RAPPORTAGE VERGELIJKBAARHEID EINDTOETSEN EN INVOERING VAN GEZAMENLIJK ANKER

Versie 1 december 2018

Expertgroep Toetsing PO

Inhoudsopgave

Samenvatting 2
1. Inleiding 4
1.1 Aanleiding 4
1.2 Begrippenlijst 5
2. De Afname van de Eindtoetsen in 2018 6
2.1 De Beoogde Ideale Procedure 6
2.2 De Gerealiseerde Procedure 6
2.3 Het functioneren van het Anker 7
2.4 De Geschatte Referentieniveaus 7
3. Evaluatie van de Vergelijkbaarheid van de Referentieniveaus en Toetsadviezen 9
3.1 Referentieniveaus 9
3.2 Toetsadviezen 10
3.3 Analyses m.b.t. het Gezamenlijk Anker 12
3.4 Schaling van de Referentieniveaus 14
3.5 Effecten op Itemniveau van Digitale versus Papieren afname 16
4. Mogelijke Scenario's voor Toetsadviezen 18
5. Conclusies en Verder Ontwikkeltraject 20
5.1 Conclusies 20
5.2 Verder ontwikkeltraject 20
6. Literatuurlijst 21
Bijlage A Verschil tussen Geschatte Referentieniveaus voor Advies en in de Kalibratie 22
Bijlage B Statistisch en Psychometrische Aspecten van de Kalibratie-procedure 23
Bijlage C IRT Schattingen van Populatie Kenmerken op de Latente Schaal 25

Samenvatting

Als de normen voor de verschillende eindtoetsen niet meer vergelijkbaar zijn, worden gelijke prestaties van leerlingen die verschillende eindtoetsen maken niet meer gelijk beoordeeld. Om de vergelijkbaarheid van de eindtoetsresultaten beheersbaar te houden is in 2018 een gezamenlijk anker van opgaven geïntroduceerd. De voorliggende rapportage is een interim rapport over de effecten van de invoering van dat gezamenlijk anker in de eindtoetsen. Het rapport richt zich vooral op de vergelijkbaarheid van de conclusies met betrekking tot beheersing van de referentieniveaus. In het rapport komen de volgende onderwerpen aan bod.

1. De uitvoering van de procedure voor het vaststellen van beheersing van de referentieniveaus en het geven van toetsadviezen tijdens de eindtoetsafname in 2018;

2. Een evaluatie van deze procedure en een evaluatie van de vergelijkbaarheid van de eindtoetsen na het beschikbaar komen van alle afnamegegevens;

3. Een vooruitblik op de procedure zoals die in 2019 en de daarop volgende jaren voorzien is.

Ad 1. De eindtoetsen van de verschillende aanbieders (de DIA-Eindtoets van Diataal B.V., de AMN Eindtoets van AMN, de IEP van Bureau ICE, de Route 8 van A-Vision en de Centrale Eindtoets, aangeboden onder auspiciën van de CvTE) bevatten gezamenlijke ankeropgaven voor de onderwerpen Lezen, Rekenen en Taalverzorging. De gezamenlijke ankers bestonden per onderwerp uit ongeveer 20 opgaven. Omdat de procedure voor de eerste keer werd uitgevoerd, moest een aantal compromissen worden gesloten. Zo waren er voor de ankeropgaven nog geen definitieve, in een operationele toetssituatie (high-stakes situation) geschatte, itemparameters beschikbaar en door de grote verschillen in afnamemomenten van de verschillende eindtoetsen was het niet mogelijk om alle eindtoetsen tegelijk te normeren. Verder moest er omwille van de tijdsdruk een eenvoudig statistisch model gebruikt worden om de referentiecesuren te berekenen. Oplossingen voor al deze problemen zijn inmiddels voor handen, hierover hieronder meer. Naar aanleiding van de afname van de eindtoetsen kan het volgende worden geconcludeerd.

a) De opgaven van het anker hebben in de verschillende eindtoetsen kwalitatief goed en vergelijkbaar gefunctioneerd. Technisch: de itemparameters zoals geschat bij de verschillende aanbieders correleerden hoog. Dat wil zeggen dat de items zodanige psychometrische eigenschappen hebben en dat ze volgend jaar opnieuw in een anker gebruikt kunnen worden.

b) De door de Expertgroep gegeven adviezen voor de referentiecesuren zijn door AMN, Diataal, en Bureau ICE slechts gedeeltelijk gevolgd. Over het algemeen kozen zij voor wat mildere cesuren. Overigens was met hen afgesproken dat ze hiervoor ook de ruimte hadden. Voor volgend jaar is dit een punt van discussie.

c) Onder andere omdat de procedure voor het eerst werd uitgevoerd en de nauwkeurigheid van de procedure nog niet rigoureus was vastgesteld, nam de Centrale Eindtoets in 2018 wel deel aan het invoeren van het gezamenlijke anker, maar werden er aan de Centrale Eindtoets in dat jaar nog geen adviezen gegeven.

d) Bij de datacommunicatie traden nog vele kleine problemen aan het licht, die in een verbeterd informatiesysteem overigens goed zijn op te lossen.

Ad 2. In de operationele fase werd een eenvoudig statistisch model gebruikt dat gebaseerd is op een beperkte beschrijving van de data. Na de campagne werden alle data verzameld en kon een geavanceerder statistisch model gebruikt worden om de nauwkeurigheid van de procedure te evalueren. De conclusies van deze evaluatie zijn als volgt.

a) Opnieuw kon worden vastgesteld dat de opgaven van het anker goed hebben gefunctioneerd en een goede basis vormen voor het vergelijkbaar maken van de eindtoetsen.

b) Omdat de eindtoetsen via het gezamenlijke anker op een en dezelfde schaal konden worden geplaatst, konden gezamenlijke referentiecesuren worden berekend. Door toepassing van deze referentiecesuren op de beschikbare data werd duidelijk dat de percentages beheerste referentiecesuren per eindtoets sterk verschilden. Dit is toe te schrijven aan het feit dat de

vaardigheidsverdelingen van de populaties van de verschillende eindtoetsen inmiddels significant uiteenlopen. Overigens is dit op zich niet onverwacht; het staat scholen vrij hun eigen keuze voor een eindtoets te maken en verschillende typen scholen maken verschillende keuzes. Een tweede resultaat was dat de percentages beheerste referentieniveaus ook duidelijk afwijken van de percentages zoals die nu in BRON zijn ingevoerd. Dit is problematischer, omdat deze gegevens nu geen goed beeld geven van de verschillen tussen de populaties van de toetsen.

c) Een deel van de ankeropgaven werd zowel in een digitale als een papieren versie afgenomen. Bij het vaststellen van referentiecesuren is ervan uitgegaan dat er geen verschil was in het functioneren van de items in de beide versies. Analyse van de beschikbare data bevestigen dat deze aanname goed verdedigbaar was.

d) Bij het vaststellen van de referentieniveaus is ervan uitgegaan dat voor ieder onderwerp de cesuren 1F, 2F en eventueel 1S op een schaal lagen. Ook hier bevestigen de analyses van de beschikbare data dat deze aanname goed verdedigbaar was.

Ad 3. Aangezien het anker goed gefunctioneerd heeft en alle eindtoetsen nu op een gezamenlijke schaal kunnen worden afgebeeld via een anker waarvan de eigenschappen geschat zijn in een operationele setting (een high-stakes situation), is besloten het huidige anker ook in 2019 toe te passen. Daarna zal het anker periodiek gedeeltelijk worden ververst. Daardoor kan de ontwikkeling van de referentieniveaus ook over de jaren heen worden bewaakt, zodat de geschatte referentieniveaus van de eindtoetsen vergelijkbaar worden. Ook de Centrale Eindtoets zal aan deze procedure deelnemen. De procedure wordt als volgt verder geoptimaliseerd.

a) Doordat de psychometrische eigenschappen van de opgaven nauwkeurig en in een operationele setting geschat zijn, is het uiteenlopen van de afnamemomenten van de verschillende eindtoetsen in 2019 geen probleem meer. Dit omdat de geschatte itemparameters vooraf bekend zijn en gelden voor alle eindtoetsen. Er is dus geen verdere afstemming tijdens de afnameperiode meer nodig.

b) Een aantal aanbieders van eindtoetsen gaat de nieuwe opgaven voor het anker van 2020 zaaien in de eindtoets van 2019. Daardoor kunnen de parameters van deze nieuwe opgaven ook in een operationele setting (high-stakes) nauwkeurig worden geschat. Hiermee ontstaat er een sterke koppeling tussen het anker zoals dat in 2018 en 2019 is gebruikt en het anker van 2020. Ook het probleem van de verschillende afnamemomenten is hiermee verder opgelost, want opnieuw zijn de eigenschappen, in de vorm van de itemparameters van het anker, vooraf bekend.

c) Er worden betere ICT faciliteiten gecreëerd voor de datacommunicatie en de statistische analyses. De aanbieders krijgen een ICT-applicatie waarin ze ver voor de toetsafname hun itemlabels en datadesign moeten definiëren, zodat kleine, maar tijdrovende, vergissingen, zoals die in 2018 voorkwamen, worden verhinderd. Verder zal de ICT applicatie ook geavanceerdere en nauwkeuriger statistische modellen ondersteunen, zoals die in de evaluatie al gebruikt zijn.

Het geven van adviezen m.b.t. de cesuren voor de toetsadviezen is gecompliceerder dan het geven van adviezen voor de cesuren voor de referentieniveaus. In de eerste plaats voert iedere aanbieder een andere weging over onderwerpen uit. Deze weging wordt, over het algemeen, bepaald door het toetsadvies statistisch te vergelijken met het schooladvies. Bij de normering van de Centrale Eindtoets worden ook resultaten van doorstroomonderzoek betrokken. In de praktijk blijken de gewichten die door de aanbieders gebruikt worden echter nauwelijks te verschillen en de invloed op de resultaten is verwaarloosbaar. Als die gewichten uiteen zouden gaan lopen ontstaat er echter wel een probleem. In de tweede plaats heeft een aantal aanbieders extra unieke onderwerpen die ook in de weging worden meegenomen. Het advies over de cesuren voor de toetsadviezen van de Expertgroep wordt gebaseerd op dezelfde data en statistische analyses als die voor de referentiecesuren. Vanwege de genoemde problemen zal dat advies echter een ruimere marge hebben dan het advies m.b.t. de referentieniveaus. In dit rapport wordt echter wel een methode voorgesteld waarmee de aanbieders ook eindtoetsen met extra onderwerpen aan het gezamenlijke anker kunnen linken.

1. Inleiding

1.1 Aanleiding

De wetswijziging 'Eindtoetsing PO' (Stb. 2014, 13) verplicht basisscholen om vanaf het schooljaar 2014/2015 bij alle leerlingen in groep 8 een eindtoets voor taal en rekenen af te nemen. De overheid heeft de publieke taak om een centrale eindtoets beschikbaar te stellen. Het College voor Toets en Examens (CvTE) is wettelijk belast met deze taak. In opdracht van het CvTE wordt door stichting Cito de Centrale Eindtoets ontwikkeld. Naast deze toets staat het scholen vrij andere eindtoetsen bij hun leerlingen af te nemen. De eindtoets die een basisschool gaat gebruiken, moet van goede kwaliteit zijn. Daarom is door de minister een onafhankelijke commissie (de Expertgroep Toetsen PO, hierna: Expertgroep) aangewezen, die de minister adviseert over het toelaten van andere toetsen als eindtoets voor het PO.

In 2014 zijn er, op basis van een positief advies van de Expertgroep, twee andere eindtoetsen in het PO toegelaten; dit betreft Route 8, ontwikkeld en uitgegeven door A-VISION B.V., en IEP Eindtoets, ontwikkeld en uitgegeven door Bureau ICE. Voor schooljaar 2016/2017 hebben de scholen naast de drie eerder genoemde eindtoetsen ook de keuze uit de Dia-Eindtoets, ontwikkeld en uitgegeven door Diataal B.V., de CESAN Eindtoets, ontwikkeld en uitgegeven door SM&C Internet services B.V., en de AMN Eindtoets, ontwikkeld en uitgegeven door AMN B.V.. Voor het schooljaar 2017/2018 werd de CESAN Eindtoets door de uitgever teruggetrokken. Hierdoor bleven er 5 eindtoetsen over waar de scholen uit konden kiezen. Het voorliggende rapport heeft betrekking op de eindtoetsen die in 2017/2018 konden worden afgenomen.

Eindtoetsen in het basisonderwijs kennen twee primaire doelen. Ten eerste, het geven van advies welk brugklastype voortgezet onderwijs het best bij de leerling past (toetsadvies). Dit advies geldt als objectieve tweede gegeven naast het advies van de school. Indien het toetsadvies aangeeft dat de leerling een hoger niveau schoolniveau aan kan dan dat het schooladvies aangeeft, dan dient de school het oorspronkelijke schooladvies te *heroverwegen*. Het schooladvies kan, maar hoeft niet, op basis van het toetsadvies naar boven worden bijgesteld. Het schooladvies is uiteindelijk doorslaggevend voor de toelating tot het voortgezet onderwijs.

Het tweede primaire doel is om te bepalen of leerlingen de referentieniveaus beheersen. Referentieniveaus geven voor Nederlandse taal en rekenen het gewenste niveau aan op verschillende momenten in het onderwijs (Stb. 2010, 265). De referentieniveaus worden uitgesplitst in de fundamentele niveaus (1F, 2F, 3F en 4F) en de streefniveaus (1S, 2S en 3S). Voor het onderdeel Nederlandse Taal worden vier domeinen onderscheiden: Mondelinge Taalvaardigheid, Lezen, Schrijven en Begrippenlijst en Taalverzorging. Van deze domeinen zijn alleen Lezen en Taalverzorging toetsing met een eindtoets verplicht. Op beide domeinen wordt afzonderlijk een uitspraak gedaan per referentieniveau. Bij de referentieniveaus Rekenen worden de vier domeinen Getallen, Meten en Meetkunde, Verbanden en Verhoudingen onderscheiden en samengevoegd tot één uitspraak per referentieniveau. Alle eindtoetsen rapporteren op de referentieniveaus Lezen 1F en 2F, Taalverzorging 1F en 2F en Rekenen 1F en 1S. De IEP rapporteerde in 2018 daarnaast ook 2F Rekenen. Vanaf 2019 rapporteert de IEP Eindtoets voor Rekenen op de niveaus 1F en 1S.

Beide primaire uitkomsten worden door de onderwijsinspectie meegewogen bij stelselevaluaties van het primair onderwijs als geheel en er is sprake van dat de primaire uitkomsten ook een rol zouden kunnen spelen bij de evaluatie van de eindresultaten van individuele scholen als onderdeel van het instellingstoezicht.

Door de keuzevrijheid worden er binnen het basisonderwijs verschillende eindtoetsen gebruikt. Hoewel de vijf eindtoetsen zijn geconstrueerd op basis van verschillende uitgangspunten, met inachtneming van de vastgestelde kaders (Algemeen deel Toetswijzer (CvTE, 2014), en voor de vier nieuwe eindtoetsen het Beoordelingskader Expertgroep PO (Folmer et al., 2018) en het Beoordelingssysteem van de COTAN (Evers et al. 2010)), is hun gemeenschappelijke doelstelling het doen van uitspraken met betrekking tot het best passende brugklastype in het voortgezet onderwijs

en de beheersing van de referentieniveaus. Een belangrijke vraag daarbij is de mate waarin verschillende eindtoetsen gelijke toetsadviezen geven bij gelijke prestaties en vergelijkbare schattingen van de gehaalde referentieniveaus hanteren. Daartoe heeft de staatssecretaris van OCW de Expertgroep PO in 2016 gevraagd om de resultaten van de verschillende eindtoetsen uit 2015-2016 naast elkaar te zetten en de vergelijkbaarheid met betrekking tot toetsadviezen en bepaling beheersing referentieniveaus nader te onderzoeken; dit onderzoek heeft geleid tot de Rapportage Vergelijkbaarheid eindtoetsen (Emons et al., 2016). In 2016 is er tevens in opdracht van het OCW een rapport geschreven met drie acties die het proces van ankering aan de referentieniveaus en normering van de toetsadviezen beter in de hand kunnen houden (Glas et al., 2016). De drie voorgestelde acties waren:

1. Het ontwikkelen van een blueprint voor de toetsverantwoording, zodat de toetsverantwoordingen van de verschillende aanbieders beter met elkaar en over de jaren heen vergeleken kunnen worden.
2. Het invoeren van een gezamenlijk anker voor alle eindtoetsen in 2018 en 2019, zodat het mogelijk wordt een zodanig advies met betrekking tot referentiecesuren en toetsadviezen te geven, dat deze dichter bij elkaar komen te liggen.
3. Het opbouwen van een database van alle relevante toetsafnamen die het mogelijk maakt de nauwkeurigheid van de cesuren van referentieniveaus en de toetsadviezen te volgen en analyses en simulaties uit te voeren om de adviezen m.b.t. het gezamenlijke anker te ondersteunen.

De drie hierboven genoemde acties zijn in de afgelopen twee jaar ingezet en op termijn moeten deze leiden naar een verhoogde vergelijkbaarheid van de vaardigheidsmetingen van de verschillende eindtoetsen.

De voorliggende rapportage heeft betrekking op de effecten van de invoering van dat gezamenlijk anker in de eindtoetsen in 2018 en doet suggesties voor de procedure voor het geven van adviezen voor de cesuren in 2019. In dit rapport komen de volgende onderwerpen aan bod.

1. De uitvoering van de procedure voor het vaststellen van de referentieniveaus en toetsadviezen tijdens de eindtoetsafname in 2018;

2. Een evaluatie van deze procedure en een evaluatie van de vergelijkbaarheid van de eindtoetsen na het beschikbaar komen van alle resultaten;

3. Een vooruitblik op de procedure zoals die in 2019 en de daarop volgende jaren voorzien is.

1.2 Begrippenlijst

Toetsadvies:	doorstroomadvies met betrekking tot het best passende brugklastype in het voortgezet onderwijs dat op basis van de resultaten op een eindtoets wordt gegeven
Schooladvies:	doorstroomadvies met betrekking tot het best passende brugklastype in het voortgezet onderwijs dat door het onderwijsteam op de basisschool van de betreffende leerling wordt opgesteld.
Eindtoets PO:	de eindtoets die binnen het PO afgenomen dient te worden.
Itemparameter	eigenschap van een item in een statistisch of psychometrisch model, gecorrigeerd voor de vaardigheidsparameters van de leerlingen.
Vaardigheidsparameter	kortweg vaardigheid. Eigenschap van een leerling in een statistisch of psychometrisch model, gecorrigeerd voor de itemparameters van de items in toetsen.

2. De Afname van de Eindtoetsen in 2018

2.1 De Beoogde Ideale Procedure

De initiële beoogde procedure voor het geven van adviezen voor cesuren ging uit van de volgende aannamen:

- Alle data van alle eindtoetsen zijn tegelijk beschikbaar, zodat alle referentiecesuren in een keer en in relatie tot elkaar kunnen worden vastgesteld.
- Een eenvoudig psychometrisch model (het Rasch model) is voldoende om de toetsen per onderwerp op een schaal te plaatsen. Daardoor zijn er maar beperkte data nodig (afdoende grootheden), namelijk:
  - Voor lineaire papieren toetsen: p-waarden van de items en frequentie verdeling van de totaalscores van de leerlingen
  - Voor digitale toetsen, zoals adaptieve toetsen en multistage toetsen: schattingen van de vaardigheidsparameters van de leerlingen en p-waarden anker items
- Hiermee kunnen de itemparameters van het gezamenlijk anker geschat worden: de vaardigheidsverdelingen van de verschillende toetspopulaties op de referentie-onderwerpen en de vaardigheidsverdelingen op de gehele toets. En daarmee kunnen de gezamenlijke cesuren voor de referentieniveaus en cesuren voor de toetsadviezen voorgesteld worden.

Technische details staan in Methoden voor Normhandhaving van Eindtoetsen, Expertgroep, December 2017.

2.2 De Gerealiseerde Procedure

De gerealiseerde procedure verliep niet geheel volgens de beoogde ideale procedure. De volgende problemen deden zich voor en de volgende oplossingen werden gevonden.

- In de invoer van de aanbieders klopten de itemlabels niet altijd, er zaten enkele fouten in de invoer van hun data-design, de volgorde van de labels van de onderwerpen (Lezen, Rekenen, Taalverzorging) werd vaak gemuteerd en het format van de invoer data werd niet altijd gevolgd.
- De data waren niet allemaal gelijk beschikbaar en ook de momenten waarop de adviezen beschikbaar moesten zijn, verschilden sterk.
- Voor de eerste adviezen waren alleen voorlopige itemparameters van het Cito beschikbaar. Deze itemparameters waren alleen in een pretest geschat en nog niet in een high-stakes situatie.

Daarom werd er voor een andere aanpak gekozen.

- Route 8 heeft in 2018 een zeer sterk anker met de afname in 2017: ongeveer 70% van de itembank niet veranderd. De geschatte vaardigheidsverschillen tussen 2017 en 2018 waren klein. Daarom werden de cesuren van 2017 gehandhaafd en de bijhorende nieuwe percentages referentieniveaus en toetsadviezen voor Route 8 geschat.
- De data van de AMN, DIA en IEP kwamen daarna ongeveer gelijktijdig binnen.
- De ankerparameters werden apart geschat per toets en vergeleken met toen inmiddels beschikbare definitieve itemparameters van de Centrale Eindtoets. Conclusie: correlaties waren hoog, dus het anker als zodanig heeft goed gefunctioneerd. Hierover meer in de volgende paragraaf.
- Alle data werden gepoold (Route8, AMN, IEP en DIA) en daarna werden gezamenlijke ankerparameters geschat op gepoolde data en met Centrale Eindtoets meegewogen, dus alles gewogen naar steekproefgrootte.
- Daarbij zijn de vaardigheidsverdelingen per toets en onderwerp geschat en hieruit volgen de geschatte percentages beheerste referentieniveaus voor de onderwerpen van de AMN, DIA en IEP.

2.3 Het functioneren van het Anker

Hieronder staan de correlaties tussen de ankeritems zoals ze in de operationele fase geschat zijn. De correlaties zijn voldoende hoog om ervan uit te kunnen gaan dat dezelfde itemparameters op alle aangeboden eindtoetsen van toepassing zijn.

Tabel 1: Correlaties tussen Ankeritems

Papieren Toetsen Centrale Eindtoets en IEP

Onderwerp	Correlatie	Aantal Items
LEZEN	0.84	22
REKENEN	0.93	20
TAALVERZORGING	0.89	20

LEZEN, DIGITAAL, 22 Items

	Centrale Eindtoets	DIA	AMN	Route 8
Centrale Eindtoets	1.00	0.78	0.67	0.86
DIA		1.00	0.82	0.80
AMN			1.00	0.72
Route 8				1.00

REKENEN, DIGITAAL, 20 Items

	Centrale Eindtoets	DIA	AMN	Route 8
Centrale Eindtoets	1.00	0.95	0.89	0.88
DIA		1.00	0.89	0.92
AMN			1.00	0.85
Route 8				1.00

TAALVERZORGING, DIGITAAL, 20 Items

	Centrale Eindtoets	DIA	AMN	Route 8
Centrale Eindtoets	1.00	0.89	0.94	0.88
DIA		1.00	0.94	0.94
AMN			1.00	0.92
Route 8				1.00

2.4 De Geschatte Referentieniveaus

In tabel 2 en 3 staan de referentieniveaus zoals die geschat zijn in de operationele fase en de referentieniveaus zoals ze in BRON zijn ingevoerd. Daarbij moet worden opgemerkt dat de Centrale Eindtoets vanwege de wettelijke taak van het CvTE een eigen normering kent. Bij Tabel 3 moet in de eerste plaats worden opgemerkt dat er bij de berekeningen van de Expertgroep vanuit is gegaan dat per onderwerp alle referentie-items en de daarbij behorende cesuren op een latente schaal lagen. Er was dus geen sprake van aparte schalen voor 1F, 2F en 1S niveaus. Hieronder wordt die aanname empirisch aannemelijk gemaakt. Bij de IEP werden, in tegenstelling tot bij de andere eindtoetsen, de 1S en 2F niveaus apart gerapporteerd. De scholen zijn echter verantwoordelijk voor de rapportage in BRON, en zij hebben dat onderscheid dit jaar onvoldoende gemaakt, wat heeft geleid tot onjuiste registraties in BRON.

In Tabel 4 staan de verschillen tussen de adviezen van de Expertgroep en de registraties in BRON. Over het algemeen zijn de niveaus in BRON hoger. De aanbieders hebben de referentiecesuren dus lager gelegd dan de adviezen. Vooral bij Lezen is dit het geval, met als uitzondering DIA. Bij Rekenen valt vooral op dat Route 8 de cesuur aanzienlijk verlaagde.

Tabel 2 Referentieniveaus volgens door de Expertgroep gegeven adviezen

		AMN	DIA	IEP	Route 8	TOTAAL
LEZEN	Lager dan 1F	6.7%	6.3%	7.0%	7.0%	6.7%
	1F	93.3%	93.7%	93.0%	93.0%	93.3%
	2F	65.3%	61.8%	57.9%	66.0%	62.8%
REKENEN	Lager dan 1F	6.1%	6.5%	8.7%	13.5%	8.7%
	1F	93.9%	93.5%	91.3%	86.5%	91.3%
	1S/2F	50.8%	45.8%	55.4%	58.0%	52.5%
TAAL	Lager dan 1F	5.8%	8.0%	5.5%	14.0%	8.3%
VERZORGING	1F	94.2%	92.0%	94.5%	86.0%	91.7%
	2F	57.6%	56.9%	57.9%	55.0%	56.8%

Tabel 3 Referentieniveaus zoals ingevoerd in BRON

		AMN	Centrale Eindtoets	DIA	IEP	Route 8	TOTAAL
LEZEN	Lager dan 1F	4.6%	2.0%	4.8%	0.8%	4.4%	2.0%
	1F	95.4%	98.0%	98.0%	95.2%	99.2%	95.6%
	2F	67.4%	75.0%	57.0%	76.9%	70.7%	74.7%
REKENEN	Lager dan 1F	4.9%	7.1%	12.5%	7.7%	5.7%	7.2%
	1F	92.9%	94.3%	94.3%	92.3%	87.5%	95.1%
	1S/2F	45.2%	64.7%	50.9%	35.2%	50.4%	49.0%
TAAL	Lager dan 1F	3.6%	3.7%	4.1%	0.6%	8.2%	4.0%
VERZORGING	1F	96.4%	96.3%	96.3%	95.9%	99.4%	91.8%
	2F	59.8%	59.1%	52.7%	60.0%	57.8%	59.1%

Tabel 4 Verschil Referentieniveau Advies en BRON

		AMN	DIA	IEP	Route 8	TOTAAL
LEZEN	Lager dan 1F	2.1%	1.5%	6.3%	2.6%	4.7%
	1F	-2.1%	-4.3%	-2.2%	-6.2%	-2.4%
	2F	-2.1%	4.8%	-19.0%	-4.7%	-12.0%
REKENEN	Lager dan 1F	1.2%	-5.9%	1.0%	7.8%	1.5%
	1F	1.0%	-0.8%	-1.0%	-1.0%	-3.8%
	1S/2F	5.6%	-5.1%	20.2%	7.6%	3.5%
TAAL	Lager dan 1F	2.2%	4.0%	4.9%	5.8%	4.3%
VERZORGING	1F	-2.2%	-4.3%	-1.4%	-13.4%	-0.1%
	2F	-2.2%	4.2%	-2.1%	-2.8%	-2.3%

3. Evaluatie van de Vergelijkbaarheid van de Referentieniveaus en Toetsadviezen

3.1 Referentieniveaus

In juli 2018 zijn de data van alle eindtoetsen, inclusief de Centrale Eindtoets, door de Expertgroep verzameld en aan elkaar gekoppeld. In de operationele fase werd alleen gebruik gemaakt van beperkte data (afdoende grootheden), maar in de evaluatie werden de complete data op het niveau van responsies van (geanonimiseerde) individuele leerlingen op individuele items geanalyseerd met een geavanceerder statistisch model (i.e. het 2-parameter logistisch model). Per onderwerp werden eerst de itemparameters geschat op een gezamenlijke schaal met behulp van alle aanwezige data. Vervolgens werden de vaardigheidsparameters van iedere leerling geschat op hun gehele antwoordpatroon. Deze schattingen zijn ook op een gezamenlijke schaal gekalibreerd en uiteraard veel nauwkeuriger dan de schattingen zoals ze in de operationele fase gemaakt zijn. Op deze manier konden de percentages beheerste referentieniveaus opnieuw worden berekend.

Een probleem bij deze aanpak is dat er geen zekerheid is over waar de referentiecesuren objectief gezien behoren te liggen. Alle aanbieders hebben hun cesuren een aantal jaren geleden eenmalig aan de referentiesets gekoppeld, maar het handhaven van de cesuren van jaar naar jaar bood geen garantie dat de cesuren ook over toetsen heen vergelijkbaar bleven. De gekozen oplossing voor dit probleem is er van uit te gaan dat de marginale verdeling van de beheerste referentieniveaus, dat wil zeggen, de gemiddeld over alle toetsen heen beheerste referentieniveaus, een redelijke indicatie is voor het vaststellen van de gezamenlijke cesuren. De gemiddelden zijn gewogen met de grootten van de populaties van de verschillende eindtoetsen. Bij de op deze manier berekende percentages worden er dan cesuren vastgesteld op de latente referentie schaal. Vervolgens kan er worden vastgesteld in welke mate aanbieders met hun gekozen cesuren afwijken van de cesuren op de gezamenlijke schaal, rekening houdend met de vaardigheidsverdeling van hun leerlingen op die gezamenlijke schaal. In Bijlage B wordt een technische samenvatting van de procedure gegeven.

In Tabel 5 staan de resultaten van de analyses van de referentieniveaus. Gezien het goed passende statistische model (zie de paragrafen 3.3 en 3.5) kan er vanuit worden gegaan dat de percentages de niveauverschillen van de populaties goed reflecteren; hierdoor kunnen er ook een aantal conclusies getrokken worden. Merk op dat deze resultaten dus gecorrigeerd zijn voor de moeilijkheid van de items in iedere eindtoets, ook de niet-anker items. AMN heeft op Rekenen en Taalverzorging de minst vaardige populatie, op de voet gevolgd door Route 8. Bij Lezen heeft Route 8 een minder vaardige populatie dan de AMN. De populatie van de IEP scoort op Lezen en Rekenen het hoogst. De Centrale Eindtoets heeft de meest vaardige populatie op het gebied van Taalverzorging.

Tabel 5 Behaalde Referentieniveaus zoals die geschat zijn nadat alle eindtoetsen en vaardigheidsverdelingen van de verschillende populaties op een schaal zijn gekalibreerd

		AMN	Centrale Eindtoets	DIA	IEP	Route 8	TOTAAL
LEZEN	Lager dan 1F	7.1%	1.7%	2.6%	0.8%	5.7%	2.0%
	1F	92.9%	98.3%	97.4%	99.2%	94.3%	98.0%
	2F	54.3%	76.9%	64.4%	82.6%	47.7%	74.7%
REKENEN	Lager dan 1F	17.3%	7.6%	7.8%	4.4%	10.8%	7.2%
	1F	82.7%	92.4%	92.2%	95.6%	89.2%	92.8%
	1S/2F	19.5%	49.9%	45.0%	59.2%	23.0%	49.0%
TAAL VERZORGING	Lager dan 1F	8.7%	2.5%	3.9%	6.2%	6.2%	4.0%
	1F	91.3%	97.5%	96.1%	93.8%	93.8%	96.0%
	2F	38.4%	65.8%	54.6%	52.8%	42.3%	59.1%

Overigens, doordat meer dan de helft van de leerlingen de Centrale Eindtoets maakten, is de kolom met de totaal behaalde referentieniveaus bijna gelijk aan de kolom met de behaalde referentieniveaus op de Centrale Eindtoets. De Centrale Eindtoets had dus ook eventueel als ijkpunt voor de cesuren gekozen kunnen worden zonder dat dat veel verschil had gemaakt.

Om deze resultaten te vergelijken met de resultaten zoals deze in BRON zijn ingevoerd, wordt in Tabel 6 het verschil tussen de resultaten in Tabel 5 en Tabel 3 weergegeven. Een Tabel met de verschillen tussen de referentieniveaus resulterend uit de adviezen van de Expertgroep (Tabel 2) en de behaalde Referentieniveaus zoals die geschat zijn nadat alle eindtoetsen en vaardigheidsverdelingen van de verschillende populaties op een schaal zijn gekalibreerd, is opgenomen in Bijlage A.

Tabel 6 Verschillen tussen de behaalde Referentieniveaus zoals die geschat zijn na kalibratie van alle eindtoetsen en vaardigheidsverdelingen van de verschillende populaties op één schaal en de gegevens zoals die in BRON beschikbaar zijn

		AMN	Centrale Eindtoets	DIA	IEP	Route 8
LEZEN	Lager dan 1F	2.5%	-0.3%	-2.2%	0.0%	1.3%
	1F	-2.5%	0.3%	-0.6%	4.0%	-4.9%
	2F	-13.1%	1.9%	7.4%	5.7%	-23.0%
REKENEN	Lager dan 1F	12.4%	0.5%	-4.7%	-3.3%	5.1%
	2F	-10.2%	-1.9%	-2.1%	3.3%	1.7%
	1S/2F	-25.7%	-14.8%	-5.9%	24.0%	-27.4%
TAAL VERZORGING	Lager dan 1F	5.1%	-1.2%	-0.2%	5.6%	-2.0%
	1F	-5.1%	1.2%	-0.2%	-2.1%	-5.6%
	2F	-21.4%	6.7%	1.9%	-7.2%	-15.5%

De verschillen, dat wil zeggen de percentages in BRON min de percentages volgens het anker, zijn aanzienlijk. AMN en Route 8 overschatten het niveau van de leerlingen die deze toetsen maakten voor alle drie de onderwerpen. Dit is te zien aan het feit dat het hoogste niveau in BRON steeds door meer leerlingen gehaald wordt, dan na de schatting via het gekalibreerde gezamenlijke anker. Bij de Centrale Eindtoets geldt dit alleen voor Rekenen. De IEP onderschat het niveau van de leerlingen bij Rekenen.

De conclusie is dat er grote verschillen zijn tussen de toetsen onderling en de resultaten zoals ze in BRON staan. Het eerste probleem, het feit dat toetsen onderling verschillen, hoeft geen probleem te zijn zolang het een verklaarbare, niet aan de moeilijkheid van de eindtoets toe te schrijven, effect is. Het is een verklaarbaar en zelfs bedoeld effect, zolang het is toe te schrijven aan het feit dat iedere school de toets kan kiezen die bij de school past. De resultaten zoals die via het anker in Tabel 5 zijn weergegeven geven inzicht in de verschillen tussen toetsen zoals die inmiddels ontstaan zijn. Het tweede is wel een probleem, omdat vertekeningen in de globale landelijke resultaten een vertekend beeld geven van de ontwikkeling van het niveau van onderwijs.

3.2 Toetsadviezen

De toetsadviezen zijn niet eenvoudig te vergelijken. In de eerste plaats voert iedere aanbieder een andere weging over onderwerpen uit. In de tweede plaats heeft een aantal aanbieders extra unieke onderwerpen. De DIA heeft als extra onderwerp Woordenschat. Route 8 heeft als extra onderwerpen

Woordenschat, Begrippenlijst, Dictee (optioneel), Kijk- en Luistervaardigheid (optioneel) en Functioneren (optioneel). De Centrale Eindtoets heeft als extra onderwerp Schrijven. Daardoor moeten de resultaten met terughoudendheid worden beoordeeld.

Om toch tot een vergelijking te komen, is ervoor gekozen om bij de berekening van de toetsadviezen op basis van het anker alleen gebruik te maken van de drie onderwerpen waaruit het anker bestaat en de onderwerpen voor alle toetsen gelijk te wegen. Deze weging is gebaseerd op de gemiddelden van de gewichten over de verschillende toetsaanbieders, waarbij de bijdragen van de toetsaanbieders gewogen zijn met het aantal leerlingen dat iedere toets maakte. In de praktijk blijken de gewichten van de aanbieders waarmee de drie gezamenlijke onderwerpen worden getoetst nauwelijks te verschillen. Grofweg komen die gewichten, en daarmee ook de gemiddelde gewichten, uit op 0.25 voor Lezen en Taalverzorging en 0.50 voor Rekenen.

Tabel 7 geeft de resultaten zoals die in BRON gearchiveerd zijn.

Tabel 7 Toetsadviezen geregistreerd in BRON (na herziening)

Advies	AMN	Centrale Eindtoets	DIA	IEP	Route 8	Totaal
PRO	1.3%	0.9%	1.3%	0.9%	0.8%	0.9%
VMBO BL	6.5%	5.7%	6.5%	6.2%	6.1%	5.9%
VMBO BL / VMBO KL	2.8%	3.4%	2.8%	3.6%	3.8%	3.5%
VMBO KL	11.1%	8.9%	11.1%	9.4%	9.9%	9.2%
VMBO KL / VMBO (G)TL	4.4%	3.4%	4.4%	3.7%	4.1%	3.6%
VMBO (G)TL	17.6%	17.3%	17.6%	18.4%	18.2%	17.7%
VMBO (G)TL / HAVO	8.5%	8.6%	8.5%	9.1%	9.1%	8.8%
HAVO	19.1%	18.6%	19.1%	18.8%	18.0%	18.6%
HAVO / VWO	7.1%	10.5%	7.1%	9.5%	9.6%	10.1%
VWO	20.9%	22.1%	20.9%	20.2%	19.8%	21.2%

Een probleem dat al gesignaleerd is bij het kiezen van de referentiecesuren is waar de ware cesuren moeten liggen. Voor dit rapport is ervoor gekozen om de marginale percentages schooladviezen (na herziening) uit BRON als richtsnoer te nemen. Deze staan in de laatste kolom van Tabel 7. De motivatie hiervoor is dat de uitslag van de eindtoets een objectief tweede gegeven is bij het schooladvies. Tabel 8 geeft de adviezen weer zoals ze via het anker gegeven zouden zijn, op basis van de drie onderwerpen in het anker en een gezamenlijke weging over alle toetsen. Verder zijn de adviezen, net als bij de referentieniveaus, berekend onder aanname dat de marginale verdeling van de percentages niet verandert. Overigens geldt ook hier dat, door het grote aantal leerlingen dat de Centrale Eindtoets maakte, de kolom met de totale resultaten en de kolom met de resultaten op de Centrale Eindtoets bijna gelijk zijn. De Centrale Eindtoets had dus ook eventueel als ijkpunt voor de cesuren gekozen kunnen worden, zonder dat dat veel verschil had gemaakt.

De Centrale Eindtoets is in Tabel 8 uitgesplitst naar de papieren toets en de digitale multistage toets.

Tabel 8 Toetsadviezen met behulp van het gezamenlijke anker, met een gezamenlijke weging over drie onderwerpen

Advies	AMN	Centrale Eindtoets Totaal	Centrale Eindtoets Papier	Centrale Eindtoets Digitaal	DIA	IEP	Route 8	Totaal
PRO	1.3%	0.8%	0.1%	1.0%	0.5%	0.2%	1.8%	0.9%
VMBO BL	7.4%	5.5%	3.2%	6.8%	4.9%	3.0%	10.0%	5.9%
VMBO BL / VMBO KL	4.1%	3.2%	2.4%	4.4%	3.2%	2.5%	5.2%	3.5%
VMBO KL	11.9%	8.6%	7.0%	11.1%	10.8%	7.6%	12.4%	9.2%
VMBO KL / VMBO (G)TL	3.8%	3.4%	3.1%	4.5%	4.6%	3.2%	4.5%	3.6%
VMBO (G)TL	19.1%	16.8%	16.6%	20.6%	22.0%	17.9%	20.4%	17.7%
VMBO (G)TL / HAVO	9.9%	8.7%	8.9%	9.4%	9.8%	9.7%	9.1%	8.8%
HAVO	18.8%	18.6%	21.2%	19.3%	19.9%	22.3%	17.5%	18.6%
HAVO / VWO	9.4%	10.3%	12.1%	9.0%	10.9%	12.7%	7.8%	10.1%
VWO	14.1%	24.2%	25.3%	14.0%	13.2%	20.9%	11.2%	21.2%

Zoals gemeld moeten de resultaten voorzichtig worden beoordeeld. Toch valt het op dat de niveaus van de adviezen voor de papieren toetsen (de papieren versie van de Centrale Eindtoets en de IEP) hoger zijn dan de adviezen voor de adaptieve toetsen (de AMN, de DIA, Route 8 en de digitale multistage Centrale Eindtoets). Voor de Centrale Eindtoets geldt in ieder geval dat scholen met een zwakkere populatie leerlingen geadviseerd zijn om de digitale versie te kiezen.

De analyses zoals die voor Tabel 8 gemaakt zijn, zijn herhaald met voor iedere aanbieder de eigen gewichten. De resultaten waren volledig vergelijkbaar.

3.3 Analyses m.b.t. het Gezamenlijk Anker

Een deel van de analyses die zijn uitgevoerd om de drie onderwerpen van de toetsen op drie schalen te brengen, is toegevoegd als een bijlage in de vorm van een zip-file. Alle analyses zijn uitgevoerd met een item response (IRT) model l, zowel met het een- als met het twee-parameter logistisch model. Het een-parameter logistisch model staat ook wel bekend als het Rasch model. De in dit rapport gerapporteerde analyseresultaten zijn uitgerekend met het twee parameter logistisch model.

De berekeningen zijn gemaakt met behulp van Lexter (Software en Handleiding beschikbaar op https://www.utwente.nl/nl/bms/omd/Medewerkers/medewerkers/glas/). Voor details m.b.t. de uitvoer zij men verwezen naar deze manual. Het programma is publiek domein, zodat de aanbieders de analyses zelf kunnen repliceren en eventueel uitbreiden (open source benadering). Hier worden vooral de belangrijkste details besproken. De procedure van de analyses bestond uit drie stappen:

1. Per onderwerp een analyse van de itemresponsies op de ankeritems van alle aanbieders gezamenlijk. Het doel van deze drie analyses is om anker-itemparameters op drie gezamenlijke schalen te krijgen.

2. Deze gezamenlijke anker-itemparameters worden vervolgens voor iedere aanbieder als constanten ingevoerd in analyses per onderwerp. Deze analyses zijn bedoeld om voor iedere aanbieder en onderwerp de parameters van de overige items en de parameters van de populaties van de aanbieders op de te schatten. Omdat de anker-itemparameters gefixeerd zijn, staan, per onderwerp, de schattingen van alle item en populatieparameters op drie gezamenlijke schalen.

3. Daarna werden met behulp van de schattingen uit de vorige stap voor iedere leerling vaardigheidsparameters geschat. Deze schattingen werden gebruikt om de cesuren bij de percentages overall behaalde toetsresultaten op referentieniveaus en toetsadviezen te bepalen, en om daarna met behulp van die cesuren de gerealiseerde percentages voor alle aanbieders per onderwerp apart te schatten.

In Bijlage B wordt een technische samenvatting gegeven van de statistisch/psychometrische procedure die is gebruikt, en in Bijlage C wordt een overzicht gegeven van de grootte van de verschillende populaties en de IRT Schattingen van de populatie kenmerken op latente schaal.

@Stap 1. De files TOTAAL_LEZEN.TXT, TOTAAL_REKENEN.TXT en TOTAAL_TAAL.TXT bevatten de analyses van de ankeritems over alle toetsaanbieders. Onder het kopje "MML-PARAMETER ESTIMATION BIRNBAUM-MODEL" staan de schattingen van de parameters van de ankeritems zoals die berekend zijn door gebruik te maken van alle data van alle aanbieders. Onder het kopje "ESTIMATION OF POPULATION PARAMETERS" staan de gemiddelden en standaard deviaties van de vaardigheidsverdelingen van de verschillende populaties op een gezamenlijke schaal. Onder het kopje "BOOKLET MARG E(8) VAR(E(8|X)) E(VAR(8|X)) VAR(8) REL" staan in de laatste kolom de betrouwbaarheden van de verschillende toetsen. Deze betrouwbaarheden zijn voor de IEP, Route 8 en de digitale multistage Centrale Eindtoets laag, maar dat komt door het design waarmee de ankeritems in die toetsen zijn afgenomen of door het feit dat iedere leerling daar slechts een beperkt aantal items maakte. In de volgende analyses wordt een realistischere schatting van de betrouwbaarheid van de eindtoetsen gemaakt.

Onder de kopjes "Lagrange tests DIF for BIRNBAUM-MODEL" staan voor iedere toets indices voor "differential item functioning". Het gaat hierbij om de vraag of de items zich per toets volgens het model gedragen. De "Focal Group" is steeds de toets die in het kopje vermeld staat, de "Reference" zijn de andere eindtoetsen. Door de grote steekproeven is de statistische Lagrange multiplier test niet relevant: de power is te groot. Wel relevant is de effectgrootte, die onder het kopje "Dif." staat. Deze effectgrootte is gebaseerd op het verschil tussen geobserveerde en onder het model verwachte scores. De norm is dat de absolute waarde van deze effectgrootte niet groter dan 0.10 mag zijn. Onder het kopje "Lagrange multipliers tracelines for BIRNBAUM MODEL" staan geobserveerde en verwachte scores op drie scoreniveaus. Ook hier is de norm dat de absolute waarde van de effect-grootte niet meer dan 0.10 mag zijn.

In Tabel 9 staat een voorbeeld m.b.t. de AMN toets voor het onderwerp Lezen. Onder de kopjes "LM", "df" en "Prob" staan per item respectievelijk de uitkomst van de statistiek voor modelpassing (een Lagrange Multiplier statistiek, die chi-kwadraat verdeeld is), het aantal vrijheidsgraden, en de significantiekans. Door de steekproefgrootte zijn deze getallen weinig informatief; alleen de relatieve waarden onder LM geven enig inzicht, waarbij grote waarden duiden op gebrek aan modelpassing, c.q. DIF. In de kolommen daarachter staan voor de "Focal Group" (c.q. AMN) en de "Reference Group" (alle andere aanbieders) respectievelijk de geobserveerde itemscore en de onder de schattingen van het model verwachte itemscore. Grote afwijkingen tussen die twee duiden op DIF. Tenslotte staat in de laatste kolom een effectgrootte voor DIF: het verschil tussen geobserveerd en verwacht voor de "Focal Group". Een effectgrootte met een absolute waarde groter dan 0.10 wijst op DIF. Dat geldt hier dus voor het item met het label "OP00016A".

@Stap 2. De overige files, van AMN_LEZEN.TXT tot en met ROUTE8_TAAL.TXT, bevatten analyses waarbij de in de vorige stap geschatte itemparameters als gefixeerde constanten zijn ingevoerd. Daardoor is voor ieder onderwerp iedere analyse uitgevoerd op dezelfde schaal. De interpretatie van de statistische toetsing is analoog aan die bij de vorige analyses. De analyses per combinatie van aanbieder en onderwerp zijn gebruikt om parameters voor de leerlingen te schatten (via weighted maximum Likelihood) en deze schattingen zijn vervolgens gebruikt om de percentages beheerste referentieniveaus en toetsadviezen te berekenen.

De conclusie van al deze analyses is dat het anker goed gefunctioneerd heeft en dat de itemparameterschattingen goed bruikbaar zijn voor het schalen van de eindtoetsen in 2019.

Tabel 9: Voorbeeld van DIF analyse AMN: functioneren de ankeritems bij AMN Vergelijkbaar als bij de andere eindtoetsen

				Focal Group		Reference Group		
Item	LM	df	Prob	Obs	Exp	Obs	Exp	Dif.
2219	0.0	1	0.98	0.70	0.70	0.79	0.79	0.00
2221	5.0	1	0.02	0.58	0.55	0.69	0.69	0.03
2222	6.6	1	0.01	0.66	0.69	0.79	0.79	-0.03
BL00099	4.2	1	0.04	0.86	0.88	0.90	0.90	-0.02
BL00100	35.9	1	0.00	0.74	0.68	0.72	0.72	0.06
BL00101	2.0	1	0.16	0.47	0.48	0.52	0.52	-0.02
BL00103	13.4	1	0.00	0.65	0.60	0.64	0.64	0.04
BL00045	0.4	1	0.52	0.64	0.65	0.68	0.68	-0.01
OP00016A	236.4	1	0.00	0.47	0.66	0.73	0.70	-0.19
OP00177	20.1	1	0.00	0.76	0.71	0.73	0.73	0.05
BL00524	16.9	1	0.00	0.63	0.58	0.70	0.70	0.05
BL00525	8.6	1	0.00	0.50	0.54	0.63	0.63	-0.04
OP00032	140.7	1	0.00	0.73	0.60	0.61	0.61	0.14
BL00372	32.9	1	0.00	0.69	0.62	0.64	0.65	0.06
BL00159	9.2	1	0.00	0.72	0.69	0.70	0.70	0.03
BL00160	7.4	1	0.01	0.59	0.62	0.62	0.62	-0.03
SV00041	62.3	1	0.00	0.52	0.62	0.59	0.59	-0.10
SV00042	8.0	1	0.00	0.65	0.61	0.59	0.59	0.04
SV00043	8.7	1	0.00	0.59	0.62	0.59	0.59	-0.04
SV00086	5.0	1	0.03	0.54	0.57	0.54	0.54	-0.03
SV00045	1.1	1	0.30	0.73	0.74	0.71	0.71	-0.01

3.4 Schaling van de Referentieniveaus

Bij alle hiervoor beschreven IRT analyses is er van uitgegaan dat voor ieder van de drie onderwerpen aparte schalen gedefinieerd zijn. Het alternatief is om per onderwerp steeds twee schalen te veronderstellen, een schaal die bestaat uit de items die als 1F zijn geclassificeerd en een schaal die voor de items die als 2F/1S zijn geclassificeerd. Een nadeel van die benadering is dat de analyses gecompliceerder worden en dat de betrouwbaarheid van de schattingen van de cesuren minder betrouwbaar wordt, omdat de schalen minder items bevatten.

Om na te gaan of de aanname van slechts drie schalen in plaats van zes schalen een redelijke is, is de correlatie tussen de schalen met 1F items en schalen met 2F/1S items berekend (technische opmerking: het gaat hierbij om via IRT geschatte latente correlaties). Als de correlatie hoog is, is dit een indicatie dat er sprake is van één schaal en dat het veronderstellen van meerdere schalen niet nodig is. Verder is de globale betrouwbaarheid van de schalen berekend om na te gaan hoeveel de betrouwbaarheid daalt als er meerdere schalen worden verondersteld.

De op basis van de beschikbare data berekende resultaten zijn weergegeven in Tabel 10.

Tabel 10 Betrouwbaarheid en correlaties van referentieniveaus

	Betrouwbaarheid			Correlatie
	Geheel	1F	2F/1S	1F en 2F
LEZEN				
AMN	0.90	0.81	0.80	0.96
Centrale Eindtoets Papier	0.83	nb	nb	nb
Centrale Eindtoets Digitaal	0.87	nb	nb	nb
DIA	0.72	nb	nb	nb
IEP	0.80	0.64	0.73	0.97
ROUTE8	0.73	0.62	0.66	0.95
REKENEN				
AMN	0.95	0.88	0.90	0.99
Centrale Eindtoets Papier	0.92	nb	nb	nb
Centrale Eindtoets Digitaal	0.96	nb	nb	nb
DIA	0.80	nb	nb	nb
IEP	0.91	0.78	0.81	0.98
ROUTE8	0.91	0.80	0.77	0.95
TAALVERZORGING				
AMN	0.94	0.77	0.90	0.86
Centrale Eindtoets Papier	0.85	nb	nb	nb
Centrale Eindtoets Digitaal	0.78	nb	nb	nb
DIA	0.83	nb	nb	nb
IEP	0.83	0.73	0.79	0.86
ROUTE8	0.69	0.58	0.61	0.66

nb staat voor niet beschikbaar

Geconcludeerd kan worden dat de correlatie tussen de 1F en 2F schalen is steeds hoog tot zeer hoog: bij Lezen en Rekenen altijd dicht bij 1.00 en bij Taalverzorging rond de 0.85.

3.5 Effecten op Itemniveau van Digitale versus Papieren afname

Bij alle hiervoor beschreven IRT analyses is er steeds van uitgegaan dat de eigenschappen van items die zowel op papier als digitaal worden aangeboden voor deze twee formats gelijk zijn. De als bijlage bijgevoegde files TOTAAL_LEZEN_DIF.TXT, TOTAAL_REKENEN_DIF.TXT en TOTAAL_TAAL_DIF.TXT bevatten DIF analyses waarbij voor alle combinaties van twee aanbieders is getoetst of de parameters van de ankeritems de gemiddelde p-waarden van de items voldoende nauwkeurig voorspelden. De Centrale Eindtoets is in dit opzicht het meest interessant, omdat daar items zowel op papier als digitaal werden aangeboden. De uitkomsten zijn in de uitvoer te vinden onder het kopje "CET_PAPIER versus CET_MULT". Deze tabellen worden hieronder, enigszins aangepast, herhaald als Tabel 11a, Tabel 11b en Tabel 11c.

Tabel 11a DIF test Centrale Eindtoets papier versus digitaal: Lezen

				Papier		Digitaal		
Item	LM	Df	Pr	Obs.	Exp.	Obs.	Exp.	Dif
BL00099	6.6	1	0.01	0.93	0.94	0.88	0.89	0.01
BL00100	36.5	1	0.00	0.77	0.81	0.67	0.70	0.03
BL00101	33.0	1	0.00	0.64	0.60	0.50	0.49	0.03
BL00103	46.4	1	0.00	0.68	0.73	0.61	0.61	0.03
BL00045	61.1	1	0.00	0.70	0.76	0.66	0.66	0.03
OP00177	2.0	1	0.16	0.81	0.81	0.72	0.72	0.00
BL00033	13.4	1	0.00	0.77	0.79	0.69	0.68	0.02
BL00035	76.3	1	0.00	0.72	0.66	0.47	0.49	0.04
BL00036	2.3	1	0.13	0.75	0.76	0.66	0.66	0.01
BL00524	63.4	1	0.00	0.79	0.73	0.60	0.59	0.03
BL00525	3.9	1	0.05	0.64	0.66	0.54	0.55	0.01
OP00032	5.2	1	0.02	0.69	0.71	0.60	0.60	0.01
BL00372	6.4	1	0.01	0.72	0.74	0.62	0.63	0.01
SV15237	36.4	1	0.00	0.67	0.63	0.36	0.40	0.04
SV15238	4.0	1	0.05	0.68	0.70	0.53	0.52	0.01
SV15239	2.4	1	0.12	0.80	0.81	0.71	0.69	0.01
SV15240	28.0	1	0.00	0.89	0.91	0.79	0.79	0.02
SV15241	4.6	1	0.03	0.76	0.77	0.57	0.56	0.01

Opnieuw zijn de uitkomsten van de Lagrange multiplier test (onder de kopjes "LM", "Df" en "Pr") alleen van relatief belang. Bijvoorbeeld: in Tabel 11a zijn de verschillen tussen papier en digitaal het grootste voor item BL00035 en het kleinste voor OP00177. Deze waarden zijn gebaseerd op de geobserveerde en verwachte p-waarden onder het kopje "Papier" en de geobserveerde en verwachte p-waarden onder het kopje "Digitaal". De gemiddelde absolute verschillen tussen deze waarden staan onder het kopje "DIF' in de laatste kolom. In deze kolom is te zien dat geen van deze effect waarden de grens van 0.10 overschrijdt. Hetzelfde geldt ook voor de resultaten voor de onderwerpen Rekenen en Taalverzorging, zoals die in Tabel 11b en Tabel 11c zijn weergegeven. Verder is in de als bijlage bijgevoegde files TOTAAL_LEZEN_DIF.TXT, TOTAAL_REKENEN_DIF.TXT en TOTAAL_TAAL_DIF.TXT te zien dat ook voor andere combinaties van eindtoetsen de data goed gerepresenteerd zijn onder de hypothese dat de parameters voor de ankeritems gelijk zijn voor papieren en digitale versies.

Tabel 11b DIF test Centrale Eindtoets papier versus digitaal: Rekenen

				Papier		Digitaal		
Item	LM	Df	Pr	Obs.	Exp.	Obs.	Exp.	Dif
RD515814	0.1	1	0.71	0.62	0.62	0.46	0.49	0.02
RD515731	1.7	1	0.19	0.73	0.72	0.65	0.65	0.01
RD514054	22.6	1	0.00	0.61	0.65	0.52	0.52	0.02
RD515449	19.8	1	0.00	0.83	0.86	0.82	0.80	0.03
RD514574	5.3	1	0.02	0.84	0.83	0.77	0.75	0.02
RD515108	10.0	1	0.00	0.85	0.86	0.80	0.79	0.02
RD514218	60.4	1	0.00	0.83	0.78	0.66	0.67	0.03
RD515700	9.0	1	0.00	0.56	0.58	0.51	0.49	0.02
RD515062	17.7	1	0.00	0.83	0.85	0.77	0.77	0.01
RD515148	10.9	1	0.00	0.73	0.75	0.66	0.65	0.02
RD515703	16.5	1	0.00	0.83	0.81	0.73	0.72	0.02

Tabel 11c DIF test Centrale Eindtoets papier versus digitaal: Taalverzorging

				Papier		Digitaal		
Item	LM	Df	Pr	Obs.	Exp.	Obs.	Exp.	Dif
IP00034	6.4	1	0.01	0.80	0.79	0.74	0.74	0.01
IP00087	1.1	1	0.29	0.68	0.69	0.60	0.58	0.01
IP01428	39.5	1	0.00	0.76	0.80	0.73	0.72	0.03
SN00031	0.1	1	0.81	0.74	0.74	0.64	0.70	0.03
SN02480	9.3	1	0.00	0.88	0.90	0.85	0.88	0.02
SN02955	1.9	1	0.17	0.83	0.82	0.76	0.76	0.01
SW00031	0.0	1	0.89	0.63	0.63	0.55	0.55	0.00
SW00043	32.8	1	0.00	0.64	0.59	0.46	0.50	0.04
SW02285	25.4	1	0.00	0.71	0.75	0.69	0.68	0.02
IP01090	14.7	1	0.00	0.83	0.80	0.74	0.73	0.02
IP01118	9.1	1	0.00	0.74	0.72	0.67	0.63	0.03
IP01373	16.9	1	0.00	0.74	0.72	0.64	0.63	0.02
SN02487	42.1	1	0.00	0.76	0.81	0.77	0.77	0.02
SN02928	1.2	1	0.27	0.79	0.78	0.70	0.73	0.02
SN02944	1.4	1	0.24	0.69	0.70	0.63	0.64	0.01
SW02282	66.3	1	0.00	0.72	0.67	0.55	0.59	0.05
SW02561	0.3	1	0.60	0.64	0.64	0.58	0.57	0.01
SW02985	73.6	1	0.00	0.61	0.67	0.63	0.62	0.04
IP00022	15.4	1	0.00	0.72	0.75	0.70	0.67	0.03

4. Mogelijke Scenario's voor Toetsadviezen

Het voorstellen van adviezen voor de cesuren van toetsadviezen is niet eenvoudig. Zo heeft iedere aanbieder heeft een andere weging over onderwerpen. Het blijkt echter dat, zoals hierboven al aangegeven, de gewichten voor de drie gezamenlijk onderwerpen dicht bij elkaar liggen en de verschillen die ze veroorzaken zijn verwaarloosbaar. Een belangrijker verschil tussen de aanbieders vormen de unieke onderwerpen die ook in de weging worden meegenomen. En verder wordt het aantal adviescategorieën in 2019 teruggebracht van 10 naar 6 categorieën.

Eerst worden de effecten van het terugbrengen van het aantal categorieën van 10 naar 6 onderzocht. Daarna wordt, aan de hand van een voorbeeld, onderzocht wat het effect is van extra onderwerpen en wordt een suggestie gedaan hoe hiermee om te gaan bij het vaststellen van cesuren.

De analyses die tot Tabel 7 leidden, zijn herhaald om na te gaan hoe de resultaten van de afname van 2018 eruit gezien zouden kunnen hebben als er al 6 in plaats van 10 categorieën geweest zouden zijn. Er zijn twee scenario's onderzocht. Het eerste scenario is ontleend aan het rapport "Adviescategorieën voor eindtoetsen 2019" (Cito, maart 2018). In dat rapport worden verschillende scenario's onderzocht. De beste resultaten in termen van het percentage correcte toetsadviezen, het percentage heroverwegingen en het percentage terechte heroverwegingen werden behaald door het zogenaamde Brede Categorieën scenario. Hierbij werden enkelvoudige advies-categorieën gecombineerd met meervoudige adviezen, alleen de categorie VWO blijft enkelvoudig. De resultaten van de her-analyse staan in Tabel 12. Uit de vergelijking van de resultaten in Tabel 7 met die in Tabel 12 kan men concluderen dat de verschillen tussen de aanbieders enigszins worden gedempt.

Tabel 12 Toetsadviezen in 6 categorieën met behulp van het gezamenlijke anker, onder het Brede Categorieën scenario.

Advies	AMN	Centrale Eindtoets Totaal	Centrale Eindtoets Papier	Centrale Eindtoets Digitaal	DIA	IEP	Route 8	Totaal
PRO / VMBO BL	1.3%	0.8%	0.1%	1.0%	0.5%	0.2%	1.8%	0.9%
VMBO BL / VMBO KL	10.4%	8.7%	5.7%	11.2%	8.2%	5.5%	15.1%	9.4%
VMBO KL / VMBO (G)TL	15.9%	11.9%	10.1%	15.6%	15.4%	10.8%	17.0%	12.8%
VMBO (G)TL / HAVO	28.3%	25.5%	25.5%	30.0%	31.9%	27.5%	29.5%	26.5%
HAVO / VWO	28.6%	28.9%	33.3%	28.2%	30.8%	35.0%	25.3%	28.7%
VWO	15.5%	24.2%	25.3%	14.0%	13.2%	20.9%	11.2%	21.2%

Tabel 13 Toetsadviezen in 6 categorieën met behulp van het gezamenlijke anker, met een alternatieve normering

Advies	AMN	Centrale Eindtoets Totaal	Centrale Eindtoets Papier	Centrale Eindtoets Digitaal	DIA	IEP	Route 8	Totaal
PRO / VMBO BL	1.3%	0.8%	0.1%	1.0%	0.5%	0.2%	1.8%	0.9%
VMBO BL / VMBO KL	16.2%	12.9%	9.1%	16.8%	14.0%	9.2%	21.5%	14.0%
VMBO KL / VMBO (G)TL	19.3%	16.0%	14.8%	20.6%	21.2%	15.6%	21.3%	17.1%
VMBO (G)TL / HAVO	28.9%	26.5%	28.0%	29.3%	30.8%	29.9%	28.2%	27.0%
HAVO / VWO	18.9%	19.7%	22.9%	18.3%	20.4%	24.2%	16.1%	19.4%
VWO	15.4%	24.1%	25.2%	13.9%	13.2%	20.8%	11.1%	21.2%

Als alternatief is een scenario onderzocht waarbij de leerlingen in enkelvoudige categorieën (behalve VWO) voor de helft aan een onderliggende meervoudige advies-categorie en voor de helft aan een

bovenliggende advies categorie worden toebedeeld. De resultaten staan in Tabel 13. In dit scenario is het niveau van de eindtoetsadviezen wat lager.

Een scenario voor adviezen voor eindtoetsen met een gezamenlijk anker zal altijd een compromis zijn. Het doel van het gezamenlijk anker is om gezamenlijke cesuren vast te stellen die de moeilijkheidsgraad van de verschillende eindtoetsen en de vaardigheidsverdeling van de verschillende eindtoetspopulaties goed representeren. Daartoe wordt voor de drie gezamenlijke onderwerpen via een gezamenlijk anker op een latente schaal een vijftal cesuren bepaald die leiden tot percentages toetsadviezen die de vaardigheid van de leerlingen goed representeren. De extra onderwerpen staan echter niet automatisch op deze schaal. Voor zo'n onderwerp kan wel een IRT schaal worden geschat, maar die schaal is niet automatisch gelinkt aan de gezamenlijke schaal. Het normeringsonderzoek van de aanbieder levert echter wel informatie over de relatie tussen de vaardigheidsverdelingen van zowel de gezamenlijke onderwerpen als de extra onderwerpen. Deze samenhang is in het normeringsonderzoek van de aanbieder bepaald door de verdelingen te linken aan de schooladviezen of door doorstroom-onderzoek, of door een combinatie van beiden. Het uitgangspunt bij de linking tussen de drie gezamenlijke onderwerpen en de extra onderwerpen is dat de marginale verdeling van toetsadviezen door het extra onderwerp niet mag veranderen. Voor individuele leerlingen heeft het extra onderwerp wel effect. Voor een individuele leerling kan een extra onderwerp compenserend werken. Omdat de vaardigheidsverdelingen op de drie gezamenlijke onderwerpen en de extra onderwerpen gelijk wordt verondersteld, en daarmee dus ook de distributie van adviezen op de drie gezamenlijke onderwerpen en de extra onderwerpen gelijk wordt, kunnen alle onderwerpen via op een gezamenlijke schaal worden afgebeeld. De procedure is in detail uitgewerkt in Bijlage C, in de stappen 8 en 9.

Hieronder wordt dat aan de hand van een voorbeeld verder onderzocht. Het voorbeeld is ontleend aan de DIA eindtoets met als extra onderwerp Woordenschat. In de kolom onder het kopje "Zonder Woordenschat" worden de resultaten uit Tabel 12 herhaald. Het gaat dus om de resultaten van de Brede Categorieën, maar zonder het onderwerp Woordenschat. De gewichten staan in de laatste rijen. Onder het kopje "Woordenschat Versie 1" staan de resultaten met de weging zoals die voor de DIA eindtoets is gebruikt. De gewichten sommeren steeds tot 1.0. De verschillen met de versie zonder woordenschat zijn erg klein. Dat komt vooral doordat de correlaties tussen de onderwerpen erg hoog zijn. Woordenschat correleert 0.669 met Lezen, 0.476 met Rekenen en 0.482 met Taalverzorging. Daardoor verandert de toevoeging van Woordenschat met een gewicht van 0.10 nauwelijks iets aan de globale resultaten. De correlaties tussen de drie gezamenlijke onderwerpen zijn te vinden in Bijlage C.

Tabel 14 Effecten van extra onderwerpen en variaties in weging op toetsadviezen: voorbeeld met Woordenschat als extra onderwerp bij DIA

Advies	Zonder Woordenschat	Woordenschat Versie 1	Woordenschat Versie 2	Woordenschat Versie 3
PRO / VMBO BL	0.5%	0.4%	0.4%	0.7%
VMBO BL / VMBO KL	8.2%	8.5%	8.9%	9.6%
VMBO KL / VMBO (G)TL	15.4%	15.4%	15.4%	15.9%
VMBO (G)TL / HAVO	32.0%	31.9%	32.5%	31.1%
HAVO / VWO	30.8%	30.4%	29.7%	27.2%
VWO	13.2%	13.4%	13.2%	15.6%
Weging				
Lezen	0.262	0.300	0.250	0.100
Rekenen	0.458	0.400	0.250	0.300
Taalvaardigheid	0.279	0.200	0.250	0.100
Woordenschat		0.100	0.250	0.500

Om de invloed van het gewicht na te gaan, zijn twee alternatieven onderzocht. De alternatieven zijn met opzet extreem gekozen en dus voor de praktijk niet realistisch. In het eerste voorbeeld zijn alle vier de onderwerpen gelijk gewogen. In het tweede voorbeeld heeft het extra onderwerp evenveel gewicht als de drie gezamenlijke onderwerpen samen, waarbij Rekenen wel meer gewicht heeft gekregen dan de twee talige ontwerpen. De resultaten van de twee voorbeelden staan respectievelijk onder de kopjes "Woordenschat Versie 2" en "Woordenschat Versie 3". Opnieuw zijn de veranderingen in de percentages klein. De hoge correlaties tussen de onderwerpen zijn hier debet aan.

5. Conclusies en Verder Ontwikkeltraject

5.1 Conclusies

De eerste conclusie die uit de analyse van de resultaten van de toetsafname in 2018 getrokken moet worden, is dat de normering van de verschillende eindtoetsen ver uit elkaar loopt en dat de resultaten daardoor niet goed te vergelijken zijn. Dit ondanks het feit dat alle aanbieders hun normering zorgvuldig hebben onderbouwd met koppelingen aan de referentiesets en schooladviezen. De reden voor de divergentie is, hangt waarschijnlijk samen met het feit dat de koppeling aan de referentiesets eenmalig was, en dat de referentiecesuren vervolgens van afname naar afname werden doorgegeven met moeilijk te controleren statistische foutenmarges. Het uiteenlopen van de toetsadviezen is waarschijnlijk toe te schrijven aan het feit dat het erg moeilijk blijkt een perfecte normeringssteekproef te trekken, waardoor de schooladviezen waarop iedere aanbieder normeert geen zuivere steekproef is uit de landelijke verzameling schooladviezen. De invoering van een gezamenlijk anker zal het probleem van een gezamenlijke schaal voor referentieniveaus en schooladviezen oplossen.

Een tweede conclusie uit de analyse van de resultaten van de toetsafname in 2018 is dat de adviezen van de expertgroep niet erg exact bleken. Dit is toe te schrijven aan de minder ideale randvoorwaarden waarbinnen gewerkt moest worden en het feit dat de Expertgroep uiteraard ook geen zicht had op de mate van divergentie tussen de verschillende eindtoetsen. Doordat het anker van 2018 hergebruikt wordt in 2019 en er daarna een stevige koppeling tussen de opeenvolgende versies zal plaatsvinden, zullen de adviezen van de Expertgroep in de toekomst veel betrouwbaarder zijn.

Een derde conclusie is dat het anker erg goed gefunctioneerd heeft. De eigenschappen van de ankeritems waren over de verschillende eindtoetsen heen erg stabiel en ook het verschil tussen papieren en digitale versies van de items waren verwaarloosbaar. Hergebruik van het anker is dus goed verdedigbaar.

5.2 Verder ontwikkeltraject

Naar aanleiding van de ervaringen die in 2018 met het gezamenlijke anker zijn opgedaan, worden voor 2019 de volgende acties ondernomen.

- Er wordt een ICT-applicatie ontwikkeld met de volgende functionaliteiten:
    a. Het invoeren van itemlabels, itemattributen (item beschrijving, eigenschappen etc.) en het data design van een eindtoets. Bij het laatste gaat het er bijvoorbeeld om hoeveel boekjes er zijn, welke items er in de verschillende boekjes zitten en welke deelpopulaties de boekjes krijgen, of bij een adaptieve toets: wie welk item beantwoordt.
    b. Het uitvoeren van statistische analyses om toetsen aan elkaar te linken en om toetsen over tijd te kunnen linken.
- Voor de aanbieders wordt er een workshop ontwikkeld en in de periode rond de afname van de toets zal er ondersteuning zijn bij eventuele problemen.
- Het anker van 2018 wordt ook gebruikt in 2019. De itemparameters, itemlabels en het toets-design worden ver voor de operationele eindtoetsafname in de applicatie ingevoerd en uitgetest, zodat de Expertgroep snel adviezen kan geven.

- Een aantal aanbieders van eindtoetsen gaat de nieuwe opgaven voor het anker van 2020 zaaien in de eindtoets van 2019. Daardoor kunnen de parameters van deze nieuwe opgaven ook in een operationele setting (in een high-stakes situatie) nauwkeurig worden geschat. Hiermee ontstaat er een sterke koppeling tussen het anker zoals dat in 2018 en 2019 is gebruikt en het anker van 2020. Ook het probleem van de verschillende afnamemomenten is hiermee verder opgelost, want opnieuw zijn de eigenschappen c.q. de itemparameters van het anker vooraf bekend.

6. Literatuurlijst

- CvTE (2014a). *Algemeen deel Toetswijzer voor eindtoets PO. Inhoudelijke kwaliteitseisen aan eindtoetsen PO*. Utrecht: CvTE.
- CvTE (2015). *Wetenschappelijke verantwoording Centrale Eindtoets 2014-2015*. Utrecht: CvTE.
- CvTE (2016). *Rapportage referentieniveaus 2015-2016.* Utrecht: CvTE.
- Cito (2018). *Adviescategorieën voor eindtoetsen 2019.* Arnhem: Cito. .
- Egberink, A., Verweij, J. & Termaat, B.R. (2016). *Verantwoording van Route 8*. Apeldoorn: A-VISION Holding B.V.
- Emons, W.H.M., Glas, C.A.W. & Berding-Oldersma, P.K. (2016). *Rapportage Vergelijkbaarheid eindtoetsen.* Utrecht: Expertgroep PO.
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests.* Utrecht: Nederlands Instituut voor Psychologen.
- Folmer, E., Boswinkel, N., Prenger, J. & Vorle, R. van de (2018). *Beoordelingskader onderwijskundige en organisatorische aspecten andere eindtoetsen.* Utrecht: Expertgroep toetsen PO.
- Glas, C.A.W., Emons, W.H.M. Emons & Berding-Oldersma, P.K. (2016). *Scenario's voor ijking van de eindtoetsen op de referentieniveaus*. Utrecht: Expertgroep Toetsen PO.
- Langeveld, E.A., Bezdan, E., Binsbergen, M., Haitjema, T., Laarhuis, R., Ebert-Flikweert, L., Van den Ouden, L., Helsloot, J. & Bonouvrie, N. (2016). *IEP Eindtoets verantwoording 2016-2017*. Culemborg: Bureau ICE.
- Ministerie van Onderwijs, Cultuur en Wetenschappen (2010). *Besluit referentieniveaus Nederlandse taal en rekenen.* Den Haag: Ministerie van OCW.
- Ministerie van Onderwijs, Cultuur en Wetenschappen (2014). *Toetsbesluit PO.* Den Haag: Ministerie van OCW.
- Wools, S. & Béguin, A. (2013). *Handleiding referentieset*. Arnhem: Cito.

Bijlage A Verschil tussen Geschatte Referentieniveaus voor Advies en in de Kalibratie

Tabel A. Verschil tussen adviezen Expertgroep en de behaalde Referentieniveaus zoals die geschat zijn nadat alle eindtoetsen en vaardigheidsverdelingen van de verschillende populaties op een schaal gekalibreerd zijn

		AMN	DIA	IEP	Route 8
LEZEN	Lager dan 1F	-0.4%	4.6%	4.4%	6.2%
	1F	0.4%	-4.6%	-4.4%	-6.2%
	2F	11.0%	-15.1%	-6.5%	-16.6%
REKENEN	Lager dan 1F	-11.2%	-1.1%	0.9%	9.1%
	1F	11.2%	1.1%	-0.9%	-9.1%
	1S/2F	31.3%	-4.1%	10.4%	-1.2%
TAAL	Lager dan 1F	-2.9%	5.5%	1.6%	7.8%
VERZORGING	1F	2.9%	-5.5%	-1.6%	-7.8%
	2F	19.2%	-8.9%	3.3%	2.2%

Bijlage B Statistisch en Psychometrische Aspecten van de Kalibratie-procedure

We introduceren de volgende definities.

Eindtoetsen	*t = 1,…,T; T = 6*
Onderwerpen	*m = 1,…,M; M=3*
Data Ankeritems	A_{tm}
Data Eindtoets	X_{tm}
Itemparameters Anker per onderwerp	β_{0m}
Itemparameters eindtoets per onderwerp	β_{tm}
Latente schaal onderwerp	θ_m
Cesuren referentieniveaus	τ_{ms} , *s=1,..,S; S=2.*
Geobserveerde marginale verdeling referentieniveaus	P_{ms}
Weging onderwerpen voor toetsadvies	$\zeta = \sum_{m=1}^{M} w_m \theta_m, \quad \sum_{m=1}^{M} w_m = 1$
Cesuren toetsadviezen	ζ_{0d} , *d = 1,…,D*
Geobserveerde marginale verdeling toetsadviezen	P_d

Stap 1. Schat de parameters van de ankeritems op een gezamenlijke schaal door het maximaliseren van

$$L\left(\beta_{om}, \mu_{1m}, ..., \mu_{Tm}, \sigma_{1m}^2, ..., \sigma_{Tm}^2\right) = \prod_t P\left(A_{tm}; \beta_{0m}, \mu_{tm}, \sigma_{tm}^2\right) ,$$

waarbij $P\left(A_{tm}; \beta_{0m}, \mu_{tm}, \sigma_{tm}^2\right)$ de kans op de anker-responsies onder het 2-parameter logistisch model is, gegeven de gezamenlijk parameters van de anker-items en de vaardigheidsparameters van populatie *t*.

Stap 2. Schat de parameters van alle toets-items en herschat de vaardigheidsverdelingen van de eindtoetsen op een gezamenlijk schaal gegeven de geschatte parameters van de ankeritems, dus schat

$$P\left(\theta_m; \beta_{om}, ..., \beta_{Tm}, \mu_{1m}, ..., \mu_{Tm}, \sigma_{1m}^2, ..., \sigma_{Tm}^2\right).$$

Stap 3. Schat referentiecesuren τ_{ms} per onderwerp op de schaal θ_m, door oplossen van de vergelijking

$$P_{ms} = P\left(\theta_m \mid \tau_{ms-1} < \theta_m < \tau_{ms}; \beta_{om}, ..., \beta_{Tm}, \mu_{1m}, ..., \mu_{Tm}, \sigma_{1m}^2, ..., \sigma_{Tm}^2\right)$$

.

Stap 4. Schat voor iedere eindtoets de verdeling van referentieniveaus:

$$P\left(\theta_m \mid \tau_{ms-1} < \theta_m < \tau_{ms}; \beta_{om}, \beta_{tm}, \mu_{tm}, \sigma^2_{tm}\right)$$

Stap 5. Schat de parameters van alle toets-items en herschat de vaardigheidsverdelingen van de eindtoetsen op een gezamenlijk schaal gegeven de geschatte parameters van de ankeritems, dus schat

$$P\left(\zeta; \beta_{o1}, ..., \beta_{TM}, \mu_{11}, ..., \mu_{TM}, \sigma^2_{11}, ..., \sigma^2_{TM}\right).$$

Stap 6. Schat cesuren voor toetsadviezen ζ_{0d} per onderwerp op de schaal ζ , door oplossen van de vergelijking

$$P_d = P\left(\zeta \mid \zeta_{d-1} < \zeta < \zeta_d; \beta_{o1}, ..., \beta_{TM}, \mu_{11}, ..., \mu_{TM}, \sigma^2_{11}, ..., \sigma^2_{TM}\right)$$

Stap 7. Schat voor iedere eindtoets de verdeling van toetsadviezen:

$$P\left(\zeta \mid \zeta_{d-1} < \zeta < \zeta_d; \beta_{ot}, ..., \beta_{tM}, \mu_{t1}, ..., \mu_{tM}, \sigma^2_{t1}, ..., \sigma^2_{tM}\right)$$

Afhandeling Extra onderwerpen.

Stap 8. Nieuwe gewichten. Stel dat er 2 extra onderwerpen zijn. Stel dat het normeringsonderzoek van de aanbieder leidde tot gewichten v_1, v_2, v_3, v_4, v_5 . Het idee achter het berekenen van de nieuwe gewichten is dat het aandeel van de extra onderwerpen naar rato van hun gewicht in het normeringsonderzoek wordt, en dat de drie gezamenlijke onderwerpen een gewicht krijgen naar rato van de oorspronkelijke gezamenlijke gewichten maar gewogen met hun aandeel in het normeringsonderzoek. Dus, de gewichten z_1, z_2, z_3, z_4, z_5 zijn gedefinieerd als

$$z_i = v_i \left(\sum_{j=1}^{3} w_i \right) \quad \text{voor } i = 1,...,3 \quad \text{en} \quad z_i = v_i \quad \text{voor } i = 4,...,5.$$

Stap 9. De vaardigheidsverdelingen en de itemparameters uit het normeringsonderzoek voor de twee extra onderwerpen worden herschaald via $\mu_4 = \sum_{j=1}^{3} \mu_j$, $\mu_5 = \sum_{j=1}^{3} \mu_j$, $\sigma_4 = \sum_{j=1}^{3} \sigma_j$, en $\sigma_5 = \sum_{j=1}^{3} \sigma_j$.

En met deze IRT schalen wordt vervolgens Stap 7 uitgevoerd, maar nu met *M=5* en de gewichten z_1, z_2, z_3, z_4, z_5 zoals gedefinieerd.

Bijlage C IRT Schattingen van Populatie Kenmerken op de Latente Schaal

Tabel B IRT schattingen van populatie kenmerken op de latente schaal

		Schatting: Anker		Schatting: Aanbieder		Correlation Matrix		
		Gemiddelde	Standaard Deviatie	Gemiddelde	Standaard Deviatie	Lezen	Rekenen	Taalvaardigheid
AMN	Lezen	-0.26	1.05	-0.25	1.04	1.00	0.70	0.80
N=1263	Rekenen	-0.36	1.04	-0.34	1.08	0.70	1.00	0.74
	Taalvaardigheid	-0.34	1.02	-0.35	0.97	0.80	0.74	1.00
CET_PAPIER	Lezen	0.52	1.12	0.53	1.08	1.00	0.65	0.60
N=76891	Rekenen	0.40	1.05	0.25	1.00	0.65	1.00	0.65
	Taalvaardigheid	0.50	1.12	0.33	0.95	0.60	0.65	1.00
CET_MULTISTAGE	Lezen	-0.24	0.96	-0.25	0.75	1.00	0.74	0.73
N=24222	Rekenen	-0.10	1.09	0.05	1.09	0.74	1.00	0.63
	Taalvaardigheid	0.03	0.99	0.01	0.98	0.73	0.63	1.00
DIA	Lezen	-0.05	0.97	-0.06	0.94	1.00	0.63	0.59
N=2407	Rekenen	0.07	1.03	0.05	0.98	0.63	1.00	0.59
	Taalvaardigheid	0.00	0.94	0.00	0.91	0.59	0.59	1.00
IEP	Lezen	0.44	0.97	0.42	0.93	1.00	0.60	0.56
N=47535	Rekenen	0.32	0.91	0.32	0.95	0.60	1.00	0.62
	Taalvaardigheid	0.02	1.02	0.02	1.02	0.56	0.62	1.00
ROUTE8	Lezen	-0.43	0.95	-0.45	0.81	1.00	0.56	0.45
N=20583	Rekenen	-0.34	0.89	-0.32	0.83	0.56	1.00	0.51
	Taalvaardigheid	-0.21	0.93	-0.24	0.76	0.45	0.51	1.00