SCENARIO'S VOOR IJKING VAN DE EINDTOETSEN OP DE REFERENTIENIVEAUS

December 2016

C.A.W. Glas, W.H.M. Emons, P.K. Berding-Oldersma

Inhoudsopgave

Inhoudsopgave 2
Samenvatting 3
1. Aanleiding 4
2. Ankering aan referentieniveaus en normering van de toetsadviezen 5
3. Analyse van de huidige methoden voor het ankeren van de eindtoetsen en de totstandkoming van de cesuren 6
4. Verschillen tussen de aanbieders 8
5. Acties en mogelijke scenario's voor normhandhaving van referentiecesuren 10
6. Literatuurlijst 14

Samenvatting

De wetswijziging 'Eindtoetsing PO' (Stb. 2014, 13) verplicht basisscholen om vanaf het schooljaar 2014/2015 bij alle leerlingen in groep 8 een eindtoets af te nemen met als minimale eis dat de toets taal- en rekenonderdelen bevat die gerelateerd zijn aan de verplichte referentieniveaus. Bij de afname in 2016 is gebleken dat de drie tot dan toe toegelaten eindtoetsen (de Centrale Eindtoets, de Route 8 Eindtoets en IEP Eindtoets) onderling verschillen in de resultaten met betrekking tot behaalde referentieniveaus (zie, Emons, et al, 2016). Ook voor de afname van 2015 waren sterke aanwijzingen dat de eindtoetsen onderling verschilden, de afnamedata waren echter van dusdanige grootte dat verdere analyse niet mogelijk bleek. Het ministerie OCW verzocht de Expertgroep om in nauw overleg met alle toetsaanbieders met een toegelaten eindtoets van schooljaar 2017/2018 enkele scenario's te ontwikkelen, aan de hand waarvan de toetsaanbieders hun eindtoets kunnen relateren aan de referentieniveaus, op een zodanige manier dat het voor de mate van beheersing van de referentieniveaus niet uitmaakt welke eindtoets een leerling maakt.

In een werkgroep hebben de toetsaanbieders en de Expertgroep de verschillen in werkwijze van de ankering besproken en aan de hand van deze informatie enkele aanbevelingen geformuleerd die op korte en lange termijn zullen moeten leiden tot een verbetering van de ankering en vergelijkbaarheid van de eindtoetsen.

Er wordt aanbevolen een standaardformulier voor de psychometrische verantwoording te ontwikkelen, waarbij ook de gebruikte statistische methoden en software worden gestandaardiseerd. Verder wordt een vervolgonderzoek aanbevolen, waarbij alle beschikbare data met betrekking tot de reeds uitgevoerde ankering ter beschikking wordt gesteld aan de Expertgroep, waarmee de toetsen op één schaal gekalibreerd worden. Daarnaast zal in dit onderzoek door middel van statistische simulatie de nauwkeurigheid van de cesuren worden onderzocht op basis waarvan aanbevelingen ten aanzien van de aantallen items en indicaties voor de minimaal gewenste nauwkeurigheid kunnen worden gegeven. Deze aanbevelingen zullen op korte termijn leiden tot adviezen aan de toetsaanbieders die verbetering van de ankering en vergelijkbaarheid tot resultaat hebben.
Als laatste wordt aanbevolen een pilot te starten om te komen tot een proces waarbij gewerkt wordt met een gezamenlijk anker in de operationele toetsen. Door te werken met een gezamenlijk anker kan de normering van de eindtoetsen verbeterd worden.

1. Aanleiding

De eindtoets die een basisschool gaat gebruiken, moet van goede kwaliteit zijn. Daarom is door de minister een onafhankelijke commissie (de Expertgroep Toetsen PO) aangewezen die de minister adviseert over het toelaten van 'andere' toetsen als eindtoets voor het primair onderwijs (PO). De term 'andere' heeft betrekking op het feit dat een onderdeel van de genoemde wetswijziging is dat de overheid de publieke taak heeft om een centrale eindtoets beschikbaar te stellen. Het College voor Toets en Examens (CvTE) is belast met deze taak. In opdracht van het CvTE wordt door stichting Cito de Centrale Eindtoets (CET) ontwikkeld.

In 2014 zijn er, op basis van een positief advies van de Expertgroep Toetsen PO (hierna Expertgroep) twee andere eindtoetsen in het PO toegelaten, dit betreft de Route 8 Eindtoets, ontwikkeld en uitgegeven door A-VISION B.V., en de IEP Eindtoets, ontwikkeld en uitgegeven door Bureau ICE. Beide eindtoetsen zijn in 2015 opnieuw beoordeeld, waaruit opnieuw een positief advies volgde, zodat beide ook in 2016 konden worden afgenomen. Daarnaast zijn voor 2017 de CESAN Eindtoets (van SM&C), de Dia Eindtoets (van Diataal) en de AMN Eindtoets (van AMN) toegelaten.

Bij een globale bestudering van de afnamedata van 2015 en 2016 zijn sterke aanwijzingen gevonden dat de resultaten met betrekking tot de referentieniveau's en de toetsadviezen voor vervolgonderwijs op de drie eindtoetsen(de Centrale Eindtoets, de Route 8 Eindtoets en de IEP Eindtoets) onderling verschillen. De verschillen doen zich op twee manieren voor. Ten eerste blijkt dat de eindtoetsen verschillen in de mate waarin het 'toetsadvies' overeenkomt met het advies van de basisschool. Ten tweede verschillen de resultaten tussen de eindtoetsen afgemeten aan de beheersing van de referentieniveaus, d.w.z. de percentages leerlingen die een bepaald referentieniveau behalen lopen erg uiteen tussen de verschillende toetsen. Echter, de wijze waarop eindtoetsen zijn geijkt aan de referentieniveaus verschilt ook per toets. Het is daardoor op dit moment onduidelijk of de gevonden verschillen zijn toe te schrijven aan daadwerkelijke niveauverschillen van de populaties van de verschillende eindtoetsen, dan wel aan verschillen in normering tussen de eindtoetsen zelf.

De bovengenoemde verschillen tussen de eindtoetsen waren voor het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) aanleiding om aan de Expertgroep Toetsen PO twee extra, doch samenhangende opdrachten te geven:

OCW heeft de Expertgroep ten eerste verzocht om de geconstateerde niveauverschillen tussen de eindtoetsen in 2016 nader te analyseren conform een nauwkeurig omschreven onderzoeksopdracht op basis van de afnamegegevens van 2016.

Met betrekking tot het tweede punt, de ijking van de eindtoetsen op de referentieniveaus verzoekt OCW de Expertgroep om in nauw overleg met alle zes toetsaanbieders voor 2017 enkele scenario's te ontwikkelen, aan de hand waarvan de toetsaanbieders hun eindtoets kunnen relateren aan de referentieniveaus, op een zodanige manier dat het voor de mate van beheersing van de referentieniveaus niet uitmaakt welke eindtoets een leerling maakt. Ook hierbij is het verzoek om

OCW te adviseren welk scenario gevolgd zou moeten worden, om ervoor te zorgen dat de eindtoetsen in 2017, en daarna, beter en meer eenduidig zijn geijkt op de referentieniveaus.

Het voorliggende rapport bevat de met de toetsaanbieders ontwikkelde scenario's voor het ankeren van de eindtoetsen aan de hand van de referentieniveaus.

2. Ankering aan referentieniveaus en normering van de toetsadviezen

De eindtoetsen dienen een uitspraak te doen over het voor de leerling best passend vervolgonderwijs (het zogenoemde 'toetsadvies') en een uitspraak over de beheersing van de referentieniveaus Nederlandse taal en rekenen. Om een valide uitspraak te kunnen doen over het best passend vervolgonderwijs en de beheersing van de referentieniveaus moeten de eindtoetsen geankerd en genormeerd worden.

Uitspraak over best passende vervolgonderwijs

De meeste aanbieders normeren hun toetsadviezen via de schooladviezen zoals ze gegeven worden door leerkrachten in een hiërarchische steekproef van leerlingen binnen scholen. De populatie is daarbij formeel gedefinieerd als alle leerlingen in Nederland (in het vervolg de algemene populatie genoemd), maar in de praktijk komt het vaak neer op een steekproef uit de eigen doelpopulatie van de toets (hierna de toetspopulatie genoemd). Toetspopulaties kunnen mogelijk systematisch afwijken van de algemene populatie van groep 8 leerlingen. Formeel is dat een probleem voor de normering, hoewel dat relatief is. In de eerste plaats kan men via statistische weging tot op zekere hoogte corrigeren voor verschillen in de samenstelling van de toetspopulaties, mits er voldoende achtergrondvariabelen beschikbaar zijn om die correctie te ondersteunen. In de tweede plaats is het op dit ogenblik niet waarschijnlijk dat er in de verschillende toetspopulaties van de verschillende eindtoetsen door leerkrachten wezenlijk verschillende schooladviezen gegeven worden.

Op de lange termijn kan het echter zo zijn dat de normering van eindtoetsen en de oordelen van leerkrachten in een toets-populatie steeds verder naar elkaar toe trekken: leerkrachten kennen de schooladviezen en toetsadviezen uit voorgaande jaren en passen hun eigen schooladviezen daar (eventueel) onbewust op aan. Mogelijk ook om een groot aantal heroverwegingen te voorkomen. Als een leerkracht bijvoorbeeld met zijn advies vaak te hoog zit, dan zou de leerkracht in het vervolg gemiddeld genomen lagere adviezen kunnen geven; dit heeft dan weer consequenties voor de normering van de adviezen in het jaar er na. Door dit soort ongewenste processen kan er scheefgroei tussen de eindtoetsen ontstaan, omdat sommige toetsen dan voor vergelijkbare leerlingen systematisch hogere schooladviezen geven dan anderen.

Uitspraak over de beheersing van de referentieniveaus

In 2010 is de *Wet referentieniveaus Nederlandse taal en rekenen* van kracht geworden. In deze wet worden de referentieniveaus voor Nederlandse taal en rekenen omschreven en wordt vastgesteld welke inhouden bij de verschillende referentieniveaus beheerst moeten worden. De referentieniveaus zijn vervolgens onder verantwoording van het CvTE door stichting Cito geoperationaliseerd en omgezet in referentiesets. De referentiesets zijn een verzameling items die tezamen de operationalisering van de referentieniveaus vormen met als doel een vergelijking tussen de niveaus van verschillende toetsen en verschillende onderwijsniveaus mogelijk te maken.

Aanbieders van eindtoetsen gebruiken in eerste instantie de referentiesets om hun toetsen te ankeren aan de referentieniveaus. Omdat op deze referentiesets de cesuur voor beheersing van de referentieniveaus is vastgesteld, kan worden bepaald in hoeverre een leerling de referentieniveaus beheerst. Omdat de referentiesets openbaar zijn, wordt aangeraden na deze eerste ankering in de daaropvolgende jaren de referentiecesuren over te brengen via de eindtoetsen en niet meer via afname van een ankerset uit de referentiesets. Wanneer deze opgaven als oefenmateriaal worden gebruikt, zal hergebruik tot vertekening leiden. De kans dat de openbare referentiesets voor oefenen worden gebruikt, wordt echter erg laag ingeschat. Daarom kiezen sommige aanbieders er voor om de referentiesets te hergebruiken.

3. Analyse van de huidige methoden voor het ankeren van de eindtoetsen en de totstandkoming van de cesuren

Het beoordelingskader van de Expertgroep en het COTAN beoordelingssysteem (Evers et al. 2010) formuleren een aantal aspecten waarop de toetsverantwoordingen van de aanbieders worden beoordeeld, maar de criteria waaraan een toetsverantwoording moet voldoen zijn erg ruim geformuleerd en bieden de aanbieders erg veel ruimte voor een eigen invulling van de ijking van hun eindtoetsen aan de referentieniveaus en de normering van de toetsadviezen. Hoewel de eindtoetsen die nu zijn toegelaten stuk voor stuk aan de criteria voldoen, blijken de verschillen dusdanig te zijn, dat meer regie nodig is. Het gaat hierbij om de psychometrische aspecten van de eindtoetsen, het gaat niet om de inhoudelijke en onderwijskundige oriëntatie van de verschillende eindtoetsen. Wanneer de inhoudelijke en onderwijskundige oriëntatie geüniformeerd zou worden, zou er voor de scholen geen reële keuzemogelijkheid meer zijn.

Voor het ankeren van een eindtoets aan de referentieniveaus moet een toetsaanbieder een ankeronderzoek uitvoeren. Voor dit ankeronderzoek is in 2014 een handleiding geschreven (Wools & Beguin, 2014). De richtlijnen in de handleiding zijn niet bindend, het zijn aanwijzingen om tot een betrouwbare ankering te komen. Om een eindtoets te ankeren aan de referentieniveaus moeten de aanbieders per domein (d.w.z. Taal-Lezen, Taal-Taalverzorging en Rekenen) en per niveau (1F, 2F en 1S), een anker uit de referentieset samenstellen. De richtlijn is dat het anker minimaal 20 opgaven bevat. De geselecteerde opgaven moeten representatief zijn met betrekking tot de sub-domeinen,

opgaventype en afnamesituaties. Verder geldt dat de sub-domeinen, opgaventypen en afnamesituaties moeten passen bij de te ankeren toets. Om opgaven te kunnen ankeren aan de centraal verzamelde data is het nodig dat er gegevens verzameld zijn van minimaal 200 leerlingen (als het Rasch model als psychometrisch model voor de ankering gebruikt wordt) tot 400 leerlingen (als het OPLM of het 2PLM als psychometrisch model gebruikt wordt). De data moeten worden verzameld in een steekproef die representatief is voor de doelpopulatie van de eindtoets. Om aannemelijk te maken dat de verzamelde data tot een stabiele ankering leidt, verdient het de aanbeveling de meetfout van de ankering te bepalen. Dat wil zeggen dat gerapporteerd moet worden hoe nauwkeurig de gekozen cesuur is bepaald.

Bij deze (niet bindende) richtlijnen worden de volgende opmerkingen geplaatst:

1) Er is geen documentatie voor handen waarin wordt aangetoond dat de keuze van 20 opgaven en 200-400 leerlingen voldoende is voor een nauwkeurige vaststelling van referentieniveaus. De richtlijn is ontleend aan een richtlijn uit het COTAN beoordelingssysteem (Evers, et al., 2010) die betrekking heeft op het schatten van item response modellen in het geval van een integraal afgenomen toets. Het toepassen van deze richtlijn op een ankering aan referentiesets is niet onderbouwd, zeker niet in relatie tot de in deze situatie mogelijke data-verzamelingsdesigns (intern anker, extern anker, of een combinatie van die twee).

2) Voor het ankeren van de opgaven aan de centraal verzamelde data kunnen verschillende psychometrische modellen gebruikt worden, hiervoor gebruikt men het eenvoudige Rasch model of een complexere model zoals het OPLM of het 2PLM. Het 2PLM heeft van deze modellen de voorkeur omdat dit model beter bij onderwijsdata past. Wat echter opvalt, is dat, volgens de voorschriften in de Handleiding voor het gebruik van de referentiesets (Wools & Beguin, 2014) voor het Rasch model gegevens moeten worden verzameld van minimaal 200 leerlingen per opgave, maar dat er bij het gebruik van het OPLM of het 2PLM gegevens van 400 leerlingen per opgave verzameld moeten worden. Deze richtlijn is tevens ontleend aan het COTAN beoordelingssysteem (Evers, et al., 2010). Hierdoor worden de aanbieders die het nauwkeuriger OPLM of 2PLM gebruiken dus benadeeld omdat ze meer observaties nodig hebben.

3) In de Handleiding (Wools & Beguin, 2014) wordt aangegeven dat de onzekerheid van de ankering kan worden bepaald door het betrouwbaarheidsinterval van de equivalente cesuur te bepalen. Geen van de zes toetsaanbieders (CvTE en aanbieders van 'andere' eindtoetsen) verstrekt echter informatie met betrekking tot de nauwkeurigheid van de schatting van de cesuur per referentieniveau. De nauwkeurigheid van de cesuur is afhankelijk van het aantal gebruikte opgaven, leerlingen en de relatie tussen itemparameters en populatieparameters. Met dat laatste wordt bedoeld dat de gebruikte items voldoende informatief moeten zijn rondom de cesuur, dus niet te moeilijk en niet te gemakkelijk. Hoewel het aantal benodigde opgaven en het aabtak leerlingen nog moeten worden vastgesteld, is uit onderzoek (Kolen en Brennan, 2004) al wel duidelijk dat het aantal leerlingen voor een betrouwbare cesuur minimaal 1000 moet zijn. Als er minder dan 1000 leerlingen beschikbaar zijn, is de ruis zo groot dat er geen uitspraak kan worden gedaan over de locatie van de cesuur. We gaan er dan ook vanuit dat alle eindtoetsen betrekking

hebben op meer dan 1000 leerlingen. Overigens wordt hier niet de betrouwbaarheid van classificatie van leerlingen rond de cesuur bedoeld. Deze informatie wordt door de aanbieders wel gegeven; informatie over de nauwkeurigheid van de overgebrachte cesuur als zodanig echter niet.

4) Toetsaanbieders ankeren hun toetsen de eerste keer aan de referentiesets, maar na deze eerste ankering worden in de volgende jaren de referentiecesuren door de meeste aanbieders overgebracht via een koppeling aan een vorige toetsafname. Sommigen kiezen ervoor de referentiesets nogmaals te gebruiken. Voor deze latere koppelingen zijn nauwelijks eisen gedefinieerd en/of onderzocht. Hierdoor is onduidelijk op welke wijze deze koppeling moet worden gerealiseerd en hoe deze beoordeeld moet worden. Hierdoor is de kans groot dat de nauwkeurigheid van de cesuren over de jaren sterk afneemt: bij de eerste koppeling ontstaat er al ruis (c.q. een schattingsfout), deze ruis wordt in de daaropvolgende jaren alleen maar groter. Doordat opeenvolgende cesuren zijn bepaald op basis van onbetrouwbare bepaalde cesuren op eerdere eindtoetsen cumuleert de ruis niet alleen, maar er kunnen uiteindelijk ook systematische vertekeningen gaan ontstaan. Dit geldt in nog sterkere mate voor de vergelijkbaarheid van de cesuren tussen de eindtoetsen van de verschillende aanbieders. Binnen een eindtoets is er sprake van het ontstaan van ruis op de cesuren door matige ankering, tussen de toetsen is nauwelijks meer sprake van ankering en wordt de ruis op de cesuren inherent nog groter.

5) Na het Head Start project (project waarbij de toetsaanbieders de mogelijkheid hadden hun toets te ankeren aan de tot dan toe nog niet publiek toegankelijke referentiesets) zijn de referentiesets openbaar gemaakt. Als het zo zou zijn dat deze opgaven in het onderwijs als oefenmateriaal gebruikt worden, kan dit hergebruik van deze set opgaven voor ankering tot vertekening leiden. De kans dat referentiesets voor oefenen worden gebruikt, wordt echter erg laag ingeschat. Het probleem is dat hier niet echt goed zicht op is.

4. Verschillen tussen de aanbieders

Bij de bestudering van de wetenschappelijke verantwoordingen van de zes verschillende aanbieders blijkt dat er veel verschillen tussen de aanbieders zijn in de werkwijze waarop de cesuren voor referentieniveaus zijn bepaald, de achterliggende inhoudelijke overwegingen, en de rapportages daarover. Het is belangrijk op te merken dat het hier niet gaat om tekortkomingen. De richtlijnen zijn ruim en deels vrijblijvend en iedere aanbieder heeft binnen de kaders van de richtlijnen keuzes gemaakt die het best passen bij hun eindtoetsen. Dit heeft echter geleid tot verschillen die de vergelijkbaarheid van de eindtoetsen niet bevorderd heeft. In het volgende hoofdstuk worden voorstellen gedaan om de verschillen te beperken en de vergelijkbaarheid te bevorderen. Hieronder volgt eerst een weergave van de huidige situatie.

Bij de ontwikkeling van de eindtoetsen gaan de aanbieders uit van hun eigen inzichten, opvattingen ideeën over toetsen en onderwijs, dit resulteert in eindtoetsen die onderling verschillen in opzet,

inhoud en opmaak. Er zijn toetsen die digitaal (via computer, laptop of tablet) afgenomen moeten worden. Bij sommige scholen kan dit klassikaal, maar meestal gebeurt dit in kleine groepen. Andere toetsen worden op papier afgenomen, welke geheel klassikaal worden afgenomen. Er zijn adaptieve toetsen, de items die leerlingen aangeboden krijgen worden geselecteerd op basis van eerder beantwoorde items, hierdoor maken leerlingen een toets op hun eigen niveau. Er zijn lineaire toetsen, alle items zijn voor alle leerlingen gelijk en worden door alle leerlingen gemaakt. Enkele toetsaanbieders hebben het besluit genomen enkel de onderdelen Taal en Rekenen te toetsen, terwijl anderen er nog andere onderdelen aan toevoegen. Aanbieders maken daarnaast ook keuzes in de opmaak van de toets en items. De diversiteit in de eindtoetsen die hierdoor is ontstaan, wordt gewaardeerd door de scholen en nagestreefd door OCW, maar dit bemoeilijkt de vergelijkbaarheid van de toetsen omdat de toetsaanbieders ieder eigen en verschillende afwegingen moeten maken in het ankeronderzoek, pre-test en het psychometrische onderzoek.
Omdat het van belang is dat de ankeropgaven die gekozen worden uit de referentiesets zo goed mogelijk passen bij de toets die geankerd moet worden, zijn de ankersets voor alle toetsaanbieders anders geworden. Daarbij is er ook nog een interpretatieverschil bij de dekking van de (sub)domeinen in het anker. Bij de ontwikkeling van de ankersets is geen bilateraal overleg tussen de toetsaanbieders geweest, waardoor de overlap tussen de gebruikte ankersets klein is.

Voor het daadwerkelijke ankeronderzoek hebben enkele toetsaanbieders meegedaan aan het Head Start project, waarbij zij de op dat moment niet openbare referentieset konden gebruiken voor de ankering. Andere toetsaanbieders hebben niet meegedaan met het Head Start project, waardoor zij de reeds openbare referentiesets moesten gebruiken voor de ankering.
Verder worden de toetsitems getest in een zogenaamde pre-test, in deze pre-test wordt gekeken hoe de items zich psychometrisch gedragen. Op basis van de gegevens uit de pre-test wordt een beslissing genomen over de items die in de operationele toets gaan komen. Verder worden de ankeritems ook meegenomen in deze pre-test, zodat op basis van deze gegevens de cesuur van de toets vastgesteld kan worden. De pre-test vindt meestal plaats in een ‘low-stakes’ afnameconditie, dit houdt in dat de resultaten geen consequenties hebben voor de leerlingen, hierdoor is hun motivatie waarschijnlijk lager dan in een ‘high-stakes’ afnameconditie. De operationele toets is een ‘high-stakes’ afname, immers het resultaat heeft voor de leerling consequenties en daardoor zal de leerling waarschijnlijk beter zijn best doen. De gegevens die verzameld worden in de pre-test kunnen vertekend worden door het verschil in afname-conditie. Voor de toetsaanbieders zijn er wel mogelijkheden om de ankeritems in een ‘high-stakes’ afname mee te laten doen, bijvoorbeeld door deze toe te voegen aan de operationele toets (dit wordt ook wel zaaien genoemd), maar niet te laten meetellen bij de bepaling van het toetsadvies. Deze mogelijkheid wordt op dit moment weinig gebruikt, maar zou wel bij de aanbevelingen een rol kunnen spelen.

Naast het ontwikkelen van een toets en het samenstellen van een ankerset moeten de toetsaanbieders psychometrisch onderzoek doen. Bij dit psychometrische onderzoek worden door de toetsaanbieders binnen de aangeboden kaders en met in achtneming van eigen opvattingen en expertise keuzes gemaakt over de inrichting van de pre-test, gebruikte psychometrische procedures, wijze van steekproeftrekking en wijze van de dataverzameling. Hoewel de gebruikte procedures

erkende procedures zijn, hebben ze allen voor- en nadelen. Het gebruik van verschillende psychometrische procedures werkt de vergelijkbaarheid van de toetsen tegen.

Als laatste zijn ook de psychometrische verantwoordingen zeer divers, voor de toetsaanbieders is het soms onduidelijk welke informatie opgenomen moet worden in de verantwoording. Hierdoor gebeurt het bij de beoordeling vaak dat er nog extra informatie, die vaak wel al voor handen is, moet worden aangeleverd. De diversiteit van de verantwoordingen bemoeilijkt de beoordelingen.

5. Acties en mogelijke scenario's voor normhandhaving van referentiecesuren

Hierboven is een aantal kritische opmerkingen bij de richtlijnen voor de huidige procedure geplaatst en een aantal verschillen tussen de gevolgde procedures van de aanbieders geschetst. De huidige procedure kan er toe leiden dat de referentiecesuren van de verschillende aanbieders, maar ook de cesuren van de toetsadviezen van de verschillende aanbieders, op den duur onaanvaardbaar uiteen gaan lopen. Het in opdracht van OCW parallel aan dit rapport geschreven rapport over de niveauverschillen tussen de eindtoetsen in 2016 geeft duidelijk aanwijzingen in die richting (Emons et al., 2016). Daarom worden hieronder acties voorgesteld om het proces van ankering aan de referentieniveaus en normering van de toetsadviezen voor vervolgonderwijs beter in de hand te kunnen houden. Overigens heeft de Expertgroep alleen een formele relatie met de 'andere' eindtoetsen. Uit raadpleging van het CvTE en stichting Cito blijkt dat ook zij positief tegenover de voorstellen staan.

1. In samenwerking met de COTAN, beoordeelt de Expertgroep de 'andere' eindtoetsen op inhoudelijk/onderwijskundige aspecten, zoals de toetsmatrijs en de kwaliteit van de items, en op psychometrische aspecten. Het beoordelingskader van de Expertgroep en het beoordelingssysteem van de COTAN schrijven weliswaar criteria voor waaraan de verslaglegging moet voldoen, maar deze voorschriften geven in de praktijk veel ruimte voor variatie. Standaardisering, uiteraard conform de genoemde voorschriften van de COTAN, kan het proces overzichtelijker en beter beheersbaar maken. Zowel met betrekking tot verschillen tussen toetsen als met betrekking tot de ontwikkeling over de jaren. Voor de psychometrische verantwoording moet een standaard sjabloon worden ontwikkeld, met een logische opeenvolging van onderwerpen en benodigde tabellen. Verder verdient het aanbeveling om gebruik te maken van één en hetzelfde psychometrische model en het daarbij behorende statistische pakket, zodat uitkomsten van verschillende aanbieders beter vergelijkbaar worden. De standaardisering omvat de volgende elementen:
    - Uniforme opbouw van de psychometrische verantwoording:
        1. Normering schooladvies
            1.1. Design dataverzameling

1.2. Steekproef
1.3. Analyse
1.4. Scoring, gewichten domeinen
1.5. Betrouwbaarheid globaal over de hele doelgroep
1.6. Betrouwbaarheid lokaal voor individuele leerlingen
1.7. Betrouwbaarheid cesuren en betrouwbaarheid ankering cesuren.

2. Normering referentieniveaus
2.1. Design dataverzameling
2.2. Steekproef
2.3. Analyse
2.4. Betrouwbaarheid globaal over de hele doelgroep
2.5. Betrouwbaarheid lokaal voor individuele leerlingen
2.6. Betrouwbaarheid cesuren en betrouwbaarheid ankering cesuren

3. Functioneren als eindtoets
3.1. Frequentieverdelingen adviezen en referentieniveaus
3.2. Betrouwbaarheid lokaal
3.3. Betrouwbaarheid globaal
3.4. Relatie proefafname en operationele afname
3.5. Betrouwbaarheid cesuren en betrouwbaarheid ankering cesuren

- Uniforme statistische methoden, gebaseerd op het 2PLM.
- Uniforme statistische software.

De ontwikkeling van dit sjabloon kan al op korte termijn gerealiseerd worden, waardoor toetsaanbieders hier al in maart 2017 beschikking kunnen hebben.

N.B.: Voor de verantwoording van de inhoudelijke en onderwijskundige aspecten wordt geen standaardformat ontwikkeld. De reden hiervoor is dat dit te zeer zou raken aan de vrijheid van de aanbieders om hun eigen eindtoets inhoudelijk vorm te geven. Teveel sturing op dit gebied zou de door de wetgever bedoelde keuzevrijheid van de scholen teveel beperken.

2. Hierboven is beargumenteerd dat er onvoldoende inzicht is in de vergelijkbaarheid en nauwkeurigheid van de referentiecesuren en toetsadviezen van de verschillende aanbieders. Het in opdracht van OCW parallel aan dit rapport geschreven rapport over de niveauverschillen tussen de eindtoetsen geeft wel indicaties, maar geen aanknopingspunten op toets- en item-niveau. Om meer inzicht te krijgen in de oorzaken van de verschillen, en om goed gemotiveerde bijsturing te realiseren, moeten alle toetsen op een gezamenlijke schaal worden gekalibreerd om na te gaan hoe groot de spreiding van de referentiecesuren en schooladviezen op dit ogenblik is. Echter, de enige koppeling tussen toetsen loopt op dit ogenblik via de referentiesets. Om meer grip te krijgen op de huidige nauwkeurigheid van de procedure van cesuurbepaling is onderzoek naar de huidige stand van zaken m.b.t. verschillen in, en de nauwkeurigheid van de referentiecesuren en schooladviezen van de verschillende aanbieders nodig. Hiervoor worden twee activiteiten uitgevoerd.

2.1. Alle aanbieders stellen hun relevante data vanaf de eerste koppeling aan de referentiesets tot en met de laatste relevante afname ter beschikking aan de Expertgroep. Het gaat hierbij om pretest data en/of (steekproeven uit) de operationele eindtoets data van de drie al operationele eindtoetsen. De Expertgroep kalibreert de data van de referentiesets en alle toetsen op een gezamenlijke schaal. Vervolgens wordt de spreiding van de cesuren en hun nauwkeurigheid berekend. De resultaten worden aan alle aanbieders ter beschikking gesteld, vergezeld van adviezen over hoe zij hun cesuren in het vervolg zo kunnen kiezen dat ze goed met de andere aanbieders in de pas blijven lopen.

2.2. Het uitvoeren van een statistische simulatie van de nauwkeurigheid van de cesuren bij verschillende designs van dataverzameling, steekproefgrootten van leerlingen, aantallen opgaven in het design en de moeilijkheidsparameters van de gekozen opgaven. In principe is het aantal mogelijke designs onuitputbaar. Daarom kunnen designs en resultaten uit de vorige stap als uitgangspunt dienen. Het doel van dit onderzoek is om aan de aanbieders verdere adviezen te geven m.b.t. hun toekomstige pretest activiteiten.

3. Een verbeterde procedure voor ankering. De huidige procedure heeft verschillende zwakke punten. In de eerste plaats is de koppeling tussen de verschillende eindtoetsen van de verschillende aanbieders erg zwak. In de huidige procedure gaan de meeste aanbieders er van uit dat een eenmalige koppeling aan de referentieniveaus via de referentiesets voldoende is en dat de koppeling daarna kan plaatsvinden door het overbrengen van de cesuren via de operationele eindtoetsen zonder dwarsverbanden met eindtoetsen van andere aanbieders. Een tweede zwak punt in de huidige procedure is het feit dat de nieuwe opgaven en de ankeritems niet in een "high-stakes" situatie worden afgenomen en dat beïnvloedt zowel de normering van de referentiecesuren als de normering van de schooladviezen. Hierboven is beargumenteerd dat dit tot een opeenstapeling van onzuiverheid en meetfouten kan leiden.

   Het voorstel is om een aantal scenario's te onderzoeken die deze problemen in meerdere of mindere mate oplossen, en om vervolgens één van die scenario's te implementeren. Overeenkomstig aan alle scenario's is dat wordt voorgesteld om te gaan werken met een anker dat uniform is voor alle aanbieders. Dit gezamenlijke anker zou moeten zijn samengesteld uit opgaven van alle aanbieders. Dat wil zeggen dat de aanbieders overeenstemming moeten bereiken over een gezamenlijk anker waar zij allen items aan bijdragen. In de tweede plaats moet het probleem van de vertekening door afnamen in 'low-stakes' situaties worden aangepakt. Idealiter zou het anker kunnen worden opgenomen in alle operationele eindtoetsen. Daardoor wordt het anker automatisch afgenomen in een "high-stakes" situatie. Een andere oplossing is om gebruik te maken van het feit dat sommige aanbieders hun items al in een "high-stakes" situatie pre-testen door ze in een operationele toets te "zaaien". Door een gezamenlijk anker komt er informatie beschikbaar over conditie-effecten zodat daarvoor bij de andere aanbieders statistisch gecorrigeerd kan worden.

   Onduidelijkheid bestaat nog over de vraag of de eindtoetsen ieder jaar 100% nieuw moeten zijn. Als dit het geval is, kan een gezamenlijke anker in een operationele eindtoets steeds maar één jaar gebruikt worden. De koppeling over de jaren heen moet dan dus in de pretest van iedere aanbieder plaatsvinden. De pretest behelst dan dus oude en nieuwe eindtoetsitems, inclusief de

anker items van de vorige en nieuwe eindtoets. Gegeven deze overwegingen kunnen de volgende scenario's op haalbaarheid onderzocht worden.

3.1. Scenario 1: Ankeren via pretesten waarbij het gezamenlijke anker vooraf is samengesteld. Dit betekent dat het gezamenlijke anker in de pretest van iedere aanbieder wordt meegenomen. Daarna kan ook een gezamenlijke set vergelijkbare voorlopige cesuren voor referentieniveaus en schooladviezen voor alle eindtoetsen worden vastgesteld. De Expertgroep kan hiervoor een software applicatie aanbieden, c.q. aanbevelen. Na afname van de operationele eindtoetsen kan worden geëvalueerd of er op basis van de afnamegegevens een bijstelling van de cesuren voor referentieniveaus en schooladviezen nodig is. Dit is, in principe, alleen relevant voor de aanbieders die niet in een "high-stakes" situatie pretesten. Bijstelling is eenvoudig te doen door de resultaten op het anker van de pretest en de operationele eindtoets te vergelijken.

3.2. Scenario 2: Gezamenlijk anker zowel in pretests als in operationele eindtoetsen, waarbij het anker na de pretesten van de verschillende aanbieders maar voor de operationele afname is samengesteld. Dus als Scenario 1, maar hier pretesten alle aanbieders apart hun nieuwe items. Het voordeel hiervan is dat uit de pretest de best functionerende items voor het gezamenlijke anker gekozen kunnen worden. Er is echter ook een belangrijk nadeel: de items uit het gezamenlijke anker zijn nog niet gezamenlijk gepretest. Immers, iedere aanbieder pretest alleen zijn eigen items. Dit betekent dat het gezamenlijke anker ook in de operationele eindtoetsen opgenomen moet zijn en de normering pas onmiddellijk na de afname van alle eindtoetsen kan plaatsvinden. Dit kan logistiek problematisch zijn. Echter, voor de psychometrisch ondersteuning kan de Expertgroep zorgen door een applicatie beschikbaar te hebben die onmiddellijk na de afnamen een cesuur vaststelt.

3.3. Scenario 3. Gezamenlijk anker in de operationele eindtoetsen. In principe wordt het gezamenlijke anker niet meer gepretest. De normering vind plaats als in scenario 2, dus onmiddellijk na de afname van de operationele eindtoetsen via een applicatie van de Expertgroep. Omdat de eindtoetsen ook over de jaren moeten worden geijkt (c.q. normhandhaving), is dit alleen mogelijk als de regel van 100% nieuwe items in een eindtoets wordt verlaten. Het probleem van het bekend worden van de items kan worden opgelost door te werken met een tweejarige cyclus, waarbij steeds de helft van de items in het anker wordt ververst. Na de afname is dan niet bekend welke items het anker vormden, en de items komen maar één keer terug.

De eerste twee actiepunten, i.e., uniforme methoden en verslaglegging, en een onderzoek om de nauwkeurigheid van de cesuren vast te stellen en te optimaliseren, kunnen op relatief korte termijn worden uitgevoerd en kunnen al voor de afname van 2017 effect hebben. Het derde punt behelst een bijsturing van de procedure op langere termijn en zal naar verwachting het grootste effect sorteren. Uit de commentaren van de toetsaanbieders wordt duidelijk dat het invoeren van een van deze scenario's voor ieder specifieke problemen oplevert en dat invoering van deze scenario's op heel korte termijn niet mogelijk is. Bij sommige aanbieders zijn de boekjes voor de pretest voor de

2018 editie al gedrukt. Daarom wordt aanbevolen om de procedure, c.q. een van de drie scenario's, als een pilot project uit te voeren in de pretest van 2018 voor de eindtoets van 2019. Op basis van de resultaten kan dan worden beslist of deze procedure de standaard wordt voor het normeren van referentieniveaus en schooladviezen. Op korte termijn kan al wel een begin worden gemaakt met het ontwikkelen van een gezamenlijk anker en het ontwikkelen de procedures die nodig zijn om met een gezamenlijk anker cesuren voor referentieniveaus en toetsadviezen vast te stellen.

6. Literatuurlijst

- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests.* Utrecht: Nederlands Instituut voor Psychologen.

- Emons, W.H.M., Glas, C.A.W. & Berding-Oldersma, P.K. (2016). *Rapportage Vergelijkbaarheid eindtoetsen.* Utrecht: Expertgroep PO.

- Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: methods and practices.* Springer

- 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. (2014). *Toelichting Ankeronderzoek met referentiesets*. Arnhem: Cito.