIP model moeilijk te meten in een dergelijke korte toets. De vraag is ook wel in hoeverre het hier meetbaar te onderscheiden vaardigheden betreft. Zou het zo kunnen zijn dat de SOmS en TIP kenmerken alle sterk samenhangen met wiskundige vaardigheid in het algemeen? Of zou een betere toets de verschillende componenten wel onderscheiden? Dit is een vraag voor vervolgonderzoek.

6.3.3 Resultaten op scholen

6.3.3.1 De afname in praktijk

In praktijk is de afname op scholen in grote lijnen goed verlopen. Technische moeilijkheden hebben zich niet voorgedaan. Leerlingen hebben in het algemeen serieus aan de toets gewerkt, zo blijkt uit de rapportage van docenten. In de vragenlijst achteraf gaven enkele docenten aan dat het opstarten met de hele klas wat traag verliep. Bij een enkele leerling leken er problemen met afsluiten of opslaan van de resultaten op te treden; er zijn voor zover bekend echter geen data verloren gegaan.

Een punt van aandacht was de beschikbare toetstijd. De effectieve tijd dat leerlingen aan de toets konden werken varieerde van klas tot klas, maar was in het algemeen te kort om de volledige toets te kunnen afmaken. Praktische factoren zoals inlogtijd leidden tot vertraging en de inschatting van de toetsontwikkelaars over de benodigde tijd wat te positief. Een en ander heeft geleid tot spanning bij sommige leerlingen en tot ontbrekende responsen in de data. Ook heeft het feit dat de toetstijd van klas tot klas nogal uiteenloopt de analyse bemoeilijkt.

Over de inhoud van de toets zijn de meeste docenten tevreden. De moeilijkheidsgraad wordt beoordeeld als goed tot te hoog. De spreiding van de moeilijkheidsgraad is als voldoende beschouwd. In enkele items kwam voorkennis aan de orde die bij leerlingen van sommige klassen nog niet behandeld is geweest.

Enkele docenten rapporteren lokale technische moeilijkheden van leerlingen. Het gaat dan met name om het werken met de formule-editor van de DWO en met de meetkundige mogelijkheden van Geogebra. Met behulp van het vooraf opgestuurde document met te verwachten technische vragen konden de docenten deze moeilijkheden in de klas hanteren. Ondanks deze technische drempels meldt een docent: "Hoewel mijn leerlingen nog niet eerder met de DWO hadden gewerkt, konden ze prima overweg met de bediening van het programma."

De meeste van de deelnemende docenten vonden de toets wel een meerwaarde hebben en vonden de meeste items zinvol. Op de vraag wat die meerwaarde dan is, worden genoemd het inzicht in wat leerlingen wel en niet kunnen, de herhaling van de hele onderbouwstof en de ervaring om weer eens in een computerlokaal aan het werk te zijn.

6.3.3.2 Analyse en rapportage

Ten behoeve van de rapportage aan scholen zijn de gegevens op leerlingniveau en klassenniveau geanalyseerd. Docenten hebben informatie ontvangen over

- de gemiddelde klassenresultaten als indicatie van het relatieve vaardigheidsniveau van de klas als geheel;
- het gemiddlede percentage van het maximaal aantal te behalen punten van de gemaakte items per domein in de klas;
- het vaardigheidsniveau per leerling;
- het percentage van het maximaal aantal te behalen punten van de gemaakte items per leerling per domein;
- informatie over de interpretatie van deze gegevens.

De vorm van deze rapportage en de begeleidende brief zijn opgenomen in Bijlage 6-3. Zoals opgemerkt in sectie 6.2 gaat het hier om een zeer verkennende en tentatieve gedaante van de rapportage.

Tabel 6-5 geeft een overzicht van de vaardigheidsniveaus per klas, links voor de klassen die toets A hebben gemaakt en rechts voor de klassen die aan toets B hebben gewerkt. In het rechter deel lopen de vaardigheids-scores op. De vwo-klassen scoren beter dan de havo-klassen, maar omdat de niveauaanduiding per leerweg is bepaald, scoren de vwo-klassen alle vier gemiddeld. In het linker deel is het beeld anders. Ten eerste valt op dat de klassen die in het Engels wiskundeles krijgen en ook een Engelstalige variant van de toets hebben gemaakt (herkenbaar aan de toevoeging 'EN' in de klassencode), niet hoog scoren. Een mogelijke verklaring hiervoor ligt in de problemen met de vertaling en het wiskundige vocabulaire waarmee leerlingen van de tweetalige klassen niet altijd vertrouwd waren. Ten tweede valt in het linker deel van Tabel 6-5 op dat de twee best presterende havo-klassen een hoge vaardigheid hebben en vijf vwo-klassen achter zich laten. Hiervoor is geen verklaring bekend.

Tabel 6-5 Resultaten per klas, links voor toets A en rechts voor toets B

Klas	Theta	Niveauaanduiding	Klas	Theta	Niveauaanduiding
havoA1	-0.43	Onder gemiddeld	havoB1	-0.51	Onder gemiddeld
havoA2	-0.37	Onder gemiddeld	havoB2	-0.43	Onder gemiddeld
havoA3	-0.37	Onder gemiddeld	havoB3	-0.20	Gemiddeld
havoA4	-0.29	Gemiddeld	havoB4	-0.14	Gemiddeld
vwoA5	-0.28	Gemiddeld	havoB5	-0.13	Gemiddeld
vwoA6	-0.26	Gemiddeld	havoB6	-0.03	Boven gemiddeld
vwoA7	-0.22	Gemiddeld	havoB7	-0.06	Boven gemiddeld
vwoA8	-0.15	Gemiddeld	havoB8	-0.04	Boven gemiddeld
vwoA9	-0.05	Gemiddeld	vwoB9	0.03	Gemiddeld
havoA10	-0.01	Boven gemiddeld	vwoB10	0.10	Gemiddeld
havoA11	0.05	Boven gemiddeld	vwoB11	0.16	Gemiddeld
vwoA12	0.23	Boven gemiddeld	vwoB12	0.22	Gemiddeld

6.3.3.3 Samengevat

Samengevat concluderen we uit deze bevindingen ten eerste dat de afname van een DTT met de DWO in praktijk realiseerbaar is. De deelnemende scholen en docenten zijn voldoende geëquipeerd om een dergelijke toets te organiseren en de omgeving is bestand tegen een dergelijke afname. De voorbereiding van een dergelijke afname door middel van schriftelijke informatie lijkt, althans voor deze groep van deelnemende docenten, voldoende te zijn.

Ten tweede lijken docenten vrij positief te reageren op deze DTT. Als meerwaarde zien ze vooral de herhaling van de hele onderbouwstof door leerlingen en de informatie hierover aan de docent. Hierbij moet wel in ogenschouw genomen worden dat het hier niet om een aselecte groep docenten gaat, maar om docenten die zich hebben aangemeld voor een afname die nadrukkelijk als informele try-out is aangekondigd. Verder valt op dat slechts één docent op de rapportage heeft gereageerd, dus dat onduidelijk is in hoeverre deze in hun behoefte heeft voorzien.

Een derde conclusie is dat de tijdinschatting van een digitale toets zonder pretest een moeilijke zaak is. Daar komt bij dat een digitale afname altijd wat extra starttijd vraagt, waardoor de effectieve toetstijd korter is.

Ten vierde concluderen we dat enige technische oefening voor leerlingen gewenst is, met name waar het gaat om het gebruik van de formule-editor en van de meetkundige constructietools.

Ten vijfde blijkt dat het vertalen van items in het Engels niet zo triviaal als gedacht. Zorgvuldige afstemming met het gebruikelijke vocabulaire in de tweetalige klassen blijkt nodig te zijn.

6.4 Conclusies en aanbevelingen

6.4.1 Conclusies

De *eerste onderzoeksvraag* was welke itemtypes voor DTT wiskunde ontstaan door de beschikbaarheid van meer vakspecifieke opties in de auteursomgeving. Op basis van de in het vorige sectie geschetste resultaten concluderen we dat een aantal mogelijkheden van de DWO, die Questify/Facet nog niet biedt, frequent gebruikt is

in de ontworpen set items. Met name gaat het om het intelligent lezen van formules en deze beoordelen op equivalentie (domeinen D en E) en om het construeren en verslepen van meetkundige objecten (domein D). Andere belangrijke mogelijkheden zijn het beoordelen van tussenstappen (bij het oplossen van vergelijkingen), het handig invoeren en beoordelen van breuken, de mogelijkheid van voorwaardelijke beoordeling en het tekenen van grafieken.

In het algemeen gesproken maakt een omgeving zoals de Digitale Wiskunde Omgeving het mogelijk om itemtypes te gebruiken die tot gevarieerde en rijke items leiden die een productief karakter hebben en grote constructieruimte bieden aan leerlingen, en daarmee in principe ook een meer genuanceerde diagnose haalbaar maken. Dit laatste is vanzelfsprekend ook afhankelijk van de inhoud van het item zelf en van mogelijkheden voor automatische beoordeling.

De *tweede onderzoeksvraag* was of de afname van deze items tot een adequate en efficiënte toetsing van het leerlingmodel leidt. De vakinhoudelijke domeinen van het leerlingmodel zijn naar het oordeel van de toetsdeskundigen adequaat getoetst, al is dit niet in detail onderzocht en is de betrouwbaarheid van de toets matig te noemen. Voor de vakdidactische modellen van het leerlingmodel, het SOmS model en het TIP schaduwmodel, geldt echter dat in dit onderzoek niet is aangetoond dat de componenten hiervan onderscheiden factoren zijn. De vakdidactische leerlingmodellen, die voor de diagnostische doelen van DTT van belang zijn, zijn in dit onderzoek dus niet empirisch gevalideerd. Mogelijke oorzaken hiervan zijn

- de onvolkomendheden in de toets (te lage validiteit, te veel items voor de beschikbare tijd, items met te lage *p*-waarden, vertaalproblemen);
- dat het in werkelijkheid niet gaat om te onderscheiden vaardigheden;
- dat de toetstijd te kort was om de verschillende vaardigheden te onderscheiden;
- dat de toetsitems onvoldoende laden op één van de drie componenten van het SOmS en het TIP model;
- dat de DWO als toetsomgeving zo veel vraagt van de ICT-vaardigheden van de leerling, dat deze vaardigheid de belangrijkste is die de toets meet, waardoor de wiskundige vaardigheden naar de achtergrond verschuiven. De reacties van de betrokken docenten wijzen hier overigens niet op.

6.4.2 Aanbevelingen

Op basis van deze conclusies doen we de volgende aanbevelingen voor vervolgstappen.

1. De haalbaarheid van de vakdidactische leerlingmodellen SOmS en TIP moet nader worden onderzocht door deze verder uit te werken, te operationaliseren en uit te testen. In concreto zou de in dit onderzoek ontwikkelde toets gereviseerd kunnen worden door het aantal items te verminderen, psychometrisch zwakke items te verwijderen, overige items beter op SOmS en TIP in te richten, en geen Engelstalige variant te maken. Naar verwachting zal de betrouwbaarheid dan toenemen en de vraag of SOmS en TIP dan empirisch onderbouwd kunnen worden kan dan beter worden beantwoord. Een kleinschalige afname in 2015 zou op basis van dit vooronderzoek efficiënt kunnen worden uitgevoerd.
2. Voor een adequate diagnose binnen DTT-wiskunde is een aantal wiskundige mogelijkheden van groot belang. Denk aan het hanteren en beoordelen van formules, het beoordelen van equivalentie, voorwaardelijke beoordeling, beoordeling van tussenstappen, en gebruik en beoordeling van grafische invoer en van meetkundige constructies. Ontwikkeling van deze mogelijkheden in Questify Builder en Facet verdient hoge prioriteit en kan behalve ‘native’ ook gerealiseerd worden door bestaande software te integreren. Dit onderzoek heeft aangetoond dat de DWO hiervoor een geschikte kandidaat is.

6.5 Referenties

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

Drijvers, P. (2013). *Voorstudie automatische scoring in DTT wiskunde*. Intern rapport. Arnhem: Cito.

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.

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 Educatio*n (pp. 721-751). New York/Berlin: Springer.

Williamson, D. M., Mislevy, R. J., & Bejar, I. (Eds.) (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

6.6 Bijlagen

Bijlage 6-1 Dankwoord

Onze dank gaat uit naar de leerlingen en docenten die hebben meegewerkt aan deze try-out. Verder dank aan het DWO-team van het Freudenthal Instituut, en met name aan Peter Boon en Sietske Tacoma, voor het ter beschikking stellen van de Digitale Wiskunde Omgeving en voor de ondersteuning bij de toetsconstructie en -afname. Ten slotte dank aan alle collega's van het DTT-team binnen Cito voor het meedenken over opzet en uitvoering van deze try-out.

Bijlage 6-2 Vragenlijst voor docenten

Enquête DTT DWO Try-out

Hartelijk dank voor uw deelname aan de informele try-out van de Diagnostische Tussentijdse Toets in de Digitale Wiskunde Omgeving. Om een beter beeld te krijgen van het verloop van de afname en de mogelijke problemen die daarbij naar voren kwamen, willen we u vriendelijk vragen onderstaande enquête in te vullen.

Na het invullen kunt u het bestand opslaan op uw computer en over de mail versturen naar Saskia Visser (Saskia.Visser@cito.nl)

Bij meerkeuzevragen zijn meerdere antwoordopties mogelijk.

Gegevens:

Naam school:

Naam leerkracht:

Klas(sen):

Klasniveau: ☐Havo ☐Vwo ☐ Beide

Technisch:

1.1 Welke technische problemen bent u of zijn uw leerlingen tijdens de afname tegengekomen?

☐ Problemen bij het inloggen

☐ Bepaalde meetkundetools (in Geogebra) werkten niet

☐ Het formulepalet werkte niet naar behoren

☐ Andere technische problemen/foutmeldingen (s.v.p. toelichten)

Toelichting/opmerkingen:

1.2 Heeft u/uw klas te maken gehad met traagheid van het systeem en zo ja, op welke momenten?

☐ Nee

☐ Ja, voornamelijk gedurende het inloggen

☐ Ja, voornamelijk bij meetkunde items (Geogebra)

☐ Ja, gedurende de gehele afname

☐ Anders (s.v.p. toelichten)

Toelichting/opmerkingen:

1.3 Welke vragen stelden leerlingen tijdens de toetsafname?

☐ Vragen over werking van de meetkunde tools (van Geogebra)

☐ Vragen over de werking van het formulepalet

☐ Over de vraagstelling

☐ Vragen over termen/begrippen uit de items,

☐ Iteminhoudelijke vragen

☐ Andere vragen (s.v.p. toelichten)

☐ Geen vragen

Toelichting/opmerkingen:

Opvallende of vaak voorkomende vragen/onduidelijkheden die bij de leerlingen opkwamen graag noteren.

Praktisch:

2.1 Welke lesmethode gebruikt u?

☐ Getal en Ruimte

☐ Moderne Wiskunde

☐ Netwerk

☐ Wageningse methode

☐ Anders (s.v.p. toelichten)

Toelichting/opmerkingen:

Bijlage 6-3 Rapportageformat

De rapportage aan scholen bestond uit een standaardbrief en een Excelbestand met de specifieke gegevens van de klas. Hieronder staat de standaardbrief en een voorbeeld van een geanonimiseerd Excelbestand van één van de deelnemende klassen.

Brief

Geachte docent,

Allereerst nogmaals hartelijk dank voor uw deelname aan de informele try-out van de Diagnostische Tussentijdse Toets wiskunde (DTT) in de Digitale Wiskunde Omgeving (DWO). Met dit document informeren we u over de resultaten van uw klas.

Hierbij aangehangen ontvangt u een spreadsheet dat deze resultaten bevat. Voor we ingaan op de interpretatie daarvan lichten we eerst de inhoud kort toe. In de bovenste vier regels vindt u de naam van uw school en uw klas, het schooltype, de variant A of B van de opgaven, en de taal van de opgaven. Dan volgt informatie op de volgende punten.

1. Onder de kop ‘Gemiddelde klassenresultaten’ staat allereerst een indicatie van het relatieve vaardigheids-niveau van uw klas als geheel, aangeduid als Onder gemiddeld, Gemiddeld of Boven gemiddeld.
   Deze indicatie beschrijft het vaardigheidsniveau ten opzichte van andere klassen van hetzelfde schooltype (dus havo of vwo) die aan deze try-out hebben meegedaan. Aan boekje A hebben 6 havo klassen en 6 vwo klassen deelgenomen. Boekje B is door 8 havo klassen en 4 vwo klassen gemaakt. De score Onder gemiddeld wijst op een significant lager niveau van uw klas ten opzichte van de andere deelnemende klassen en de score Boven gemiddeld staat voor een significant hogere score. Omdat items waaraan een leerling niet is toegekomen (en dus geen of een te verwaarlozen hoeveelheid tijd aan heeft besteed) van de analyse zijn uitgesloten, is bij deze vaardigheidsindicatie rekening gehouden met de tijd die uw klas aan de toets heeft kunnen besteden; de effectieve toetstijd loopt om diverse redenen namelijk uiteen voor de verschillende klassen.
2. De items van deze try-out hebben betrekking op de domeinen Meten en meetkunde en Verbanden en formules. Voor elk van deze domeinen is in de rijen 10-13 aangegeven welk percentage van het maximaal aantal te behalen punten van de gemaakte items gemiddeld in uw klas is behaald en hoeveel items er gemiddeld zijn gemaakt. Vervolgens is aangegeven wat de gemiddelde tijdsduur is die uw leerlingen aan de toets hebben besteed.
3. Dan volgt in het spreadsheet de lijst met leerlingen van uw klas die aan de try-out hebben deelgenomen.
   De kolom Bijzonderheden zal in de meeste gevallen voor het grootste deel leeg zijn, maar als er zich bij de betreffende leerling iets opmerkelijks heeft voorgedaan dan staat dat hier beschreven. Vervolgens ziet u voor elke leerling het vaardigheidsniveau op dezelfde manier als bovenaan voor de klas als geheel is gegeven. Ook hierbij gaat het om een relatief niveau en de termen Onder gemiddeld, Gemiddeld en Boven gemiddeld geven aan of het vaardigheidsniveau van de leerling onder, op, of boven het niveau ligt van de andere leerlingen van hetzelfde schooltype die aan de try-out hebben deelgenomen. Er hebben in totaal 139 havisten en 135 vwo-leerlingen aan boekje A deelgenomen en respectievelijk 201 en 103 aan boekje B. Beide boekjes bestaan uit 21 items, wat veel bleek te zijn voor de beoogde toetstijd van 50 minuten.
4. Ook is per leerling voor elk van de domeinen Meten en meetkunde en Verbanden en formules het percentage van het maximale aantal te behalen punten berekend en het gemaakte aantal items. Aan de percentages kunnen geen directe uitspraken over het niveau gekoppeld worden. In de meeste gevallen correspondeert het vaardigheidsniveau met de percentages van het maximale aantal te behalen punten. Dit kan echter ook niet het geval zijn, aangezien bij de vaardigheidsscore tevens rekening is gehouden met de moeilijkheid van de gemaakte items in plaats van enkel het gemaakte aantal. Wanneer twee leerlingen evenveel items hebben gemaakt en hetzelfde aantal punten hebben behaald, kunnen zij dus toch een ander vaardigheidsniveau hebben gekregen, doordat ze andere items hebben gemaakt. De individuele rapportage sluit af met de tijd die de betreffende leerling aan de toets heeft besteed.

Hoe kunt u deze scores interpreteren? Om te beginnen willen we benadrukken dat grote voorzichtigheid bij de interpretatie in acht moet worden genomen. Door verschillende oorzaken kan van betrouwbare conclusies in deze try-out geen sprake zijn. Zo zijn de items nog onvoldoende uitgekristalliseerd of gekalibreerd, is het totale aantal leerlingen dat heeft deelgenomen te klein om stellige uitspraken te kunnen doen en zijn er praktische en technische problemen die de afname sterk beïnvloed kunnen hebben. Met dit voorbehoud lopen we de bovenstaande punten opnieuw langs.

1. Het vaardigheidsniveau van uw klas als geheel geeft een indicatie van het niveau vergeleken met de andere klassen van hetzelfde schooltype (dus havo of vwo) die aan deze try-out hebben meegedaan. Ongeveer een kwart van de klassen heeft de indicatie Onder gemiddeld en ongeveer een kwart Boven gemiddeld. Behalve de werkelijke wiskundevaardigheid van de klas kunnen verschillende factoren dit resultaat hebben beïnvloed, zoals technische vaardigheden, vertrouwdheid met de DWO en al dan niet behandelde lesstof.
   Bij Engelstalige klassen heeft bovendien de gebruikte wiskundige terminologie mogelijk niet voldoende aangesloten bij het taalgebruik in de les, waardoor de scores mogelijk zijn vertekend. In het algemeen hebben de Engelstalige klassen bij deze try-out laag gescoord ten opzichte van de reguliere klassen. De omvang van de Engelstalige klassen is echter te klein om ze als aparte groep te kunnen beschouwen.
2. De gemiddelde percentages van het maximaal aantal te behalen punten voor de domeinen Meten en meetkunde en Verbanden en formules geven een beeld van het beheersingsniveau van uw klas per domein. Om een indicatie te geven van de betrouwbaarheid van deze percentages is daarnaast het gemiddeld aantal gemaakte items per domein te zien. Er kunnen verschillende verklaringen zijn voor de hoogte van deze percentages en de verschillen daartussen. Het is mogelijk dat in het onderwijs meer aandacht aan het ene dan aan het andere domein is besteed. Ook kunnen de items van het ene domein technisch (denk aan meetkundige constructies of aan het gebruik van de formule-editor) of inhoudelijk moeilijker zijn uitgevallen dan die van het andere. Over alle leerlingen heen is het gemiddelde percentage van het maximaal aantal te behalen punten voor Verbanden en formules iets hoger dan dat van Meten en meetkunde.
3. De interpretatie van het vaardigheidsniveau per leerling is een relatieve: de leerling doet het minder goed, even goed, of beter dan de gemiddelde leerling van hetzelfde schooltype. Net als bij de klassen heeft ongeveer een kwart van de leerlingen het oordeel Onder gemiddeld, een kwart Boven gemiddeld en vormt dus ongeveer de helft de middengroep.
4. Als de percentages van een leerling over de domeinen Meten en meetkunde en Verbanden en formules sterk verschillen, kan dit wijzen op achterblijvende kennis in het specifieke domein waarop lager is gescoord. Belangrijk is echter om in deze overweging wel het aantal gemaakte items per domein te betrekken. Een kleine verzameling items kan minder representatief zijn voor het hele domein en een meting hierbinnen heeft een minder grote betrouwbaarheid.

Mocht deze rapportage aanleiding zijn tot vragen, dan kunt u die vanzelfsprekend per email (paul.drijvers@cito.nl) aan ons melden. Tevens houden we ons aanbevolen voor suggesties voor aanvullende rapportages bij toekomstige afnames van soortgelijke toetsen. Als u belangstelling heeft om ook in het vervolg aan dergelijke proeftoetsen deel te nemen, laat het ons dan weten!

Met vriendelijke groet,

Saskia Visser

Paul Drijvers

Excel sheet met leerlingresultaten

Resultaten DTT-DWO Try-out januari 2014							
School:	<NAAM>						
Klas:	<NAAM>						
Schooltype:	havo						
Boekje:	B (Nederlands)						
Gemiddelde klassenresultaten							
Vaardigheidsniveau:		Boven gemiddeld					
Percentage correct Meten en meetkunde:		41%					
Aantal items gemaakt Meten en meetkunde:		8,08					
Percentage correct Verbanden en formules:		53%					
Aantal items gemaakt Verbanden en formules:		8,8					
Tijdsduur toets:		0:40:35					
Leerling	**Bijzonderheden**	**Vaardigheidsniveau**	**% correct Meten en Meetkunde**	**Aantal items gemaakt Meten en meetkunde**	**% correct Verbanden en formules**	**Aantal items gemaakt Verbanden en formules**	**Tijdsduur toets**
<naam leerling>		Boven gemiddeld	22%	7	80%	8	0:42:24
<naam leerling>		Gemiddeld	60%	5	42%	5	0:42:50
<naam leerling>		Boven gemiddeld	38%	7	55%	8	0:40:38
<naam leerling>		Gemiddeld	40%	8	54%	9	0:41:39
<naam leerling>		Onder gemiddeld	25%	10	36%	11	0:36:48
<naam leerling>		Gemiddeld	50%	10	29%	11	0:39:58
<naam leerling>		Onder gemiddeld	36%	9	21%	10	0:41:23
<naam leerling>		Gemiddeld	30%	8	54%	9	0:40:07
<naam leerling>		Gemiddeld	27%	9	64%	9	0:41:37
<naam leerling>		Boven gemiddeld	42%	10	50%	11	0:28:58
<naam leerling>		Gemiddeld	80%	5	33%	5	0:42:52
<naam leerling>		Gemiddeld	33%	4	48%	5	0:41:35
<naam leerling>		Gemiddeld	80%	5	33%	5	0:42:09
<naam leerling>		Gemiddeld	38%	10	60%	11	0:34:50
<naam leerling>		Boven gemiddeld	70%	8	86%	9	0:43:09
<naam leerling>		Boven gemiddeld	17%	4	100%	4	0:42:53
<naam leerling>		Boven gemiddeld	38%	10	67%	11	0:39:26
<naam leerling>		Boven gemiddeld	71%	6	56%	7	0:44:48
<naam leerling>		Gemiddeld	55%	9	31%	10	0:39:53
<naam leerling>		Boven gemiddeld	42%	10	60%	11	0:42:08
<naam leerling>		Gemiddeld	33%	10	51%	11	0:41:56
<naam leerling>		Onder gemiddeld	25%	10	32%	11	0:35:25
<naam leerling>		Gemiddeld	36%	9	66%	9	0:42:18
<naam leerling>		Onder gemiddeld	21%	10	44%	11	0:42:00
<naam leerling>		Boven gemiddeld	27%	9	85%	9	0:42:40

Bijlage 6-4 Informatie voor docenten

Aanwijzingen voor de uitvoering van de DTT-DWO try-out

De uitvoering behelst dat u in de periode 6-17 januari in een les van 50 minuten uw leerlingen individueel met een computer aan een digitale proefversie van de DTT in de DWO laat werken. Hieronder gaan we in op de voorbereiding, de inleiding voor leerlingen, de afname en de afsluiting.

Voorbereiding

Als voorbereiding op de toetsafname raden we de volgende stappen aan.

- *Controle logins*
  U heeft een leerlingenlijst met loginaccounts van uw leerlingen ontvangen. Ga s.v.p. na of deze lijst compleet is. De loginnamen zijn samengesteld uit voor- en achternaam van de leerling (zonder spaties en tussenvoegsels), gevolgd door _cito. Een leerling met de naam 'Ingrid van Groot Nuland' krijgt dus de loginnaam IngridGrootNuland_cito. Hoofdletters zijn hierbij niet van belang. Ook in de leerlingenlijst vindt u voor elke leerling een wachtwoord dat bestaat uit vier cijfers. Dit wachtwoord kunt u mondeling aan de leerling meedelen. In de leerlingenlijst vindt u ook uw eigen naam zodat u ook zelf als leerling kunt inloggen. (Let op: wanneer u vóór de officiële afname periode items beantwoordt, kan het zich voordoen dat deze resultaten verloren gaan, doordat wij eventueel nog kleine wijzigingen in de modules kunnen aanbrengen.) Tevens ziet u een fictieve leerling met voornaam 'extra' of de naam van de klas als voornaam. Deze login kan gebruikt worden als reserve, bijvoorbeeld voor een leerling die niet op de lijst stond.
- *Controle afnameplatform*
  Om onaangename verrassingen bij de afname te voorkomen, raden we aan om met uw eigen leerlinglogin de configuratie in het lokaal waar de afname plaats vindt vooraf te testen. Het kan bijvoorbeeld voorkomen dat Java niet up-to-date is. In zo'n geval kan de systeembeheerder helpen, maar dat is op het moment zelf niet wenselijk.
- *Inhoudelijke en technische oriëntatie*
  Het is mogelijk dat leerlingen tijdens de afname technische vragen hebben. Het is prettig als u zelf de toets vooraf globaal doorneemt en daardoor dergelijke vragen kunt beantwoorden. Het is immers niet de bedoeling dat leerlingen een vraag niet kunnen beantwoorden door computer technische obstakels. Verderop wordt een aantal van deze obstakels benoemd.

Informatie voor leerlingen

Voorafgaand aan de afname, bijvoorbeeld een les ervoor, is het goed om de leerlingen te informeren over het doel van de Diagnostische Tussentijdse Toets (DTT). Deze informatie kan de onderstaande elementen bevatten.

- De DTT is een toets die mogelijk ingevoerd gaat worden om leerlingen van klas 3 hv (en klas 2 vmbo) duidelijk te maken waar ze staan op weg naar het examen. De toets die de leerlingen nu gaan maken is daarvoor een vingeroefening; we willen graag weten welke vragen goed werken en welke niet. Het is dus op dit moment eerder een toets van opgaven dan een toets van de leerlingen.
- Toch zullen we op basis van het leerlingenwerk feedback geven over de beheersing van bepaalde vaardigheden. De leerling krijgt er geen cijfer voor maar deze feedback kan natuurlijk alleen maar zinvol zijn als er serieus aan wordt gewerkt. Dus doe je best!
- De tijdsduur is 50 minuten. Laat je echter in geval van tijdnood niet opjagen, maar werk gewoon rustig en geconcentreerd door.
- In de toets staan allerlei opgaven door elkaar: moeilijke en makkelijke, algebra en meetkunde. Laat je niet ontmoedigen. Je kunt vooruit en achteruit bladeren en springen tussen de opgaves en de twee deelmodules. In totaal zijn er 21 opgaven.
- Bij technische problemen kun je de docent om hulp vragen.

Afname

De afname begint met het uitdelen van de logins en wachtwoorden. Uw leerlingen gaan naar www.fisme.uu.nl/dwo. Afhankelijk van de school zien ze dan twee modules A1 en A2, dan wel B1 en B2. Voor de tweetalige klassen zijn er Engelstalige A-modules.

Om afkijken te voorkomen en om ervoor te zorgen dat alle opgaven door een aantal leerlingen worden beantwoord, raden we aan om de leerlingen om en om met module 1 dan wel 2 te laten beginnen. Leerlingen kunnen zelf overstappen van bijvoorbeeld A2 naar A1.

Bij de meeste opgaven is geen rekenmachine nodig. Indien gewenst mogen leerlingen wel gebruik maken van hun rekenmachine en van kladpapier.

Tijdens de afname kunnen er, afhankelijk van de ervaring die leerlingen met de DWO hebben, technische vragen zijn. Die kunt u als docent beantwoorden, want de toets gaat immers niet om de technische beheersing. Hieronder een aantal mogelijke technische vragen en antwoorden.

Over *formules*:

- Hoe krijg ik een formule-venster? Door in een formule-invoer veld te klikken. Als de cursor al in het formule-invoer veld staat, kan het nodig zijn om nog een keer extra te klikken.
- Hoe krijg ik een formule-venster weg? Het kan verschoven worden. Het wordt gesloten met het kruisje rechts boven.
- Hoe krijg ik een x in een formule? Gewoon via het toetsenbord. Let op dat je voor variabelen kleine letters gebruikt en geen hoofdletters.
- Hoe krijg ik een vermenigvuldiging? Met de x in de formule-editor, of met het * (shift-8) van het toetsenbord.
- Hoe kom ik bij het invoeren van een formule ‘onder het wortelteken uit’? Met de pijltjestoets naar rechts op het toetsenbord.
- Hoe krijg ik het π-teken? Door ‘pi’ in te typen of met het formulepalet, dat je opent door in het invoervak te klikken.
- Hoe kan ik meerdere oplossingen van een vergelijking invoeren? Verschillende oplossingen kunnen worden gescheiden door het woord ‘*of*’, dus bijvoorbeeld ‘x=3 of x=5’.

Over *meetkunde*:

- Hoe zie ik wat de knoppen boven het meetkundescherm betekenen? Door er met de muis boven te gaan staan zonder erop te klikken (‘mouse-over’).
- Waarom kan ik in het meetkundevenster een punt niet verslepen? Als er meerdere knoppen beschikbaar zijn, moet er vóór je een punt versleept eerst op het pijltje linksboven worden geklikt.
- Hoe kan ik in een meetkundevenster de laatste bewerking ongedaan maken? Rechtsboven zie je een undo-knop en een reset-knop.

Afsluiting

Als de toetsafname klaar is, is het prettig als u dat per email aan ons laat weten (saskia.visser@cito.nl en paul.drijvers@cito.nl). Om te voorkomen dat de leerlingen ’s avonds thuis hun werk nog gaan verbeteren, sluiten we dan namelijk de toegang tot de toets af.

Vervolgens worden de resultaten binnen Cito geanalyseerd. De resultaten daarvan zult u per email ontvangen. We vermoeden dat deze eind februari beschikbaar zullen zijn.

7 Beoordeling open vragen schrijfvaardigheid Nederlands

7.1 Inleiding

In de try-out van voorjaar 2014 zijn voor het schoolvak Nederlands items getest voor de DTT leesvaardigheid en voor de DTT schrijfvaardigheid. De basis voor beide toetsen wordt gevormd door een leerlingmodel (Cito, 2012). Het leerlingmodel beschrijft de aard van de vaardigheid die getoetst wordt in de vorm van een samenhangende structuur van hoofd-, deel- en subattributen. De attributen verwijzen naar deelvaardigheden, strategieën en kenniselementen zoals regels, procedures, begrippen en principes die deel uitmaken van de vaardigheid of sterk samenhangen met presteren op de totale vaardigheid. Het leerlingmodel schrijfvaardigheid Nederlands bestaat uit vier hoofdattributen en respectievelijk twaalf (vmbo) of dertien (havo/vwo) deelattributen (zie Figuur 7-1)

Hoofdattribuut	Deelattribuut
1. retorische vaardigheden (doel en publiek)	1.1. voorkennis en informatiebehoefte bij de lezer inschatten
	1.2 toonzetting afstemmen op de lezer (o.a. formeel - informeel)
	1.3 schrijfdoel bepalen (bijv. informeren, overtuigen, uitleggen, uitnodigen)
2. tekststructurele vaardigheden (structuur)	2.1 tekstelementen kiezen, rekening houdend met het genre
	2.2 passende volgorde van tekstelementen bepalen en correcte indeling en lay-out aanbrengen
	2.3 samenhang tussen tekstelementen aanbrengen (coherentie)
	2.4 een standpunt weergeven en van passende argumenten voorzien (alleen h/v)
3. linguïstische vaardigheden (woord- en zinsniveau)	3.1 een correcte zinsbouw hanteren
	3.2 een passende en cohesieve schrijfstijl hanteren
	3.3 passend en gevarieerd woordgebruik laten zien
4. orthografische vaardigheden (spelling en interpunctie)	4.1 correct spellen van werkwoorden
	4.2 correct toepassen van overige regelgeleide spelling
	4.3 leestekens en hoofdletters correct hanteren

Figuur 7-1 Leerlingmodel schrijfvaardigheid (CvTE, 2014)

De deelattributen worden systematisch getoetst in het gesloten deel van de schrijftoets, met behulp van verschillende vraagvormen. Naast de traditionele meerkeuze-items krijgen leerlingen ook opgaven voorgelegd waarbij ze bijvoorbeeld een tekst in alinea's moeten verdelen, tekstonderdelen in de juiste volgorde moeten slepen of fouten in teksten moeten aanklikken en verbeteren. Naast deze gesloten items, bevatte de try-out DTT ook enkele open schrijfopdrachten. In deze geleide stelopdrachten worden leerlingen uitgelokt zelf een tekst te schrijven over een gegeven onderwerp. Het toevoegen van open stelopdrachten aan de DTT schrijfvaardigheid zou van belang kunnen zijn voor de (indruks-)validiteit, het draagvlak in het veld en de diagnostische waarde van de DTT (Cito, 2012).

In de try-out van voorjaar 2014 zijn twee open schrijfopdrachten voorgelegd aan de leerlingen.
De schrijfopdrachten zijn weergegeven in Bijlage 7-1.

Het doel van de afname van deze opdrachten was tweeledig. Enerzijds was de afname bedoeld om informatie te verzamelen over het geautomatiseerd beoordelen van leerlingteksten. Door in de toekomst (een deel van de) leerlingteksten automatisch te beoordelen met programma's als T-Scan (tekscomplexiteit) en Valkuil (spelling), wordt docenten veel werk uit handen genomen, gezien het tijdrovende karakter van het laten beoordelen van schrijftaken door docenten. Op basis van de resultaten van de try-out 2013 bleek dat aanvullend onderzoek noodzakelijk is om de exacte werking van de maten voor tekstcomplexiteit in T-Scan te doorgronden, te controleren en waar nodig te verbeteren. Daarnaast bleek een nauwkeurigere automatische correctie van de spelling noodzakelijk (Cito, 2013). De verzamelde leerlingteksten in de try-out 2014 worden gebruikt om verder

onderzoek te verrichten naar het geautomatiseerd beoordelen. Daarnaast is op basis van de try-out 2013 geconcludeerd dat het naar verwachting op termijn mogelijk zal zijn om met behulp van automatische tekstanalyse een betrouwbare en betekenisvolle uitspraak te doen over de beheersing van een aantal aspecten van schrijfvaardigheid. De verwachting is echter dat voor de beoordeling van met name 'doel en publiek' en 'tekststructuur' menselijke beoordelaars nodig blijven (Cito, 2013). Om scholen tegemoet te komen bij de beoordeling van de leerlingteksten in het open deel van de toets, zijn de leerlingteksten die nu verzameld zijn in de try-out 2014 ook gebruikt voor een studie naar verschillende beoordelingsmodellen die door docenten gehanteerd zouden kunnen worden. Uit diverse onderzoeken blijkt dat de beoordeling van leerlingteksten een tijdrovende taak is waarbij het lastig is om een voldoende mate van overeenstemming en betrouwbaarheid te bereiken (o.a. Meestringa & Ravesloot, 2012, Schoonen & De Glopper, 1992). In een studie hebben wij daarom drie potentieel geschikte modellen getest op hun betrouwbaarheid en op hun bruikbaarheid voor eventuele opname in de DTT.

In de volgende paragraaf wordt verslag gedaan van de studie naar de drie beoordelingsmodellen. Eerst wordt de achtergrond van de verschillende beoordelingsmodellen beschreven, vervolgens gaan we in op de gehanteerde onderzoeksmethode, waarna we de resultaten beschrijven.

7.2 De beoordelingsmodellen

In de studie zijn drie potentiële beoordelingsmodellen getest: een 1) globaal beoordelingsmodel, een 2) beoordelingsmodel waarbij gebruikgemaakt wordt van ankeropstellen en een 3) analytisch beoordelingsmodel. Hieronder wordt elk van de modellen toegelicht.

7.2.1 Het globale beoordelingsmodel

Bij het globale model wordt de beoordelaar gevraagd een oordeel over een tekst te vellen aan de hand van een of enkele beoordelingspunten. Het globale oordeel verschilt van het impressieoordeel, waarbij er enkel een oordeel op basis van indruk gegeven wordt, zonder beoordelingspunten aan te geven. In dit onderzoek is ervoor gekozen om de globale beoordelingspunten overeen te laten komen met de vier hoofdattributen die in het leerlingmodel onderscheiden worden, zodat het gegeven oordeel zo nauw mogelijk kan aansluiten bij en zo goed mogelijk vergeleken kan worden met de uitkomsten van het gesloten deel van de DTT-schrijftoets. Dit betekent dat de beoordelaars een oordeel geven over 1) aansluiting bij doel & publiek, 2) de structuur van de tekst, 3) het woordgebruik en de zinsbouw en 4) de spelling en interpunctie. De gebruikte schaal sluit ook aan bij het gesloten deel van de toets: beoordelaars wordt gevraagd om aan te geven of een leerling voor het betreffende hoofdattribuut onder, op of boven het leerlingniveau presteert.

Een van de verwachte voordelen van het globale beoordelingsmodel is dat de beoordeling relatief snel kan plaatsvinden, doordat er maar op vier punten beoordeeld hoeft te worden. Dit in tegenstelling tot meer analytische modellen die als relatief tijdrovend te boek staan. Eerder onderzoek laat bovendien zien dat het globale oordeel even betrouwbaar kan zijn als meer analytische beoordelingen. Door de hoofdattributen te bevragen wordt daarnaast gewaarborgd dat de hoofdpunten uit het leerlingmodel in de beoordeling terugkomen.

7.2.2 Het model met gebruik van ankeropstellen

Een van de moeilijke punten bij het beoordelen van leerlingteksten is het leggen van de grens tussen verschillende punten op een beoordelingsschaal. In dit onderzoek wordt gevraagd om aan te geven of een tekst op een bepaald beoordelingspunt onder, op of boven het betreffende onderwijsniveau gesitueerd moet worden. Maar waar precies de grens ligt tussen onder en op, en tussen op en boven wordt bij het globale model aan de beoordelaar overgelaten. Wanneer verschillende beoordelaars eenzelfde tekst moeten beoordelen, kunnen zij verschillen in het oordeel over waar die grens precies ligt. Bij het gebruik van ankeropstellen wordt geprobeerd aan deze moeilijkheid tegemoet te komen. Ankeropstellen zijn leerlingteksten die een bepaald punt op een schaal of een bepaalde grens markeren. In dit onderzoek is ervoor gekozen om per hoofdattribuut twee ankeropstellen te selecteren: een ankeropstel dat voor een bepaald onderwijsniveau (vmbo-tl, havo of vwo) de grens tussen onder en op niveau markeert en een ankeropstel dat de grens tussen op en boven niveau aangeeft. Voor vmbo-bb en vmbo-kb zijn er geen ankeropstellen geselecteerd gezien het te geringe aantal leerlingteksten dat voor dit niveau verzameld was.

Om tot een goede selectie van ankeropstellen te komen, zijn per onderwijsniveau 100[3] willekeurige leerlingteksten geselecteerd. Een toetsdeskundige heeft deze teksten doorgelezen en verdeeld in vijf stapels:

1. leerlingteksten die duidelijk onder niveau waren voor het betreffende hoofdattribuut
2. leerlingteksten die zich voor het betreffende hoofdattribuut bevonden op de grens onder/op niveau
3. leerlingteksten die op niveau waren voor het betreffende hoofdattribuut
4. leerlingteksten die zich voor het betreffende hoofdattribuut bevonden op de grens op/boven niveau
5. leerlingteksten die duidelijk boven niveau waren voor het betreffende hoofdattribuut

Vervolgens zijn de teksten die in stapel 2 en stapel 4 terechtgekomen waren opnieuw doorgenomen en heeft de toetsdeskundige drie teksten uit deze stapels geselecteerd die volgens hem of haar het best de grens onder/op markeerden en drie teksten die het best de grens op/boven niveau markeerden.

Nadat op deze wijze voor ieder onderwijsniveau per hoofdattribuut en per grens drie teksten geselecteerd waren, zijn ze naast elkaar gelegd en bekeken door drie toetsdeskundigen. Die hebben vervolgens uit de geselecteerde drie teksten telkens een tekst gekozen die het uiteindelijke ankeropstel zou vormen. Hierbij is ernaar gestreefd om ankerteksten te selecteren die enerzijds een duidelijk beeld gaven van het betreffende hoofdattribuut en anderzijds een oplopend niveau lieten zien tussen de onderwijsniveaus.

7.2.3 Het analytische model

Het derde beoordelingsmodel dat getest is in dit onderzoek is het analytische model. Bij een analytisch model wordt een tekst beoordeeld aan de hand van meerdere beoordelingspunten die bij elkaar een zo nauwkeurig mogelijk beeld geven van de schrijfvaardigheidsniveau van een leerling. In dit onderzoek is ervoor gekozen om het analytische model te laten aansluiten bij de deelattributen van het leerlingmodel. Voor vmbo betreft dit 12 beoordelingspunten en voor havo/vwo door de toevoeging van deelattribuut 2.4 13 beoordelingspunten. Deelattributen die bestaan uit meerdere elementen zijn opgesplitst in meerdere beoordelingspunten. Zo is deelattribuut 3.3. passend en gevarieerd woordgebruik beoordeeld aan de hand van twee punten: de vraag of het woordgebruik passend is en de vraag of het woordgebruik gevarieerd is. Daarnaast is, nadat de deelattributen die behoren tot hetzelfde hoofdattribuut beoordeeld waren, ook een oordeel gegeven op hoofdattribuutniveau. Hiermee is het mogelijk om een vergelijking te maken tussen het analytische model en de overige modellen waarbij ook een oordeel op hoofdattribuutniveau gevraagd wordt.

Het analytische model is het model dat naar verwachting het meest gebruikt wordt in het onderwijsveld. Het idee achter het analytische model is dat het door zijn gedetailleerdheid volgens velen het meest recht doet aan de schrijfvaardigheid van leerlingen. Tegelijkertijd is het door zijn gedetailleerdheid echter ook een tijdrovend model.

7.3 Methode

Voor de beoordeling van de leerlingteksten zijn 8 beoordelaars geworven die in september 2014 de verzamelde leerlingteksten beoordeeld hebben. Bij de werving van de beoordelaars was een van de vereisten dat de beoordelaars minimaal enige leservaring hadden in het vak Nederlands in het voortgezet onderwijs, hetzij als docent Nederlands, hetzij studerend aan de lerarenopleiding. Op die manier werd gewaarborgd dat de modellen getest zouden worden door de doelgroep voor wie ze ontwikkeld zijn.

De 480 verzamelde leerlingteksten zijn beoordeeld volgens het design dat weergegeven is in Figuur 7-2. Hierbij is ervoor gezorgd dat iedere beoordelaar een tekst beoordeelt volgens twee modellen: globaal & analytisch, globaal & anker, anker & analytisch, waardoor overeenstemming en verschillen tussen modellen vastgesteld kunnen worden. Daarnaast wordt iedere tekst voor ieder model minimaal door twee personen beoordeeld, waardoor interbeoordelaarsovereenstemming voor de verschillende modellen bepaald kan worden.

[3] Voor vmbo-tl is de selectie gebaseerd op alle 86 verzamelde leerlingteksten.

	Opstel 1-60	Opstel 61-120	Opstel 121-180	Opstel 181-240	Opstel 241-360	Opstel 301-360	Opstel 361-420	Opstel 421-480
Analytisch	1	2	7	8	1	2	5	6
	5	6	3	4	7	8	3	4
Anker	1	3	2	4	5	8	7	6
	4	2	3	1	7	6	5	8
Globaal	5	3	7	1	1	6	7	4
	4	6	2	8	5	2	3	8

Figuur 7-2 Design beoordeling leerlingteksten

Uiteindelijk heeft iedere beoordelaar op basis van dit design 3 sets (geel, groen, oranje) = 360 leerlingteksten beoordeeld. De beoordeling vond plaats van globaal naar specifiek: eerst beoordeelde de beoordelaar de teksten volgens het globale model, vervolgens volgens het ankermodel en ten slotte volgens het analytische model. Op die manier werd voorkomen dat de gedetailleerde analytische beoordeling de beoordelingen volgens het globale en het ankermodel zou beïnvloeden.

Voordat er gestart werd met het beoordelen van de leerlingteksten, woonden alle beoordelaars een introductie-bijeenkomst bij. Tijdens deze bijeenkomst werden de achtergronden van het project geschetst, werd ieder beoordelingsmodel kort toegelicht en werd er kort geoefend met de verschillende modellen op basis van teksten die niet meededen aan het daadwerkelijke onderzoek. Het beoordelen zelf vond achter de computer plaats, waarbij de leerlingteksten zowel op scherm als op papier beschikbaar waren voor de beoordelaars. Indien een tekst niet beoordeelbaar werd geacht, bijvoorbeeld omdat de lengte te kort was of de uitwerking wezenlijk verschilde van de opdracht, werd deze gecodeerd als ‘niet beoordeelbaar’. De teksten waarvoor dit gold, zijn niet meegenomen in analyses.

Tijdens het beoordelen zijn ook gegevens verzameld over de ervaringen met het werken met het model, met het oog op de vaststelling van de bruikbaarheid. Nadat een beoordelaar gereed was met de beoordeling volgens een bepaald model is een vragenlijst afgenomen door een van de onderzoekers. Met behulp van die vragenlijst werden gegevens verzameld over de ervaringen met het werken met de specifieke modellen en de tijdsinvestering. Aan het eind zijn ook enkele vragen voorgelegd waarmee de modellen vergeleken werden op hun kwaliteit, bruikbaarheid en efficiëntie.

7.4 Resultaten

7.4.1 Betrouwbaarheid van de modellen

In Tabel 7-1 is weergegeven hoeveel teksten volgens ieder beoordelingsmodel beoordeeld zijn. Enkele teksten zijn afgevallen omdat ze niet beoordeelbaar bleken te zijn gezien de tekstlengte (te kort) of omdat ze te veel afweken van de opgegeven schrijfopdracht.

Tabel 7-1 Aantal beoordeelde leerlingteksten

	vmbo	havo/vwo
A. Globaal beoordelingsmodel	109	340
B. Beoordeling met ankerteksten	110	341
C. Analytisch beoordelingsmodel	108	344

Voordat de betrouwbaarheid van de verschillende modellen in kaart gebracht is, is gecontroleerd of er geen opvallende verschillen tussen de beoordelaars optraden, dat wil zeggen of een bepaalde beoordelaar niet opvallend strenger of milder oordeelde dan de andere beoordelaars. Dit bleek niet het geval te zijn.

Hieronder geven we de eerste resultaten van de betrouwbaarheid van de verschillende modellen weer. We beperken ons tot weergave van de resultaten per model op hoofdattribuutniveau en tot vergelijkingen tussen modellen op hetzelfde hoofdattribuutniveau. Het gaat om eerste analyses die met de nodige voorzichtigheid geïnterpreteerd moeten worden. De komende tijd zullen er verdere analyses uitgevoerd worden waarmee de resultaten nauwkeuriger in beeld gebracht kunnen worden. Dan zullen de resultaten van het analytische model ook op deelattribuutniveau in kaart gebracht worden en kunnen verdere vergelijkingen tussen de verschillende modellen gemaakt worden.

7.4.1.1 Uitkomsten van het globale model

Bij de beoordeling van de vmbo-teksten is gebruikgemaakt van een zevenpuntsschaal. De gebruikte schaal voor havo/vwo bestond uit een vijfpuntsschaal.

	bb		**kb**		**gt**	
○	○	○	○	○	○	○
1	2	3	4	5	6	7

	havo		**vwo**	
○	○	○	○	○
1	2	3	4	5

Figuur 7-3 Gebruikte schalen bij het globale model

In Tabel 7-2 zijn de gemiddelde waarden weergegeven per hoofdattribuut. Bij vmbo zijn teksten op alle attributen gemiddeld beoordeeld boven het niveau vmbo-kb en onder het niveau vmbo-gt. Hierbij moet in ogenschouw genomen worden dat er slechts een beperkt aantal vmbo-bb en vmbo-kb teksten beoordeeld zijn, waardoor het verklaarbaar is dat het gemiddelde op vmbo-gt-niveau ligt. Bij de havo/vwo-teksten ligt het gemiddelde net boven het havo-niveau. Interessanter is echter dat alle hoofdattributen ongeveer op hetzelfde niveau beoordeeld zijn. Volgens de beoordelaars presteren de leerlingen op alle attributen gemiddeld genomen ongeveer even goed en is er geen attribuut dat er in positieve of in negatieve zin tussenuit springt.

Tabel 7-2 Gemiddelden (*gem*), standaarddeviatie (*sd*), samenhang (*corr*) en overeenstemming (*kappa*) globale model

	vmbo (N=109)				**havo/vwo (N=340)**			
	gem.	sd	corr	kappa	gem.	sd	corr	kappa
Attribuut 1: doel en publiek	5.17	1.43	0.81	0.66	2.60	1.19	0.75	0.51
Attribuut 2: structuur	5.05	1.53	0.80	0.57	2.53	1.15	0.77	0.55
Attribuut 3: woord- en zinsniveau	4.94	1.34	0.82	0.55	2.46	1.14	0.79	0.61
Attribuut 4: spelling en interpunctie	4.90	1.35	0.86	0.66	2.55	1.27	0.86	0.64

Om de betrouwbaarheid in kaart te brengen is per hoofdattribuut de Pearson's correlatie en de gewogen coëfficiënt kappa (lineair) berekend. De correlatie tussen de beoordelaars varieert van 0.75 tot 0.86 (zie Tabel 7-2). De kappa varieert van 0.55 tot 0.66. Er is dus sprake van een sterke samenhang en een matige tot redelijke overeenstemming. Van alle hoofdattributen is de overeenstemming het hoogst op attribuut 4: spelling en interpunctie.

7.4.1.2 Uitkomsten van het ankermodel

Bij de beoordeling van het ankermodel is er voor alle onderwijsniveaus gebruik gemaakt van onderstaande driepuntsschaal.

onder niveau	op niveau	boven niveau
○	○	○
1	2	3

Figuur 7-4 Gebruikte schaal bij het ankermodel

Uit Tabel 7-3 valt op te maken dat op alle attributen gemiddeld tussen 'onder niveau' en 'op niveau' beoordeeld is. Zowel bij vmbo als bij havo/vwo valt attribuut 3 'woord- en zinsniveau' het laagst uit. Bij vmbo zijn de oordelen het hoogst op attribuut 1 'doel en publiek', bij havo/vwo op attribuut 4 'spelling en interpunctie'.

Tabel 7-3 Gemiddelden (*gem*), standaarddeviatie (*sd*), samenhang (*corr*) en overeenstemming (*kappa*) ankermodel

	vmbo (N=110)				havo/vwo (N=341)			
	gem.	sd	corr	kappa	gem.	sd	corr	kappa
Attribuut 1: doel en publiek	1.72	0.69	0.43	0.30	1.69	0.64	0.41	0.33
Attribuut 2: structuur	1.62	0.70	0.42	0.31	1.68	0.67	0.38	0.28
Attribuut 3: woord- en zinsniveau	1.58	0.63	0.57	0.36	1.66	0.63	0.35	0.25
Attribuut 4: spelling en interpunctie	1.68	0.71	0.54	0.43	1.84	0.75	0.60	0.46

Kijken we naar de samenhang en interbeoordelaarsovereenstemming (zie Tabel 7-3) voor het ankermodel dan valt op dat de waarden wat lager liggen dan bij het globale model. De vergelijking tussen de twee modellen kan echter niet zo maar getrokken worden: omdat de schaal van het ankermodel slechts uit drie punten bestaat is het verklaarbaar dat de waarden lager uitvallen dan bij het globale model. Er is slechts een beperkt aantal meetpunten, waardoor de coëfficiënt kappa op voorhand al lager zal uitvallen. Dit in ogenschouw nemend, is de overeenstemming tussen de beoordelaars bij dit model ook matig tot redelijk te noemen. Om echter een goede vergelijking te kunnen maken tussen het ankermodel en het globale model, dient de vijf- of zevenpuntsschaal van het globale model eerst getransformeerd te worden naar een driepuntsschaal. Dat is hieronder gebeurd voor de niveaus havo en vwo in Tabel 7-4.

Tabel 7-4 Vergelijking van de samenhang en interbeoordelaarsovereenstemming bij het globale en het ankermodel

		havo (N=183)		vwo (N=157)	
	Model	corr	kappa	corr	kappa
Attribuut 1: doel en publiek	Globaal	0.20	0.13	0.14	0.17
	Anker	0.44	0.34	0.37	0.30
Attribuut 2: structuur	Globaal	0.40	0.31	0.29	0.24
	Anker	0.36	0.26	0.41	0.30
Attribuut 3: woord- en zinsniveau	Globaal	0.28	0.22	0.12	0.17
	Anker	0.32	0.24	0.37	0.26
Attribuut 4: spelling en interpunctie	Globaal	0.54	0.47	0.41	0.28
	Anker	0.54	0.40	0.65	0.52

Uit deze Tabel 7-4 valt op te maken dat de interbeoordelaarsovereenstemming bij het globale model na gelijktrekking van de schalen over het algemeen iets lager ligt dan de overeenstemming bij het ankermodel. Alleen bij het attribuut structuur bij havo valt de overeenstemming bij het globale model (iets) hoger uit.

Verder onderzoek zal moeten uitwijzen of door de inzet van de ankerteksten de beoordelaars toch iets meer overeenkomen in hun oordelen dan zonder de ankerteksten.

7.4.1.3 Uitkomsten van het analytische model

Bij het analytische model is dezelfde zeven- en vijfpuntsschaal ingezet als bij het globale model (Figuur 7-5). De resultaten voor het analytische model zijn weergegeven op hoofdattribuutniveau. Nadere analyses op deelattribuutniveau volgen in een vervolgonderzoek.

	bb		kb		gt	
○	○	○	○	○	○	○
1	2	3	4	5	6	7

	havo		vwo	
○	○	○	○	○
1	2	3	4	5

Figuur 7-5 Gebruikte schalen bij het analytische model

Net als bij het globale model liggen de gemiddelde waarden boven het niveau vmbo-kb en onder het niveau vmbo-gt. Bij havo-vwo liggen ze gemiddeld net boven het havo-niveau. Ook hier is er geen attribuut dat in positieve of in negatieve zin opvalt (zie Tabel 7-5).

Tabel 7-5 Gemiddelden (*gem*), standaarddeviatie (*sd*), samenhang (*corr*) en overeenstemming (*kappa*) analytische model

	vmbo (N=108)				havo/vwo (N=344)			
	mean	sd	corr	kappa	mean	sd	cor	kappa
Attribuut 1: doel en publiek	5.11	1.33	0.79	0.59	2.67	1.12	0.76	0.59
Attribuut 2: structuur	4.99	1.37	0.80	0.64	2.63	1.11	0.81	0.65
Attribuut 3: woord- en zinsniveau	4.99	1.35	0.78	0.57	2.70	1.11	0.82	0.69
Attribuut 4: spelling en interpunctie	5.04	1.37	0.77	0.54	2.71	1.20	0.84	0.71

Kijken we naar de interbeoordelaarsovereenstemming, dan valt op dat de samenhang (redelijk) sterk is en de overeenstemming matig tot redelijk is. Bij vmbo is de overeenstemming het hoogst op het beoordelingspunt 'structuur', bij havo/vwo op het punt spelling.

Vergelijken we de interbeoordelaarsbetrouwbaarheid tussen het globale en het analytische model (Tabel 7-6), dan valt op dat er voor vmbo geen duidelijke verschillen aan te wijzen zijn tussen het globale en het analytische model. Alleen bij havo/vwo lijkt het analytische model een iets hogere overeenstemming te geven dan het globale model.

Tabel 7-6 Vergelijking van de samenhang en interbeoordelaarsovereenstemming bij het globale en het analytische model

	Model	vmbo		havo/vwo	
		corr	kappa	corr	kappa
Attribuut 1: doel en publiek	globaal	0.81	0.66	0.75	0.51
	analytisch	0.79	0.59	0.76	0.59
Attribuut 2: structuur	globaal	0.80	0.57	0.77	0.55
	analytisch	0.80	0.64	0.81	0.65
Attribuut 3: woord- en zinsniveau	globaal	0.82	0.55	0.79	0.61
	analytisch	0.78	0.57	0.82	0.69
Attribuut 4: spelling en interpunctie	globaal	0.86	0.66	0.86	0.64
	analytisch	0.77	0.54	0.84	0.71

7.4.2 Bruikbaarheid van de beoordelingsmodellen

Na iedere beoordelingsronde zijn de beoordelaars bevraagd over hun ervaringen met het werken met het specifieke model. In de eerste vragenlijst is de beoordelaars gevraagd naar hun ervaringen met het computer-programma waarmee beoordeeld moest worden. De beoordelaars waren positief over het werken met dit programma: ze werden niet afgeleid door niet goed functionerende onderdelen en ze gaven ook aan het fijn te vinden de teksten zowel op papier als op scherm ter beschikking te hebben. Ook bij het beoordelen van de teksten aan de hand van de overige modellen zijn op dit punt geen problemen gesignaleerd. Praktische zaken hebben de beoordeling dus niet beïnvloed. Hieronder zullen de ervaringen met de verschillende modellen afzonderlijk besproken worden.

7.4.2.1 Ervaringen met het globale model

De beoordelaars zijn bewust redelijk vrij gelaten in hun aanpak van het werken met het globale model. Iedere beoordelaar gaf echter aan als voorbereiding op het beoordelen dezelfde aanpak gehanteerd te hebben.
Ze hadden allen een bepaald aantal leerlingteksten per onderwijsniveau doorgelezen, alvorens ze teksten van het betreffende niveau gingen beoordelen.

De beoordelaars gaven aan de vier gehanteerde beoordelingspunten van het globale model duidelijk geformuleerd te vinden. Echter, op de vraag in hoeverre de vier beoordelingspunten toepasbaar waren op de te beoordelen teksten, gaven drie personen het antwoord 'redelijk goed' en vijf personen het antwoord 'niet zo goed'. Dit werd volgens de meeste beoordelaars veroorzaakt door het feit dat in alle beoordelingspunten behalve structuur

meerdere punten samengenomen waren. Hierdoor gaven beoordelaars aan vaak op een oordeel 'op niveau' uitgekomen te zijn, terwijl ze eigenlijk het ene punt als 'onder of op niveau' en het andere als 'op of boven niveau' hadden willen beoordelen. Dit speelde met name bij punt 4 'spelling en interpunctie', omdat in de te beoordelen teksten de interpunctie volgens de beoordelaars meestal ver achterbleef bij de spelling. Daarnaast was bij het attribuut 'doel en publiek' niet goed beoordeelbaar doordat het beoogde publiek in de schrijfopdracht zelf niet duidelijk genoeg benoemd werd. Het was voor de leerlingen niet duidelijk genoeg of ze hun reactie moesten richten aan de schoolleiding of aan de gymleraren (vmbo) of de wiskundedocenten (havo/vwo). Verder gaven vijf beoordelaars aan dat ze graag nog beoordelingspunten aan het beoordelingsmodel toegevoegd hadden willen zien met betrekking tot de inhoud/argumentatie en tekstlengte. Hierbij werd door vier beoordelaars genoemd dat de leerlingen soms een inhoudelijk niet kloppend of qua onderwerp sterk afwijkend verhaal konden schrijven, en toch hoog konden scoren op de meeste punten van het beoordelingsmodel. En hoewel de beoordelaars geïnstrueerd waren om te korte teksten niet te beoordelen, blijkt dit toch een grijs gebied te zijn. Sommige teksten waren wel beoordeelbaar, maar ook kort, waarbij er door de korte lengte weinig fouten gemaakt konden worden in de spelling en interpunctie. Bij leerlingen die langere teksten geschreven hadden, is de kans op fouten groter.

Wat betreft de schaal *onder, op* en *boven niveau*, gaven vier beoordelaars aan de schaal handig tot erg handig te vinden, terwijl drie beoordelaars aangaven hem niet zo handig te vinden en een persoon hem helemaal niet handig vond. De beoordelaars die minder positief waren, vonden het met name lastig te bepalen wat het gemiddelde niveau moest zijn.

Het laatste punt dat bevraagd werd, was de tijdsbesteding per tekst. Hierbij werd de beoordelaars gevraagd in te schatten hoe lang ze dachten aan de beoordeling van een tekst besteed te hebben inclusief het lezen van de tekst. Het aangegeven gemiddelde ligt op 6,6 minuten per tekst. Drie beoordelaars gaven aan geen goede inschatting te kunnen maken, omdat ze de tijd niet hadden bijgehouden.

De beoordelaars werd ook gevraagd wat ze van de bestede tijd per tekst vonden. Zes beoordelaars gaven aan de tijdsinvestering precies goed te vinden. Een beoordelaar gaf aan liever wat sneller te willen beoordelen, maar weet dit aan haar eigen langzame leestempo. De laatste beoordelaar had graag wat meer tijd aan een tekst besteed door meer beoordelingspunten aan het model toe te voegen.

7.4.2.2 Ervaringen met het ankermodel

Net als bij het globale model gaven de beoordelaars aan als voorbereiding opnieuw enkele teksten van een bepaald onderwijsniveau doorgelezen te hebben. Ook had iedereen de ankerteksten zelf van tevoren doorgelezen en sommige beoordelaars hadden opvallende passages in deze teksten aangestreept.

Aangezien de beoordelingspunten van het ankermodel overeenkwamen met die van het globale model, is de duidelijkheid van de beoordelingspunten niet opnieuw bevraagd. Wel is gevraagd naar de bruikbaarheid en toepasbaarheid van de ankerteksten zelf. De beoordelaars oordeelden wisselend over hun ervaringen met de ankerteksten. In sommige gevallen vond men de ankerteksten van een te hoog niveau, waardoor ze te veel teksten als 'onder niveau' moesten beoordelen. Dit gold met name voor de havo-teksten voor hoofdattribuut 2 (structuur) en voor de vmbo- en havo/vwo-teksten voor hoofdattribuut 3 (woordgebruik en zinsniveau). Daarnaast gaven vier beoordelaars aan de verschillen tussen de ankertekst voor de grens onder/op-niveau en de tekst voor de grens op/boven niveau niet altijd even duidelijk genoeg was. Ze zouden in een vervolg graag zien dat bij de ankerteksten een korte toelichting zou worden gegeven met de belangrijkste criteria voor de selectie van deze teksten. Dit is in dit onderzoek overigens bewust niet op voorhand gedaan, om te voorkomen dat de beoordelaars zich zouden laten leiden door de criteria en niet door de ankertekst zelf.

Bij het globale model en het analytische model konden de beoordelaars hun oordeel plaatsen op een schaal van respectievelijk zeven (vmbo) en vijf (havo/vwo) punten (zie paragraaf 7.4). Bij het ankermodel was dit niet mogelijk, omdat voor ieder onderwijsniveau andere ankerteksten geselecteerd waren. Enkele beoordelaars gaven aan het jammer te vinden dat ze hierdoor 'uitschieters', teksten die veel hoger of juist veel lager waren dat het gegeven onderwijsniveau, niet konden plaatsen op een ander onderwijsniveau.

Ook bij het ankermodel is gevraagd naar een inschatting van de bestede tijd per tekst. De bestede tijd ligt hoger dan bij het globale model. Het gemiddelde in minuten per beoordelaar komt hier uit op 10.80 minuten per tekst. Zes van de acht beoordelaars gaven aan dit 'iets te veel' of 'best wel veel' te vinden. Daarbij speelde mee dat de tekst op vier beoordelingspunten beoordeeld moest worden en dat iedere tekst dus vergeleken moest worden met in totaal acht ankerteksten (twee per hoofdattribuut).

7.4.2.3 Ervaringen met het analytische model

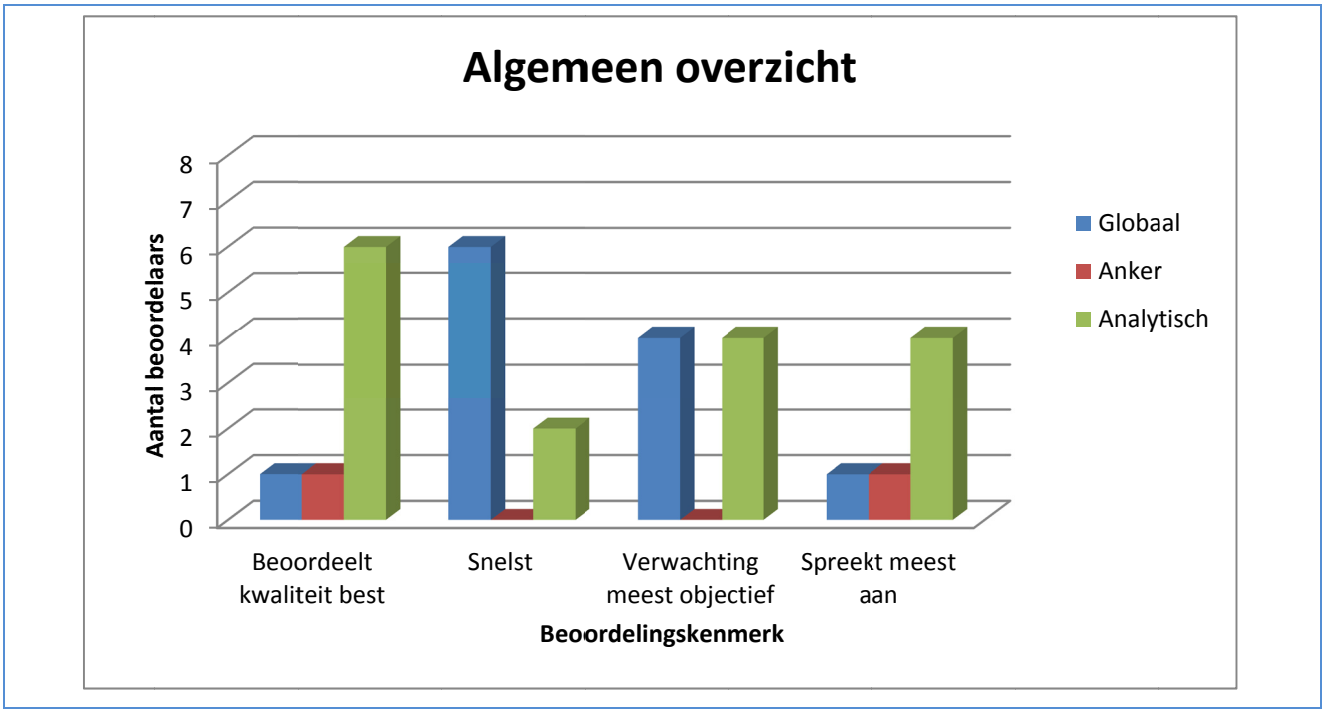
De manier van voorbereiding op het werken met dit laatste model was vergelijkbaar met de voorbereiding bij de andere modellen. Wel gaven de beoordelaars aan dat ze inmiddels zo veel teksten gezien hadden dat ze wel een beeld hadden van het gemiddelde niveau. Hierdoor werden er minder teksten doorgelezen als voorbereiding.

Het analytische model bestaat uit een groter aantal beoordelingspunten dan de andere modellen. Daarom zijn de beoordelaars net als bij het globale model bevraagd naar de duidelijkheid van de formulering van de beoordelingspunten. Alle beoordelaars waren van mening dat de beoordelingspunten duidelijk geformuleerd waren, waarbij wel kanttekeningen geplaatst werden bij enkele punten. Zo was het de beoordelaars niet helemaal duidelijk wat er bedoeld werd met 'tekstelementen' (deelattribuut 2.1). Daarnaast was het onderscheid tussen deelattribuut 3.3 'passend woordgebruik' en 3.2 'passende schrijfstijl' niet voor alle beoordelaars even duidelijk, evenals het verschil tussen 3.2 en 1.2 'stemt toon af op de lezer'. Over de toepasbaarheid van de beoordelingspunten waren de beoordelaars tevreden. Wel gaf een beoordelaar aan dat de vraag naar de lay-out van de tekst (onderdeel van deelattribuut 2.2) wellicht overbodig is, omdat vrijwel iedere tekst door de afname op de computer dezelfde lay-out had. Daarnaast gaven twee beoordelaars aan dat het concentratievermogen van hen gedurende het beoordelen afnam door de hoeveelheid beoordelingspunten.

Opnieuw werd gevraagd naar de tijdsbesteding per tekst. De gemiddelde beoordelingstijd per tekst komt bij dit model neer op 7.75 minuten, wat meer is dan de ingeschatte tijd bij het globale model, maar minder dan bij het ankermodel. Hierbij moet wel een kanttekening geplaatst worden: doordat het analytische model het laatste model was, hadden de beoordelaars alle te beoordelen teksten al een keer langs zien komen. Daardoor kan de geschatte tijdsinvestering geflatteerd zijn en zou deze waarschijnlijk hoger uitgevallen zijn wanneer men als eerste met het analytische model had gewerkt. Zes van de acht beoordelaars gaven aan de beoordelingsduur precies goed te vinden, de overige twee vonden de duur aan de lange kant.

7.4.2.4 Vergelijking van de ervaringen met de verschillende modellen

Nadat de beoordelaars hun beoordelingen afgerond hadden, is hun ook gevraagd om hun ervaringen met de verschillende modellen op enkele punten te vergelijken. De resultaten daarvan zijn weergegeven in Figuur 7-6.



Figuur 7-6 Vergelijkende oordelen over de verschillende modellen

Op de vraag welk beoordelingsmodel het meest recht doet aan de kwaliteit van de tekst, kiezen zes beoordelaars voor het analytische model. Een beoordelaar gaat voor het globale model en een beoordelaar voor het ankermodel. Het beoordelen met het globale model is volgens de meeste beoordelaars echter het minst tijdrovend. Twee beoordelaars vinden het analytische model echter het snelst gaan, omdat de 'kleine' deelpunten die daarbij onderscheiden werden snel te beoordelen waren. Bij het globale model gaven zij aan langer te moeten nadenken over hun oordeel. Ook werd de beoordelaars gevraagd aan te geven bij welk model zij verwachtten dat de onderlinge oordelen het meest overeen zouden komen. Het ankermodel werd hierbij niet genoemd.
Vier beoordelaars kozen voor het globale model en vier beoordelaars voor het analytische model.

Tot slot werd aan de beoordelaars gevraagd met welk model zij het liefst zouden willen werken. Het analytische model geniet de voorkeur van vier beoordelaars. Een beoordelaar kiest voor het globale model en een beoordelaar voor het ankermodel. Twee beoordelaars zouden het liefst een combinatie van twee modellen terugzien. Een beoordelaar gaat daarbij voor een combinatie van het globale en het analytische model, omdat het eerste model volgens haar te weinig en het laatste model te veel beoordelingspunten had. De andere beoordelaar gaf aan een combinatie van het anker- en het analytische model het best te vinden.

7.5 Conclusie en discussie

De DTT schrijfvaardigheid Nederlands bestaat idealiter uit een open en een gesloten deel. In het open deel, dat onder andere uit het oogpunt van validiteit in de toets opgenomen is, krijgen leerlingen de opdracht zelf een tekst te schrijven. Dit onderdeel wordt in de toekomst zo mogelijk (deels) geautomatiseerd gescoord. Twee van deze schrijfopdrachten zijn in de try-out van 2014 afgenomen. Deze afname diende ten eerste om informatie te verzamelen die bruikbaar is voor het ontwikkelen van het geautomatiseerd beoordelen. Daarnaast zijn de verzamelde teksten ook gebruikt om informatie te verzamelen over de betrouwbaarheid en bruikbaarheid van drie verschillende beoordelingsmodellen die wellicht aangeboden kunnen worden als handvat aan de docent tot het moment dat geautomatiseerd beoordelen volledig mogelijk is. Het betreft een globaal model, een model waarbij gebruik gemaakt wordt van ankerteksten en een analytisch model. Acht beoordelaars hebben de verzamelde leerlingteksten beoordeeld aan de hand van deze modellen.

In dit verslag zijn de eerste resultaten van deze beoordelingsronde weergegeven. De resultaten moeten met de nodige voorzichtigheid beschouwd worden: verdere analyses zijn nodig om goede conclusies te kunnen trekken over (de vergelijking van) de betrouwbaarheid van de verschillende modellen.

De eerste analyses laten zien dat er bij het globale model en het analytische model geen attributen zijn die opvallend lager of hoger dan de andere attributen beoordeeld worden. Bij het ankermodel lijkt attribuut 3 (woord- en zinsniveau) iets lager beoordeeld te worden dan de andere attributen. Dit kan wellicht verklaard worden door de geselecteerde ankerteksten: de beoordelaars gaven aan dat ze de ankerteksten voor dit attribuut van een (te) hoog niveau vonden, waardoor veel leerlingteksten als 'onder niveau' beoordeeld zijn. De interbeoordelaarsovereenstemming lijkt voor alle modellen redelijk goed te zijn. Na gelijktrekking van de schalen van het globale en het ankermodel, blijkt de overeenstemming bij het ankermodel nog iets hoger te liggen dan de overeenstemming bij het globale model. Wellicht zorgen de ankerteksten ervoor dat de beoordelaars iets uniformer beoordelen. Verdere analyses moeten echter uitwijzen of de verschillen significant zijn. Bij vergelijking van het globale en het analytische model lijken er voor vmbo geen verschillen te zijn in interbeoordelaarsovereenstemming. Bij havo/vwo neigt het ernaar dat de overeenstemming bij het analytische model iets hoger is dan bij het globale model. Verdere analyses zullen echter moeten uitwijzen of dit inderdaad het geval is.

De beoordelaars zijn ook bevraagd naar hun ervaringen met het werken met de verschillende modellen. Hoewel de betrouwbaarheid van het ankermodel zeker niet lijkt onder te doen voor de andere modellen (en misschien nog wel beter is), heeft dit model niet de eerste voorkeur van de beoordelaars. Ze vinden het beoordelen aan de hand van deze teksten nog niet zo eenvoudig en plaatsen ook hun vraagtekens bij de selectie van sommige teksten. Daarnaast wordt het werken met het ankermodel als tijdrovend ervaren gezien de vele ankerteksten. De beoordeling met het globale model gaat volgens de beoordelaars het efficiëntst: dit model kost het minste tijd. Het analytische model spreekt de beoordelaars echter het meest aan. Door de vele beoordelingspunten hebben ze het idee dat ze het meest recht doen aan de leerlingteksten. De tijdsinvestering die ingeschat wordt, ligt hoger dan het globale model, maar lager dan het ankermodel. Deze inschatting is echter wellicht wat geflatteerd: de beoordelaars hadden de leerlingteksten allemaal al een keer gezien, waardoor de beoordeling vermoedelijk versneld werd. Bij alle modellen gaven de beoordelaars overigens verbeterpunten aan. Voordat de modellen uitgezet zouden worden in het veld is het raadzaam te onderzoeken wat de effecten van deze aanpassingen zullen zijn.

7.6 Referenties

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

Cito (2013). *Diagnostische tussentijdse toets: onderzoek*. Arnhem: Cito.

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

Meestringa, T., & Ravesloot, C. (2012). *Het schoolexamen Nederlands in de tweede fase vo. Uitkomsten van een enquête.* Enschede: SLO.

Schoonen, R., & De Glopper, K. (1992). Toetsing van schrijfvaardigheid: problemen en mogelijkheden. *Levende Talen*, (470), 187-196.

8 Beoordeling open vragen schrijfvaardigheid Engels

8.1 Inleiding

Net zoals bij Nederlands het geval was (zie hoofdstuk 7), zijn in de try-out 2014 voor schrijfvaardigheid Engels ook enkele open schrijfopdrachten voorgelegd aan leerlingen. Engelse schrijfproducten zijn in de toekomst nodig voor het ontwikkelen en evalueren van computerprogramma's voor automatische beoordeling. Op dit moment is automatische beoordeling echter nog niet in ontwikkeling is voor Engels. Indien besloten wordt om voor die tijd open schrijfopdrachten in te zetten bij de DTT zullen de schrijfproducten door docenten beoordeeld moeten worden. In dit onderzoek willen wij nagaan of - en welke - beoordelingsschema's goed werken.

Vorig jaar, in de try-out 2013, is gewerkt met een analytisch beoordelingsmodel, waarbij alle deelattributen uit het leerlingmodel voor schrijfvaardigheid werden beoordeeld. Dit model bleek niet optimaal te zijn. Zo waren sommige deelattributen niet op alle niveaus even relevant (bijvoorbeeld wanneer een deelattribuut niet bleek uit een schrijfproduct). Soms werden ook beoordelingspunten overgeslagen door de beoordelaars doordat de punten niet herkenbaar waren (bijv. idiomatische uitdrukkingen). Andere beoordelingspunten bleken lastig te meten in de afgenomen open schrijfopdrachten zoals 'afstemming op doel en publiek' en 'compenserende strategieën' (Verslag Try-out, Cito 2013).

Een tweede reden voor de open schrijfopdrachten is om de kwaliteit van een beoordelingsschema voor communicatieve effectiviteit na te gaan. De toetswijzercommissie heeft geadviseerd om communicatieve effectiviteit op te nemen binnen het hoofdattribuut 'afstemming op doel en publiek' (Verslag Try-out, Cito 2013). Bij het beoordelen van communicatieve effectiviteit gaat het puur om de effectiviteit van het schrijfproduct: bereikt de leerling hiermee zijn of haar doel. Communicatieve effectiviteit is niet echter opgenomen in het leerlingmodel (zie CvTE, 2014), omdat het lastig toetsbaar lijkt te zijn met gesloten of kortere opgaven. Aangezien het wel als een belangrijk aspect wordt gezien, wordt in dit onderzoek nagegaan of communicatieve effectiviteit globaal beoordeeld kan worden bij schrijfproducten.

8.2 Methode en procedure

8.2.1 De open schrijftaken

In de try-out leesvaardigheid zijn een drietal open schrijftaken aangeboden aan leerlingen van alle leerwegen aan het einde van de onderbouw (2 vmbo en 3 havo/vwo). De open schrijftaken zijn bedoeld om vrij schrijfgedrag uit te lokken waarmee effectiviteit en kwaliteit in kaart gebracht kunnen worden. De drie schrijftaken omvatten verschillende thema's, waarbij er wordt gevraagd naar een beschrijving van een vakantie, naar een aangifte van diefstal en als laatste een gedetailleerde beschrijving van een afbeelding.

De eerste schrijftaak is alleen voor 3 havo/vwo. Het betreft een opgave waarin de leerlingen een verhaal moeten schrijven over hun laatste vakantie. Er wordt expliciet gevraagd naar waar de vakantie was, met wie ze gegaan zijn, hoe lang ze er verbleven. Verder moest men twee gebeurtenissen beschrijven en aangeven wat ze van de vakantie vonden. Er staan afbeeldingen bij de opgave om te voorkomen dat de fantasie van de leerling de resultaten zouden beperken.

De andere twee schrijftaken zijn voor 2 vmbo. Deze opgaven zijn wel open schrijftaken maar zijn minder uitgebreid dan de eerste opgave. Het betreft een aangifte van een diefstal, waar de leerlingen gevraagd wordt een formulier in te vullen op het politiebureau. Zo moeten de leerlingen hun naam invullen, de datum, een beschrijving van de gebeurtenis en van de dief en afsluiten met contactgegevens. Wederom wordt de opdracht vergezeld van een afbeelding van een cartoon. De leerling zou de afbeelding kunnen gebruiken om zo een mogelijke dief te beschrijven.

De vmbo-leerlingen beschrijven ook een gedetailleerde afbeelding. Het betreft een foto van een picknicktafel in het park, waar een kraai op zit die kijkt naar etensresten die op de tafel liggen. Op de achtergrond is een klein terras, waar men eten en drinken kan bestellen. Ook is een bord met prijzen zichtbaar en staan er koffiekannen op de balie. Het is aan de leerlingen om dit plaatje te beschrijven in drie volzinnen waarbij zoveel mogelijk details genoemd dienen te worden.

8.2.2 De beoordelingsmodellen en beoordelingsprocedure

Het beoordelingsmodel op basis waarvan de afgenomen herkansingsschrijfitems in de try-out van leesvaardigheid in 2014 zijn nagekeken, is weergegeven in de bijlagen.

In dit globale beoordelingsmodel geeft de beoordelaar op twee wijzen een globale beoordeling: door middel van een rapportcijfer en door het niveau van het schrijfproduct te classificeren (zie Figuur 8-1). Dit model wordt ingezet om zowel communicatieve effectiviteit en globale kwaliteit te beoordelen.

Figuur 8-1 Gebruikte schalen voor de classificatie van het niveau

In totaal zijn 571 schrijfproducten beoordeeld door twee beoordelaars. Elke beoordelaar kreeg in totaal 400 schrijfproducten (uit alle leerwegen) aselect toegewezen en gaven tweemaal een oordeel van de globale kwaliteit en vervolgens tweemaal een oordeel van de communicatieve effectiviteit. Bij globale kwaliteit wordt het schrijfproduct beoordeeld door een inschatting te maken van de correctheid van de tekst in het achterhoofd houdend dat er holistisch naar de attributen gekeken dient te worden. Er wordt een rapportcijfer gegeven gevolgd door een niveau inschatting per leerweg, waarbij er gelet moet worden of de tekst onder, op, of boven niveau is van de huidige leerweg waarin de leerling zich bevindt. Bij communicatieve effectiviteit wordt eveneens gebruik gemaakt van een rapportcijfer en een niveau inschatting per leerweg, maar moet de beoordelaar niet letten op de correctheid van de tekst, maar op de effectiviteit.

Bijna 60% van de schrijfproducten zijn voor dit onderzoek door beide beoordelaars beoordeeld (N=229). Om vertekening te voorkomen zijn de 22 vmbo-schrijfopdrachten die door de havo2 leerlingen zijn gemaakt niet meegenomen in de analyses (havo2 komt ook niet voor in de beoordelingsschaal, zie Figuur 8-1).

Tabel 8-1 Aantal schrijfproducten die door beide beoordelaars beoordeeld zijn.

	Opdracht geschikt voor niveau leerling						Opdracht voor ander niveau	Totaal aantal
	vmbo-bb2	vmbo-gt2	vmbo-kb2	havo3	vwo3	Totaal geschikt voor niveau	havo2	
Beoordeeld	**10**	**19**	**14**	**69**	**86**	**198**	20	218
Invalide[4]	3	2	3	1		9	2	11
Toegewezen	13	21	17	70	86	207	22	229

8.2.3 Analyses

De kwaliteit van de beoordelingsmodellen wordt ten eerste nagegaan door de intercodeursovereenstemming en -samenhang na te gaan. Voorts wordt een eerste indruk van de validiteit verkregen door een Multitrait Multimethod vergelijking. Een Multitrait Multimethod Matrix (MTMM) geeft waarden weer in een matrix die gebruikt kan worden om te zien of de manier van beoordelen (*method*) en hetgeen dat beoordeeld wordt (*trait* of aspect) ook daadwerkelijk tot de juiste beoordelingen leidt. De waarden zijn correlaties tussen de twee methoden (rapportcijfer en niveau inschatting) en de aspecten (globale kwaliteit en communicatieve effectiviteit). Het is de bedoeling dat de correlatie tussen de methoden hoog is (en 1,0 nadert) en de correlatie tussen aspecten laag is (nadert 0). Een hoge correlatie tussen de methoden betekent dat er een sterke samenhang is tussen het cijfer en het niveau van de leerling (bijv. hogere cijfers voor leerlingen die op of boven niveau scoren van hun leerweg). Een lage correlatie tussen de aspecten betekent dat er geen samenhang is tussen de te meten aspecten wat aangeeft dat er tijdens de beoordeling een duidelijk onderscheid is tussen globale kwaliteit en communicatieve effectiviteit.

[4] Invalide: geen Engelstalig geschreven tekst, die een uitwerking van de gegeven schrijfopdracht lijkt

Het is wenselijk dat deze samenhang afwezig is, zodat duidelijk is dat de twee aspecten onafhankelijk van elkaar beoordeeld kunnen worden. Wanneer dit niet het geval is kan men zich afvragen of er wel een herkenbaar (en vooral meetbaar) verschil is tussen kwaliteit en effectiviteit.

8.3 Resultaten

8.3.1 Beoordelaarsovereenstemming en samenhang

Bij de niveau-inschatting (zie Tabel 8-2) zien we sterke samenhang tussen de niveau-inschatting van beoordelaar1 en beoordelaar2 voor communicatieve effectiviteit (γ=.91) en een goede overeenstemming (κ_w=.80). Bij de globale kwaliteit zien we hetzelfde beeld: een sterke samenhang tussen de niveau-inschatting van beoordelaar1 en beoordelaar2 (γ=.98) en de goede overeenstemming (κ_w=.89).

Bij de cijfers zijn de resultaten minder goed. We zien een redelijke samenhang tussen het cijfer van beoordelaar1 en beoordelaar2 voor communicatieve effectiviteit (γ=.73), maar een matige overeenstemming (κ_w=.46). Ook bij de globale kwaliteit zien we een redelijke samenhang tussen het cijfer van beoordelaar1 en beoordelaar2 (γ=.76), maar een matige overeenstemming (κ_w=.50).

Tabel 8-2 Beoordelaarsovereenstemming en samenhang (*N*=197)

	Meetniveau	Samenhang		Overeenstemming	
Niveau communicatieve effectiviteit	Ordinaal	Gamma	0,91	Gewogen kappa	0,80
	Interval	Pearson's R	0,95	Intraklasse correlatie	0,98
Niveau globale kwaliteit	Ordinaal	Gamma	0,98	Gewogen kappa	0,89
	Interval	Pearson's R	0,98	Intraklasse correlatie	0,99
Cijfer communicatieve effectiviteit	Ordinaal	Gamma	0,73	Gewogen kappa	0,46
	Interval	Pearson's R	0,73	Intraklasse correlatie	0,84
Cijfer globale kwaliteit	Ordinaal	Gamma	0,76	Gewogen kappa	0,50
	Interval	Pearson's R	0,73	Intraklasse correlatie	0,84

8.3.2 Validiteit (Multimethod-Multitrait vergelijking)

De beoordelingen lijken geen goede convergente en divergente validiteit te hebben. De twee methoden van beoordelen (niveau-inschatting en cijfer) convergeren niet erg. Bij beoordelaar 1 (zie Tabel 8-3) is de samenhang tussen de niveau-inschatting en het cijfer voor de globale kwaliteit matig (γ=.68) en voor de communicatieve effectiviteit redelijk (γ=.84). Bij beoordelaar2 (zie Tabel 8-4) is de samenhang tussen de niveau-inschatting en het cijfer voor zowel de globale kwaliteit (γ=.56) als voor de communicatieve effectiviteit (γ=.63) matig.

De beoordelingen van de verschillende aspecten, globale kwaliteit en communicatieve effectiviteit, zijn niet divergent. We zien bij beide beoordelaars een zeer sterke samenhang tussen de niveau-inschattingen van de globale kwaliteit en van de communicatieve effectiviteit (respectievelijk γ=.95 en γ=.90). Blijkbaar kunnen de beoordelaars bij de niveau-inschattingen geen onderscheid maken tussen de communicatieve effectiviteit en de globale kwaliteit.

Ook bij de cijfers zien we bij beide beoordelaar1 een vrij sterke samenhang tussen het cijfer voor globale kwaliteit en het cijfer voor de communicatieve effectiviteit (respectievelijk γ=.84), bij de tweede beoordelaar vinden we een matige samenhang (γ=.68). Bij de cijfermatige beoordeling slagen de beoordelaars er dus beter in om een verschil te maken tussen de globale kwaliteit en de communicatieve effectiviteit, maar (zie de eerste paragraaf) de overeenstemming tussen de beoordelaars is matig.

Tabel 8-3 Multimethod-Multitrait vergelijking (MMMT) beoordelaar1

	Meet-niveau		Niveau globale kwaliteit	Cijfer globale kwaliteit	Niveau communica-tieve effectiviteit	Cijfer communica-tieve effectiviteit
Niveau globale kwaliteit	Ordinaal	Gamma		0,68	0,95	0,72
	Interval	Pearson's R		0,63	0,97	0,71
		N		198	198	198
Cijfer globale kwaliteit	Ordinaal	Gamma	0,68		0,74	0,84
	Interval	Pearson's R	0,63		0,63	0,82
		N	198		198	198
Niveau communicatieve effectiviteit	Ordinaal	Gamma	0,95	0,74		0,84
	Interval	Pearson's R	0,97	0,63		0,74
		N	198	198		198
Cijfer communicatieve effectiviteit	Ordinaal	Gamma	0,72	0,84	0,84	
	Interval	Pearson's R	0,71	0,82	0,74	
		N	198	198	198	

Tabel 8-4 Multimethod-Multitrait vergelijking (MMMT) beoordelaar2

	Meet-niveau		Niveau globale kwaliteit	Cijfer globale kwaliteit	Niveau communica-tieve effectiviteit	Cijfer communica-tieve effectiviteit
Niveau globale kwaliteit	Ordinaal	Gamma		0,56	0,90	0,46
	Interval	Pearson's R		0,46	0,96	0,46
		N		197	196	196
Cijfer globale kwaliteit	Ordinaal	Gamma	0,56		0,56	0,68
	Interval	Pearson's R	0,46		0,45	0,63
		N	197		196	196
Niveau communicatieve effectiviteit	Ordinaal	Gamma	0,90	0,56		0,63
	Interval	Pearson's R	0,96	0,45		0,57
		N	196	196		197
Cijfer communicatieve effectiviteit	Ordinaal	Gamma	0,46	0,68	0,63	
	Interval	Pearson's R	0,46	0,63	0,57	
		N	196	196	197	

8.4 Conclusie

De resultaten van het onderzoek wijzen uit dat het model nog niet in alle opzichten erg geschikt is voor het beoordelen van open schrijfopdrachten. Positief is dat het gebruik van de niveau-inschatting als beoordelings-schaal betrouwbare resultaten geeft gezien de goede intercodeursovereenstemming. Dit geeft aan dat de meerdere mensen in staat zijn om het eens te worden over het globale niveau (bijvoorbeeld onder, op of boven vmbo-niveau) van de schrijfproducten.

Weliswaar is de niveau-inschatting heel betrouwbaar, maar beide beoordelaars slagen er niet in om een onderscheid te maken tussen communicatieve effectiviteit en globale kwaliteit van de schrijfproducten met behulp van de niveau inschatting. Deze bevinding zou kunnen liggen aan de onervarenheid van de beoordelaars, die geen docenten zijn. Een andere oorzaak kan zijn dat het verschil tussen de globale kwaliteit en communicatieve effectiviteit niet helder is uitgelegd en de beoordelaars altijd beide aspecten achter elkaar moesten beoordelen.

Ook de cijfermatige beoordeling lijkt weinig valide en geeft beduidend minder betrouwbare resultaten dan de niveau-inschatting. De reden hiervoor is dat de rapportcijfer-methode wellicht verkeerd toegepast is. Beide beoordelaars hebben een absolute schaal toegepast. Dit houdt in dat havo-leerlingen vrijwel altijd een rapportcijfer hebben gekregen van zes, zeven of acht, waarbij de bovengemiddelde leerling een acht kreeg, en de onder-gemiddelde leerling een zes. Vmbo-leerlingen hebben scores gekregen van drie, vier of vijf en vwo-leerling kregen acht, negen en tien. De beoordelaars hebben te kennen gegeven dat ze de beoordelingen verkeerd toegepast hebben.

Het verdient de aanbeveling om verder onderzoek te doen naar de beoordelingsmodellen. Bij vervolgonderzoek zouden ervaren docenten ingezet moeten worden om de bruikbaarheid en validiteit van de modellen te onderzoeken.

8.5 Referenties

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

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

8.6 Bijlagen

Bijlage 8-1 Beoordelingsmodel try-out opgaven Engels: Globale indruk van het schrijfproduct

De docent geeft een globale beoordeling van het schrijfproduct op basis van punten als afstemming op de lezer, structurering, woordgebruik en spelling en grammatica. Hij/zij geeft een rapportcijfer en een inschatting van het schoolniveau, waarbij de aandacht uitgaat naar correctheid enerzijds en naar het volgen van de opdracht anderzijds.

Beoordelingsmodel voor docenten - globaal - vmbo

Vooraf invullen:

Naam leerling:…….

Leerjaar: 2 Niveau: bb / kb / gt / …….

Datum: …….

Kenmerk schrijfopdracht: …….

Preconditie

Gaat het om een Engelstalig geschreven tekst, die een uitwerking van de gegeven schrijfopdracht lijkt?	ja / nee*

*Indien nee: stop met de beoordeling van de schrijfopdracht.

	Globale indruk van het schrijfproduct *(Denk aan zaken als: afstemming op de lezer, structurering, woordgebruik en spelling en grammatica)*												
1.	Rapportcijfer:	1	2	3	4	5	6	7	8	9	10		
2.	Inschatting niveau:	○	bb ○		○		kb ○		○		gt ○		○
	Opmerkingen:												

Beoordelingsmodel voor docenten - globaal - havo/vwo

Vooraf invullen:

Naam leerling: ……

Leerjaar: 3 Niveau: havo/vwo / ..……

Datum: ……

Kenmerk schrijfopdracht: ……

Preconditie

Gaat het om een Engelstalig geschreven tekst, die een uitwerking van de gegeven schrijfopdracht lijkt?	ja / nee*

*Indien nee: stop met de beoordeling van de schrijfopdracht.

	Globale indruk van het schrijfproduct *(Denk aan zaken als: afstemming op de lezer, structurering, woordgebruik en spelling en grammatica)*								
1.	Rapportcijfer:	1 2 3 4 5 6 7 8 9 10							
2.	Inschatting niveau:	○	havo ○		○		vwo ○		○
	Opmerkingen:								

Bijlage 8-2 Beoordelingsmodel try-out opgaven Engels: Communicatieve effectiviteit

De docent geeft een inschatting van de communicatieve effectiviteit van het schrijfproduct. Het gaat daarbij niet om de 'correctheid' van de opdracht maar puur om de effectiviteit van het schrijfproduct: bereikt de leerling hiermee zijn of haar doel?

Beoordelingsmodel voor docenten - communicatieve effectiviteit - vmbo

Vooraf invullen:

Naam leerling: …….

Leerjaar: 2 Niveau: bb / kb / gt / …….

Datum: …….

Kenmerk schrijfopdracht: …….

Preconditie

Gaat het om een Engelstalig geschreven tekst, die een uitwerking van de gegeven schrijfopdracht lijkt?	ja / nee*

*Indien nee: stop met de beoordeling van de schrijfopdracht.

	Communicatieve effectiviteit van het schrijfproduct *(Hoe effectief is het product? Zal de leerling hiermee zijn doel bereiken?)*												
1.	Rapportcijfer:	1	2	3	4	5	6	7	8	9	10		
2.	Inschatting niveau:	○	bb ○		○		kb ○		○		gt ○		○
	Opmerkingen:												

Beoordelingsmodel voor docenten - communicatieve effectiviteit - havo/vwo

Vooraf invullen:

Naam leerling: …….

Leerjaar: 3 Niveau: havo/vwo / ..…….

Datum: …….

Kenmerk schrijfopdracht: …….

Preconditie

Gaat het om een Engelstalig geschreven tekst, die een uitwerking van de gegeven schrijfopdracht lijkt?	ja / nee*

*Indien nee: stop met de beoordeling van de schrijfopdracht.

	Communicatieve effectiviteit van het schrijfproduct *(Hoe effectief is het product? Zal de leerling hiermee zijn doel bereiken?)*					
1.	Rapportcijfer:	1 2 3 4 5 6 7 8 9 10				
2.	Inschatting niveau:	○	havo ○	○	vwo ○	○
	Opmerkingen:					

Cito helpt je inzicht te krijgen in je ontwikkeling en mogelijkheden. Door kennis, vaardigheden en competenties objectief meetbaar te maken en de ontwikkeling er van te volgen, kun je het beste uit jezelf halen, verantwoorde keuzes maken en beter richting geven aan je toekomst. Cito draagt daaraan bij door wereldwijd werk te maken van goed en eerlijk toetsen, vanuit de kernwaarden kundig, toonaangevend, integer, innovatief en betrokken.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
www.cito.nl

Fotografie: Ron Steemers