RCEC
Het expertisecentrum voor de kwaliteit van examinering

Ministerie OCW
DGPV – Directie Primair Onderwijs

Onderzoek naar de kwaliteit van de normering eindtoetsen primair onderwijs 2022

Arnold J. Brouwer, Bernard P. Veldkamp

Ministerie OCW
DGPV – Directie Primair Onderwijs

Onderzoek naar de kwaliteit van de normering eindtoetsen primair onderwijs 2022

Arnold J. Brouwer, Bernard P. Veldkamp

Apeldoorn, maart 2022

RCEC

RCEC, *Research Center voor Examinering en Certificering*, is het expertisecentrum voor het borgen en bevorderen van de kwaliteit van examinering.

Als auditorganisatie beoordelen wij de kwaliteit van studietoetsen en examens. Wij zijn betrokken bij de beoordeling en certificering van grote Nederlandse examenstelsels. Zowel binnen het reguliere onderwijs ressorterend onder het ministerie van Onderwijs, Cultuur en Wetenschap, als voor het niet-reguliere onderwijs, zoals de door de ministeries ingestelde zelfstandige bestuursorganen (ZBO's).

Verder zijn wij een volledig onafhankelijke adviesorganisatie. Wij ontwikkelen geen eigen instrumenten, maar zijn er exclusief om instellingen in het regulier onderwijs en de beroepspraktijk te ondersteunen met integraal toetsdeskundig advies. Met praktijkgericht onderzoek, data-gedreven innovatie en psychometrische dienstverlening richten wij ons op de volledige toetscyclus. Van constructie en afname tot beoordeling en analyse.

De kennis en ervaring die wij in ons bestaan hebben opgedaan, en nog steeds opdoen, delen we via onze academie. Wij bieden verschillende vormen van opleiding, gericht op kwalitatieve examinering en toegepaste psychometrie. Dit doen we via open inschrijving, in-company maatwerkopleidingen en online via ons learning management systeem. Zo doen wij recht aan beslissingen door, over en voor talent.

RCEC B.V. • Villa de Regentes • Regentesselaan 5 • 7316 AA Apeldoorn
Postbus 71 • 8170 AB Vaassen
T +31 (0)55 - 303 31 24 • **I** www.rcec.nl • **E** info@rcec.nl

Dependance:
Universiteit Twente • Faculteit BMS – OMD • Gebouw Cubicus • Kamer B323
Drienerlolaan 5 • 7522 NB Enschede

Managementsamenvatting

In opdracht van het Ministerie van Onderwijs, Cultuur en Wetenschap, Directeur-Generaal Primair en Voortgezet Onderwijs (DGPV), Directie Primair Onderwijs, heeft RCEC onderzoek gedaan naar de kwaliteit van de normering eindtoetsen primair onderwijs 2022.

RCEC heeft een bureauonderzoek gedaan naar de vraag of de IRT1 en IRT2 methode valide normeringsmethoden zijn. De centrale onderzoeksvraag is daarbij gespecificeerd in vier separate deelvragen, waarmee de validiteit, de bruikbaarheid en de rechtsgeldigheid van de IRT1 en IRT2 normeringsmethode zijn onderzocht. Aansluitend is onderzocht onder welke voorwaarden de IRT2 normeringsmethode kan worden losgelaten.

RCEC heeft geconcludeerd dat de redenering dat met de IRT1 normeringsmethode de toetsscores op een valide manier omgezet kunnen worden in cesuren voor referentieniveaus en toetsadviescategorieën in bepaalde situaties weerlegd kan worden. RCEC heeft ook kunnen vaststellen dat de IRT1 normeringsmethode psychometrisch gezien voldoende valide is om toe te passen bij de normering van de eindtoetsen voor 2022. Dit onder de voorwaarde dat er gekeken wordt of er geen grote verschillen tussen de voorlopige schooladviezen en de toetsadviezen zijn ontstaan, bijvoorbeeld ten gevolge van de coronapandemie.

Voor wat betreft de IRT2 normeringsmethode heeft RCEC vastgesteld dat de resultaten van individuele leerlingen over de jaren heen op valide wijze kunnen worden vergeleken, mits wordt voldaan aan de aannames die ten grondslag liggen aan de IRT2 methode. De methode lijkt daarmee geschikt voor het bepalen van de schooljaar specifieke toetsadviezen. En de methode lijkt eveneens geschikt om via de koppeling met de PCET afname van dit jaar uitspraken te doen over de behaalde referentieniveaus van individuele leerlingen over de jaren heen. De scores van de individuele leerlingen ten aanzien van de referentieniveaus dienen daarbij geïnterpreteerd te worden alsof de leerlingen de PCET hebben gemaakt.

RCEC heeft vervolgens geconcludeerd dat voor de normering van de eindtoetsen 2022 de huidige primaire processen voldoende bruikbaar zijn. Dit onder de voorwaarde dat er een vierogen principe wordt ingericht waarmee toezicht wordt gehouden op een correcte uitvoering van alle cruciale processtappen waarin data wordt aangeleverd, data wordt geanalyseerd en informatie wordt gegenereerd, geïnterpreteerd en geïmplementeerd.

Aansluitend heeft RCEC vastgesteld dat, mits de genoemde adviezen en aanbevelingen worden opgevolgd, er voor wat betreft de normering van de eindtoetsen 2022 in voldoende mate aansluiting is met de centrale eisen van de eindtoets, zoals vermeld in het Toetsbesluit PO.

Tot slot heeft RCEC enkele inhoudelijke en procesmatige aanbevelingen gegeven voor specifiek de normering van de eindtoetsen 2022. Het opvolgen van deze aanbevelingen kan de opmaat zijn naar een IRT1 normeringsmethode die inhoudelijk, procesmatig en wettelijk valide is en waarbij de IRT2 normeringsmethode op termijn kan worden losgelaten.

Inhoudsopgave

Lijst van Figuren

Lijst van Tabellen

1. Inleiding

1.1. Aanleiding

De achterliggende periode heeft Stichting Cito Control (SCC) in samenspraak met de Expertgroep Toetsen PO (EPO) en de aanbieders van een eindtoets gewerkt aan het normeringshandboek 2022. Het normeringshandboek 2022 betreft een verzameling documenten waarin onder andere een beschrijving van de toetsoverstijgende IRT1 en IRT2 normeringsmethoden[1] wordt gegeven. In de documenten wordt daarnaast een voorstel gepresenteerd om stapsgewijs de tijdelijk toegevoegde IRT2 normeringsmethode los te kunnen laten, zodat er weer volledig met de IRT1 normeringsmethode kan worden gewerkt.

De IRT1 methode is een normeringsmethode waarbij voor de equivalering[2] gewerkt wordt met een itemkalibratie[3] met de gezamenlijke ankeropgaven. Uit de toepassing van de IRT1 methode op data van voorafgaande jaren volgt dat dit een psychometrisch verantwoorde methode lijkt te zijn. De methode geeft echter nog niet voor alle eindtoetsen een betrouwbaar beeld (c.q. een valide vergelijking), wat vermoedelijk samenhangt met de wijze waarop de gezamenlijke ankeropgaven in de verschillende eindtoetsen worden gebruikt. De locatie van de ankeropgaven in de verschillende eindtoetsen lijkt van invloed te zijn op de kwaliteit van de kalibratie. De IRT2 methode is een normeringsmethode waarbij voor de equivalering gebruik wordt gemaakt van de voorlopige schooladviezen en bijbehorende populatieverdeling en van de schaalwaarden van de Centrale Eindtoets (CET). Uit analyses van SCC volgt dat deze methode een betrouwbaar beeld (c.q. een valide vergelijking) voor alle eindtoetsen geeft. De methode heeft er echter de schijn van dat het een vorm is van relatief normeren[4], in plaats van absoluut normeren[5].

Naast een beschrijving van de inhoud van beide methoden en de daarmee berekende resultaten voor de referentieniveaus en de toetsadviezen, bevatten de documenten een beschrijving van een voorgestelde route om stapsgewijs voor alle eindtoetsen terug te keren van de IRT2 naar de IRT1 methode.

Binnen het ministerie van Onderwijs, Cultuur en Wetenschap (OCW) zijn omtrent beide normeringsmethoden enkele inhoudelijke vragen geformuleerd, die om een nadere onderbouwing (contra expertise) vragen. Een onderzoek naar deze vragen is gewenst om extra aanvullend inzicht te krijgen in de validiteit van de IRT1 en IRT2 methode voor wat betreft de normering van de eindtoetsen primair onderwijs 2022. Daarnaast heeft het ministerie OCW gevraagd om onderzoek te doen naar de

[1] Normeringsmethode betreft de procedures die tot doel hebben de scores, behaald op de onderdelen Lezen, Rekenen en Taalverzorging van de verschillende eindtoetsen, op dezelfde wijze te waarderen en om te zetten in de behaalde referentieniveaus en het berekende toetsadvies.
[2] Equivalering betreft de procedure om de scores van de verschillende onderdelen van de eindtoetsen op eenzelfde schaal te brengen.
[3] Itemkalibratie is het schatten (berekenen) van de psychometrische eigenschappen (parameters) van een verzameling opgaven (items) met het doel om deze items op dezelfde schaal te brengen.
[4] Relatief normeren is het na afname van de eindtoetsen omzetten van scores in de behaalde referentieniveaus en toetsadviezen gebaseerd op een onderlinge vergelijking van de schooladviezen van groepen van leerlingen.
[5] Absoluut normeren is het vooraf bepalen van de grenzen (c.q. cesuren) voor de referentieniveaus en toetsadviezen.

bruikbaarheid van de daarbij passende voorgestelde route waarin stapsgewijs wordt teruggekeerd naar de IRT1 methode.

1.2. Centrale onderzoeksvraag

Het door RCEC uit te voeren onderzoek heeft als doel antwoord te geven op de volgende centrale onderzoeksvragen:

"Zijn de IRT1 en IRT2 methode, zoals beschreven in het normeringshandboek 2022, valide normeringsmethoden en sluiten deze aan bij de eisen voor eindtoetsen, zoals beschreven in het Toetsbesluit PO? En in welke mate is het voorgestelde implementatieplan haalbaar en uitvoerbaar voor de betrokken partijen?"

De centrale onderzoeksvragen beantwoordt RCEC door het doen van een bureauonderzoek naar de volgende vier deelvragen.

1. In welke mate zijn de IRT1 en IRT2 normeringsmethode psychometrisch inhoudelijk juist en volledig beschreven? Het doel van de eerste onderzoeksvraag is om de mate van validiteit van: (1) de inhoud van de IRT1 en IRT2 methode, (2) de resultaten die beide methoden opleveren en (3) de conclusies die eruit getrokken worden, te evalueren.

2. In welke mate voldoet de huidige werkwijze van de IRT1 en IRT2 normeringsmethode aan de proceseisen voor valide normering? Het doel van de tweede deelvraag is om de mate van bruikbaarheid van beide methoden te evalueren.

3. In welke mate sluiten de IRT1 en IRT2 normeringsmethode aan bij de centrale eisen van de eindtoets zoals vermeld in artikel 3 uit het Toetsbesluit PO? Het doel van de derde deelvraag is om de mate van rechtsgeldigheid van beide methoden te evalueren.

4. Onder welke voorwaarden is de voorgestelde stapsgewijze terugkeer van de IRT2 naar de IRT1 methode uitvoerbaar voor de toetsaanbieders? Het doel van de vierde en laatste deelvraag is om te evalueren op welk moment het haalbaar is om de IRT2 normeringsmethode los te kunnen laten.

Onderdeel van het RCEC onderzoek is het (in de vorm van een kascontrole) door RCEC nalopen van de door Stichting Cito uit te voeren berekeningen in de bijbehorende software. RCEC gaat er voor dit onderzoek van uit dat de software zelf foutloos functioneert. RCEC doet derhalve geen onderzoek naar de wijze waarop de software en algoritmes door Stichting Cito zijn geprogrammeerd. RCEC richt zich daarnaast in dit onderzoek specifiek op de normering van de eindtoetsen in 2022. Zij doet daarbij geen uitvoerig onderzoek naar de wijze waarop de bestaande werkafspraken tussen de betrokken partijen tot nu toe in de praktijk zijn uitgevoerd.

1.3. Leeswijzer

Hoofdstuk 2 van dit rapport start met een beschrijving van de totstandkoming en verantwoording van de eindtoetsen 2022 aan de hand van het Beoordelingskader Eindtoetsen (EPO, 2020). Deze beschrijving bevat de centrale aspecten die fungeren als de bron voor de toetsoverstijgende normering

met de IRT1 en de IRT2 methode. In hoofdstuk 3 worden de resultaten van het eerste deelonderzoek beschreven. Centraal staat hier de vraag in welke mate de IRT1 en IRT2 normeringsmethode psychometrisch inhoudelijk juist en volledig zijn beschreven. Voor dit onderzoek is gewerkt volgens de principes van de argumentgerichte benadering van validiteit (Kane, 1992; Kane, 2004; Wools, 2012).

Hoofdstuk 4 gaat vervolgens in op de resultaten van het tweede deelonderzoek. Hierin staat de vraag centraal in welke mate de processen van de IRT1 en IRT2 normeringsmethode bruikbaar zijn. Hiertoe worden de stappen van de primaire processen in kaart gebracht volgens de indeling van de procesmodellering (Aguilar-Saven, 2004). Vervolgens worden deze stappen geëvalueerd met behulp van dezelfde principes van de argumentgerichte benadering als in hoofdstuk 3. In hoofdstuk 5 wordt beschreven in welke mate de IRT1 en IRT2 normeringsmethode aansluiten bij de centrale eisen van de eindtoets zoals vermeld in het Toetsbesluit PO[6]. Daartoe worden de bevindingen uit de hoofdstukken 3 en 4 individueel getoetst aan de losse leden van Artikel 4 uit het Toetsbesluit PO.

Aansluitend wordt in hoofdstuk 6 onderzocht onder welke inhoudelijke en procesmatige voorwaarden het voor de toetsaanbieders haalbaar is om volledig terug te kunnen keren naar de IRT1 normeringsmethode en daarbij de IRT2 normeringsmethode los te kunnen laten. Tot slot bevat hoofdstuk 7 de conclusies en aanbevelingen van dit onderzoek.

Namens RCEC,

Dr. Arnold J. Brouwer
Apeldoorn, 2 maart 2022

[6] https://wetten.overheid.nl/BWBR0035216/2019-08-01

2. Stelsel eindtoetsen 2022

2.1. Totstandkoming eindtoetsen

Vanaf schooljaar 2014 – 2015 zijn scholen in het primair onderwijs (po) verplicht in groep 8 een eindtoets Rekenen en Taal af te nemen. Zoals gepresenteerd in Tabel 1, kunnen scholen in het po voor het schooljaar 2021 – 2022 scholen kiezen uit één van de volgende, tot het stelsel toegelaten, eindtoetsen (in alfabetische volgorde):

Tabel 1.
Inhoud en samenstelling eindtoetsen 2022 (Stichting Cito Control, 2021b)

Toets	**AMN**	**CET (ACET)**	**CET (PCET)**	**DIA**	**IEP**	**ROUTE8**
Type	CAT	MST + Digitaal Lineair	Papier Lineair	MST + Digitaal Lineair	Papier Lineair	CAT
Onderdelen (*verplicht*)	Lezen Rekenen Taal-verzorging	Lezen (MST) Rekenen (MST) Schrijven Taal-verzorging (MST)	Lezen Rekenen Schrijven Taal-verzorging	Lezen (MST) Rekenen (MST) Taal-verzorging Woorden-schat	Lezen Rekenen Taal-verzorging	Lezen Rekenen Taalverzorging Begrippenlijst Woordenschat
Onderdelen (*facultatief*)	x	Wereld-oriëntatie	Wereld-oriëntatie	x	x	Dictee Luister-vaardigheid Persoonlijk functioneren

Iedere toetsaanbieder, zowel de publieke aanbieder van de publieke Centrale Eindtoets (CET)[7] als de private aanbieders van één van de overige eindtoetsen, is verantwoordelijk voor het initieel ontwikkelen en vervolgens jaarlijks onderhouden, afnemen, scoren en rapporteren van de eigen eindtoets. Elk van de private aanbieders dient daarbij te voldoen aan de kwaliteitseisen, zoals vastgelegd in het Beoordelingskader Eindtoetsen (EPO, 2020). De CET heeft binnen het huidige stelsel een eigenstandige positie, en valt onder toezicht van de Onderwijsinspectie. Het Beoordelingskader Eindtoetsen (EPO, 2020) beschrijft de kwaliteitseisen voor de onderwijskundige inhoud, de organisatorische aspecten, de psychometrische aspecten en de beveiligingsaspecten. De inhoud van het beoordelingskader is gebaseerd op de kwaliteitseisen in het Toetsbesluit PO (2014)[8] en het document Algemeen deel Toetswijzer voor Eindtoets Taal en Rekenen (CvTE, 2014)[9].

[7] https://www.centraleeindtoetspo.nl/publicaties/vragen-en-antwoorden/wie-is-verantwoordelijk-voor-de-centrale-eindtoets
[8] https://www.rijksoverheid.nl/documenten/besluiten/2014/01/20/toetsbesluit-po.
[9] https://www.centraleeindtoetspo.nl/binaries/centraleeindtoets/documenten/publicaties/2015/05/01/toetswijzer-algemeen-deel-voor-eindtoets-po/Toetswijzer_eindtoets_po_algemeen_deel_27_aug_2014.pdf

De Kerndoelen (OCW, 2006) en het Referentiekader (OCW, 2010)[10] voor Taal en Rekenen voor het einde basisonderwijs vormen het wettelijk kader waarbinnen de toetsinhouden voor eindtoetsen Rekenen en Taal worden beschreven. In de kerndoelen staan de reken- en taalinhouden beschreven die leerkrachten de leerlingen moeten aanbieden, opdat zij deze doelen in voldoende mate kunnen bereiken voor of aan het einde van het basisonderwijs. Daarbij gaat het om aanbodsdoelen. De referentieniveaus beschrijven wat leerlingen moeten kennen, kunnen en begrijpen aan het einde van de basisschool. Hierbij gaat het om beheersingsdoelen. De referentieniveaus dekken de kerndoelen, ze bevatten geen nieuwe leerstof. Voor het einde van de basisschool zijn twee referentieniveaus van belang, het fundamenteel- en het streefniveau. Voor Rekenen zijn dat niveau 1F en 1S. Voor Taal zijn dat niveau 1F en niveau 2F.

Iedere eindtoets dient per individuele leerling een standaardscore te berekenen. De standaardscore representeert de gemeten vaardigheid van de individuele leerling, wat wordt uitgedrukt in het toetsadvies voor het bij hem of haar best passende brugklastype. Sinds het schooljaar 2018 – 2019 worden individuele leerlingen bij alle toetsaanbieders ingedeeld in dezelfde zes adviescategorieën:

- pro/vmbo bb
- vmbo bb/vmbo kb
- vmbo kb/vmbo gl-tl
- vmbo gl-tl/havo
- havo/vwo
- vwo

Een belangrijk uitganspunt in het stelsel van de door de minister voor Basis- en Voortgezet Onderwijs en Media toegelaten eindtoetsen is dat een individuele groep 8 leerling, onafhankelijk van welke eindtoets hij of zij door de school krijgt aangeboden, eenzelfde uitkomst moet krijgen van de behaalde referentieniveaus Rekenen en Taal en het voorgestelde toetsadvies voor het best passende brugklastype.

Om deze vereiste onderlinge vergelijkbaarheid van eindtoetsen te waarborgen, wordt er gebruik gemaakt van het gezamenlijk anker[11]. Het gezamenlijk anker kan worden beschouwd als een mini-versie van de eindtoets, waarbij een gelijke prestatie op deze minitoets een vaste koppeling heeft via de PCET met een prestatie op de referentieniveaus Rekenen en Taal en het geven van een passend toetsadvies (Stichting Cito Control, 2020). Elke eindtoets bevat gezamenlijke ankeropgaven voor het onderdeel Rekenen en voor de domeinen Lezen en Taalverzorging van het onderdeel Taal. Om een representatieve set ankeropgaven in de verschillende type eindtoetsen te kunnen inbouwen, zijn er in 2018 drie aparte subsets ontwikkeld: een set voor papieren toetsen, bestaande uit een subset papier en een subset papier-digitaal, en een set voor digitale toetsen, bestaande uit een subset digitaal en een subset papier-digitaal. Door verschillende subsets van ankeropgaven aan te bieden, wordt rekening gehouden met specifieke randvoorwaarden voor wat betreft toetsvorm en presentatiemogelijkheden van opgaven. Het gezamenlijk anker bestond bij de start in 2018 uit in totaal 82 ankeropgaven, waarvan 27 Lezen, 28 Rekenen en 27 Taalverzorging opgaven.

Het initiële gezamenlijke anker uit 2018 is samengesteld uit geproeftoetste opgaven van de CET. De ankeropgaven zijn toentertijd in samenspraak met de toetsaanbieders door inhoudsdeskundigen van

[10] https://wetten.overheid.nl/BWBR0027879/
[11] https://www.rijksoverheid.nl/documenten/rapporten/2018/12/01/vergelijkbaarheid-eindtoetsen-en-invoering-van-gezamenlijk-anker

de EPO beoordeeld op de onderwijskundige aspecten, zoals gedocumenteerd in het Beoordelingskader Eindtoetsen (EPO, 2020). Vervolgens zijn de ankeropgaven in datzelfde jaar door Stichting Cito gekalibreerd met zowel een één-parameter logistisch model (1PLM; Rasch) als met het One Parameter Logistic Model (OPLM). In een 1PL model wordt de moeilijkheidsgraad (c.q. de β-parameter) van de opgaven populatie-onafhankelijk geschat. In het OPLM model wordt daarnaast ook het onderscheidend vermogen (c.q. de α-parameter) van de opgave geïmputeerd. Als schattingsmethode is gebruik gemaakt van Marginal Maximum Likelihood (MML). Kenmerkend voor MML is dat er bij het schatten van de itemparameters van uitgegaan wordt dat de persoonsparameters (c.q. de vaardigheid; θ) random (c.q. willekeurig) getrokken zijn uit een populatieverdeling. Met behulp van deze aanname kunnen de itemparameters uniek worden geïdentificeerd. In 2019 is de itemkalibratie van het gezamenlijk anker door de EPO uitgevoerd in de Lexter software op basis van circa 1.000 observaties per opgave. Voor het schooljaar 2022 zal SCC het gezamenlijk anker opnieuw kalibreren met alle beschikbare data sinds 2018, waarna de EPO deze kalibratie zal controleren.

Sinds 2018 wordt het gezamenlijk anker periodiek door de toetsaanbieders gezamenlijk uitgebreid en ververst. Hiertoe construeert elke toetsaanbieder een deel van de nieuwe ankeropgaven, waarna iedere toetsaanbieder deze opgaven als zaai-item pretest in zijn eigen eindtoets. De goed functionerende ankeropgaven worden vervolgens aanvullend door de EPO gekalibreerd met het huidige anker. Na goedkeuring worden de nieuwe ankeropgaven toegevoegd aan het zogenaamde operationele gezamenlijke anker. Vanaf dat moment kunnen de ankeropgaven worden gebruikt om de verschillende eindtoetsen aan elkaar te verbinden. Het gezamenlijk anker wordt tevens gebruikt om de cesuurpunten voor de referentieniveaus en toetsadviezen (c.q. plaatsingsadviezen) voor de private toetsaanbieders (niet zijnde de CET) vast te stellen.

Bovengenoemd procedé heeft geresulteerd in een actueel operationeel gezamenlijk anker van 255 opgaven, waarvan 100 Lezen, 89 Rekenen een 66 Taalverzorging. Door inhoudelijk experts van de EPO is in samenspraak met de toetsaanbieders vastgesteld welk deel van de opgaven welk referentieniveau representeert. Op dit moment (situatie 2021) is deze verdeling als volgt: 35 opgaven Lezen 1F, 38 opgaven Lezen 2F, 31 opgaven Rekenen 1F, 30 opgaven Rekenen 1S, 24 opgaven Taalverzorging 1F, en 33 opgaven Taalverzorging 2F. Ieder schooljaar selecteert elke toetsaanbieder de subsets van ankeropgaven zoals geselecteerd door de EPO, en voegt deze toe aan de eindtoets van het betreffende schooljaar.

Aanvullend op de (periodiek ververste) ankeropgaven, ontwikkelt elke toetsaanbieder jaarlijks eigen opgaven voor de domeinen en onderdelen in zijn toets[12]. Het Beoordelingskader Eindtoetsen (EPO, 2020)[13] schrijft voor dat voor papieren en/of digitale lineaire toetsen jaarlijks alle eigen opgaven moeten worden vernieuwd. Voor adaptieve toetsen die worden samengesteld uit een itembank, geldt dat jaarlijks 20 – 30% van de eigen opgaven moet worden vernieuwd. Alle nieuw geconstrueerde opgaven worden eerst door inhoudsdeskundigen van de EPO beoordeeld op de onderwijskundige aspecten, voordat deze mogen worden ingebouwd in de eerstvolgende eindtoets. In het huidige

[12] Het Toetsbesluit PO (2014) schrijft voor dat opgaven van zowel de CET als van de eindtoetsen van de private aanbieders dienen te worden vernieuwd.
[13] De EPO adviseert over het toelaten van de eindtoetsen van de private aanbieders, niet zijnde de publieke CET. Dit gebeurt op basis van het Beoordelingskader Eindtoetsen (EPO, 2020).

Beoordelingskader Eindtoetsen zijn vooralsnog alleen inhoudelijke eisen voor opgaven voor de wettelijk verplichte domeinen Lezen, Rekenen en Taalverzorging en voor de optionele wettelijke domeinen Begrippenlijst, Mondelinge taalvaardigheid en Schrijven van het onderdeel Taal opgesteld.

Iedere toetsaanbieder is vervolgens verantwoordelijk voor het pretesten van de eigen ontwikkelde opgaven. Hiertoe moet worden voldaan aan de eisen voor de steekproefomvang en -samenstelling, zoals beschreven in het Beoordelingskader Eindtoetsen (EPO, 2020). Het pretesten mag plaatsvinden in de vorm van een minder high-stakes proeftoets op vrijwillig deelnemende scholen. En het pretesten mag ook plaatsvinden door de nieuwe opgaven te zaaien in de huidige operationele (high-stakes) afname van de eindtoets. Voor pretesten in een proeftoets omgeving gelden, met het oog op model fit, minder stringente psychometrische eisen dan voor het pretesten in een operationele setting.

Na afronding van de pretest procedure stelt de toetsaanbieder zijn eigen eindtoets samen. Elke eindtoets bevat de nieuw ontwikkelde opgaven voor de onderdelen en domeinen uit de toets. Voor de wettelijk verplichte onderdelen Rekenen en Taal dient iedere eindtoets in ieder geval opgaven te bevatten met een moeilijkheidsgraad tussen de door de EPO vastgestelde cesuren voor de referentieniveaus 1F en 2F / 1S. Daarnaast dient iedere eindtoets minimaal de geselecteerde subset ankeropgaven te bevatten. Iedere subset bestaat bij voorkeur uit minimaal 20 gezamenlijke ankeropgaven per wettelijk verplicht domein (c.q. Lezen, Rekenen en Taalverzorging). En iedere subset is representatief voor wat betreft de referentieniveaus. De samengestelde toets wordt eerst ter goedkeuring aangeboden aan de EPO, voordat deze op de scholen wordt ingezet.

Vervolgens vindt de operationele afname van alle eindtoetsen plaats in de periode april / mei. In het schooljaar 2022 is hiervoor de afnameperiode 15 april – 15 mei 2022 aangewezen. De toetsaanbieders kijken vervolgens zelf hun eigen toetsen na. De ruwe scores van de individuele opgaven worden hierna aangeboden aan Stichting Cito Control (SCC).

Met behulp van deze ruwe scores van alle toetsaanbieders gezamenlijk voert SCC een toetsoverstijgende normeringsprocedure uit. Het doel van deze procedure is om de ruwe scores, behaald op de opgaven van de wettelijk verplichte domeinen Lezen, Rekenen en Taalverzorging van de verschillende eindtoetsen, op dezelfde wijze te waarderen en om te zetten in cesuren voor de referentieniveaus en deze, voor zover van toepassing aangevuld met de optionele en facultatieve toetsonderdelen, om te zetten in cesuren voor de toetsadviescategorieën. Voorafgaande aan de normeringsvergadering voert de EPO een second opinion uit.

Voor dit procedé zijn in het schooljaar 2021 – 2022 een tweetal normeringsmethoden beschikbaar, de IRT1 normering en de IRT2 normering. De IRT1 methode is een normeringsmethode waarbij voor de equivalering (d.w.z. de procedure om de scores van de verschillende onderdelen van de eindtoetsen op dezelfde schaal te brengen) gewerkt wordt met een itemkalibratie met de gezamenlijke ankeropgaven (d.w.z. het schatten van de psychometrische eigenschappen van een verzameling opgaven met het doel om deze opgaven op dezelfde schaal te brengen). De IRT2 methode is een normeringsmethode waarbij voor de equivalering gebruik wordt gemaakt van de schooladviezen en bijbehorende populatieverdeling en van de schaalwaarden van de Centrale Eindtoets (CET). Voor de IRT1 methode maakt SCC gebruik van de DexterMML software. En voor de IRT2 methode maakt SCC gebruik van een aanvullend algoritme, door SCC geprogrammeerd in de open source statistische programmeertaal R.

Tot slot ontvangen de private toetsaanbieders de resultaten (c.q. de cesuren) van de toetsoverstijgende normeringsprocedure voor SCC. De private toetsaanbieders voeren deze normen na akkoord in de stuurgroep[14] door in hun eindtoetsen. De CET gebruikt deze resultaten als onafhankelijke controle voor haar normering. Met behulp van deze cesuren berekent de toetsaanbieder de standaardscore. Hiertoe kiest iedere toetsaanbieder zelf een eigen wiskundige functie (bv. een lineaire transformatiefunctie) waarmee de toetsaanbieder de Gemiddelde Latente Vaardigheid (GLV)[15] transformeert naar de standaardscore (c.q. eindscore). De bandbreedtes van standaardscores representeren tevens de toetsadviescategorieën op basis waarvan iedere leerling een eigen toetsadvies (c.q. plaatsingsadvies) krijgt. De toetsaanbieder stelt zelf de leerlingrapporten samen en deelt deze met de scholen. Vervolgens stelt de private aanbieder van een eindtoets een verantwoordingsverslag op en deelt dat met de EPO. In het Beoordelingskader Eindtoetsen (EPO, 2020) is beschreven waar dit verslag inhoudelijk aan dient te voldoen. De EPO beoordeelt de rapporten, om zo de kwaliteit van de eindtoetsen over de jaren heen te kunnen monitoren.

Nadat de scholen de resultaten van de eindtoets hebben ontvangen, delen zij deze met hun leerlingen en hun ouders of verzorgers. Conform het Toetsbesluit PO (2014) is de school verplicht het voorlopige schooladvies van een individuele leerling te heroverwegen wanneer het toetsadvies daar aanleiding toe geeft. Daarbij kan de school het schooladvies naar boven bijstellen. De scholen zijn tevens verantwoordelijk voor het doorgeven van de definitieve schooladviezen aan de Dienst Uitvoering Onderwijs (DUO) van het ministerie van Onderwijs, Cultuur en Wetenschap. De DUO resultaten worden tevens gedeeld met de Onderwijsinspectie, zodat zij haar wettelijke taak om de kwaliteit van het onderwijs op scholen te beoordelen kan uitvoeren.

Bovenstaande procesbeschrijving maakt inzichtelijk dat het jaarlijks onderhouden, afnemen, scoren, normeren en rapporteren van de eindtoetsen uit vele stappen bestaat en laat zien dat er qua toezicht een verschil bestaat tussen de publieke aanbieder en de private aanbieders. Bij de verschillende stappen hebben de toetsaanbieders meerdere keuzemogelijkheden. Om de vier deelvragen van dit onderzoek (zie paragraaf 1.2) te kunnen beantwoorden, is het belang om eerst deze verschillende keuzemogelijkheden van de processtappen in kaart te brengen. De verschillende opties kunnen mogelijk van invloed zijn op de mate van validiteit, bruikbaarheid en rechtsgeldigheid van de IRT1 en IRT2 normeringsmethode. Daarom wordt in paragraaf 2.2 eerst een overzicht van de keuzemogelijkheden gegeven, alvorens vanaf hoofdstuk 3 de onderzoeksvragen te beantwoorden.

2.2. Keuzemogelijkheden eindtoetsen

Aan de toetsaanbieders wordt de ruimte geboden om een eigen onderscheidende eindtoets te ontwikkelen die past bij de visie en identiteit van de aanbieder. Daartoe biedt het huidige stelsel in ieder geval de volgende keuzemogelijkheden:

[14] Besluiten over de normeringen van de eindtoetsen worden sinds het schooljaar 2019 – 2020 genomen door een stuurgroep onder voorzitterschap van het ministerie van OCW.
(https://www.nieuwsbrievenminocw.nl/actueel/nieuws/2019/10/17/eindtoetsen-po-verbeterstappen-na-incident)

[15] De Gemiddelde Latente Vaardigheid (GLV) is de gewogen latente vaardigheid. Op deze vaardigheidsschaal worden de cesuren voor de toetsadviezen bepaald.

Toetsvorm

De toetsaanbieders kunnen kiezen uit in ieder geval de volgende toetsvormen:

- Papieren lineaire toets
- Digitaal lineaire toets
- Digitaal adaptieve toets op itemniveau (CAT)
- Digitaal adaptieve toets op moduleniveau (MST)
- Combinatie van digitaal lineair en adaptief (CAT of MST)

Toetsinhoud

Naast de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen, kan de aanbieder extra toevoegen:

- Optionele domeinen van het wettelijk verplichte onderdeel Taal (Begrippenlijst, Mondelinge taalvaardigheid, Schrijven).
- Vrij te kiezen optionele (of facultatieve) onderdelen en domeinen, zoals bijvoorbeeld Persoonlijk functioneren of Wereldoriëntatie. Hiervoor worden in het Beoordelingskader Eindtoetsen (EPO, 2020) geen inhoudelijke kwaliteitscriteria vermeld.

Doelgroep

Iedere toetsaanbieder biedt zijn toets aan voor de basisschoolleerlingen in groep 8. Historisch gezien lijkt de samenstelling van de populaties van de verschillende aanbieders te kunnen verschillen, voor wat betreft het aandeel (c.q. percentage) speciaal basisonderwijs (SBO) en speciaal onderwijs (SO) leerlingen.

Pretesten van opgaven

- De toetsaanbieder bepaalt zelf of het pretesten plaatsvindt in een proefopstelling of plaatsvindt door middel van het zaaien van de opgaven in een operationele setting.
- De toetsaanbieder bepaalt zelf het psychometrisch model waarmee de opgaven zo deugdelijk mogelijk kunnen worden gekalibreerd.
- De toetsaanbieder bepaalt zelf de methode waarmee de itemparameters worden geschat.
- De toetsaanbieder bepaalt zelf in welke software de itemkalibratie wordt uitgevoerd.
- Voor het kalibreren van opgaven voor de wettelijk verplichte onderdelen Rekenen en Taal (bestaande uit de domeinen Lezen en Taalverzorging), gebruiken de toetsaanbieders dezelfde subsets ankeropgaven met gefixeerde itemparameters. Voor het kalibreren van opgaven voor optionele wettelijke domeinen en/of vrij te kiezen domeinen kan de toetsaanbieder geen gebruik maken van gezamenlijke ankeropgaven. Hier mag de toetsaanbieder kiezen voor het gebruik maken van bijvoorbeeld een eigen ontwikkeld intern anker.

Toetssamenstelling

- De toetsaanbieder bepaalt zelf de totale lengte van de toets. De wet schrijft geen minimale en/of maximale toetslengte voor.

- De toetsaanbieder bepaalt zelf de verhouding van het aantal toetsvragen voor de wettelijk verplichte onderdelen en domeinen, de optionele domeinen van wettelijk verplichte onderdelen, en de vrij te kiezen onderdelen en domeinen.

Ankeropgaven

- De toetsaanbieder selecteert één van de subsets van ankeropgaven voor gebruik binnen de eigen eindtoets (papier en papier-digitaal binnen een papieren eindtoets, digitaal en papier-digitaal binnen een digitale eindtoets).
- De toetsaanbieder kiest zelf de posities in de toets, waar de ankeropgaven terecht komen.
- Nieuwe ankeropgaven die in de loop van de tijd worden ontwikkeld, worden door de toetsaanbieder zelf geconstrueerd en als zaai-item gepretest in de eigen eindtoets.
- De toetsaanbieder kiest er zelf voor om wel of geen aanvullend eigen ontwikkeld anker in de toets in te bouwen.

Referentieniveaus

- De toetsaanbieder kiest zelf of hij binnen zijn eigen eindtoets voor de referentieniveaus 1F en 2F / 1S één vaardigheidsschaal of twee separate vaardigheidsschalen (een 1F schaal en een 2F / 1S schaal) gebruikt. Sinds de invoering van het gezamenlijk anker in 2018 gebruiken de private aanbieders één vaardigheidsschaal voor de referentieniveaus 1F en 2F / 1S.
- De toetsaanbieder kiest zelf of de referentieniveaus aanvullend ook gekoppeld worden aan andere bronnen, zoals bijvoorbeeld het leerlingvolgsysteem.

Standaardscore en toetsadvies

- De toetsaanbieder bepaalt zelf of (en welke van) de ankeropgaven wel of niet meetellen in het berekenen van de ruwe somscore van de leerling bij een papieren of een digitale lineaire toets dan wel de vaardigheid van de leerling bij een adaptieve toets (c.q. CAT of MST).
- De toetsaanbieder kiest zelf of optionele domeinen van wettelijk verplichte onderdelen en/of vrij gekozen onderdelen wel of niet meetellen voor de standaardscore.
- De toetsaanbieder kiest zelf een eigen wiskundige functie (bv. een lineaire transformatiefunctie) om de standaardscore te berekenen.
- De toetsaanbieder hanteert een eigen tabel met intervallen van de standaardscores, welke de toetsadviescategorieën representeren.

Verantwoordingsverslag

- De toetsaanbieder bepaalt zelf met welke software en algoritmes hij de vereiste berekeningen voor het verantwoordingsverslag maakt.

In het volgende hoofdstuk van dit onderzoeksrapport wordt de mate van validiteit van de IRT1 en de IRT2 methode onderzocht. De hierboven in paragraaf 2.1 beschreven totstandkoming van de eindtoetsen en de uiteenzetting van de keuzemogelijkheden in paragraaf 2.2 dienen als bron voor dit onderzoek.

3. Validiteit IRT1 en IRT2 normeringsmethode

3.1. Inleiding

In dit hoofdstuk worden de resultaten van het eerste deelonderzoek beschreven. Centraal staat hier de vraag in welke mate de IRT1 en IRT2 normeringsmethode psychometrisch inhoudelijk juist en volledig zijn beschreven. Het doel van de eerste onderzoeksvraag is om de mate van validiteit van: (1) de inhoud van de IRT1 en IRT2 methode, (2) de resultaten die beide methoden opleveren en (3) de conclusies die eruit getrokken worden, te evalueren.

Voor dit onderzoek is gewerkt volgens de principes van de argumentgerichte benadering (c.q. argument-based approach) van validiteit (Kane, 1992; Kane, 2004; Wools, 2012). In deze benadering van validiteit wordt een redenering onderzocht, zoals bijvoorbeeld de redenering dat de antwoorden die leerlingen geven op de vragen in de eindtoetsen op een betrouwbare en valide manier gebruikt kunnen worden om toetsadviezen te geven en af te leiden of en welke referentieniveaus ze behaald hebben. Bij het toepassen van de argumentgerichte benadering wordt allereerst de redenering nauwkeurig geformuleerd, vervolgens wordt er gekeken of argumenten uit de redenering ontkracht kunnen worden doordat er vraagtekens bij geplaatst kunnen worden. Tenslotte wordt bewijsmateriaal gezocht en geëvalueerd om de redenering te ondersteunen en de vraagtekens te ontkrachten.

In de volgende paragraaf worden de principes van de argumentgerichte benadering van validiteit toegepast om de mate van validiteit van de IRT1 normeringsmethode te beoordelen en te evalueren.

3.2. Normeringsmethode en item-responstheorie

De IRT1 methode is een op de item-responstheorie gebaseerde normeringsmethode waarbij voor de equivalering gewerkt wordt met een itemkalibratie met de gezamenlijke ankeritems (Stichting Cito Control, 2021a). Kenmerkend voor de item-responstheorie is dat de vaardigheid van de leerling nauwkeurig wordt geschat, rekening houdend met het verschil in moeilijkheid en, in het geval van een 2PLM, het onderscheidend vermogen van de toetsopgaven. De schatting van de vaardigheid van de individuele leerling is daarbij onafhankelijk van de moeilijkheid van de toets en van de gebruikte opgaven. Bovendien kan met de item-responstheorie een individuele schatting van de meetfout worden berekend.

Ook met betrekking tot de itemparameters kent de item-responstheorie enkele voordelen. Door de itemparameters separaat te schatten, ontstaat er inzicht in de kwaliteit van de individuele opgaven. De itemparameters kunnen bovendien onafhankelijk van de steekproef van leerlingen geschat worden. Daarnaast liggen de itemmoeilijkheid en de vaardigheid van de leerling op eenzelfde vaardigheidsschaal, wat veel inzicht geeft. Het gebruik van de item-responstheorie kan daarom leiden tot eerlijke en accurate toetsing en normering.

Ten slotte maakt de item-responstheorie het mogelijk om de eindtoetsen adaptief af te nemen. Dat houdt in dat tijdens de afname van een toets of examen al een inschatting van de vaardigheid van de leerling gemaakt wordt en dat de moeilijkheid van de opgaven afgestemd wordt op het niveau van de leerling. Dit leidt tot kortere toetsen en voorkomt frustratie vanwege het moeten beantwoorden van veel te makkelijke of veel te moeilijke opgaven.

Binnen het stelsel eindtoetsen wordt van genoemde kenmerken van de item-responstheorie onder andere gebruik gemaakt tijdens het pretesten van de opgaven. Tevens is door het gebruik van de item-responstheorie af te leiden dat de itemparameters van de ankeropgaven die worden geschat op basis van een eerdere afname, toegepast kunnen worden in een volgend schooljaar. Op deze manier wordt het ook mogelijk om toetsopgaven uit verschillende leerjaren en uit verschillende eindtoetsen op eenzelfde vaardigheidsschaal te kalibreren. Dit is van belang voor het onderbouwen van de onderlinge vergelijkbaarheid van de verschillende eindtoetsen.

Om de item-responstheorie en daarmee ook de IRT1 normeringsmethode op verantwoorde wijze te kunnen toepassen, gelden er wel enkele voorwaarden waar aan dient te zijn voldaan.

Ten eerste, de interpretatie van de scores. Binnen de item-responstheorie wordt de vaardigheid van leerlingen gepresenteerd op een onderliggende vaardigheidsschaal. De schattingen zijn daarbij begrensd en lopen veelal van -4,0 tot +4,0 met een gemiddelde van 0,0. De vaardigheidsschaal is bovendien niet lineair. Voor de leerlingen is daarom een vertaling nodig van de vaardigheidsscore naar een schaalscore. Dit gebeurt door middel van de normeringsprocedure. In deze procedure wordt bepaald bij welke vaardigheidsscore een leerling een bepaald referentieniveau heeft gehaald en bij welke vaardigheidsscores welk toetsadvies voor het best passende brugklastype horen. Zonder deze transformatie zijn de vaardigheidsscores nauwelijks tot niet te interpreteren voor de toetsgebruikers.

Een tweede voorwaarde is dat de steekproef van leerlingen voldoende groot moet zijn om de itemparameters nauwkeurig genoeg te kunnen schatten. Daardoor is de item-responstheorie alleen toepasbaar bij toetsen en examens met veel kandidaten. Mede hierom is afgesproken dat iedere eindtoets jaarlijks bij minimaal 1.000 leerlingen moet worden afgenomen, om in aanmerking te komen voor de door de overheid aangeboden subsidieregeling.

Een derde voorwaarde, ten slotte, is dat de item-responsmodellen wel een goede fit moeten laten zien bij de responsdata. Als de geobserveerde score erg afwijkt van de item-karakteristieke curve (c.q. een grafiek waarin het verband uitgedrukt wordt tussen de kans op het goed beantwoorden van een item en de vaardigheid van een leerling), dan is er sprake van misfit en zijn item-responsmodellen niet toepasbaar.

Uit de eerdere analyses van EPO en SCC, zoals gedocumenteerd in de in dit onderzoek aangehaalde referenties, volgt dat de IRT1 methode een psychometrisch verantwoorde normeringsmethode lijkt te zijn. Uit de resultaten van EPO en SCC volgt echter ook dat de methode nog niet voor alle eindtoetsen een betrouwbaar beeld (c.q. een valide vergelijking) oplevert. Het vermoeden bestaat dat dit samenhangt met de wijze waarop de gezamenlijke ankeropgaven in de verschillende eindtoetsen worden gebruikt. De locatie van de ankeropgaven in de verschillende eindtoetsen lijkt van invloed te zijn op de kwaliteit van de kalibratie. Daarom vervolgt dit hoofdstuk met een beschrijving van het nader uitgevoerde onderzoek naar de mate van validiteit van de IRT1 methode.

3.3. IRT1 normeringsmethode

De argumentgerichte benadering start met het formuleren van de redenering (c.q. de opbouw) van de te onderzoeken normeringsmethode. De IRT1 methode maakt gebruik van de mogelijkheid om scores op de verschillende eindtoetsen onderling vergelijkbaar te maken, door het gezamenlijk anker te

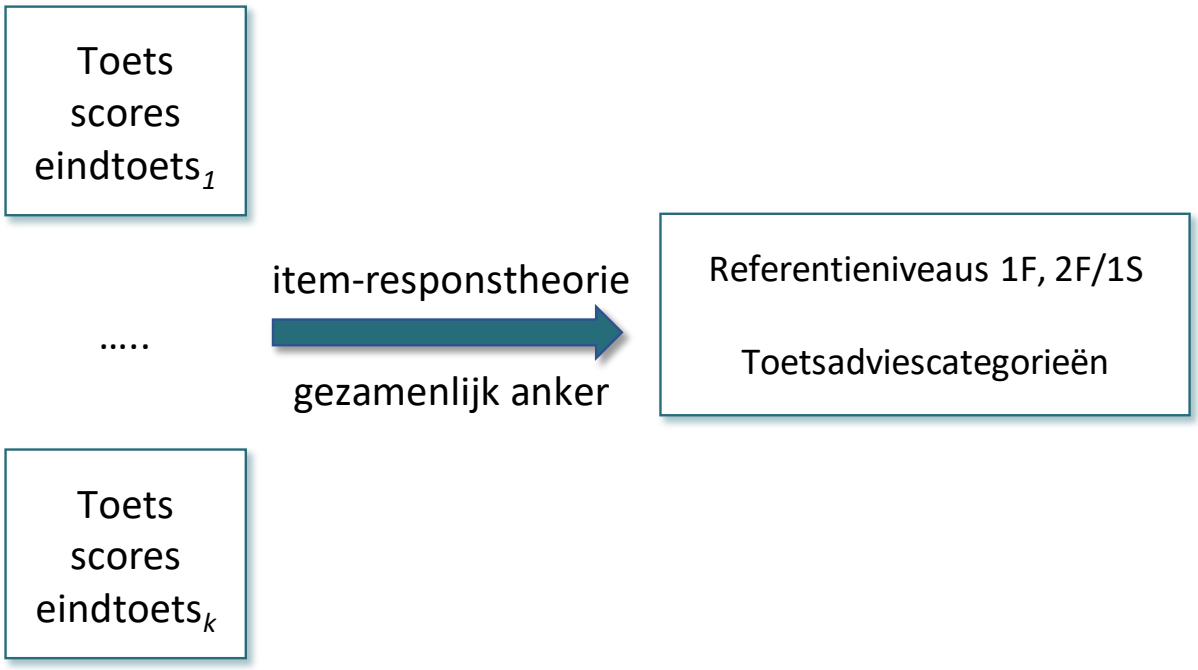
implementeren en door te werken met de item-responstheorie. Zoals beschreven in paragraaf 2.1, wordt bij de eindtoetsen gebruik gemaakt van een drietal subsets van gezamenlijke ankeropgaven.

De ankeropgaven worden niet alleen gebruikt om de resultaten van de verschillende eindtoetsen uit hetzelfde jaar onderling vergelijkbaar te maken, maar ook om de resultaten van de eindtoetsen vergelijkbaar te maken met die van eerdere schooljaren. Per de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen kunnen de cesuren die ten grondslag liggen aan de referentieniveaus en toetsadviescategorieën van het ene jaar ook overgebracht worden naar het volgende jaar.

Door gebruik te maken van de link die het (initiële) gezamenlijke anker heeft met de referentieniveaus, kunnen de scores op de eindtoetsen gelinkt worden met deze zelfde referentieniveaus. De link van het initiële gezamenlijke anker naar de referentieniveaus verloopt via de CET. De CET heeft in 2018 zowel het gezamenlijk anker als de eigen opgaven (c.q. de Niet Openbare Set, NOS) afgenomen. Eerder al zijn de NOS en de openbaar beschikbare referentiesets (c.q. de Openbare Set, OS)[16] gezamenlijk afgenomen, waardoor het (initiële) gezamenlijke anker is gekoppeld aan de OS. De prestaties die de leerlingen laten zien bij de verschillende eindtoetsen, kunnen daarmee gebruikt worden om uit te rekenen of de leerlingen de verschillende referentieniveaus 1F of 2F/1S hebben behaald voor de verschillende wettelijk verplichte domeinen Lezen, Rekenen en Taalverzorging. Daarnaast worden de prestaties van de leerlingen op zowel de wettelijke als de optionele en facultatieve toetsonderdelen gebruikt om de toetsadviescategorieën van de leerlingen uit te rekenen. Dit gebeurt aan de hand van een gewogen gemiddelde van de IRT1-vaardigheidsscores op de verschillende onderdelen uit de toets. De gewichten voor de wettelijk verplichte onderdelen (c.q. de referentieonderdelen) zijn daarbij relatief gelijk voor alle eindtoetsen. De optionele en facultatieve onderdelen hebben allen een eigen gewicht binnen de verschillende eindtoetsen. Alle gewichten samen worden genormeerd tot 1.0. Daarna past iedere toetsaanbieder zijn eigen transformatie naar standaardscores toe. De onderliggende schaal wordt daarbij opgedeeld in intervallen, welke worden gebruikt om de toetsadviezen naar de scholen en leerlingen te communiceren.

Kort samengevat: De IRT1 methode maakt gebruik van bewezen psychometrische technieken, zoals item-responstheorie en het gezamenlijk anker, om de eindtoetsen te normeren, en om referentieniveaus en toetsadviescategorieën te geven (Figuur 1).

[16] http://www.toetsspecials.nl/html/referentiesets_openbaar/over_referentiesets.shtm



Figuur 1. Grafische weergave IRT1 normeringsmethode

In de tweede fase van de argumentgerichte benadering worden de potentiële risico's in kaart gebracht die kunnen ontstaan bij het volgen van de redenering dat toetsscores op een valide manier omgezet kunnen worden in bijpassende referentieniveaus en toetsadviescategorieën door gebruik te maken van de item-responstheorie en het gezamenlijk anker.

Hieruit volgt dat de resultaten die de IRT1 normeringsmethode oplevert niet of minder valide kunnen zijn op het moment dat het niet lukt om de item-responstheorie op een verantwoorde manier toe te passen. Rekening houdend met de voordelen en voorwaarden van deze theorie, zoals beschreven in paragraaf 3.2 spelen de volgende potentiële risico's hierbij een rol (Tabel 2):

Tabel 2.
Potentiële risico's IRT1 normeringsmethode

#	Potentieel risico
1.	Onvoldoende kwaliteit van het gezamenlijk anker
2.	Te weinig datapunten
3.	Ontbreken van een uniforme wijze van afname van het anker tussen de eindtoetsen
4.	Differential Item Functioning (DIF)[17]
5.	Invloed van de instellingen van de gebruikte software

[17] Differential Item Functioning (DIF) is een statistische eigenschap van een opgave die aangeeft in welke mate de opgave mogelijk verschillende vaardigheden meet voor leerlingen uit verschillende subgroepen, zoals bijvoorbeeld een verschil tussen jongens en meisjes.

Voor elke van deze potentiële risico's dient in kaart te worden gebracht in welke mate deze een rol spelen en welke maatregelen genomen zijn om te beargumenteren dat deze geen onoverkomelijke bezwaren geven.

Ad 1. Onvoldoende kwaliteit van het gezamenlijk anker

Op het moment dat het gezamenlijk anker van onvoldoende kwaliteit is, gaat dit ten koste van: (1) de koppeling tussen de verschillende eindtoetsen, en (2) van de koppeling met eerdere jaren en de referentieniveaus. Om dit risico te vermijden is er voor gekozen om gebruik te maken van een gezamenlijk anker van voldoende lengte en van voldoende inhoudelijke spreiding over de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen.

De toetsaanbieders integreren de jaarlijks geselecteerde subsets van ankeropgaven in hun eigen eindtoets. Daarnaast wordt de psychometrische kwaliteit van de individuele ankeropgaven en het anker als geheel uitgebreid gecontroleerd door zowel de EPO als door SCC (zie paragraaf 2.1 voor de beschrijving van de totstandkoming van het gezamenlijk anker).

Ad 2. Te weinig datapunten

Wanneer er te weinig datapunten (c.q. geldige afnames van eindtoetsen) zijn, dan kunnen de parameters van de IRT modellen niet nauwkeurig genoeg geschat worden voor de ankeropgaven en voor de aanbieder specifieke opgaven (zie ook paragraaf 3.2). Omdat de itemparameters op de complete dataset geschat worden en er voor de adaptieve eindtoetsen gebruik gemaakt wordt van gekalibreerde itembanken, is dit alleen een risico wanneer het marktaandeel van één van de aanbieders te klein wordt. Daar is op dit moment geen sprake van, omdat er een ondergrens van 1.000 afnames per eindtoets per schooljaar geldt.

Ad 3. Ontbreken van een uniforme wijze van afname van het anker tussen de eindtoetsen

Op het moment dat de wijze van afname van het anker tussen de verschillende toetsaanbieders sterk varieert, is het niet realistisch om te veronderstellen dat de eindtoetsen vergelijkbare vaardigheden meten. Risico's waaraan op dat moment gedacht kan worden zijn:

a. Andere presentatie van de ankeropgaven in verschillende eindtoetsen;
b. Afnamecondities die leerlingen in meer of mindere mate in staat stellen om hun potentie te tonen op de eindtoets;
c. Algehele motivatie en/of ervaring van deelnemende leerlingen;
d. Verschillen in onderwijs die tot andere prestaties op de (anker)opgaven leiden.

Met betrekking tot 3a kan opgemerkt worden dat er wel gebruik wordt gemaakt van drie verschillende subsets van ankeropgave. Er zitten echter verschillen in de manieren waarop de ankeropgaven aangeboden worden in de eindtoetsen. Zo zijn bepaalde eindtoetsen (deels) adaptief en/of (deels) lineair. Daarnaast worden de ankeropgaven op verschillende plaatsen in de toetsen gepositioneerd. Zie hiervoor ook de omschrijvingen in de paragrafen 2.1 en 2.2. Het is nog niet duidelijk of en eventueel welke gevolgen deze keuzemogelijkheden hebben voor het normeringsproces. Er worden in dit stadium nog geen specifieke maatregelen genomen om hier rekening mee te houden.

Voor 3b geldt dat de ankeropgaven bij sommige eindtoetsen wel meetellen en bij andere niet. Soms worden deze opgaven tijdens de toets, maar soms ook aan het einde gepresenteerd. Bovendien speelt het bij adaptieve afnames een rol dat verschillende leerlingen verschillende ankeropgaven voorgelegd krijgen en dat het aantal afnames per ankeropgave daardoor kan variëren. Ook speelt de instructie die de leerlingen krijgen bij het beantwoorden van de ankeropgaven een rol. Dit soort informatie en de mogelijke effecten ervan wordt op dit moment niet meegenomen bij de normering.

Bij het kalibreren van de ankeropgaven wordt er van uitgegaan dat de leerlingen alle opgaven uit de toets gemotiveerd beantwoorden en dat ze op hun werkelijke niveau presteren. Mochten de leerlingen minder gemotiveerd zijn bij het beantwoorden van specifiek de ankeropgaven of bijvoorbeeld minder ervaring hebben kunnen opdoen met het toets-format van de specifieke toetsaanbieder (3c) dan heeft dat tot gevolg dat zij minder goed presteren en dat de moeilijkheid van de ankeropgaven te hoog en hun vaardigheid te laag wordt ingeschat, als gevolg van verschillen in afnamecondities. Hiervoor worden nog geen specifieke maatregelen genomen.

Ten aanzien van 3d kan opgemerkt worden dat mogelijke verschillen in onderwijs, mede als gevolg van de coronapandemie en de daarmee samenhangende tijdelijke schoolsluitingen, in 2021 geleid hebben tot problemen met de normering. Daarom is er voor gekozen om in 2021 te werken met een aangepaste normering. De coronapandemie heeft er ook in schooljaar 2022 voor gezorgd dat niet meer gegarandeerd kan worden dat de leerlingen dit jaar vergelijkbaar onderwijs hebben gehad met de leerlingen in 2018 en 2019. Daarbij zijn sommige regio's in Nederland zwaarder getroffen door de coronapandemie dan andere regio's (IvhO, 2021; CBS, 2021). Omdat het marktaandeel van de toetsaanbieders varieert over de regio's, kan dit gevolgen hebben voor de verschillende normeringsresultaten. Met de NPO gelden[18] wordt weliswaar geprobeerd om mogelijke leerachterstanden zoveel mogelijk te verkleinen. Het is echter niet uit te sluiten dat dit een probleem geeft voor de toetsoverstijgende normering van de eindtoetsen.

Aanvullend dient opgetekend te worden dat veel van de verschillen in de wijze van afname die bij 3a tot en met 3d genoemd zijn, gaan over verschillen tussen de toetsaanbieders. Daarbij zijn over de jaren heen verschillen waarneembaar in de wijze waarop de verschillende aanbieders de eindtoetsen afnemen. Zo is door de jaren de positionering van de ankeropgaven bij bepaalde toetsaanbieders gewijzigd, en is bij andere aanbieders het design en/of het aantal opgaven per onderdeel aangepast.

Ter nuancering dient echter opgemerkt te worden dat de overige unieke opgaven in de verschillende eindtoetsen worden gekalibreerd met de gezamenlijke ankeropgaven. Wanneer de ankeropgaven uniform afgenomen zijn, dan heeft het voor de aanbieders in beginsel geen consequenties wanneer deze andere (toetsspecifieke) opgaven niet uniform afgenomen worden.

Ad 4. Differential Item Functioning (DIF)

Om de leerprestaties van het ene leerjaar te kunnen vergelijken met die van het andere leerjaar, is het van belang dat de ankeropgaven op een vergelijkbare wijze functioneren in beide leerjaren. Technisch

[18] Het Nationaal Programma Onderwijs (NPO) is een steunpakket van 8,5 miljard euro om schade als gevolg van corona te herstellen.

geformuleerd komt dit er op neer dat er geen Differential Item Functioning mag zijn met betrekking tot de verschillende leerjaren.

Theoretisch gezien zou het kunnen dat bepaalde ankeropgaven systematisch anders functioneren voor specifieke subgroepen. Een veel gebruikt voorbeeld is het verschil tussen jongens en meisjes bij specifieke onderwerpen zoals ballet of paardrijden. Dit zijn specifieke onderwerpen die over het algemeen meer leven bij meisjes dan bij jongens. Wanneer er (anker)opgaven opgenomen worden over dergelijke onderwerpen, kunnen meisjes in het voordeel zijn en zal de itemmoeilijkheid voor meisjes op dat moment vermoedelijk lager zijn dan die voor jongens. Iets dergelijks zou ook bij de ankeropgaven kunnen spelen wanneer bijvoorbeeld de populaties van de aanbieders onderling verschillen en dit verschil van invloed is op de itemparameter schattingen, wanneer de itemparameters per aanbieder los geschat zouden worden. Om dit te voorkomen, worden de itemparameters en toetsoverstijgende normen vastgesteld op alleen de BO populatie. De subpopulatie van S(B)O leerlingen wordt echter wel gescoord op deze itemparameters en normen.

Door EPO is een DIF analyse uitgevoerd voor de afnamejaren 2018, 2019 en 2021 (EPO, 2021). Dit om te controleren of de opgaven in 2021 anders waren gaan functioneren als gevolg van de invloed van de coronapandemie op het onderwijs. Uit deze analyse kwam naar voren dat er nauwelijks sprake was van DIF. Op basis hiervan is de verwachting verwachten dat DIF ook in 2022 geen grote rol zal spelen.

Ad 5. Invloed van de instellingen van de gebruikte software

Software om item-responsmodellen te schatten kent vele instellingen voor de gehanteerde schattingsmethoden. Om te voorkomen dat verschillen in instellingen impact hebben op de normering is het van belang om exact dezelfde instellingen te hanteren. Door gebruik te maken van één softwarepakket en daarbinnen met één schattingsmethode, met voor elke berekening dezelfde instellingen kan dit risico voorkomen worden. Bij het gebruik van de IRT1 methode is het daarom van belang dat de gebruikte instellingen, en de keuze voor de gebruikte schattingsmethode, goed gedocumenteerd worden, om te kunnen garanderen dat mogelijke verschillen tussen eindtoetsen of over verschillende leerjaren niet veroorzaakt worden door verschillende software instellingen.

Los van deze punten speelt nog mee dat er door de aanbieders gebruik gemaakt kan worden van verschillende item-responsmodellen (1PLM, OPLM of 2PLM). De IRT1 methode kan voor alle drie de modellen separaat toegepast worden. Dat garandeert alleen niet dat de normering voor de verschillende modellen resulteert in een identieke ordening van de leerlingen en in een identieke score van de leerlingen ten opzichte van de referentieniveaus (Stichting Cito Control, 2020). Voor 2021 is er voor gekozen om verschillen tussen de modellen toe te staan, zolang de verschillen in referentiescores en toetsadviescategorieën kleiner zijn dan 4%. Afhankelijk van de spreiding die in 2022 empirisch bepaald kan worden, zijn hogere waarden onder bepaalde omstandigheden niet ondenkbaar.

Daarnaast worden met de IRT1 normeringsmethode allereerst de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen genormeerd. Zoals beschreven in paragraaf 3.3, worden daarnaast eveneens de optionele en facultatieve onderdelen gekalibreerd met de IRT1 methode. Daarbij wordt het gemiddelde van de schaal voor deze optionele en facultatieve onderdelen gelijk gesteld aan het gemiddelde van de drie wettelijk verplichte onderdelen. Het achteraf door de EPO uitgevoerde onderzoek heeft laten zien dat de invloed van de extra door de aanbieders

toegevoegde onderdelen op de weging van de verschillende onderdelen en uiteindelijk op de toetsadviezen (c.q. plaatsingsadviezen) in het verleden relatief klein was (Stichting Cito Control, 2020; Stichting Cito Control, 2021a). Aangenomen wordt daarom dat dit ook in 2022 geen noemenswaardige invloed zal hebben.

In de derde en laatste fase van de argumentgerichte benadering worden de genoemde potentiële risico's geëvalueerd. Daartoe zijn in dit onderzoek de risico's, een inschatting van de kans op deze risico's en de reeds voorgenomen maatregelen weergegeven in Tabel 3.

Tabel 3.
Risicoanalyse IRT1 normeringsmethode

Risico	Kans op het risico	Voorgenomen maatregelen
1	Minimaal	Expliciete toewijzing ankeropgaven en psychometrische controle
2	Minimaal	Gebruik complete dataset en gekalibreerde itembanken
3a	Reëel	Geen maatregelen
3b	Reëel	Geen maatregelen
3c	Reëel	Geen maatregelen
3d	Reëel	Extra ondersteuning onderwijs met NPO gelden
4	Minimaal	Geen maatregelen, check achteraf
5	Minimaal	Duidelijke specificatie gebruikte instellingen
IRT modellen	Klein	Geen maatregelen, jaarlijks te bepalen maximale afwijking toegestaan
Extra onderdelen	Klein	Geen maatregelen, check achteraf

De samenvatting in Tabel 3 laat zien dat niet alle bezwaren tegen de redenering dat met de IRT1 normeringsmethode de toetsscores op een valide manier omgezet kunnen worden in referentieniveaus en toetsadviescategorieën, weggenomen kunnen worden. De openstaande bezwaren die mogelijk een rol spelen hebben allemaal te maken met de wijze van afnemen van de ankeropgaven. Strikt genomen wordt daarmee niet voldaan aan de eisen voor validiteit.

Omdat de aanbieders de wijze van afname over de jaren heen wel zo veel mogelijk gestandaardiseerd hebben en het niet mogelijk is om de gevolgen van de coronapandemie voor het onderwijs goed in te schatten concludeert RCEC dat de IRT1 normeringsmethode voldoende valide is om toe te passen bij de normering van de eindtoetsen voor 2022, mits gekeken wordt of er geen grote verschillen tussen de schooladviezen en de toetsadviezen zijn ontstaan, bijvoorbeeld ten gevolge van de coronapandemie.

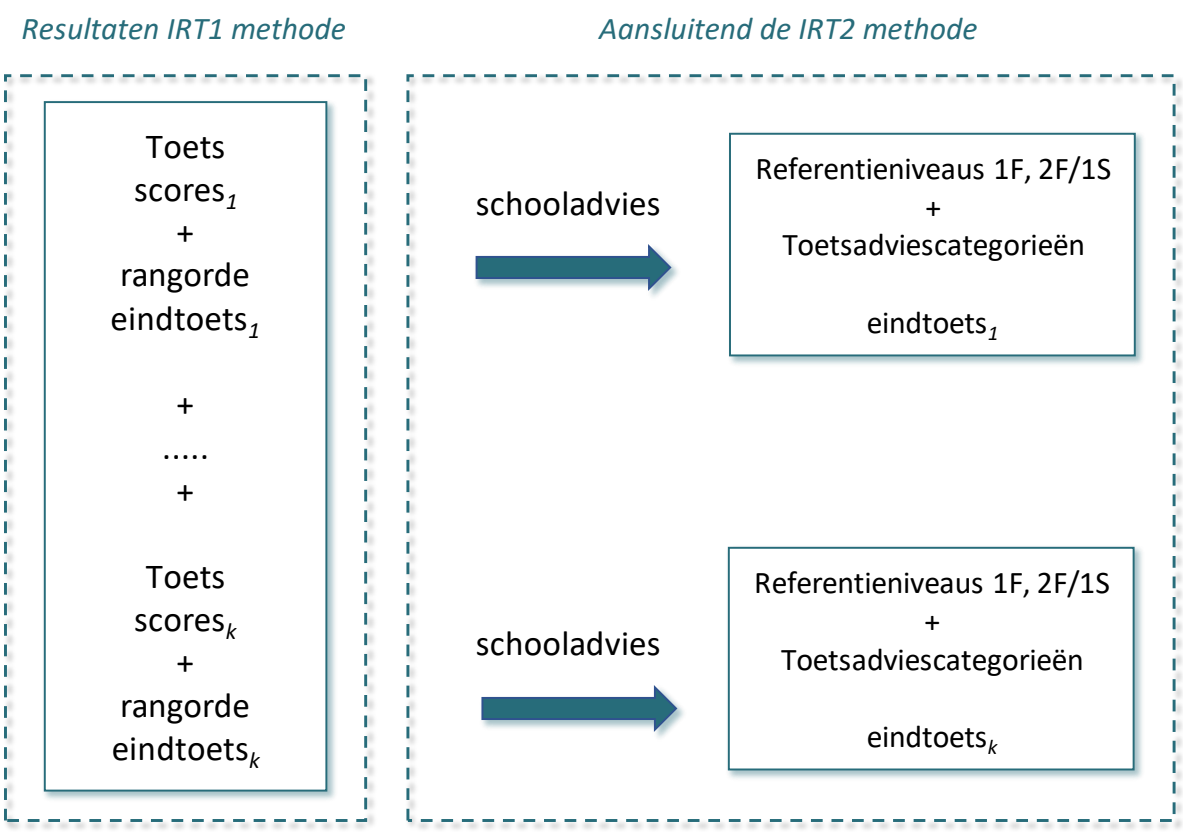
3.4. IRT2 normeringsmethode

Volgend op de IRT1 methode wordt binnen het huidige stelsel eindtoetsen de toetsoverstijgende IRT2 normeringsmethode toegepast. De aanleiding voor de IRT2 methode is dat uit de resultaten van EPO en SCC volgt dat de IRT1 methode nog niet voor alle eindtoetsen een betrouwbaar beeld (c.q. een valide vergelijking) geeft. Zoals beschreven in paragraaf 3.2 bestaat het vermoeden dat dit samenhangt met de wijze waarop de gezamenlijke ankeropgaven in de verschillende eindtoetsen worden gebruikt. De locatie van de ankeropgaven in de verschillende eindtoetsen lijkt van invloed te zijn op de kwaliteit van de kalibratie.

Om een voor alle eindtoetsen zo betrouwbaar mogelijk beeld van de resultaten te bereiken, is de IRT2 normeringsmethode geïntroduceerd. Zoals hierboven reeds genoemd, wordt de IRT2 na afronding van de IRT1 methode uitgevoerd. De IRT2 methode is derhalve geen zelfstandige methode, maar een methode welke voortbouwt op de resultaten van de IRT1 methode. De IRT2 methode is te kenmerken als een normeringsmethode waarbij voor de equivalering gebruik wordt gemaakt van de voorlopige schooladviezen en bijbehorende populatieverdeling en van de schaalwaarden van de Centrale Eindtoets (CET). Uit de eerdere analyses van SCC volgt dat deze methode een betrouwbaar beeld (c.q. een valide vergelijking) voor alle eindtoetsen geeft. De methode heeft er echter de schijn van dat het een vorm is van relatief normeren, in plaats van absoluut normeren. Daarom vervolgt dit hoofdstuk met een beschrijving van het uitgevoerde onderzoek naar de mate van validiteit van de IRT2 methode. In dit onderzoek zijn opnieuw de principes van de argumentgerichte benadering van validiteit toegepast. Zo wordt volgens dezelfde onderzoeksmethode als gebruikt voor het evalueren van de IRT1 methode, de mate van validiteit van de IRT2 normeringsmethode beoordeeld en geëvalueerd.

De eerste fase van de argumentgerichte benadering betreft het formuleren van de redenering (c.q. de opbouw) van de te onderzoeken normeringsmethode. De IRT2 normeringsmethode maakt gebruik van de vaardigheidsschattingen die resulteren uit de IRT1 normering. Daarbij komt de IRT2 methode op een eigen wijze tot de cesuren voor de referentieniveaus en de toetsadviescategorieën. De rangorde van de leerlingen blijft daarbij hetzelfde als bij de IRT1 normering. De grenswaarden, op basis waarvan de referentieniveaus en de toetsadviescategorieën worden bepaald, komen bij de IRT2 methode tot stand door gebruik te maken van de verdeling van de schooladviezen van de verschillende aanbieders. Daarvoor wordt met behulp van software gesimuleerd hoe de populatie van de aanbieder gepresteerd zou hebben op de PCET van het jaar van afname. Uitgaande daarvan worden de cesuurpunten per aanbieder bepaald. Dit betekent dat toepassing van de IRT2 normeringsmethode unieke cesuurpunten per eindtoets oplevert, die alleen geldig zijn binnen één afnamejaar, en alleen a posteriori (c.q. achteraf beschouwd) te bepalen zijn.

Kort samengevat: De IRT2 methode maakt gebruik van de vaardigheidsschattingen en rangorde van leerlingen uit de IRT1 normering, en bepaalt per eindtoets apart de grenswaarden voor de referentieniveaus en de toetsadviescategorieën, rekening houdend met de prestaties van de populatie per aanbieder (Figuur 2).



Figuur 2. Grafische weergave IRT2 normeringsmethode

Volgend op bovenstaande beschrijving van de redeneringen van de IRT2 methode, dienen conform de tweede fase van de argumentgerichte benadering de gevolgen en potentiële risico's van deze normeringsmethode in kaart gebracht te worden.

Het aansluitend op de IRT1 methode toepassen van de IRT2 normeringsmethode heeft twee belangrijke gevolgen. Het eerste gevolg is dat de verschillende toetsaanbieders eigen grenswaarden (c.q. cesuren) aangeleverd krijgen die passen bij de eigen populatie. Dit zal naar verwachting de acceptatiegraad van de resulterende toetsadviezen verhogen. Een tweede gevolg is dat hierdoor de interpretatie van de behaalde referentieniveaus verandert. Dit betekent dat om de IRT2 methode op een valide wijze toe te kunnen passen, er wel een aantal risico's zijn aan te wijzen waar maatregelen voor genomen dienen te worden (Tabel 4):

Tabel 4.
Potentiële risico's IRT2 normeringsmethode

#	Potentieel risico
1.	Aanname dat er geen aanbieder-specifieke bias in de schooladviezen zit
2.	Aanname dat de data waarover schooladviezen berekend zijn representatief is voor de complete data
3.	Cesuurpunten leiden tot kleine afwijkingen t.o.v. de streefpercentages vanwege het gebruikte IRT model
4.	PCET ongevoelig aan de onderkant van de vaardigheidsschaal
5.	Onder- of overschattingen van de prestaties door mismatch populatie aanbieder en PCET
6.	Afhankelijkheid van IRT1 methode

Voor elke van deze potentiële risico's dient in kaart te worden gebracht in welke mate deze een rol spelen en welke maatregelen genomen zijn om te beargumenteren dat deze geen onoverkomelijke bezwaren geven.

Ad 1. Aanname dat er geen aanbieder-specifieke bias in de schooladviezen zit

Op het moment dat er een specifieke bias (c.q. een systematische vertekening) zit in de schooladviezen van een specifieke aanbieder, leidt de IRT2 normeringsmethode tot vertekende cesuren. Om dit effect te onderzoeken zou er een doorstroomonderzoek voor alle toetsaanbieders uitgevoerd dienen te worden. Daarnaast zou het ook zo kunnen zijn dat er een aanbieder-specifieke bias is ontstaan vanwege regionale verschillen in het marktaandeel van de aanbieder en vanwege het gegeven dat er regionale verschillen zijn geweest in de impact van de coronapandemie (IvhO, 2021; CBS, 2021). Op dit moment lijken er bij de toetsoverstijgende IRT2 normering nog geen specifieke maatregelen te worden genomen om hiervoor te controleren.

Ad 2. Aanname dat data waarover schooladviezen berekend zijn representatief is voor complete data

Met name voor de IEP eindtoets wordt de IRT2 methode toegepast op een gedeelte van de dataset. Dit omdat vanwege het handmatig nakijken er, op het moment van het berekenen van de normen, slechts gegevens van een derde van het aantal respondenten beschikbaar zijn. Wanneer de aangeleverde data niet representatief mocht blijken te zijn, dan leidt dit tot vertekende cesuren en zal dat leiden tot een vertekend beeld voor de rest-populatie van de IEP eindtoets. Op dit moment is het niet duidelijk in hoeverre dit een risico is en lijken hiervoor nog geen specifieke preventieve maatregelen te worden genomen.

Ad 3. Cesuurpunten leiden tot kleine afwijkingen t.o.v. streefpercentages door gebruikte IRT model

Wanneer toetsaanbieders gebruik maken van het 1PL model, bestaat het risico op het effect dat grote clusters leerlingen met dezelfde vaardigheid zich rondom de cesuurpunten concentreren. Hierdoor kunnen mogelijk afwijkingen van de streefpercentages ontstaan. Hiervoor lijken nog geen aparte beheersmaatregelen te worden genomen.

Ad 4. PCET ongevoelig aan de onderkant van de vaardigheidsschaal

Op basis van historische gegevens, is het bekend dat er een relatief klein aantal pro-leerlingen meedoet aan de PCET. Dat heeft tot gevolg dat de cesuur tussen de toetsadviescategorieën pro/vmbo-bb en vmbo-bb/kb op een, verhoudingsgewijs, erg laag percentage uitkomt. Voor een juiste toepassing van de IRT2 methode is het daarom nodig dat voor de andere aanbieders een extra stap aan de methode wordt toegevoegd, zoals beschreven op pagina 6 in het document *Advies normering Eindtoetsen, 15 mei 2021, Ref.nr.: 21.01* (EPO, 2021).

Ad 5. Onder- of overschattingen van de prestaties door mismatch populatie aanbieder en PCET

Wanneer de populatie van een bepaalde toetsaanbieder sterk afwijkt van de populatie van de PCET, kan ook dat tot vertekeningen leiden. De IRT2 methode bepaalt de cesuren door de verdeling van leerlingen te simuleren voor de populatieverdeling van de PCET. Wanneer er relatief gezien met te veel of juist te weinig leerlingen uit de PCET populatie wordt gesimuleerd, kan dit resulteren in onevenwichtige cesuurpunten. Onderzoek van de 2021 basisonderwijs populatie laat echter zien dat dit voor het grootste gedeelte van de berekende cesuren geen probleem geeft.

Ad 6. Afhankelijkheid van IRT1 methode

De toepassing van de IRT2 methode is direct afhankelijk van de toepassing van de IRT1 methode. Op het moment dat er bij het toepassen van de IRT1 methode één of meerdere problemen ontstaan, bijvoorbeeld ten gevolge van de coronapandemie, dan zal dit direct doorwerken in de resultaten van de IRT2 methode. Wanneer er bij de IRT2 methode wordt genormeerd op basis van de PCET afname van dit jaar, verdient dit punt extra aandacht. Omdat er regionale verschillen zijn waar te nemen tussen het marktaandeel van iedere eindtoetsaanbieder en er blijkt dat sommige regio's zwaarder getroffen zijn door de coronapandemie dan andere (IvhO, 2021; CBS, 2021), zou dit tot verschillen kunnen leiden.

Afrondend worden conform de derde en laatste fase van de argumentgerichte benadering de genoemde potentiële risico's van de IRT2 normeringsmethode geëvalueerd. Daartoe zijn in Tabel 5 de risico's, een inschatting van de kans op deze risico's en de reeds voorgenomen maatregelen weergegeven.

Tabel 5.
Risicoanalyse IRT2 normeringsmethode

Risico	Kans op het risico	Voorgenomen maatregelen
1	Klein	Geen
2	Klein	Geen
3	Klein	Geen
4	Reëel	Aanpassing cesuren pro-leerlingen
5	Klein	Controle achteraf
6	Reëel	Correctie cesuren voor specifieke problemen zoals bv. corona

Bij de beschrijving van de IRT2 methode is al aangegeven dat de methode resulteert in unieke normen per toetsaanbieder die gelden voor specifiek dit schooljaar. Gegeven de aannames die ten grondslag liggen aan de IRT2 methode, zijn de cesuren en streefpercentages vergelijkbaar via de koppeling die is gelegd met de PCET. De scores van de individuele leerlingen ten aanzien van de referentieniveaus dienen geïnterpreteerd te worden alsof zij de PCET hebben gemaakt.

Bij het beoordelen van de validiteit van de IRT2 normeringsmethode dient daarom geconstateerd te worden dat de methode de resultaten van individuele leerlingen over de jaren heen op valide wijze kan vergelijken, mits wordt voldaan aan de aannames die ten grondslag liggen aan de IRT2 methode. De methode lijkt geschikt voor het bepalen van de schooljaar specifieke toetsadviezen. En de methode lijkt eveneens geschikt om via de koppeling met de PCET afname van dit jaar uitspraken te doen over de behaalde referentieniveaus van individuele leerlingen over de jaren heen.

Kenmerkend voor validering volgens de argumentgerichte benadering, is dat er wordt uitgegaan van een validering met het oog op een specifiek doel. Voor 2022 hebben de betrokken partijen het doel van de methode geformuleerd als het genereren van eindtoets specifieke toetsadviezen en cesuren voor de referentieniveaus voor het schooljaar 2022. Voor dit specifieke doel concludeert RCEC dat er sprake is van een voldoende valide methode mits er achteraf wordt gecontroleerd voor de specifieke risico’s genoemd bij ad 3 en ad 5 en mits de normering wordt aangepast voor pro-leerlingen (ad 4). Indien de resultaten er aanleiding toe geven, dient er daarnaast te worden gecorrigeerd voor de effecten van de coronapandemie.

In het volgende hoofdstuk wordt verslag gedaan van de mate waarin de huidige werkwijze van de IRT1 en IRT2 normeringsmethode voldoen aan de proceseisen voor valide normering.

4. Bruikbaarheid IRT1 en IRT2 normeringsmethode

4.1. Inleiding

Dit hoofdstuk behandelt de resultaten van het tweede deelonderzoek. Centraal daarin staat de vraag in welke mate de huidige werkwijze van de IRT1 en IRT2 normeringsmethode voldoet aan de proceseisen voor valide normering. Het doel van de tweede deelvraag is om de mate van bruikbaarheid van beide normeringsmethoden te evalueren.

Het tweede deelonderzoek concentreert zich op de voorgestelde werkwijze van SCC voor wat betreft de normering 2022. In deze studie is geen gedetailleerd onderzoek gedaan naar de wijze waarop de bestaande werkafspraken tussen de betrokken partijen tot nu toe in de praktijk zijn uitgevoerd. Ook is er geen onderzoek gedaan naar de werking en de algoritmes van de DexterMML software en van het door SCC geprogrammeerde R script voor de IRT2 normering.

Dit deelonderzoek is uitgevoerd volgens dezelfde principes van de argumentgerichte benadering, zoals beschreven in paragraaf 3.1. Het proces van de toetsoverstijgende normering is daarbij opgedeeld in drie opeenvolgende deelprocessen: (1) de processtappen van de totstandkoming van de eindtoetsen, zoals beschreven in paragraaf 2.1 en 2.2, (2) de processtappen van de IRT1 normeringsmethode, zoals beschreven in paragraaf 3.3, en (3) de processtappen van de IRT2 normeringsmethode, zoals beschreven in paragraaf 3.4.

In paragraaf 4.2 worden, conform de eerste fase van de argumentgerichte benadering, de huidige stappen van de drie deelprocessen in kaart gebracht volgens de indeling van de procesmodellering[19] (Aguilar-Saven, 2004). Een juiste redenering en een transparante beschrijving, uitvoering, controle en bijsturing draagt bij aan kwalitatieve en valide ontwikkelde, afgenomen en genormeerde eindtoetsen. Daarnaast vergroot dit het inzicht, de duidelijkheid en het overzicht voor alle betrokkenen. Dit onderzoek concentreert zich op de modellering van de primaire processen[20] op hoofdlijn. Ondersteunende en besturende processen zijn buiten beschouwing gelaten.

Na de beschrijving van de redenering, worden in paragraaf 4.3 de vijf potentiële risico's van de huidige procesbeschrijving in kaart gebracht, zoals bekend vanuit de procesmodellering. Het consistent voorkomen van deze risico's draagt bij aan een kwalitatieve structuur en cultuur binnen het stelsel eindtoetsen 2022. Tenslotte worden in paragraaf 4.4, conform de derde fase van de argumentgerichte benadering, de potentiële risico's geëvalueerd.

Samengevat resulteren de drie fasen van de argumentgerichte benadering (Kane, 1992; Kane, 2004; Wools, 2012) in een conclusie over de mate van bruikbaarheid van de drie deelprocessen.

[19] Procesmodellering (c.q. Business Process Modelling; BPM) is een verzameling methoden en technieken om bedrijfsprocessen in kaart te brengen. Deze modellen worden onder meer gebruikt in de bedrijfskunde, met name in het projectmanagement en in de analyse en ontwerp van informatiesystemen.

[20] Primaire processen zijn die processen die rechtstreekse bijdragen aan de realisatie en de normering van de eindtoetsen.

4.2. Beschrijving primaire processen

Zoals beschreven in het document *Governance, Normering Eindtoetsen 2022, versie 0.6* (Stichting Cito Control, 2022) zijn door alle betrokken partijen de volgende principeafspraken gemaakt over samenwerking in de keten (OCW, 2021): (1) iedereen telt mee, (2) iedereen doet mee, (3) we houden oog voor de gebruikers, (4) combinatie van psychometrie, onderwijskunde, beleid en praktijk, en (5) OCW voert de regie. Met deze principeafspraken wordt bewerkstelligd dat de normering voor alle eindtoetsen zoveel mogelijk op een eenzelfde manier tot stand komt. Daarbij zijn binnen het stelsel eindtoetsen 2022 onderstaande partijen met de volgende verantwoordelijkheden benoemd (Tabel 6):

Tabel 6.
Betrokken partijen en verantwoordelijkheden

Partij	Verantwoordelijkheden
Stichting Cito (SC)	1. Verantwoordelijk voor de ontwikkeling van de CET en de advisering omtrent de normering van de CET
Private Eindtoetsaanbieders (EA)	1. Opstellen van inhoudelijk valide, betrouwbare en deugdelijk genormeerde eindtoets 2. Committeert zich aan de toetsoverstijgend vastgestelde normering 3. Voert eigenstandige analyses uit (plus SC voor wat betreft de CET) 4. Met betrekking tot de toetsoverstijgende normering: a. verantwoordelijk voor het tijdig en correct aanleveren van de data volgens een van tevoren afgestemd format (plus SC voor wat betreft de CET) b. bereikbaar voor eventuele vragen van SCC (plus SC voor wat betreft de CET) c. kan tijdens de SG vergadering instemmen met het advies van de werkgroep
College voor Toetsen en Examens (CvTE)	1. Verantwoordelijk voor het aanbod en de kwaliteit van de CET 2. Verantwoordelijk voor het vaststellen van de normering van de CET
Stichting Cito Control (SCC)	1. Is als primaire rekenpartij verantwoordelijk voor de normeringsanalyse: a. Proefdraaien van de normering b. Ontwikkeling van één geharmoniseerd normeringsmodel c. Geeft een advies over de normering aan de stuurgroep (SG) 2. Stelt als planner een plan van aanpak op voor het normeringsproces 2021/2022 3. Stelt de betrouwbaarheid van de toetsoverstijgende normering vast 4. Psychometrisch adviseur voor de toetsoverstijgende normering
Expertgroep Toetsen PO (EPO)	1. Controleert en houdt toezicht op de onderwijskundige en psychometrische kwaliteit van de door EA ontwikkelde eigen opgaven en ankeropgaven 2. Verantwoordelijk voor de toelating van de eindtoetsen van private aanbieders tot het stelsel 3. Voert als controlerende rekenpartij een check uit op de normeringsanalyse van SCC: a. Vooraf accorderen van de DexterMML software b. Achteraf controleren van de SCC output uit DexterMML c. Optioneel aanvullend met eigen software controleren van DexterMML d. Delen van de bevindingen met SCC

	4. Valideert het oordeel van SCC over de betrouwbaarheid van de normering 5. Psychometrisch en onderwijskundig adviseur voor de toetsoverstijgende normering
Stuurgroep (SG)	1. Besluit over een betrouwbare en deugdelijke normering na advies van de werkgroep a. Wanneer er geen consensus is, neemt OCW het besluit
Ministerie OCW (OCW)	1. Formeel stelselverantwoordelijk en regievoerder 2. Formeel vaststellen van cesuren referentieniveaus en toetsadviescategorieën 3. Overkoepelend verantwoordelijk voor het gehele proces 4. Faciliteert de SG om tot besluitvorming te komen 5. Stelt werkafspraken op ten aanzien van verslaglegging, vaststelling en opvolging van afspraken, bijsturing en communicatie 6. Stelt aanvullende werkafspraken op voor de normering: a. Wijze van aanlevering van data en documentatie door EA b. Wijze van afstemming SCC en EA ten aanzien van data en documentatie c. Wijze waarop SCC in samenspraak met EA tot besluiten komt d. Wijze van delen van de DexterMML software door SCC met EA e. Wijze waarop EA scholing krijgt van SCC

Wanneer er binnen het stelsel geen sprake is van een happy flow, kunnen SC, CVTE, EA, EPO, SCC en/of SG dit aangeven bij OCW. OCW geeft betrokken partijen een specifieke periode de tijd om tot een verklaring / oplossing te komen. Wanneer dit niet lukt, besluit OCW als stelselverantwoordelijke hoe men verder dient te gaan in het proces.

Om te komen tot een toetsoverstijgende normering volgens achtereenvolgens de IRT1 en de IRT2 methode, doorlopen de betrokken partijen gezamenlijk een drietal opeenvolgende deelprocessen:

1. de processtappen van de totstandkoming van de eindtoetsen (paragraaf 2.1 en 2.2);
2. de processtappen van de IRT1 normeringsmethode (paragraaf 3.3), en;
3. de processtappen van de IRT2 normeringsmethode (paragraaf 3.4).

Per deelproces worden de primaire processen op hoofdlijn beschreven. De beschrijving volgt de indeling van het processchema Business Process Modelling (Aguilar-Saven, 2004), waarin achtereenvolgens aan bod komen: (1) de type en de herkomst van de bron die dient als input voor de processtap, (2) de beschrijving van de verschillende activiteiten van de betreffende processtap, (3) de wijze waarop de activiteiten van de processtap worden gecontroleerd, en (4) het type en de bestemming van de resultaten die fungeren als de output (c.q. de opbrengst) van de processtap.

Ad 1. Deelproces totstandkoming eindtoetsen

Iedere toetsaanbieder ontwikkelt eigen toetsopgaven voor de onderdelen en domeinen van de eigen eindtoets. Daarnaast ontwikkelt iedere toetsaanbieder periodiek nieuwe ankeropgaven voor het gezamenlijk anker. De eigen opgaven en de ankeropgaven dienen beide te voldoen aan de inhoudelijke eisen uit het Beoordelingskader Eindtoetsen (EPO, 2020). De toetsaanbieder legt de nieuwe (anker)opgaven ter controle voor aan de inhoudsdeskundigen van de EPO. De EPO koppelt de bevindingen van de inhoudelijke beoordeling terug aan de toetsaanbieder.

De toetsaanbieder pretest vervolgens de nieuw ontwikkelde (anker)opgaven en kiest daarbij zelf voor een proefopstelling of voor het zaaien in een operationele afname. Het Beoordelingskader Eindtoetsen (EPO, 2020) schrijft voor aan welke psychometrische eisen de pretest steekproef moet voldoen. De toetsaanbieder kalibreert aansluitend de wettelijk verplichte onderdelen en domeinen met het bestaande gezamenlijk anker.

De periodiek ontwikkelde en gepreteste ankeropgaven deelt de toetsaanbieder met de EPO. De EPO kalibreert de nieuwe ankeropgaven bij het bestaande gezamenlijke anker en controleert daarbij onder andere op DIF. Vervolgens stelt SCC, indien nodig, de cesuren voor de referentieniveaus en de toetsadviezen bij. Aanvullend selecteert elke toetsaanbieder ieder schooljaar de subsets van de ankeropgaven en voegt deze toe aan de eindtoets van het betreffende schooljaar.

De toetsaanbieder doet voor wat betreft de eigen ontwikkelde opgaven een interne kwaliteitsbeoordeling van de pretestresultaten en stelt vervolgens de nieuwe eindtoets samen. De toetsaanbieder zorgt ervoor dat de eindtoets de door de EPO geselecteerde subsets ankeropgaven bevat en dat de eindtoets de cesuren voor de referentieniveaus en de toetsadviezen kan meten. De samengestelde toets wordt eerst beoordeeld door de EPO, alvorens deze mag worden aangeboden aan de scholen.

Vervolgens laat de toetsaanbieder de eindtoets afnemen door de scholen, kijkt deze zelf na en scoort zelf de antwoorden. Van de resultaten stelt de toetsaanbieder een dataset met ruwe scores samen en verstrekt deze aan SCC. Hierna start SCC met de processtappen van de IRT1 normeringsmethode.

Ad 2. Deelproces IRT1 normeringsmethode

Uit de ontvangen schriftelijke documentatie en de op 21 januari 2022 door SCC verzorgde presentatie, heeft RCEC op hoofdlijn de processtappen van de IRT1 normeringsmethode kunnen vaststellen. Met de huidige kennis van zaken is hier het volgende primaire proces uit af te leiden.

Het IRT1 normeringsproces start met het verzamelen door de toetsaanbieder van alle ruwe scores in het format dataset. De toetsaanbieder verstrekt dit format als input aan SCC. Uit het format is direct af te lezen of het minimale aantal afnames van 1.000 leerlingen per eindtoets is behaald.

Zoals beschreven in paragraaf 3.3 maakt de IRT1 normeringsmethode gebruik van de mogelijkheid om scores op de verschillende eindtoetsen onderling vergelijkbaar te maken, door het gezamenlijk anker te implementeren en door gebruik te maken van de item-responstheorie. De hiervoor door SCC te maken berekeningen worden uitgevoerd in DexterMML. RCEC heeft niet de werking en de algoritmes van de software gecontroleerd. Wel heeft RCEC vast kunnen stellen dat de software verschillende kenmerken heeft, die het normeringsproces ten goede komen. Zo is de broncode vrij toegankelijk en daarmee reproduceerbaar. De software gebruikt standaard R methoden, wat de kans op fouten tijdens het inlezen van ruwe data beperkt. En de software heeft een hoge verwerkingssnelheid, waarmee de analyse van de ruwe data in korte tijd kan worden gedaan. Verder worden de resultaten (output) teruggegeven in zogenaamde dataframes, wat verdere bewerking van de resultaten mogelijk maakt. Ten slotte ondersteunt DexterMML negatieve discriminatieparameters in 2PLM en kent het een objectief convergentiecriterium. Hierdoor is het schattingsalgoritme niet afhankelijk van subjectieve instellingen door de gebruiker.

Uit de op 21 januari 2022 door SCC verzorgde presentatie volgt eveneens een gedetailleerde weergave van het proces van kalibreren van de itemparameters, de cijfermatige en grafische weergave van de resultaten, en de wijze van toetskalibratie en cesuurbepaling van de referentieniveaus en toetsadviescategorieën.

In het gezamenlijk proces is afgesproken dat de EPO de normeringsresultaten van SCC controleert. Met het vaststellen van de cesuren voor de referentieniveaus en de toetsadviescategorieën is de IRT1 normeringsprocedure afgerond. Ter voorbereiding op de SG vergadering wordt eveneens de IRT2 normering alvast doorgerekend. In de SG vergadering zelf wordt vervolgens op basis van een beslisprocedure gekozen voor de IRT1 of de IRT2 normeringsprocedure voor alle private toetsaanbieders.

Ad 3. Deelproces IRT2 normeringsmethode

Evenals geldt voor de IRT1 normeringsmethode, heeft RCEC uit de ontvangen schriftelijke documentatie en de op 21 januari 2022 door SCC verzorgde presentatie op hoofdlijn de processtappen van de IRT2 normeringsmethode kunnen vaststellen. Op basis van de huidige informatie is hier het volgende primaire proces uit af te leiden.

Het IRT2 normeringsproces start bij de vaardigheidsschattingen die volgen uit de IRT1 normeringsmethode. De leerlingendata van de private toetsaanbieders wordt aangevuld met de voorlopige schooladviezen van de leerkracht. Daarnaast wordt als extra bron (input) de leerlingendata van de PCET met de schooladviezen en toetsresultaten van de eigen normering toegevoegd.

In een separaat door SCC geprogrammeerd R script worden vervolgens de cesuren voor de referentieniveaus en de toetsadviescategorieën bepaald. De rangorde van de leerlingen blijft daarbij hetzelfde als bij de IRT1 normering. De werking van de software is op 21 januari 2022 door SCC gepresenteerd. RCEC heeft daarbij niet de werking en de algoritmes van de software gecontroleerd.

De output van de IRT2 normeringsmethode is te omschrijven als de verwachte prestatie van de individuele leerling, wanneer hij of zij de PCET gemaakt zou hebben, waarbij de IRT1 methode als output de getoonde prestatie op de gemaakte toets geeft. De keuze voor de IRT1 of IRT2 output is procesmatig dan ook niet zozeer een keuze op basis van psychometrie of modelfit maar een keuze voor aannemelijkheid van de aannames en een keuze voor een rechtvaardige presentatie van de vaardigheid van de leerling. Voordat de IRT2 normeringsresultaten met iedere toetsaanbieder worden gedeeld, worden de resultaten gecontroleerd door de EPO.

4.3. Potentiële risico’s primaire processen

In de tweede fase van de argumentgerichte benadering worden de potentiële risico’s van de procesmodellering in kaart gebracht. Het processchema Business Process Modelling (Aguilar-Saven, 2004) onderscheidt de volgende kritieke factoren (Tabel 7):

Tabel 7.
Potentiële risico's deelprocessen

#	Potentieel risico
1.	Het proces heeft geen duidelijk begin (input) en/of duidelijk einde (output)
2.	Het resultaat is niet gericht op een interne of externe betrokkene bij het stelsel
3.	Het proces voegt geen waarde toe voor de betrokkenen
4.	Het proces bevat geen (chronologische) ordening van activiteiten
5.	Het proces bevat geen of onvoldoende controlemomenten

Voor elk van deze potentiële risico's dient te worden aangetoond in welke mate deze een rol spelen en welke maatregelen in het proces zijn ingebouwd om te beargumenteren dat deze geen onoverkomelijke bezwaren geven. De potentiële procesrisico's worden in de tabellen 8 tot en met 10 per deelproces genoemd en voorzien van de betreffende risicocode(s).

Tabel 8.
Risicoanalyse deelproces 1: totstandkoming eindtoetsen

Potentieel risico	Classificatie
In het Beoordelingskader Eindtoetsen (EPO, 2020) zijn vooralsnog geen inhoudelijke kwaliteitseisen voor vrij te kiezen optionele (of facultatieve) onderdelen en domeinen opgenomen.	1 + 5
De toetsaanbieder kiest zelf het gewenste IRT model, de gewenste schattingsmethode en de gewenste software voor het kalibreren van gepreteste opgaven.	1
De toetsaanbieder kiest daarnaast een eigen wijze van kalibreren van opgaven van vrij te kiezen optionele (of facultatieve) onderdelen en domeinen.	1 + 4 + 5
De betrokkenen werken met een vast format voor aanlevering van ruwe scores aan SCC. Het format is op onderdelen nog niet volledig in afstemming met de voor de IRT1 en IRT2 normeringsmethode benodigde informatie (zie ook Tabel 9).	1
Er lijken nog geen procesafspraken te zijn gemaakt waarmee een uniforme afname van de verschillende eindtoetsen wordt gegarandeerd.	4

Tabel 9.

Risicoanalyse deelproces 2: IRT1 normeringsmethode

Potentieel risico	Classificatie
Op basis van de beschikbare documentatie worden in het dataformat voor het aanleveren van de ruwe scores, voor zover in dit onderzoek was vast te stellen, nog niet de volgende gegevens (informatiebronnen) verzameld:	
1. welke ankeropgaven wel en niet meetellen;	1 + 4
2. op welke plaats in de toets welk deel van de ankeropgaven is gepositioneerd;	1 + 4
3. welke instructies de leerlingen hebben gekregen ten aanzien van de ankeropgaven en in welke mate dat de motivatie kan hebben beïnvloed;	2 + 5
4. welke antwoordpatronen horen bij leerlingen uit het regulier basisonderwijs, en welke antwoordpatronen horen bij leerlingen uit het s(b)o Nb. Dit is wel al een vereiste aanlevering;	1
5. welk deel van de totale afnames per dataformat is ingeleverd en welk deel nog later komt, omdat dat later wordt nagekeken;	1 + 4
6. per leerling: het voorlopige schooladvies van de leerkracht (ten behoeve van de IRT2 normeringsmethode). Nb. Dit is wel al een vereiste aanlevering;	1
7. welk IRT model en welke schattingsmethode zijn gebruikt bij het vooraf kalibreren van de toetsopgaven;	4
8. welke software de toetsaanbieder heeft gebruikt voor de itemkalibratie en voor het berekenen van de ruwe toetsscores.	1

Tabel 10.

Risicoanalyse deelproces 3: IRT2 normeringsmethode

Potentieel risico	Classificatie
Er is nog geen verantwoordelijkheid benoemd voor wie er controleert of de voorlopige schooladviezen van de leerlingen juist en volledig zijn opgenomen in het door de toetsaanbieder aan te leveren databestand.	1 + 4
Er zijn nog geen formele afspraken gemaakt over de wijze waarop de EPO de IRT1 en/of IRT2 berekeningen en/of output van SCC controleert. Ook is nog niet beschreven of en hoe SCC bezwaar kan aantekenen en/of hoeveel tijd SCC krijgt om een of meerdere berekeningen opnieuw uit te voeren. En hoe tweede controle vervolgens plaatsvindt.	2 + 3 + 5

4.4. Evaluatie bruikbaarheid

In de laatste fase van de argumentgerichte benadering worden de in paragraaf 4.3 genoemde procesrisico's geëvalueerd. Daartoe is in dit onderzoek geturfd hoe vaak welk van de vijf potentiële risico's worden aangewezen. Deze telling, zoals gepresenteerd in Tabel 11 resulteert in een berekende kans op het voorkomen van het betreffende risico.

Tabel 11.

Risicoanalyse primaire processen

#	Potentieel risico	Observaties	Kans op het risico
1.	Het proces heeft geen duidelijk begin (input) en/of duidelijk einde (output)	11x	Reëel
2.	Het resultaat is niet gericht op een interne of externe betrokkene bij het stelsel	2x	Klein
3.	Het proces voegt geen waarde toe voor de betrokkenen	1x	Minimaal
4.	Het proces bevat geen (chronologische) ordening van activiteiten	7x	Reëel
5.	Het proces bevat geen of onvoldoende controlemomenten	4x	Klein

Uit de risicoanalyse volgt dat er in de processen op meerdere momenten een reëel risico bestaat dat het type en de herkomst van de bron die dient als input voor de processtap en/of het type en de bestemming van de resultaten die fungeren als de output (c.q. de opbrengst) van de processtap onvoldoende beschikbaar of betrouwbaar zijn. Tevens valt op dat op meerdere momenten de chronologische ordening van activiteiten onduidelijk en/of onvolledig lijkt te zijn. Formeel wordt daarmee niet voldaan aan de eisen voor bruikbaarheid.

Desalniettemin verdient deze conclusie enige nuance. In deze studie is nog geen gedetailleerd onderzoek gedaan naar de werkwijze van de EPO en van de toetsaanbieders. Wel is het RCEC bekend dat de EPO in 2021 een draaiboek heeft ontwikkeld en deze ter beschikking heeft gesteld aan alle betrokkenen. Voor het jaar 2022 zal dit draaiboek worden bijgewerkt. Beide partijen dragen in het proces daarnaast bij aan het op tijd en in de juiste volgorde aanleveren van informatie. Daarnaast zijn alle betrokkenen gezamenlijk verantwoordelijk voor het controleren van elkaars werk en resultaten. Deze controles vinden op meerdere momenten in de deelprocessen plaats. Deze nuancering bevestigt daarmee eerder dat er nader onderzoek is gewenst alvorens deelvraag 4 (zie hoofdstuk 6), waarin wordt onderzocht in welke mate de voorgestelde stapsgewijze overgang van de IRT2 naar de IRT1 methode uitvoerbaar is voor de toetsaanbieders, kan worden beantwoord.

Voor wat betreft de normering van de eindtoetsen voor 2022 concludeert RCEC dan ook dat de huidige primaire processen voldoende bruikbaar zijn, rekening houdend met het bewustzijn dat er bij alle betrokken partijen is voor wat betreft de tijdelijkheid van deze overgangsperiode. De primaire processen zijn wat RCEC betreft bruikbaar onder de voorwaarde dat er een vierogen principe wordt ingericht waarmee toezicht wordt gehouden op een correcte uitvoering van alle cruciale stappen binnen de drie deelprocessen waarin data (input) wordt aangeleverd, data wordt geanalyseerd en informatie (output) wordt gegenereerd, geïnterpreteerd en geïmplementeerd.

5. Rechtsgeldigheid IRT1 en IRT2 normeringsmethode

5.1. Inleiding

Dit hoofdstuk bespreekt de resultaten van het derde deelonderzoek. Centraal daarin staat de vraag in welke mate de IRT1 en IRT2 normeringsmethode aansluiten bij de centrale eisen van de eindtoets, zoals vermeld in artikel 4 uit het Toetsbesluit PO. Het doel van de derde deelvraag is om de mate van rechtsgeldigheid van beide normeringsmethoden te evalueren. Dit onderzoek concentreert zich daarbij specifiek op de leden *a*, *b*, *d*, *f* en *h* van Artikel 4 (Tabel 12).

Tabel 12.
Artikel 4, Toetsbesluit PO

Onverminderd artikel 9b van de Wet op het primair onderwijs of artikel 18b van de Wet op de expertisecentra, voldoet een eindtoets aan de volgende kenmerken:
Lid *a*: de eindtoets leidt, op basis van het door een leerling behaalde resultaat, tot een eenduidig advies omtrent het te volgen vervolgonderwijs en hanteert daarbij categorieën van schoolsoorten of leerwegen in het voortgezet onderwijs die gelijkluidend zijn aan de gehanteerde categorieën in andere eindtoetsen,
Lid *b*: de toets is inhoudelijk valide, betrouwbaar en heeft een deugdelijke normering,
Lid *d*: de eindtoetsen bevatten een gezamenlijke set aan opgaven Nederlandse taal en rekenen en wiskunde, die zodanig van omvang is dat daarmee de onderlinge vergelijkbaarheid van de eindtoetsen is geborgd.
Lid *f*: het toetsresultaat geeft een indicatie van de beheersing van de referentieniveaus Nederlandse taal en rekenen,
Lid *h:* de toets biedt de inspectie voldoende basis voor een oordeel over de leerresultaten, bedoeld in artikel 10a van de Wet op het primair onderwijs of artikel 19a van de Wet op de expertisecentra.

De vijf leden van Artikel 4 worden achtereenvolgens getoetst aan de eerdere bevindingen uit de hoofdstukken 2, 3 en 4 van dit onderzoeksrapport.

5.2. Evaluatie rechtsgeldigheid

Artikel 4, lid *a*

Vertaald naar de validiteit (hoofdstuk 3) en de bruikbaarheid (hoofdstuk 4) van de IRT1 en IRT2 normeringsmethode, laat de kern van Artikel 4, lid *a* zich als volgt lezen:

1. iedere eindtoets leidt tot een eenduidig advies omtrent het te volgen vervolgonderwijs, en;
2. iedere eindtoets hanteert daartoe dezelfde categorieën van schoolsoorten.

Uit de beschrijving van de toetsoverstijgende normering in hoofdstuk 3 volgt dat voor iedere toetsaanbieder geldt dat de IRT1 en IRT2 methode resulteren in cesuren welke de verschillende categorieën van schoolsoorten onderscheiden. Hiermee leidt elke eindtoets in beginsel tot een eenduidig advies omtrent het te volgen vervolgonderwijs. Daarmee is voldaan aan punt 1 van Artikel 4, lid *a*. Uit de beschrijving van de IRT2 normeringsmethode volgt daarnaast dat de gehanteerde

adviescategorieën voor alle aanbieders hetzelfde zijn. Daarmee wordt eveneens voldaan aan punt 2 van Artikel 4, lid *a*.

Artikel 4, lid *b*

Lid *b* van Artikel 4 schrijft voor dat iedere eindtoets inhoudelijk valide en betrouwbaar moet zijn en dat deze een deugdelijke normering moet hebben.

Het Beoordelingskader Eindtoetsen (EPO, 2020) bevat kwaliteitseisen voor de onderwijskundige aspecten. Als zodanig worden de toetsopgaven van de verschillende onderdelen en domeinen beoordeeld op inhoudelijke validiteit door de inhoudsdeskundigen van de EPO, voordat de opgaven worden goedgekeurd voor gebruik in de eindtoets. Daarmee wordt voldaan aan de eis voor inhoudelijke validiteit.

Iedere toetsaanbieder legt de voorgestelde toetssamenstelling ter beoordeling voor aan de EPO. De EPO beoordeelt de toetssamenstelling onder andere op betrouwbaarheid. Daarmee wordt eveneens voldaan aan de eis voor betrouwbaarheid.

Na afname van de eindtoetsen volgen de IRT1 en de IRT2 normeringsmethode. Uit de beschrijvingen in paragraaf 3.3 en 3.4 volgt dat de IRT1 een vorm is van absoluut normeren (d.w.z. het vooraf bepalen van een en dezelfde cesuren voor de referentieniveaus en toetsadviezen voor alle toetsaanbieders). De aansluitende IRT2 normeringsmethode heeft het karakter van relatief normeren (d.w.z. het na afname van de eindtoetsen omzetten van scores in eindtoetsaanbieder specifieke cesuren voor de referentieniveaus en toetsadviezen). Psychometrisch bezien is er daarmee in feite sprake van een combinatie van absoluut en relatief normeren, c.q. het vooraf bepalen en achteraf bijstellen van een absolute norm. Omdat beide normeringsmethoden wetenschappelijk zijn onderbouwd en omdat – zonder dat RCEC dit juridisch heeft kunnen (laten) toetsen – de wet geen expliciete uitspraak lijkt te doen over het al dan niet verplicht absoluut of relatief moeten normeren, wordt er in beginsel voldaan aan de eis voor een deugdelijke normering. Uit het RCEC onderzoek volgt echter ook dat beide normeringsmethoden inhoudelijk nog niet volledig voldoen aan de eisen voor validiteit. Daarmee kan nog niet onomstotelijk worden gewaarborgd dat de normeringsmethoden voldoen aan de wettelijke eis voor deugdelijkheid. RCEC concludeert in paragraaf 3.3 en 3.4 daarbij ook dat de beide methoden voor specifiek het schooljaar 2022, onder de in paragraaf 3.3 en 3.4 genoemde voorwaarden, voldoende bruikbaar zijn. Dit in ogenschouw nemende, wordt er binnen deze zelfde gedachtegang voor het schooljaar 2022 in voldoende mate voldaan aan de wettelijke eis voor een deugdelijke normering.

Artikel 4, lid *d*

Artikel 4, lid *d* schrijft voor dat de eindtoetsen een gezamenlijke set aan opgaven Nederlandse taal en rekenen en wiskunde bevatten, die zodanig van omvang is dat daarmee de onderlinge vergelijkbaarheid van de eindtoetsen is geborgd.

Uit de verschillende beschrijvingen in dit rapport volgt dat er sinds 2019 wordt gewerkt met een drietal subsets van gezamenlijke ankeropgaven welke vooraf door de EPO zijn gekalibreerd op basis van circa 1.000 observaties per opgave. De ankeropgaven beschrijven de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen. Het gezamenlijk anker wordt onder

andere gebruikt om de cesuurpunten voor de referentieniveaus en toetsadviezen voor de private toetsaanbieders (niet zijnde de CET) vast te stellen. Daarmee wordt in beginsel voldaan aan Artikel 4, lid *d*, in de zin dat iedere eindtoets een gezamenlijk set aan opgaven bevat.

Uit paragraaf 2.2 en paragraaf 3.3 volgt echter ook dat er verschillen zijn in de manieren waarop de ankeropgaven aangeboden worden in de eindtoetsen. Daarnaast worden de ankeropgaven op verschillende plaatsen in de toetsen gepositioneerd. En kiest een toetsaanbieder zelf of en welke ankeropgaven wel en niet meetellen bij het vaststellen van het vaardigheidsniveau van de leerling. Bovendien speelt het bij adaptieve afnames een rol dat verschillende leerlingen verschillende ankeropgaven voorgelegd krijgen en dat het aantal afnames per ankeropgave daardoor kan variëren. Ook speelt de instructie die de leerlingen krijgen bij het beantwoorden van de ankeropgaven een rol. Bij het kalibreren van de ankeropgaven wordt er daarnaast van uitgegaan dat de leerlingen de opgaven even gemotiveerd beantwoorden en dat ze allen op hun werkelijke niveau presteren. Daarmee wordt volgens de exacte letter ook voldaan aan de wettelijke eis voor 'omvang' van het gezamenlijk anker. Er kan echter nog niet onomstotelijk worden gesteld dat daarmee de onderlinge vergelijkbaarheid van de eindtoetsen is geborgd. Ook hier geldt echter dat dit voor wat betreft het schooljaar 2022 vooralsnog voldoende lijkt te zijn gegarandeerd.

Artikel 4, lid *f*

Dit lid schrijft voor dat het toetsresultaat een indicatie moet geven van de beheersing van de referentieniveaus Nederlandse taal en rekenen.

Zoals hierboven genoemd bij de evaluatie van Artikel 4, lid *a* hebben bij de IRT2 normeringsmethode de scores van de leerlingen op de verschillende eindtoetsen ten aanzien van de referentieniveaus een eigen interpretatie per toetsaanbieder. Daarmee is het vooralsnog niet voldoende aannemelijk te maken dat de IRT2 methode geschikt is om leerlingen over de jaren heen of met betrekking tot de referentieniveaus te beoordelen. In paragraaf 3.4 concludeert RCEC echter ook dat de IRT2 methode wel voldoende geschikt lijkt te zijn voor het bepalen van de eindtoets specifieke cesuren voor de referentieniveaus en voor de eindtoets specifieke toetsadviezen voor het schooljaar 2022. Derhalve mag worden aangenomen dat er in 2022 wordt voldaan aan de wettelijke eis dat het toetsresultaat een indicatie van de behaalde referentieniveaus geeft.

Artikel 4, lid *h*

Lid *h* van Artikel 4 schrijft voor dat de eindtoets voor de inspectie voldoende basis biedt om een oordeel te kunnen geven over de leerresultaten, bedoeld in artikel 10a van de Wet op het primair onderwijs[21] of artikel 19a van de Wet op de expertisecentra[22].

Uit de genoemde wetsartikelen volgt dat de leerresultaten van de school jaarlijks worden beoordeeld op basis van de resultaten van de afgelegde centrale eindtoetsen, op het gebied van Nederlandse taal en rekenen en wiskunde. Daarbij is er sprake van onvoldoende leerresultaten wanneer, gemeten over een periode van drie schooljaren, de resultaten op het gebied van de Nederlandse taal en rekenen en

[21] https://wetten.overheid.nl/BWBR0003420/2022-01-01
[22] https://wetten.overheid.nl/BWBR0003549/2022-01-01

wiskunde, liggen onder de normering die daarvoor geldt in vergelijking tot die leerresultaten over diezelfde schooljaren van scholen met een vergelijkbaar leerlingenbestand.

Uit het RCEC onderzoek volgt dat elke eindtoets jaarlijks per leerling het behaalde referentieniveau op de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen meet. Iedere eindtoets hanteert daartoe cesuren voor het fundamenteel niveau 1F en voor het streefniveau 2F / 1S. De IRT2 normeringsmethode draagt eraan bij dat de resultaten van de leerlingen binnen een en hetzelfde schooljaar onderling vergelijkbaar zijn, doordat de cesuren voor het referentieniveau en het toetsadvies worden gecorrigeerd voor de modellering alsof zij de PCET van dit jaar zouden hebben gemaakt. Daarmee wordt in beginsel voldaan aan Artikel 4 lid *h*.

In het volgende hoofdstuk wordt verslag gedaan van de mate waarin de voorgestelde stapsgewijze overgang van de IRT2 naar de IRT1 methode uitvoerbaar is voor de toetsaanbieders.

6. Haalbaarheid IRT1 normeringsmethode

6.1. Inleiding

Dit hoofdstuk behandelt de resultaten van het vierde deelonderzoek. Centraal staat hier de vraag onder welke voorwaarden de beoogde terugkeer naar de IRT1 methode uitvoerbaar is voor de toetsaanbieders. Het doel van deze vierde en laatste deelvraag is om te evalueren op welk moment in de tijd de IRT2 normeringsmethode kan worden losgelaten.

Uit analyses van SCC en EPO komt naar voren dat de IRT1 normeringsmethode op dit moment nog niet voor alle eindtoetsen een even betrouwbaar beeld (c.q. een valide vergelijking) geeft. Het vermoeden bestaat dat dit samenhangt met de wijze waarop de gezamenlijke ankeropgaven in de verschillende eindtoetsen worden gebruikt. In deze onderhavige studie heeft RCEC hier aanvullend onderzoek naar gedaan. De resultaten daarvan zijn gepresenteerd in de hoofdstukken 2 tot en met 5. Meerdere van de bevindingen lijken van invloed te zijn op de mate waarin de IRT1 normeringsmethode op dit moment al dan niet slaagt. Hier wordt in de volgende paragraaf op gereflecteerd.

6.2. Evaluatie haalbaarheid

Uit de beschrijving van de totstandkoming van de eindtoetsen in paragraaf 2.1 volgt dat de toetsaanbieders verschillende keuzemogelijkheden hebben bij het ontwikkelen, afnemen en scoren van de eigen eindtoets. In paragraaf 2.2 zijn deze mogelijkheden samengevat. In paragraaf 3.3 zijn vervolgens de potentiële risico's benoemd die van invloed kunnen zijn op het al dan niet slagen van de IRT1 normeringsmethode. Meerdere van deze risico's hangen samen met de genoemde keuzemogelijkheden. Hierbij vallen een aantal zaken in het bijzonder op:

- De toetsaanbieder kan kiezen uit verschillende toetsvormen;
- De toetsaanbieder kan naar eigen inzicht toetsinhoud toevoegen bovenop de wettelijk verplichte onderdelen Taal (bestaande uit de domeinen Lezen en Taalverzorging) en Rekenen;
- Als gevolg van positionering en marktaandeel is er een, tussen de toetsaanbieders, onderling verschil in samenstelling van de populaties van leerlingen waar te nemen;
- De toetsaanbieder heeft meerdere keuzemogelijkheden voor wat betreft het pretesten van nieuwe toetsopgaven;
- De toetsaanbieder kiest een eigen toetssamenstelling met een eigen toetslengte en een eigen verhouding van aantal toetsvragen per toetsonderdeel;
- De toetsaanbieder kiest een eigen berekeningswijze waarop hij tot de getransformeerde standaardscore komt van waaruit de toetsadviezen worden berekend;
- De toetsaanbieder kiest zelf met behulp van welke software en algoritmes hij het jaarlijkse verantwoordingsverslag samenstelt.

Aanvullend vallen voor wat betreft het gezamenlijk anker de volgende zaken in het bijzonder op:

- Er is sprake van drie subsets van ankeropgaven (papier, digitaal, papier-digitaal);
- De toetsaanbieders ontwikkelen eigenstandig nieuwe ankeropgaven en pretesten deze op eigen wijze binnen de eigen eindtoets (psychometrisch model en schattingsmethode);
- De ankeropgaven worden op verschillende manieren aangeboden in de eindtoetsen (lineair, adaptief);

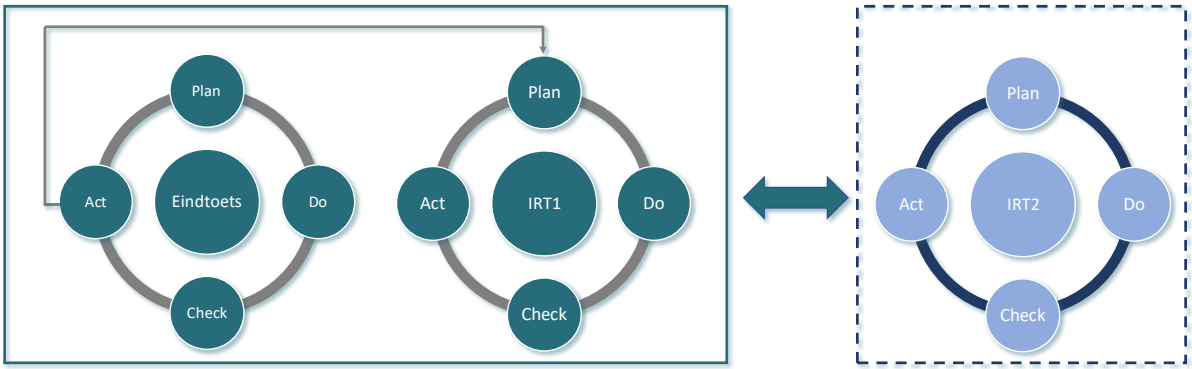
- De ankeropgaven worden op verschillende plaatsen in de toetsen gepositioneerd (start, halverwege, eind);
- De ankeropgaven tellen bij sommige eindtoetsen wel en bij andere eindtoetsen niet mee;
- Iedere toetsaanbieder hanteert een eigen instructie voor de leerlingen voor wat betreft de ankeropgaven;
- De motivatie waarmee de leerlingen de ankeropgaven invullen, kan verschillen per eindtoets.

Daarmee zijn er, ondanks dat er gebruik wordt gemaakt van bewezen psychometrische technieken, meerdere variabelen die van invloed kunnen zijn op de resultaten van de IRT1 normeringsmethode. Uit het RCEC onderzoek komt naar voren dat de mogelijke effecten al deels worden ondervangen en daarmee al deels kunnen worden uitgesloten. Zo wordt de psychometrische kwaliteit van de ankeropgaven uitgebreid gecontroleerd en wordt er gewerkt met een ondergrens van 1.000 afnames per eindtoets per schooljaar. Ook wordt er jaarlijks gecontroleerd op DIF. En zijn er voorzieningen getroffen om zo zuiver mogelijk te kunnen rekenen met het gegeven dat de toetsaanbieders met verschillende item-responsmodellen en schattingsmethoden werken. Desalniettemin zijn de mogelijke effecten van meerdere van de genoemde potentiële risico's nog niet of nog niet in voldoende mate onderzocht. RCEC concludeert in hoofdstuk 3 dan ook dat de data waar de IRT1 normeringsmethode op toegepast wordt nog een aantal onduidelijkheden en tekortkomingen kunnen bevatten, waardoor de methode nog niet op volledig valide wijze toepasbaar is.

Voor wat betreft de proceskant van de normeringsmethoden, start hoofdstuk 4 met de weergave van de onderlinge rolverdeling (Tabel 6). Hieruit is af te lezen dat er heldere afspraken zijn gemaakt over het verdelen van taken en verantwoordelijkheden. Dat biedt een goede basis voor een valide procesgang voor wat betreft de IRT1 normeringsmethode. Een nadere analyse van de deelprocessen laat echter zien dat er zich in de uitvoering meerdere potentiële risico's voordoen welke de kwaliteit en de resultaten van de IRT1 normeringsmethode kunnen remmen. Deze risico's betreffen veelal de input en/of de output van deelstappen. Daarnaast is op onderdelen de chronologische ordening van activiteiten nog onduidelijk en/of onvolledig. Meerdere van deze potentiële risico's lijken samen te hangen met de veelvoud aan keuzemogelijkheden die het stelsel biedt. Daarmee concludeert RCEC in hoofdstuk 4 dat de huidige procesinrichting nog niet voldoende is ingeregeld voor het op de kortere termijn kunnen loslaten van de IRT2 normeringsmethode.

Dit betekent derhalve dat er, voordat de validiteit van de inhoud en het proces kan worden gewaarborgd, er aanvullend onderzoek naar de nog onbekende effecten nodig is. Daarnaast adviseert RCEC om, waar mogelijk, het aantal keuzemogelijkheden voor de toetsaanbieders stapsgewijs te reguleren, dan wel de genoemde aspecten meer op één en dezelfde lijn van denken en handelen te brengen. Procesmatig dient dit te worden ondersteund met eenduidige taakafspraken, sjablonen en werkinstructies. Daarbij dient een controle- en feedbackcyclus te worden geïmplementeerd, welke het mogelijk maakt om organisatieonafhankelijk toezicht op het stelsel te behouden. Het terugdringen van de variatie en het standaardiseren van werkwijzen zal naar verwachting bijdragen aan het verminderen van potentiële ruis en aan het meer robuust maken van de IRT1 normeringsresultaten. Daarmee zal de IRT1 normeringsmethode op termijn ook (nog) beter aansluiten bij de eisen uit de wet, zoals beschreven en onderzocht in hoofdstuk 5.

Om de genoemde voorstellen stapsgewijs te implementeren, kan in de overgangsperiode op weg naar het nieuwe stelsel doorstoomtoetsen bijvoorbeeld als volgt te werk worden gegaan. Geadviseerd wordt om een gedetailleerd normeringshandboek te laten opstellen en deze door een onafhankelijke partij te laten accrediteren. In dit handboek dienen zowel de inhoudelijke als de procesmatige afspraken te worden beschreven. Daarvoor dient er in ieder geval aanvullend onderzoek te worden gedaan naar de huidige en gewenste werkwijze van de EPO en van de toetsaanbieders. Daarnaast verdient het de aanbeveling om bij dit handboek bijpassende sjablonen voor input van data en output van resultaten op te (laten) stellen. Tevens dient er een sluitende controle- en feedbackcyclus te worden ingericht, een en ander conform de indeling van de Deming Cirkel (Walton en Deming, 1988). De Deming Cirkel is te definiëren als een cyclus voor het ontwerpen en controleren van effectieve en kwalitatieve processen. De methode onderscheidt hiertoe vier stappen: Plan – Do – Check – Act, waarmee het proces wordt gepland, uitgevoerd, gecontroleerd en, waar nodig, bijgesteld. Zoals weergegeven in Figuur 3 zou voor elk van de drie in hoofdstuk 4 genoemde deelprocessen een PDCA cyclus opgesteld moeten worden.



Figuur 3. Stelsel eindtoetsen en PDCA-cyclus

Synchroon aan het ontwikkelen van een handboek, zouden de betrokken partijen, zoveel als mogelijk is, op een zo uniform mogelijke wijze het gezamenlijk anker dienen in te zetten binnen de eindtoetsen 2022. Dit betekent dat de toetsaanbieders het gezamenlijk anker zoveel mogelijk op dezelfde plaats in de toets positioneren, en alle toetsaanbieders bij voorkeur het volledige gezamenlijke anker mee laten tellen. Daarbij dienen de toetsaanbieders een zo uniform mogelijk instructievoorschrift met de scholen te delen, om zo ook de wijze van afname zoveel als mogelijk gelijk te trekken. Bij dit alles dient in ogenschouw te worden genomen dat, zolang er nog een effect is van de coronapandemie, er nog niet aan alle voorwaarden van de IRT1 normeringsmethode kan worden voldaan.

De combinatie van een handboek en een verdere standaardisatie van het gebruik van het gezamenlijk anker kan de opmaat zijn naar een inhoudelijk, procesmatig en wettelijk valide IRT1 normeringsmethode, waarbij een aanvullende IRT2 normeringsmethode niet langer benodigd is. In een dergelijke toekomstige situatie volstaan de PDCA cyclus van de eindtoets en de PDCA cyclus van de IRT1 normeringsmethode (zoals aangegeven in Figuur 3), waarbij de – nu nog tijdelijk benodigde – PDCA cyclus van de IRT2 normeringsmethode kan worden losgelaten.

7. Conclusies en aanbevelingen

7.1. Conclusies

In opdracht van het Ministerie van Onderwijs, Cultuur en Wetenschap, Directeur-Generaal Primair en Voortgezet Onderwijs (DGPV), Directie Primair Onderwijs, deed RCEC onderzoek naar de kwaliteit van de normering eindtoetsen primair onderwijs 2022. Centraal in dit onderzoek stonden de vragen: “Zijn de IRT1 en IRT2 methode, zoals beschreven in het normeringshandboek 2022, valide normeringsmethoden en sluiten deze aan bij de eisen voor eindtoetsen, zoals beschreven in het Toetsbesluit PO? En in welke mate is het voorgestelde implementatieplan haalbaar en uitvoerbaar voor de betrokken partijen?”

RCEC beantwoordde deze centrale onderzoeksvragen, door een bureauonderzoek te doen naar vier separate deelvragen. Met deze vragen werd de validiteit, de bruikbaarheid en de rechtsgeldigheid van de IRT1 en IRT2 normeringsmethode onderzocht. Aansluitend werd onderzocht onder welke voorwaarden de IRT2 normeringsmethode zou kunnen worden losgelaten.

In hoofdstuk 2 werd beschreven en vastgesteld dat de toetsaanbieders de eigen eindtoets op verschillende manieren kunnen inrichten en vormgeven en daarbij meerdere keuzemogelijkheden hebben. Uit dit deelonderzoek volgde dat, na toetsafname van de door de EPO goedgekeurde eindtoets op de scholen, iedere toetsaanbieder de eigen afnamedata inlevert bij SCC. Aansluitend werd beschreven dat SCC met de afnamedata 2022 de IRT1 en de IRT2 normeringsprocedure uitvoert.

In paragraaf 3.3. werd geconcludeerd dat de IRT1 methode gebruik maakt van bewezen psychometrische technieken, zoals de item-responstheorie en het gezamenlijk anker, om de eindtoetsen te normeren en om eenduidige cesuren voor de referentieniveaus en toetsadviescategorieën voor alle toetsaanbieders te geven. Uit het RCEC onderzoek is naar voren gekomen dat de redenering dat met de IRT1 normeringsmethode de toetsscores op een valide manier omgezet kunnen worden in cesuren voor referentieniveaus en toetsadviescategorieën in bepaalde situaties weerlegd kan worden. Strikt genomen wordt daardoor niet voldaan aan de eisen van validiteit. Dit wordt grotendeels veroorzaakt door de verschillende wijzen waarop de vijf toetsaanbieders de gezamenlijke ankeropgaven afnemen. Omdat de toetsaanbieders de wijze van afname over de jaren heen wel zo veel mogelijk gestandaardiseerd hebben en het niet mogelijk is om de gevolgen van de coronapandemie voor het onderwijs goed in te schatten, concludeerde RCEC dat de IRT1 normeringsmethode psychometrisch gezien voldoende valide is om toe te passen bij de normering van de eindtoetsen voor 2022. Dit onder de voorwaarde dat er gekeken wordt of er geen grote verschillen tussen de schooladviezen en de toetsadviezen zijn ontstaan, bijvoorbeeld ten gevolge van de coronapandemie.

Uit paragraaf 3.4 volgde dat de IRT2 methode gebruik maakt van de schooladviezen en rangorde van leerlingen uit de IRT1 normering. Daarna bepaalt de methode per eindtoets apart de grenswaarden voor de referentieniveaus en de toetsadviescategorieën, rekening houdend met de prestaties van de populatie per aanbieder. Bij de beschrijving van de IRT2 methode is eveneens aangegeven dat de methode resulteert in unieke normen per toetsaanbieder die gelden voor specifiek dit schooljaar. RCEC stelde daarbij vast dat, gegeven de aannames die ten grondslag liggen aan de IRT2 methode, de

cesuren en streefpercentages vergelijkbaar zijn via de koppeling die is gelegd met de PCET. De scores van de individuele leerlingen ten aanzien van de referentieniveaus dienen daarbij geïnterpreteerd te worden alsof de leerlingen de PCET hebben gemaakt. Daarmee concludeerde RCEC dat de IRT2 methode de resultaten van individuele leerlingen over de jaren heen op valide wijze kan vergelijken, mits wordt voldaan aan de aannames die ten grondslag liggen aan de IRT2 methode. De methode lijkt daarmee geschikt voor het bepalen van de schooljaar specifieke toetsadviezen. En de methode lijkt eveneens geschikt om via de koppeling met de PCET afname van dit jaar uitspraken te doen over de behaalde referentieniveaus van individuele leerlingen over de jaren heen.

In hoofdstuk 4 kwam vervolgens de bruikbaarheid van het toetsoverstijgende normeringsproces aan bod. Er is toegelicht dat, om te komen tot een toetsoverstijgende normering volgens achtereenvolgens de IRT1 en de IRT2 methode, de betrokken partijen gezamenlijk een drietal opeenvolgende deelprocessen doorlopen: (1) de processtappen van de totstandkoming van de eindtoetsen, (2) de processtappen van de IRT1 normeringsmethode, en (3) de processtappen van de IRT2 normeringsmethode.

Uit de risicoanalyse in paragraaf 4.4 volgde dat er in de processen op meerdere momenten een reëel risico bestaat dat het type en de herkomst van de input voor de processtap en/of het type en de bestemming van de output van de processtap onvoldoende beschikbaar of betrouwbaar zijn. Tevens is gebleken dat op meerdere momenten de chronologische ordening van activiteiten onduidelijk en/of onvolledig lijkt te zijn. Met inachtneming van de nuancering dat RCEC nog geen gedetailleerd onderzoek naar de werkwijze van de EPO en van de toetsaanbieders heeft gedaan, concludeerde RCEC dat, voor wat betreft de normering van de eindtoetsen voor 2022, de huidige primaire processen voldoende bruikbaar zijn. Als voorwaarde hiervoor noemde RCEC dat er een vierogen principe moet worden ingericht waarmee toezicht wordt gehouden op een correcte uitvoering van alle cruciale stappen binnen de drie deelprocessen waarin data (input) wordt aangeleverd, data wordt geanalyseerd en informatie (output) wordt gegenereerd, geïnterpreteerd en geïmplementeerd.

In hoofdstuk 5 deed RCEC verslag van haar onderzoek naar de derde deelvraag. Er is geëvalueerd in welke mate de IRT1 en IRT2 normeringsmethode aansluiten bij de wettelijke eisen van Artikel 4 uit het Toetsbesluit PO. RCEC concludeerde dat de beide methoden voor wat betreft de langere termijn op onderdelen nog niet volledig in overeenstemming zijn met de wetgeving. Rekening houdend met de inhoudelijke en procesmatige bevindingen uit hoofdstuk 3 en 4 concludeerde RCEC dat, mits de genoemde adviezen en aanbevelingen worden opgevolgd, er voor wat betreft schooljaar 2022 in voldoende mate aansluiting is met de centrale eisen van de eindtoets, zoals vermeld in het Toetsbesluit PO.

Tot slot werd in hoofdstuk 6 onderzocht onder welke voorwaarden het mogelijk is om op een valide wijze van de IRT1 normeringsmethode gebruik te kunnen maken, waarbij de IRT2 normeringsmethode kan worden losgelaten. RCEC evalueerde daartoe de inhoudelijke, procesmatige en wettelijke voorwaarden. Van daaruit deed RCEC concrete voorstellen voor vervolgacties en een ondersteunend vervolgonderzoek.

7.2. Aanbevelingen

Dit onderzoeksrapport bevat, samengevat, de volgende inhoudelijke aanbevelingen voor specifiek het schooljaar 2022:

- Om de IRT1 normeringsmethode in schooljaar 2022 voldoende valide toe te kunnen passen bij de toetsoverstijgende normering van de eindtoetsen voor 2022 dient er gecontroleerd te worden dat:
    - er geen grote verschillen tussen de schooladviezen en de toetsadviezen ontstaan, bijvoorbeeld ten gevolge van de coronapandemie;
    - de vastgestelde cesuurpunten voor de toetsaanbieders niet resulteren in onverantwoord grote verschuivingen in de streefpercentages, als gevolg van het gebruik van een 1PL item-responsmodel door de toetsaanbieders;
    - de verschillen in de populatiesamenstelling van de toetsaanbieders niet resulteren in onevenwichtige cesuurpunten van de verschillende eindtoetsen;
    - de normering wordt aangepast voor de pro-leerlingen.
- Daarnaast dient er bij de IRT2 normeringsmethode te worden gecorrigeerd voor de effecten van de coronapandemie, indien de normeringsresultaten daar aanleiding toe geven.
- De betrokken partijen wordt geadviseerd om, zoveel als mogelijk is, op een zo uniform mogelijke wijze het gezamenlijk anker in te zetten binnen de eindtoetsen 2022. Dit betekent dat de toetsaanbieders het gezamenlijk anker zoveel mogelijk op dezelfde plaats in de toets positioneren, en alle toetsaanbieders bij voorkeur het volledige gezamenlijke anker mee laten tellen. Daarbij dienen de toetsaanbieders een zo uniform mogelijk instructievoorschrift met de scholen te delen, om zo ook de wijze van afname zoveel als mogelijk gelijk te trekken.

Voor wat betreft het schooljaar 2022, bevat dit onderzoeksrapport, samengevat, daarnaast de volgende procesmatige aanbevelingen:

- Er dient een vierogen principe te worden ingericht waarmee toezicht wordt gehouden op een correcte uitvoering van alle cruciale stappen binnen de drie deelprocessen waarin data (input) wordt aangeleverd, data wordt geanalyseerd en informatie (output) wordt gegenereerd, geïnterpreteerd en geïmplementeerd.
- Er dient een gedetailleerd normeringshandboek opgesteld te worden welke door een onafhankelijke partij wordt geaccrediteerd. Dit handboek moet in ieder geval bestaan uit:
    - de inhoudelijke en procesmatige afspraken binnen het stelsel;
    - sjablonen voor input van data en output van resultaten;
    - een controle- en feedbackcyclus inclusief organisatieonafhankelijk toezicht.
- Er dient aanvullend onderzoek te worden gedaan naar de huidige en gewenste werkwijze van de EPO en de toetsaanbieders.

Het opvolgen van de genoemde aanbevelingen kan de opmaat zijn naar een IRT1 normeringsmethode die inhoudelijk, procesmatig en wettelijk valide is en waarbij de IRT2 normeringsmethode op termijn kan worden losgelaten.

Tot nader overleg bereid,

RCEC

8. Geraadpleegde literatuur

AERA (2014). *Standards for educational and psychological testing. Joint Committee on Standards for Educational and Psychological Testing.* Washington D.C.: American Educational Research Association.

Aguilar-Saven, R.S. (2004). Business process modelling: Review and framework. *International Journal of production economics 90.2*, 129–149.

Brennan, R.L. (2006). *Educational Measurement, fourth edition*. Westport, CT: Prager Publishers.

Centraal Bureau voor de Statistiek (2021). *Hoofdstuk 4. Regionale verschillen in oversterfte in: Oversterfte tijdens eerste golf corona-epidemie bijna dubbel zo hoog als tijdens griepepidemie. Den Haag: CBS*. Geraadpleegd van: https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/oversterfte-tijdens-eerste-golf-corona-epidemie-bijna-dubbel-zo-hoog-als-tijdens-griepepidemie.

Cito (2010). *Toetstechnische begrippenlijst*. Geraadpleegd van: https://www.cito.nl/kennis-en-innovatie/kennisbank/toetstechnische-begrippenlijst.

Eggen, T.J.H.M., & Sanders, P.F. (1993). *Psychometrie in de praktijk*. Arnhem: Cito.

Evers, A., Lucassen, W., Meijer, R.R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests, geheel herziene versie, mei 2009; gewijzigde druk mei 2010*. Utrecht: Nederlands Instituut van Psychologen.

Expertgroep Toetsen PO (2020). *Beoordelingskader Eindtoetsen, versie 20.01*. Utrecht: EPO. Geraadpleegd van: https://www.expertgroeptoetsenpo.nl/uimg/ept/b63321_att-200220-def-beoordelingskader-eindtoetsen-expertgroep-2020.pdf.

Expertgroep Toetsen PO (2021). *Advies normering Eindtoetsen, 15 mei 2021, Ref.nr.: 21.0*1. Utrecht: EPO.

Inspectie van het Onderwijs (2021). *Technisch rapport primair onderwijs. Bovensectoraal themaonderzoek 16 maanden coronacrisis*. Utrecht: IvhO.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Kane. M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2, 135-170.

Ministerie OCW (2006). *Kerndoelen Primair Onderwijs*. Den Haag: OCW. Geraadpleegd van: https://www.rijksoverheid.nl/documenten/rapporten/2006/04/28/kerndoelenboekje.

Ministerie OCW (2021). *Plan van aanpak landelijke normering eindtoetsen. Referentieniveaus en toetsadviezen*. Den Haag: OCW.

Sanders, P.F., Brouwer, A.J., Eggen, T.J.H.M, & Veldkamp, B.P. (2018). *RCEC beoordelingssysteem voor de kwaliteit van studietoetsen en (praktijk)examens*. Apeldoorn: RCEC.

Stichting Cito Control (2020). *Evaluatie functioneren gezamenlijk anker*. Arnhem: Stichting Cito Control.

Stichting Cito Control (2021a). *Harmonisering landelijke normering, versie 0.4*. Arnhem: Stichting Cito Control.

Stichting Cito Control (2021b). *Normeringsproces eindtoetsen, versie 0.4*. Arnhem: Stichting Cito Control.

Stichting Cito Control (2021c). *IRT2 normering. Een andere basis voor referentieniveaus, versie 0.4.* Arnhem: Stichting Cito Control.

Stichting Cito Control (2022). *Governance, versie 0.6.* Arnhem: Stichting Cito Control.

Walton, M., & Deming, W.E. (1988). *The Deming management method*. New York: Perigee Books.

Wools, S. (2012). Towards a comprehensive evaluation system for the quality of tests and assessments. In T.J.H.M. Eggen & B.P. Veldkamp (Eds), *Psychometrics in practice at RCEC* (pp. 95-106). Enschede: RCEC.

Apeldoorn, maart 2022

In opdracht van:

Ministerie van Onderwijs, Cultuur en Wetenschap

RCEC • Research Center voor Examinering en Certificering
Villa de Regentes • Regentesselaan 5 • 7316 AA Apeldoorn
T 055 - 303 31 24 • I www.rcec.nl • E info@rcec.nl