Verantwoording onderzoek werkgroep Meijer

Meijer, Rob R.; Egberink, Iris J.L.; Albers, Casper J.; Tendeiro, Jorge N.; Niessen, A. Susan M.

rijksuniversiteit groningen
faculteit gedrags- en maatschappijwetenschappen
psychometrie & statistiek

Verantwoording onderzoek werkgroep Meijer: Aanvullingen COTAN Beoordelingssysteem wat betreft normering referentieniveaus en computer adaptief toetsen van andere eindtoetsen (deel 1)

prof. dr. Rob R. Meijer
dr. Iris J. L. Egberink
dr. Casper J. Albers
dr. Jorge N. Tendeiro
A. Susan M. Niessen, MSc.

Voorwoord

Voor u ligt de verantwoording van het onderzoek dat is uitgevoerd voor het schrijven van aanvullingen op het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer, & Sijtsma, 2010) in het kader van medewerking van de COTAN aan de werkzaamheden van de Expertgroep Toetsen PO. Specifiek gaat het om aanvullingen die het mogelijk maken om (andere) eindtoetsen en/of leerling- en onderwijsvolgsystemen (lovs) te kunnen beoordelen op de toepassing van de referentieniveaus, computer adaptief toetsen en het volg-aspect.

Een eerste uitgangspunt bij het opstellen van deze aanvullingen was om zo dicht mogelijk te blijven bij de inhoud van bestaande documenten en notities. Enerzijds omdat deze documenten geschreven zijn door experts uit de psychometrie en onderwijsveld en anderzijds omdat – niet onbelangrijk – een aantal documenten dienden als leidraad voor de toetsconstructeurs. Verder wordt bij de aanvullingen uitgegaan van het principe dat de bewijslast bij de toetsconstructeur ligt. Dit is immers ook een van de uitgangspunten bij het huidige COTAN Beoordelingssysteem. Dit klinkt wat zwaar, maar het blijkt in veel gevallen niet eenvoudig om voor elke situatie geldende en passende vuistregels op te stellen vanwege de diversiteit aan mogelijkheden en beslissingen die genomen moeten worden bij de toetsconstructie en bij de uit te voeren (psychometrische) analyses. Het gaat er daarom bij dit uitgangspunt om dat de toetsconstructeur aannemelijk moet maken dat alle beslissingen die genomen zijn tijdens het ontwikkelingsproces uitgebreid beschreven en verantwoord worden middels argumentatie ondersteund door de psychometrie. Er is getracht waar mogelijk te werken met vuistregels, richtlijnen en/of uitgebreide informatie over de meest gangbare mogelijkheden en/of analysetechnieken. Aan de andere kant maakt het ook duidelijk dat naast ‘kunde’, toetsconstructie ook een ‘kunst’ is waarbij het gaat om overtuigd te worden - met valide argumenten - door de toetsconstructeur. Om met Abelson (1995) te spreken: "Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations" (p. 2).

Tenslotte, ervaring zal moeten uitwijzen in hoeverre deze aanvullingen een werkzaam geheel vormen. Waar nodig zijn in de toekomst - op basis van opgedane ervaringen en argumenten - wellicht aanpassingen nodig.

prof. dr. Rob R. Meijer
dr. Iris J. L. Egberink
dr. Casper J. Albers
dr. Jorge N. Tendeiro
A. Susan M. Niessen, MSc.

Groningen, juli 2015

Inhoudsopgave

Hoofdstuk 1 Inleiding en doel onderzoek

1.1 Inleiding

De ‘Wet centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs’, hierna ‘Wet Eindtoetsing PO’ (Stb. 2014, 13), verplicht basisscholen om vanaf het schooljaar 2014/2015 bij alle leerlingen in groep 8 een eindtoets voor Nederlandse taal en rekenen af te nemen. Hierbij gaat het om het bepalen van het eindniveau van leerlingen ten opzichte van de referentieniveaus (zie Wet Referentieniveaus Nederlandse taal en rekenen). Voor Nederlandse taal betreft het ten minste de twee domeinen Lezen en Begrippenlijst en taalverzorging, voor rekenen de vier domeinen Getallen, Verhoudingen, Meten en meetkunde, en Verbanden. De wet voorziet tevens in keuzevrijheid; scholen zouden naast de centrale eindtoets ook moeten kunnen kiezen voor een andere eindtoets. De centrale eindtoets wordt door de overheid beschikbaar gesteld. De diverse toetsaanbieders wordt de mogelijkheid geboden om andere eindtoetsen (hierna ‘eindtoetsen’ genoemd) te ontwikkelen. Tevens is door het ministerie aangegeven dat men toe wil naar adaptieve eindtoetsing.

Naast een verplichte eindtoetsafname schrijft de ‘Wet Eindtoetsing PO’ (Stb. 2014, 13) het gebruik van een leerling- en onderwijsvolgsysteem (lovs) voor elke leerling voor. Het gaat hierbij om het systematisch in kaart brengen van de leervorderingen, ook wel groei van kennis en vaardigheden van leerlingen. Scholen zijn vrij om te bepalen welke kennis en vaardigheden zij in kaart brengen.

Een voorwaarde die geldt voor zowel de eindtoetsen als het lovs die een basisschool gaan gebruiken, is dat deze van goede kwaliteit moeten zijn. Daarom is door de minister een onafhankelijke commissie ingesteld die enerzijds de minister adviseert over toelating van eindtoetsen en anderzijds een geschiktheidsoordeel uitspreekt over de toetsen van lovs-en, zodat basisscholen een goede, onderbouwde keuze kunnen maken uit het aanbod eindtoetsen en lovs-en. Deze commissie is de Expertgroep Toetsen PO.

1.2 Doel van het huidige onderzoek

Om de Expertgroep Toetsen PO in staat te stellen een goede afweging te maken over de psychometrische kwaliteit van de eindtoetsen en lovs toetsen, wordt bijgedragen door de Commissie Testaangelegenheden Nederland (afgekort tot COTAN) in de vorm van het beoordelen van dergelijke toetsen aan de hand van het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer, & Sijtsma, 2010). Hoewel het COTAN Beoordelingssysteem toegepast kan worden op zowel psychologische als onderwijskundige tests/toetsen en vragenlijsten, is er een lacune wat betreft het beoordelen van de rapportage en toepassing van de referentieniveaus en het gebruik van computer adaptief toetsen (CAT) van andere eindtoetsen alsmede het beoordelen van het ‘volg-aspect’ van lovs toetsen. Deze aspecten zijn van cruciaal belang bij het doen van een uitspraak over de psychometrische kwaliteit van dergelijke toetsen en/of toetssystemen. Het is complexe materie die enerzijds hoogwaardige psychometrische kennis vereist en waarvoor anderzijds kennis van het (complexe) veld nodig is.

De COTAN is een bestuurscommissie van het Nederlands Instituut van Psychologen (NIP). Vanuit die ‘relatie’ is het NIP opdrachtgever voor dit onderzoek. De opdracht voor de werkgroep is tweeledig, enerzijds beoordelingsvragen opstellen voor de aspecten ‘normering referentieniveaus’ en ‘computer adaptief toetsen’ van eindtoetsen en anderzijds beoordelingsvragen opstellen voor het ‘volg-aspect’ van lovs toetsen. Het is de bedoeling dat de COTAN deze beoordelingsvragen kan gebruiken als aanvulling op het huidige COTAN Beoordelingssysteem.
Dit document (deel 1) betreft de beoordelingsvragen aangaande de aspecten ‘normering referentieniveaus’ en ‘computer adaptief toetsen’ van eindtoetsen.

1.3 Werkwijze

Bij het formuleren van de aanvullende beoordelingsvragen is zoveel mogelijk gebruik gemaakt van de documenten die ook zijn geformuleerd voor de toetsaanbieders. Deze documenten zijn uitgebreid bestudeerd. Verder zijn interviews gehouden met diverse experts en stakeholders om een goed overzicht te krijgen van wat mogelijk is en waar op gelet dient te worden bij het beoordelen van dergelijke toetsen. Al deze informatie is samengenomen, zaken zijn toegevoegd en/of gespecificeerd, waarbij rekening is gehouden om het in lijn te houden met het huidige COTAN Beoordelingssysteem. Dit eerste basisstuk is besproken met een aantal COTAN leden voor een eerste feedback ronde. De hierna aangepaste versie is inclusief een aantal specifieke vragen voorgelegd aan Anton Béguin en Cees Glas, als auteurs van een aantal belangrijke documenten omtrent deze onderwerpen. Vervolgens is de werkwijze toegelicht in een bijeenkomst met toetsaanbieders op het ministerie van OCW en is de toetsaanbieders de gelegenheid geboden om onduidelijkheden aan te geven. De aangegeven onduidelijkheden en vragen zijn zo veel en zo goed als mogelijk verwerkt in een versie die op 2 juli 2015 is voorgelegd en besproken tijdens de COTAN vergadering. De zaken die daar besproken zijn en de feedback van COTAN leden die per mail in de week erna is ontvangen, zijn verwerkt tot deze voor de werkgroep definitieve versie van aanvullende beoordelingsvragen wat betreft normering referentieniveaus, computer adaptief toetsen en volg-aspect van andere eindtoetsen en/of leerling- en onderwijsvolgsystemen.

1.4 Relatie ten opzichte van het huidige COTAN Beoordelingssysteem

Hoewel in eerste instantie getracht is de aanvullende beoordelingsvragen zoveel mogelijk als verduidelijking van en aanvulling op bestaande beoordelingsvragen op te nemen in het huidige COTAN Beoordelingssysteem, is gebleken dat dit een te grote aanpassing van het systeem vergt. De aanvullende beoordelingsvragen zijn daarom als losse aanvullingen op het COTAN Beoordelingssysteem geschreven. Dit betekent dat alleen voor de groep toetsen die in het kader van de werkzaamheden van de Expertgroep Toetsen PO worden beoordeeld een extra beoordeling zal worden gegeven op de betreffende aanvullende aspecten, die de Expertgroep Toetsen PO kan helpen bij het uitbrengen van een advies aan de minister van OCW of een eigenstandig geschiktheidsoordeel.

Het is denkbaar dat de aanvullende beoordelingsvragen in de toekomst wel verweven zullen worden in een nieuwe editie van het COTAN Beoordelingssysteem, maar daarover zullen betrokken partijen tijdig worden geïnformeerd.
Overigens zijn wij ons ervan bewust dat het ontwikkelen van een toets een ingewikkeld en langdurig proces is en dat de toetsaanbieders tijdens dat proces nog niet op de hoogte waren van de aanvullende beoordelingsvragen zoals die in dit document zijn geformuleerd. Dit is voor ons een van de redenen geweest om zoveel mogelijk aan te sluiten bij informatie die voor de toetsaanbieders beschikbaar was.

1.5 Bepaling beoordeling per aspect

Net als bij het huidige COTAN Beoordelingssysteem blijft het belangrijk dat de verschillende keuzes in het ontwikkelingstraject uitgebreid beschreven en verantwoord worden, zodat de redenering gevolgd kan worden en de correctheid kan worden beoordeeld. Echter, onder andere vanwege de complexiteit van de materie op een groot aantal punten en het niet altijd even duidelijk te maken onderscheid tussen ‘voldoende’ en ‘goed’ aan de hand van concrete richtlijnen is ervoor gekozen om de aanvullende aspecten te beoordelen als ‘onvoldoende/voldoende’.
Als uitgangspunt voor het bepalen van het eindoordeel per aanvullend aspect geldt dat alle van toepassing zijnde beoordelingsvragen als ‘voldoende’ moeten zijn beoordeeld.

Het is verder goed om te benadrukken dat een COTAN beoordeling als geheel (d.w.z. inclusief beoordeling op de aanvullende aspecten) dient als advies voor de psychometrische beoordeling door de Expertgroep Toetsen PO en dat tevens via de Expertgroep Toetsen PO een onderwijskundige beoordeling wordt uitgevoerd.

1.6 Opbouw document

Hierna is voor elk aspect een hoofdstuk opgenomen, waarin na een korte uitleg een beschrijving zal worden gegeven van de gebruikte termen, gevolgd door een tabel met aanvullende beoordelingsvragen voor het betreffende aspect. Onder de tabel volgt per beoordelingsvraag een toelichting.

Hoofdstuk 2 Normering Referentieniveaus

2.1 Introductie

In het Toetsbesluit staan de volgende relevante passages over de normering van de referentieniveaus:

- "Sinds 1 augustus 2010 is de Wet referentieniveaus Nederlandse taal en rekenen van kracht (Stb. 2010, 194). In de WPO en de WEC is vastgelegd dat scholen vanaf deze datum de referentieniveaus als uitgangspunt moeten nemen bij het geven van taal- en rekenonderwijs (artikel 9, achtste lid, van de WPO en artikel 12, tiende lid, van de WEC). Ook is in de WPO en de WEC vastgelegd dat een school over iedere leerling in het laatste leerjaar objectieve en valide gegevens verzamelt waaruit blijkt welk eindniveau de leerling heeft behaald ten opzichte van de referentieniveaus. Deze gegevens worden opgenomen in het onderwijskundig rapport en maken zo onderdeel uit van de overdracht van gegevens aan het voortgezet onderwijs. De overdracht van deze inhoudelijke gegevens past in het streven naar een doorgaande leerlijn voor taal en rekenen." (p. 12-13)
- "Onder regie van het CvTE is een instrumentarium in ontwikkeling waarmee de beheersing van de referentieniveaus in eindtoetsen kan worden getoetst. Dit instrumentarium bestaat uit referentiesets en ankersets van op deze niveaus afgestemde toetsopgaven en een normering voor het vaststellen van de referentiecesuur. De cesuur stelt vast hoeveel opgaven een leerling goed beantwoord moet hebben om een bepaald niveau te beheersen. Het instrumentarium komt beschikbaar voor alle aanbieders van eindtoetsen. Hiermee kunnen zij ervoor zorgen dat hun eindtoets betrouwbaar de beheersing van de referentieniveaus meet. In artikel 3 van het Toetsbesluit PO wordt daaraan inhoud gegeven. In dit artikel is opgenomen welke informatie met de eindtoets ten minste beschikbaar moet komen over de beheersing van de referentieniveaus door elke leerling. Dit is uitgewerkt in het algemeen deel van de toetswijzer voor eindtoetsen taal en rekenen, die het CvTE vaststelt. Het algemene deel van deze toetswijzer beschrijft welke domeinen van de referentieniveaus een eindtoets in elk geval moet meten en welke ruimte er daarnaast is voor inhoudelijke keuzes. Daarbij gaat het om ten minste alle domeinen van het onderdeel rekenen: Getallen, Verhoudingen, Meten en meetkunde, en Verbanden. Voor Nederlandse taal gaat het vooralsnog tenminste om het domein Lezen en het domein Taalverzorging. Deze inhoudelijke eisen gelden voor alle eindtoetsen, dus zowel de centrale eindtoets die het CvTE aanbiedt, als eindtoetsen van andere toetsaanbieders." (p. 13)
- "De Expertgroep zal bij het beoordelen van de wijze waarop de referentieniveaus taal en rekenen zijn verwerkt in andere eindtoetsen, rekening houden met de beschikbaarheid van instrumenten waarmee toetsaanbieders hun eindtoets kunnen ijken aan de referentieniveaus. Het gaat hierbij in het bijzonder om de ankersets van toetsopgaven voor het domein taalverzorging en de bijbehorende normering voor het vaststellen van de referentiecesuur." (Toetsbesluit, p.16)

Het gaat dus enerzijds om een deugdelijke inhoudelijke dekking en anderzijds om het overbrengen van de bijbehorende prestatiestandaard ofwel referentiecesuur.

2.2 Terminologie

Onder *referentiesets* wordt een verzameling items verstaan die bedoeld zijn om prestatiestandaarden behorende bij de verschillende referentieniveaus over te brengen op de eindtoetsen voor de verschillende populaties.

Een *populatie* is bijvoorbeeld groep 8 basisschool maar kan ook zijn vijfde klas HAVO.

Er zijn twee referentiesets; een voor Nederlandse taal en een voor rekenen. De *referentiesets taal* hebben alleen betrekking op een van de twee verplichte domeinen, namelijk het domein ‘Lezen’. De *referentiesets rekenen* hebben betrekking op alle vier verplichte domeinen.

Bij de *referentieniveaus* wordt onderscheid gemaakt tussen een fundamenteel (F) niveau (= basisniveau) en een streef (S) niveau (= niveau voor leerlingen die meer aankunnen). Voor taal zijn er vier F-niveaus en voor rekenen drie. Daarnaast zijn er voor zowel taal als rekenen drie S-niveaus gedefinieerd. Een overzicht is te vinden in de tabellen 2.2 en 2.3 verderop in deze verantwoording. Bij taal zijn de streefniveaus gerelateerd aan de fundamentele niveaus, bij rekenen zijn of worden aparte streefniveaus gedefinieerd.
Wanneer een bepaald niveau niet sectoroverstijgend (d.w.z. voor meerdere onderwijstypen van toepassing) hoeft te zijn en alleen betrekking heeft op een onderwijstype, wordt gesproken van een ankerset in plaats van een referentieset. Voor de eenduidigheid wordt in deze verantwoording altijd gesproken over referentiesets.

De *prestatiestandaard*, ook wel referentiecesuur, geeft aan hoeveel items correct moeten worden beantwoord van een set items die een afspiegeling zijn van de referentieniveaus voor taal en rekenen om zodoende vast te kunnen stellen of een bepaald referentieniveau is gehaald. Bij de eerste versie van een eindtoets zal de koppeling aan de referentieniveaus gebeuren door gegevens te verzamelen bij leerlingen die zowel (een selectie van) items uit de referentiesets als (een deel van) de nieuwe (d.w.z. te ankeren c.q. te koppelen) eindtoets maken. Voor volgende versies van de toets kan vervolgens gekozen worden voor een standaardhandhavingsprocedure (zie hiervoor bijvoorbeeld Engelen & Eggen, 1993; Holland & Dorans, 2006; Kolen & Brennan, 2004) en hoeft niet opnieuw direct gekoppeld te worden aan de referentieniveaus (aanvullende criteria document, 2014). Het is belangrijk dat hierbij in de gaten wordt gehouden en aannemelijk wordt gemaakt dat de cesuur niet gaat en/of is gaan verschuiven door de jaren heen (zie hiervoor vraag 7).
De applicaties ‘BerekenGrensscore-Rekenen & BerekenGrensscore-Taal’ zijn ontwikkeld door het CvTE en Stichting Cito en beschikbaar gesteld aan de toetsaanbieders om bij afname van een selectie van de referentiesets de bijbehorende cesuur te bepalen.

Ook kan de referentiecesuur worden overgebracht door een toets te koppelen aan een bestaande door de minister toegelaten eindtoets waarin het referentieniveau is opgenomen. Het is hierbij van belang dat de normering van de referentieniveaus van die toegelaten eindtoets beschreven staat en van goede kwaliteit moet zijn. Verder geldt ook hier dat het belangrijk is dat hierbij in de gaten wordt gehouden en aannemelijk wordt gemaakt dat de cesuur niet gaat en/of is gaan verschuiven door de jaren heen (zie hiervoor vraag 7).

Ankeritems zijn items die in minimaal twee versies van een toets worden opgenomen zodat de minimaal twee onderscheiden groepen personen voor een groot deel dezelfde items hebben gemaakt. Bij een eerste versie van een nieuwe (andere) eindtoets gaat het om zowel (een selectie van) items uit de referentiesets als (een deel van) de nieuwe (d.w.z. te ankeren/koppelen) eindtoets die dienen als ankeritems (samen *anker* genoemd).
Bij volgende versies van de eindtoetsen gaat het om zowel (een selectie van) items uit de 'oude'/eerste versie van de toets als (een deel van) de nieuwe eindtoets die dienen als ankeritems.

2.3 Beoordelingsvragen m.b.t. normering referentieniveaus

In de volgende tabel wordt een overzicht gegeven van de vragen die van belang zijn bij het evalueren van de toepassing van de referentieniveaus.

Tabel 2.1. Beoordelingsvragen m.b.t. normering referentieniveaus

		onv.	vold.	
Vraag 1	Worden er gegevens verstrekt over het verband van de toets met de referentieniveaus?	1	2	
Vraag 2	Keuze van de ankeritems			
	a. Is het aantal ankeritems voldoende?	1	2	
	b. Zijn de ankeritems representatief voor de inhoud van de toets?	1	2	
	c. Zijn de ankeritems representatief voor de doelpopulatie?	1	2	
Vraag 3	Wat is de kwaliteit van een van de drie soorten afnamedesigns?			
	a. intern anker design	1	2	n.v.t.
	b. extern anker design	1	2	n.v.t.
	c. gecombineerd ankerdesign	1	2	n.v.t.

vv Tabel 2.1. Beoordelingsvragen m.b.t. normering referentieniveaus

Vraag 4	Wat is de kwaliteit van de gebruikte steekproef?			
	a. Komen de steekproeven zoals gebruikt in het design overeen met de doelgroep waarvoor de toets bedoeld is?	1	2	
	b. Zijn de steekproeven groot genoeg?	1	2	
	c. In het geval van computerafname: is aangetoond dat de items uit de referentieset op een adequate manier gekoppeld zijn aan de digitale afname?	1	2	n.v.t.
Vraag 5	Zijn de procedures op basis waarvan de analyses zijn uitgevoerd correct?			
	a. KTT methoden	1	2	n.v.t.
	b. IRT methoden	1	2	n.v.t.
Vraag 6	Worden er gegevens verstrekt over de meetnauwkeurigheid rond de plek van de cesuur?	1	2	
Vraag 7	Wordt aangetoond dat de toetsversies vergelijkbaar zijn over verschillende jaren?	1	2	

2.4 Toelichting beoordelingsvragen

Aanwijzingen bij vraag 1: "Worden er gegevens verstrekt over het verband van de toets met referentieniveaus?

In de handleiding moet een apart gedeelte worden besteed aan de beschrijving van de manier waarop items in de eindtoets gekoppeld zijn aan de referentieniveaus. Wanneer een beschrijving ontbreekt, dient deze basisvraag met een 'onvoldoende' te worden beoordeeld en hoeven de overige vragen betreffende referentieniveaus niet te worden beantwoord.

Bij deze vraag hoeft alleen aangegeven te worden óf er gegevens worden verstrekt over het verband van de toets met de referentiesets. Het gaat niet om de kwaliteit en/of uitgebreidheid van de gegevens, dit zal aan bod komen bij de vervolgvragen.

Aanwijzingen bij vraag 2: keuze van de ankeritems

Om de referentiecesuur over te brengen op de nieuwe (versie van de) eindtoets moeten de items in de toets ook voor een gedeelte bestaan uit items uit de referentieset of koppeling naar een door het ministerie toegelaten toets die rapporteert op de referentieniveaus. In het algemeen moet er een onderbouwing zijn voor de wijze waarop ankeritems zijn gekozen en hoeveel ankeritems zijn gekozen. Meer concreet is het van belang dat informatie gegeven wordt over de volgende zaken: het aantal ankeritems, de representativiteit van de ankeritems en de geschiktheid voor de doelpopulatie.

Deze eisen gelden per te hanteren set ankeritems en per subset van de items uit de toets die direct gekoppeld zijn aan de ankeritems.

Overigens zijn na de headstartperiode de referentieset items openbaar geworden. Bij gebruik van referentieset items na openbaarmaking, moet onderzocht worden in hoeverre items zich anders gedragen dan bij de referentiesetafname of tijdens de headstart en in hoeverre de items bekend zijn. Tevens mogen referentieset items niet in het eigenlijke instrument zitten.

In de tabellen 2.2 en 2.3 is te zien welke type opgaven zijn opgenomen in de verschillende beschikbare referentiesets. Specifieke informatie over aantallen is te vinden in Appendix A.

Tabel 2.2. Informatie met betrekking tot opgaven referentiesets Taal (domein Lezen)

Niveau	Doelpopulatie	Type opgave	Type tekst
1F	PO	open en MC	zakelijk (informatief, instructief, betogend) of fictioneel
1S (=2F)	PO	open en MC	zakelijk (informatief, instructief, betogend) of fictioneel
2F	VMBO, MBO123	open en MC	zakelijk (informatief, instructief, betogend) of fictioneel
2S (=3F)	VMBO, MBO123	open en MC	zakelijk (informatief, betogend)
3F	HAVO, MBO4	open en MC	zakelijk (informatief, betogend)
3S (=4F)	HAVO, MBO4	*nog niet beschikbaar*	
4F*	VWO	*nog niet beschikbaar*	

NB voor het andere verplichte domein *Begrippenlijst en taalverzorging* is de 'headstart ankersets taalverzorging' per 1 april 2015 van start gegaan. De eindtoetsen die ter beoordeling worden ingediend voor het schooljaar 2015/2016 kunnen daarom nog niet beoordeeld worden op het aspect 'normering referentieniveaus'.
* er wordt bij 4F gesproken van ankerset in plaats van referentieset omdat ze niet sectoroverstijgend hoeven te zijn; er hoeft niet vergeleken te worden tussen schooltypen.

Tabel 2.3. Informatie met betrekking tot opgaven referentiesets Rekenen

Niveau	Doelpopulatie	Type opgave (1)	Type opgave (2)	Rekenmachine
1F	PO	open en MC	kaal en context	zonder
1S*	PO	open en MC	kaal en context	zonder
2F	VMBO, MBO123	open en MC	kaal en context	met en zonder
2S	VMBO, MBO123	*nog niet beschikbaar*		
3F	HAVO,VWO, MBO4	open en MC	kaal en context	met en zonder
3S	HAVO,VWO, MBO4	*nog niet beschikbaar*		

* er wordt bij 1S gesproken van ankerset in plaats van referentieset omdat ze niet sectoroverstijgend hoeven te zijn; er hoeft niet vergeleken te worden tussen schooltypen.

Aanwijzingen bij vraag 2a: Is het aantal ankeritems voldoende?
We sluiten ons hier aan bij Wools en Béguin (2013) die stellen dat er minimaal 20 ankeritems in de toets aanwezig moeten zijn. Dit aantal geldt per domein en per niveau, dit is nodig om niveau en domein af te dekken. Voor Taal-Lezen geldt naast de voorwaarde van minimaal 20 ankeritems dat er minimaal 3 teksten voor ankering moeten worden gebruikt.
Richtlijnen voor de beoordeling: Voldoende = 20 of meer ankeritems; Onvoldoende = minder dan 20 ankeritems.
Wanneer een toets zich richt op meer dan één niveau in een enkele toetsversie (bijvoorbeeld 2F en 1F in één toets) dan geldt dat het aantal ankeritems per niveau minder kan zijn. Voor 'Voldoende' geldt dan 15 of meer items per niveau.

Aanwijzingen bij vraag 2b: Zijn de ankeritems representatief voor de inhoud van de toets?
De toetsconstructeur dient te zorgen dat de gekozen ankeritems een goede representatie geven voor de domeinen, type opgaven en situaties van afname (bijvoorbeeld met of zonder rekenmachine). De informatie uit de tabellen 2.2 en 2.3 kan gebruikt worden bij het beantwoorden van deze vraag.
Richtlijnen voor de beoordeling: Onvoldoende = niet alle domeinen, typen opgaven en situaties van afname zijn vertegenwoordigd in de ankeritems; Voldoende = alle domeinen, typen opgaven en situaties van afname zijn redelijkerwijs vertegenwoordigd in de ankeritems. Hierbij is het streven dat dit ook in verhouding is van alle items in de betreffende referentieset.
Tevens geldt dat de beoordeling automatisch 'Voldoende' is wanneer de volledige referentieset van de juiste doelpopulatie is gebruikt in het ankeronderzoek.
Bij koppeling aan een andere eindtoets waarin de referentieniveaus worden bepaald, geldt dat een 'Voldoende' beoordeling geldt als de volledige toets is afgenomen als anker, waarbij in de gaten moet worden gehouden en aannemelijk moet worden gemaakt dat de cesuur niet gaat en/of is gaan verschuiven door de jaren (cq. over toetsversies) heen (zie tevens vraag 7).

Aanwijzingen bij vraag 2c: Zijn de ankeritems representatief voor de doelpopulatie?
De ankeritems moeten representatief zijn qua niveau en inhoud voor de populatie van de toets. Dat wil zeggen dat er niet te veel makkelijke en/of moeilijke items in de toets mogen zitten, omdat anders de toets niet discrimineert voor de verschillende niveaus.

Aanwijzingen bij vraag 3: Wat is de kwaliteit van een van de drie soorten afnamedesigns?
Om de prestaties op de eindtoets te kunnen vergelijken met de prestaties op de referentieset, dat wil zeggen om een toets te kunnen ankeren, zijn verschillende afnamedesigns mogelijk; een ankerdesign met intern anker, een ankerdesign met extern anker en een gecombineerd design (aanvullende criteria document, 2014). Bij een intern design worden de ankeritems en de toetsitems tegelijkertijd aangeboden (in een afname), terwijl bij een extern design de ankeritems en de toetsitems apart (d.w.z. niet tegelijkertijd) worden afgenomen. Belangrijk is dan wel dat ervoor wordt gezorgd dat de afnamecondities (bijvoorbeeld low-stakes of high-stakes; individueel of klassikaal; met of zonder supervisie) vergelijkbaar zijn. Ook kan er

worden gekozen voor een gecombineerd design. In dit design zijn er meerdere toetsversies waarin items uit de referentiesets en de toets worden gecombineerd. Dit design is met name bruikbaar voor het ankeren van langere toetsen, aangezien het daar vaker voorkomt dat het niet mogelijk is dat dezelfde leerling zowel de gehele toets als het anker maakt.
Bij de beoordeling is het van belang om de volgende richtlijnen aan te houden: er is een duidelijke beschrijving van een van de drie designs opgenomen in de handleiding waarbij duidelijk wordt welke strategie is gekozen om tot een design te komen en waarom. Een intern ankerdesign heeft de voorkeur en als gebruik wordt gemaakt van een extern ankerdesign moet aannemelijk gemaakt worden dat er voldaan is aan de aanname van gelijke afnamecondities. Tevens moet bij alle designs rekening worden gehouden met vermoeidheids- en/of volgorde effecten.
Voor een van de drie afnamedesigns moet een beoordeling worden gegeven.

Aanwijzingen bij vraag 4: Wat is de kwaliteit van de gebruikte steekproef?
Wat betreft de steekproef zijn twee zaken van belang: de steekproef moet representatief zijn voor de doelpopulatie en de grootte van de steekproef.

Aanwijzingen bij vraag 4a: Komen de steekproeven zoals gebruikt in het design overeen met de doelgroep waarvoor de toets bedoeld is?
Wat betreft de representativiteit is het van belang dat de steekproeven in het design overeenkomen met de doelgroep waarvoor de toets is bedoeld, zoals in het huidige COTAN Beoordelingssysteem gedaan wordt bij het criterium *Normen*.

Aanwijzingen bij vraag 4b: Zijn de steekproeven groot genoeg?
Wanneer een 1-parameter logistisch model wordt gebruikt dan moet elk item bij minimaal 200 personen zijn afgenomen, wanneer het 2-parameter logistisch model wordt gebruikt moet elk item bij minimaal 400 personen zijn afgenomen. Voor de volledigheid wordt hierbij vermeld dat niet elke persoon alle items (uit de toets) hoeft te hebben ingevuld. Het gebruikte design moet daarvoor wel uitgebreid zijn beschreven.
Bij gebruik van de KTT methoden moet elk item door minimaal 300 personen zijn ingevuld.

Aanwijzingen bij vraag 4c: In het geval van computerafname: is aangetoond dat de items uit de referentieset op een adequate manier gekoppeld zijn aan de digitale afname?
Bij afname van digitale versies van de items uit de referentieset gaat het er bij deze vraag om dat beschreven wordt of er onderzoek is gedaan naar een eventueel verschil tussen afname op papier of digitaal. De itemparameters hoeven in dit geval niet identiek te zijn als ze maar op dezelfde latente trek schaal te plaatsen zijn.
Wanneer de computerafname van de eindtoets gekoppeld wordt aan een papieren anker naar de referentiesets hoeft niet aan deze eis te worden voldaan.

Aanwijzingen bij vraag 5: Zijn de procedures op basis waarvan de analyses zijn uitgevoerd correct?
Er zijn twee methoden om de referentiecesuur over te brengen naar de toets; een op basis van klassieke test theorie (KTT) methoden en een op basis van item respons theorie (IRT) methoden.
N.B. De applicaties 'BerekenGrensscore-Rekenen & BerekenGrensscore-Taal' zijn ontwikkeld door het CvTE en Stichting Cito en beschikbaar gesteld aan de toetsaanbieders om bij afname van een selectie van de referentiesets de bijbehorende cesuur te bepalen.

Aanwijzingen bij vraag 5a: KTT methoden
In de KTT betreft het de volgende drie methoden: de equipercentiel methode, de lineaire methode en de regressie methode. Voor meer specifieke informatie zie Engelen en Eggen (1993) en Kolen en Brennan (2004).
Wanneer gebruik wordt gemaakt van de regressie methode moet een volledig design gebruikt zijn, waarbij alle personen in de steekproef alle items gemaakt hebben.

Aanwijzingen bij vraag 5b: IRT methoden
Bij item respons theorie methoden wordt gebruik gemaakt van een vaardigheidsschaal of latente schaal. "In dit geval wordt de cesuur omgezet in een cesuur op de latente schaal. De score op de te ankeren toets en de referentietoets worden beide afgebeeld op dezelfde latente schaal. De cesuur op de te ankeren toets kan worden bepaald als de score die een verwachte latente schaalscore heeft die het dichtst bij de latente cesuur ligt. Wanneer gebruik gemaakt wordt van een statistisch model moet aangetoond worden dat het model voldoende passend is. In het geval van IRT modellen betekent dit dus dat er statistieken over de modelpassing worden geleverd. Bij afwijkingen in de modelpassing moet worden aangetoond dat de gebruikte modellering robuuste en verdedigbare resultaten oplevert." (Wools & Béguin, 2014). Dit betekent dat wordt aangetoond dat de fout in de resulterende cesuur die gerelateerd is aan de modelafwijking naar verwachting klein zal zijn.

Aanwijzingen bij vraag 6: Worden er gegevens verstrekt over de meetnauwkeurigheid rond de plek van de cesuur?
Het is belangrijk dat informatie gegeven wordt over zowel de globale als de lokale meetnauwkeurigheid rond de plek van de cesuur.

Aanwijzingen bij vraag 7: Wordt aangetoond dat de toetsversies vergelijkbaar zijn over verschillende jaren?
De eindtoets dient elk jaar ververst te worden, en de toets mag niet dezelfde items bevatten in opeenvolgende jaren. De Expertgroep Toetsen PO heeft laten weten dat voor eindtoetsen die op papier worden afgenomen in het schooljaar 2015-2016 en 2016-2017 geldt dat 50% van de opgaven vervangen moet zijn, daarna dient 100% van de opgaven vervangen te zijn.
Om te onderzoeken of de toetsversies vergelijkbaar zijn (zowel qua cesuur als inhoud) tussen de verschillende jaren dient onderzoek te worden gerapporteerd. Hiertoe dient een

standaardhandhavingsprocedure of normhandhavingsprocedure te worden uitgevoerd en te worden aangetoond dat de inhoudelijke domeinen nog steeds gedekt zijn.
Voor het uitvoeren van een dergelijk onderzoek moet aan dezelfde zaken gedacht worden als beschreven staan bij voorgaande vragen.
Verder is het belangrijk om vast te stellen of de gehanteerde referentiecesuur afwijkingen vertoont met de oorspronkelijke cesuur of met de cesuur zoals gehanteerd in andere toetsen (zoals de centrale eindtoets).

N.B. bij de eerste versie van een eindtoets is deze beoordelingsvraag niet van toepassing. Verder is het mogelijk dat deze vraag door de Expertgroep Toetsen PO zal worden beoordeeld tijdens de jaarlijkse check die uitgevoerd gaat worden door de Expertgroep Toetsen PO nadat een eindtoets is toegelaten door de minister voor een periode van vier jaar.

Hoofdstuk 3 Computer Adaptief Toetsen

3.1 Introductie

Bij de gangbare toetsen in het onderwijs krijgt elke leerling dezelfde toets. Een nadeel van deze procedure is dat deze toetsen soms te moeilijk zijn voor minder goede leerlingen waardoor het niet goed mogelijk is om het niveau van deze leerlingen nauwkeurig in kaart te brengen. Een oplossing is om een computer adaptieve toets af te nemen waarbij items worden geselecteerd uit een itembank die optimaal geschikt zijn voor het niveau van de kandidaat. Deze manier van toetsen heeft een aantal specifieke kenmerken die niet zijn opgenomen in het COTAN system en die hieronder worden geschetst.

Een andere vorm van adaptief toetsen die hier ook onder valt is ‘multistage testing’. Hierbij geldt dat de items goed gekalibreerd moeten zijn en dat de afhandeling adequaat moet zijn d.w.z. volledig gebaseerd op de observaties en niet op andere achtergrond variabelen, zodat het ignorability-principe van Rubin geldt. Verder blijven alle eisen m.b.t. een CAT van toepassing. Voor meer informatie, zie Lord (1980).

3.2 Terminologie

Bij een *adaptieve toets* krijgt niet iedere kandidaat eenzelfde toets maar worden items aangeboden die het beste passen bij het niveau van de leerling.

Een *itembank* is een verzameling gekalibreerde items.

Onder *kalibreren* wordt verstaan het schatten van de itemparameters van een model in een representatieve steekproef.

Bij *modelpassing* gaat het om het controleren of het psychometrisch model dat wordt gebruikt (bijvoorbeeld het 2-parameter logistisch model) een goede beschrijving geeft van de gegevens. Informatie kan worden verkregen door bijvoorbeeld statistische toetsen, maar ook kan gebruik worden gemaakt van figuren.

Linking heeft betrekking op de relatie tussen toetsscores op verschillende toetsen die niet eenzelfde toetsinhoud of moeilijkheid hebben. Om scores op deze toets met elkaar te vergelijken moet er een linking procedure worden uitgevoerd.

3.3 Beoordelingsvragen m.b.t. computer adaptief toetsen

Tabel 3.1. Beoordelingsvragen m.b.t. computer adaptief toetsen

		onv.	vold	
Vraag 1	Worden er gegevens verstrekt over de ontwikkeling van de CAT?	1	2	
Vraag 2	Is er sprake van een adequaat kalibratie design?			
	a. Voldoende observaties en samenhang design?	1	2	
	b. Is er voldoende informatie over algemene model passing?	1	2	
	c. Zijn item locatie parameters in de kalibratie steekproef en de operationele toets vergelijkbaar?	1	2	
	d. Worden er gegevens verstrekt over de vooraf verwachte en achteraf gerealiseerde betrouwbaarheid?	1	2	
Vraag 3	Worden beslisregels geëxpliciteerd om dekking van het inhoudelijke domein te waarborgen per individuele afname?	1	2	
Vraag 4	Wat is de kwaliteit van de item verversingsstrategie? Is er sprake van een adequate verversingsstrategie			
	a. voor niet geselecteerde items	1	2	
	b. voor vaak geselecteerde items	1	2	
	c. voor items met discrepanties in kalibratie design en operationele fase	1	2	
	d. voor items die uitgelekt lijken	1	2	
	voor items die over jaargangen heen anders functioneren			
	e. proportie van tenminste 1/T van de items wordt elk jaar ververst waarbij T de minimale levensduur van de itembank in jaren is	1	2	

3.4 Toelichting beoordelingsvragen

Aanwijzingen bij vraag 1: Worden er gegevens verstrekt over de ontwikkeling van de CAT?

Er dient een beschrijving te zijn opgenomen van de ontwikkeling van de CAT. Deze beschrijving dient zo uitgebreid mogelijk te zijn, waarbij keuzes en beslissingen zoveel mogelijk worden onderbouwd en verantwoord. Indien dit niet het geval is, kan deze vraag met ‘onvoldoende’ worden beantwoord en hoeven de overige vragen betreffende CAT niet te worden beantwoord.

Aanwijzingen bij vraag 2: Is er sprake van een adequaat kalibratie design?
Hierbij gaat het om verschillende aspecten, te weten het aantal observaties, het gebruikte design, de modelpassing en de schattingsmethoden.

Aanwijzingen bij vraag 2a: voldoende observaties en samenhang design?
Voordat een adaptieve toets kan worden afgenomen dienen items te worden gekalibreerd, dat wil zeggen er dienen parameters te worden geschat voor elk item. Dit is de kalibratiefase (of pre-test). Niet elke person krijgt elk item en vaak worden boekjes gebruikt die aan elkaar worden gelinkt. Belangrijk is nu dat er voldoende overlap is tussen de boekjes en dat het aantal personen dat elk item maakt van voldoende omvang is. Als richtlijn stelt het COTAN Beoordelingssysteem nu voor *lineaire* tests/toetsen en het 1-parameter model 200 personen en voor het 2-parameter model 400 personen. Belangrijk is dan wel dat deze aantallen gelden *per item* (Glas, 2015). Immers, in tegenstelling tot een lineaire test/toets krijgt niet elke persoon alle items. Er kan worden gesteld dat wanneer het 1-parameter logistisch model wordt gebruikt minimaal 200 personen elk item moet hebben gemaakt en wanneer het 2-parameter logistisch model wordt gebruikt dit 400 personen per item zijn. Hierbij dient aangegeven te worden dat dit minimale aantallen zijn en het beter is om meer gegevens te gebruiken. De standard errors van de itemparameters dienen zo klein mogelijk te zijn.

Verder moeten de items en boekjes voldoende worden gelinked, de kwaliteit van de link "kan worden bepaald door de schattingsfouten van verschillen tussen parameterschattingen van items in verschillende boekjes (adequaat) uit te rekenen" (Glas, 2015).

Aanwijzingen bij vraag 2b: Is er voldoende informatie over algemene model passing?
Bij deze vraag gaat het erom dat aannemelijk moet worden gemaakt dat het gekozen IRT model (voldoende) past bij de data. Dit kan bijvoorbeeld worden gedaan door het vergelijken van geobserveerde en verwachte item respons functies. Ook zijn er verschillende fit maten beschikbaar.

Aanwijzingen bij vraag 2c: Zijn item locatie parameters in de kalibratie steekproef en de operationele toets vergelijkbaar?
Ook dient te worden aangetoond dat de itemparameters in de kalibratiefase en de operationele fase overeenkomen. Dit kan bijvoorbeeld door een grafiek te maken van de locatie parameters in de kalibratiefase en de operationele fase.

Aanwijzingen bij vraag 2d: Worden er gegevens verstrekt over de vooraf verwachte en achteraf gerealiseerde betrouwbaarheid?
Informatie omtrent de adequaatheid van de itembank voor de verschillende deelpopulaties dient te worden beschreven. De vooraf verwachte betrouwbaarheid kan gedaan worden via benaderende formules, via de informatiewaarden in de pre-test of via simulatie. De achteraf gerealiseerde betrouwbaarheid kan inzichtelijk worden gemaakt door bijvoorbeeld

informatiewaarden van de toetsen van de leerlingen te middelen om zo tot een schatting van de gemiddelde betrouwbaarheid te komen (Glas, 2014).
In het algemeen is het van belang dat over de gehele vaardigheidsrange vastgesteld wordt met welke nauwkeurigheid individuele leerlingen worden gemeten.

Aanwijzingen bij vraag 3: Worden beslisregels geëxpliciteerd om dekking van het inhoudelijke domein te waarborgen per individuele afname?
Naast dat de beslisregels geëxpliciteerd dienen te zijn (zie vraag 2.9 van het COTAN Beoordelingssysteem), is het belangrijk dat tevens beslisregels geëxpliciteerd zijn om dekking van het inhoudelijke domein te waarborgen per individuele afname. Dit houdt in dat indien er vier verschillende domeinen van Rekenen en wiskunde worden bevraagd, de beslisregels van de adaptieve toets ervoor dienen te zorgen dat ook uit alle domeinen opgaven worden afgenomen.

Aanwijzingen bij vraag 4: Wat is de kwaliteit van de itembank verversingstrategie?
Een itembank bestaat uit een verzameling items waarvan de itemparameters bekend zijn. In tegenstelling tot een lineaire toets krijgt een leerling een samengestelde toets die afhankelijk is van het niveau van de leerling.

Hoeveel items moet een itembank bevatten wanneer deze wordt gebruikt voor high-stakes testing? In de literatuur (e.g., Parshall et al, 2002, p.141) wordt gesproken over minimaal vijf tot liefst tien keer de lengte van de gemiddelde CAT. Dus wanneer bijvoorbeeld een gemiddelde CAT bestaat uit 25 items, dan zal de itembank minimaal 125 items moeten bevatten. Veel zal ook weer afhangen van de type vragen die worden gesteld en van de soort CAT die wordt gebruikt. Bijvoorbeeld, bij begrijpend lezen zal het zo zijn dat er groepjes van items zijn die behoren bij eenzelfde tekstpassage en dus als geheel een testlet vormen.

Het idee is dat deze itembank niet elk jaar hoeft te worden ververst. Echter, bij herhaaldelijk gebruik van deze itembank kunnen items bekend raken, ook kan het zo zijn dat sommige items helemaal niet worden gebruikt. In de literatuur wordt dan ook aangeraden om controles uit oefenen om in de gaten te houden dat items niet bekend raken. In het algemeen (zie Glas, 2015) wordt aangeraden om

(a) items te verwijderen die (bijna) niet gebruikt worden. Bijvoorbeeld, sommige items blijken zo moeilijk of makkelijk in de populatie dat afname geen zin heeft. Of sommige items blijken niet te differentiëren tussen leerlingen met een verschillende vaardigheid.
(b) Items die erg vaak worden geselecteerd te identificeren en te vervangen. De kans dat deze items bekend raken is groot. Hierbij geldt ook dat items die vaak worden gekozen bij leerlingen met een specifiek vaardigheidsniveau (bijvoorbeeld heel hoog of heel laag) worden geïdentificeerd en vervangen.

(c) Items die verschillend functioneren in de kalibratiefase en de operationele fase te identificeren en te vervangen.
(d) Items te verwijderen die bekend zijn geworden. Dit kan bijvoorbeeld door item parameters te vergelijken van twee afnameperioden. Dus er dient een zekere vorm van exposure control procedure te zijn.
(e) Verder stelt Glas (2015) dat er standaard een proportie van de items moet worden ververst. Een vuistregel is dat de itemproductie niet minder mag zijn dan bij een lineaire toets. Dus als een itembank T jaren wordt gebruikt zal de proportie te verversen items ten minste $1/T$ zijn maar als de itembank niet goed functioneert moeten er meer items worden ververst.

N.B. bij de eerste versie van een eindtoets is deze beoordelingsvraag niet van toepassing. Verder is het mogelijk dat deze vraag door de Expertgroep Toetsen PO zal worden beoordeeld tijdens de jaarlijkse check die uitgevoerd gaat worden door de Expertgroep Toetsen PO nadat een eindtoets is toegelaten door de minister voor een periode van vier jaar.

Literatuur

Abelson, R. P. (1995). *Statistics as a principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie, mei 2009; gewijzigde herdruk mei 2010)*. Amsterdam: NIP.

Engelen, R. J. H., & Eggen, T. J. H. M. (1993). Equivaleren. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 309-348). Arnhem: Cito.

Geen auteur. *Aanvullende criteria COTAN voor andere eindtoetsen en leerlingvolgsystemen.* Versie 13-5-2014.

Glas, C. (2015). *Aandachtspunten voor aanvulling beoordelingssysteem COTAN*. 18 maart 2015.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 187-220). Westport, CT: Greenwood.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY: Springer.

Parshall, C.G., Spray, J. A., Kalohn. J.C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Toetsbesluit PO. *Staatsblad 2014,* 000. 3 juni 2014.

Wet centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs. *Staatsblad* 2014, 13. 16 januari 2014.

Wools, S., & Béguin, A. (2013). *Handleiding Referentiesets*. Juli 2013

Wools, S., & Béguin, A. (2014). Het toetsen van referentieniveaus: stellen we aan alle leerlingen dezelfde eisen? *Examens*, februari 2014, no 1.

Appendix A

Deze appendix bevat de pagina's 15 tot en met 20 uit het document "Rapportage Referentiesets Nederlandse taal (Lezen) en Rekenen: Verantwoording project" en bevat specifieke informatie over het aantal items per referentieset en type opgave, waarbij tevens onderscheid gemaakt wordt tussen de openbare en niet-openbare referentiesets.

4 Toetsmatrijs referentiesets

Aan de hand van een toetsmatrijs heeft Cito opgaven voor de referentiesets geselecteerd en ontwikkeld. Een toetsmatrijs geeft weer hoeveel en welk type opgaven een referentieset moet bevatten. De toetsmatrijs van elke referentieset is gebaseerd op de domeinindelingen van het betreffende referentieniveau. Het is van belang dat de geselecteerde opgaven het inhoudsdomein zo goed mogelijk dekken. De opgaven zijn daarom voorgelegd aan een expertpanel (zie voor meer informatie hoofdstuk 5 'inhoudelijke validering')

De samenstelling van de referentiesets moet passend moeten zijn voor de hoofdfunctie: het overbrengen van de referentiecesuur op verschillende toetsen. De toetsmatrijzen zijn hier ook op gebaseerd. De toetsmatrijzen van de openbare referentiesets (OS) laten een diversiteit aan opgaven en vraagvormen zien. Een OS dient immers bruikbaar te zijn voor andere toetsaanbieders ten behoeve van hun ankeronderzoek (zie paragraaf 2.2). De opgaven uit de niet-openbare referentiesets (NOS) worden gebruikt als anker in de centrale toetsen en examens. Daarom bevatten de NOS enkel opgaven die passend zijn voor de centrale toetsen en examens die het betreffende referentieniveau meten.

4.1 Toelichting toetsmatrijs referentiesets taal

Tabel 1 geeft een beschrijving van de samenstelling van de referentiesets taal 1F, 2F en 3F. De referentiesets taal betreffen enkel het domein 'begrijpend lezen'. De tabellen geven weer uit hoeveel opgaven en teksten de niet-openbare en openbare referentiesets bestaan, welke tekstsoorten hierin voorkomen, en welke verdeling gehanteerd wordt tussen de verschillende onderwijssectoren. De referentiesets bestaan uit de volgende tekstsoorten:

- Fictionele teksten (1F en 2F, ten behoeve van het PO)
  - Boek
  - Fabel
  - Gedicht
  - Historisch
  - Mop
  - Parabelachtig
  - Sprookje
  - Realistische fictie
- Zakelijke teksten
  - Betogend
  - Informatief
  - Instructief

De referentiesets taal bevatten hoofdzakelijk meerkeuzevragen.

Aanvullende kernmerken van de referentiesets taal:
- Er is sprake van een oververtegenwoordiging van informatieve teksten in referentieset taal 1F.
- In referentiesets taal 1F en 2F ontbreken evaluatieve opgaven.
- Een operationalisering van de vaardigheid samenvatten is niet opgenomen in referentieset taal 2F.
- Referentieset taal 3F bevat geen instructieve tekst. Het blijkt erg lastig een instructieve tekst te ontwikkelen op niveau 3F.

De expertpanels, verantwoordelijk voor de inhoudelijke validering van de referentiesets, geven aan het belangrijk te vinden dat deze inhoudelijke kenmerken bij de referentiesets benoemd worden in de communicatie over de referentiesets (zie voor meer toelichting hoofdstuk 5).

Tabel 1. Toetsmatrijs referentiesets taal 1F, 2F en 3F

1F	**Nederlandse taal (Lezen)**		
Aantal teksten	**NOS**	**OS**	**Totaal**
Fictioneel	4	6	10
Boek		3	3
Fabel	1		1
Gedicht		2	2
Historisch		1	1
Mop	1		1
Parabelachtig	1		1
Sprookje	1		1
Zakelijk	10	14	24
Betogend		2	2
Informatief	8	10	18
Instructief	2	2	4
Totaal	**14**	**20**	**34**
Aantal opgaven	**NOS**	**OS**	**Totaal**
Totaal	**50**	**85**	**135**

2F	**Nederlandse taal (Lezen)**		
Aantal teksten	**NOS**	**OS**	**Totaal**
Fictioneel	3	3	6
Fictie, historisch	1		1

Fictioneel, grap		1	1
Realistische fictie	2	2	4
Zakelijk	9	18	27
Betogend	1	7	8
Informatief	7	9	16
Instructief	1	2	3
Totaal	**12**	**21**	**33**
Aantal opgaven	**NOS**	**OS**	**Totaal**
Totaal	**69**	**89**	**158**

3F	**Nederlandse taal (Lezen)**		
Aantal teksten	**NOS**	**OS**	**Totaal**
Zakelijk	7	13	20
Betogend	4	9	13
Informatief	2	3	5
Instructief	1	1	2
Totaal	**7**	**13**	**20**
Aantal opgaven	**NOS**	**OS**	**Totaal**
Totaal	**66**	**98**	**164**

4.2 Toelichting toetsmatrijs referentiesets rekenen

Tabel 2 is een beschrijving van de samenstelling van de referentiesets rekenen. De tabel geeft weer uit hoeveel opgaven de NOS en OS bestaan en hoe de verdeling over domeinen en schooltypen eruit ziet.

De volgende domeinen komen voor in de referentiesets rekenen:

- getallen
- meten en meetkunde
- verbanden

- verhoudingen

In de referentiesets komen open vragen en meerkeuzevragen voor. Verder zijn er opgaven met rekenmachine en opgaven zonder rekenmachine.

Aanvullend kenmerk bij de referentieset rekenen 3F:

- Het expertpanel rekenen 3F geeft aan dat het enkel opnemen van niveau 3F opgaven in de referentiesets rekenen 3F ertoe heeft geleid dat het domein meten en meetkunde wat ondervertegenwoordigd is in de referentieset rekenen 3F. Het expertpanel raadt daarom aan toetsenmakers aan om, ten behoeve van het overbrengen van de referentiecesuur, een mix van opgaven uit de 2F en 3F referentieset samen te stellen om zo het zo tot een goed mogelijk inhoudelijke dekking van het domein te komen.

Tabel 2. Toetsmatrijs referentiesets rekenen 1F, 2F en 3F, ankersets rekenen 1S

1F	**Rekenen**		
	NOS	**OS**	**Totaal**
Domein: getallen	20	32	52
Domein: meten	15	24	39
Domein: verbanden	5	8	13
Domein: verhoudingen	10	16	26
Totaal	**50**	**80**	**130**
	NOS	**OS**	**Totaal**
Vraagtype: meerkeuze	33	48	81
Vraagtype: open	17	32	49
Totaal	**50**	**80**	**130**

1S	**Rekenen**		
	NOS	**OS**	**Totaal**
Domein: getallen	16	16	32
Domein: meten	12	12	24
Domein: verbanden	4	4	8

Domein: verhoudingen	8	8	16
Totaal	**40**	**40**	**80**
	NOS	**OS**	**Totaal**
Vraagtype: meerkeuze	37	23	60
Vraagtype: open	3	17	20
Totaal	**40**	**40**	**80**

2F	**Rekenen**		
	NOS	**OS**	**Totaal**
Domein: getallen	15	24	39
Domein: meten	10	16	26
Domein: verbanden	10	16	26
Domein: verhoudingen	15	24	39
Totaal	**50**	**80**	**130**
	NOS	**OS**	**Totaal**
Vraagtype: meerkeuze	19	34	53
Vraagtype: open	31	46	77
Totaal	**50**	**80**	**130**
	NOS	**OS**	**Totaal**
Hulpmiddel: rekenmachine	46	44	90
Hulpmiddel: zonder rekenmachine	4	36	40
Totaal	**50**	**80**	**130**

3F	Rekenen		
	NOS	**OS**	**Totaal**
Domein: getallen	12	21	33
Domein: meten	11	18	29
Domein: verbanden	11	15	26
Domein: verhoudingen	16	26	42
Totaal	**50**	**80**	**130**
	NOS	**OS**	**Totaal**
Vraagtype: meerkeuze	1	31	32
Vraagtype: open	49	49	98
Totaal	**50**	**80**	**130**
	NOS	**OS**	**Totaal**
Hulpmiddel: rekenmachine	44	53	97
Hulpmiddel: zonder rekenmachine	6	27	33
Totaal	**50**	**80**	**130**