1. Algemene informatie

Algemeen en meetpretentie
Het IEP LVS is een methodeonafhankelijk volgsysteem waarin leerlingen vanaf leerjaar 3 tot aan de eindtoets gevolgd kunnen worden in hun ontwikkeling. Het IEP-LVS is een leer- en criteriumgericht volgsysteem.

Met de IEP volgtoetsen Taal en Rekenen wordt gemeten op welk niveau de leerlingen de onderdelen Lezen, Rekenen, Taalverzorging en Technisch Lezen beheersen. Het inhoudelijk toetskader vormt de basis voor de inhoud van de toets Rekenen 6a.

Doelgroep
De doelgroep van het IEP LVS bestaat uit alle leerlingen in leerjaar 3 t/m 8 van het regulier basisonderwijs. Iedere IEP LVS toets 3a t/m 6a bestaat uit 30 items van het niveau van de toets, 5 items van het niveau eronder en 5 items van het niveau erboven. Zo bevat de toets IEP LVS Rekenen 6a dus 30 items van het niveau 6a, 5 items van het niveau 5b en 5 items van het niveau <1F-1F (<1F of 1F) versie 1. Bij de bepaling van de cesuur is in de standaardsetting gedegen rekening gehouden met deze keuze in toetssamenstelling. Doordat het mogelijk is om bij een leerling toetsen van verschillende niveaus af te nemen, kun goed worden aangesloten bij het individuele niveau van de leerling. Dit maakt dat het IEP LVS ook geschikt is voor leerlingen uit het SBO.

Inhoudelijke theoretische inkadering:
De IEP LVS Rekentoetsen voor leerjaar 3 tot en met 6a zijn gebaseerd op het IEP Toetskader Rekenen. In het toetskader wordt onderscheid gemaakt tussen de niveaus: 3a, 3b, 4a, 4b, 5a, 5b en 6a. De kenmerken van het toetskader hebben een cumulatief karakter: de leerling moet op een bepaald niveau ook de inhoud beheersen van de onderliggende niveaus. Zo beheerst een leerling op 4b niveau ook de inhouden van het 3a, 3b en 4a niveau. Het IEP Toetskader Rekenen is gebaseerd op de volgende bronnen:

- Tussendoelen rekenen-wiskunde voor het primair onderwijs (Noteboom, Aartsen & Lit, 2017);
- De kerndoelen voor het primair onderwijs (Ministerie van Onderwijs, Cultuur en Wetenschap, 2006);
- Het Referentiekader taal en rekenen (Meijerink et al., 2009);
- Rekenen met hele getallen op de basisschool (Veltman & Van den Heuvel-Panhuizen, 2010).

Inhoud van het toetspakket:
Het toetspakket Rekenen 6a bestaat uit de volgende documenten:

- Verantwoording LVS-toets "Rekenen" groep 6a; deze bevat informatie over:
    - De uitgangspunten van de toets (hfdst. 2),
    - De inhoud van de toets (hdfst. 3),
    - Normeringspopulatie (hfdst. 4),
    - Het design van de dataverzameling (hfdst. 5),
    - De kalibratie en kwaliteit van de items (hfdst. 6),
    - De bepaling van cesuren (hfdst. 7),
    - De constructvaliditeit (hfdst. 8),
    - Het volgaspect (hfdst. 9), en
    - Inzicht in de leervorderingen (hfdst. 10).

Daarnaast werden de volgende bijlagen beschikbaar gesteld:

- Toetswijzer (bijlage 1);
- Items (bijlage 2);
- Toelichting Rekenen (bijlage 3);
- Itemparameters (bijlage 4);
- TIA's IEP LVS Toets (bijlage 5);
- Algemene toelichting methode (bijlage 6);
- Differential Item Functioning (bijlage 7);
- Lagrange Multiplier Tracelines (bijlage 8);
- Handreiking interpreteren toetsresultaten Taal & Rekenen (bijlage 9);
- IEP LVS talentenkaart (bijlage 10);
- Leeswijzer talentenkaart – leerkrachtversie (bijlage 11).

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor instrumenten binnen leerlingvolgsystemen (LVS)', zoals opgesteld door de Expertgroep Toetsen PO. De Expertgroep Toetsen PO wordt gevormd door Prof. Dr. Cees Van der Vleuten (voorzitter), Prof. dr. Cees Glas (psychometrisch expert), Dr. Desiree Joosten-Ten Brinke (onderwijskundig expert) en Liza Kozlowska MA (secretaris).
Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording en de Handleiding veelal letterlijk vermeld.

De kwaliteit van de dataverzameling

S1 Is de steekproef van leerlingen representatief?

Bevindingen:
Uitgangspunt bij het bepalen van de kwaliteit van de items ten behoeve van de samenstelling van de definitieve toets is een steekproef die resulteert in tenminste 400 observaties per item. Hoewel voor de kalibratie van de toetsen met het hier gehanteerde 1PL-model (Rasch model) het minimale aantal observaties 200 is, is voor het evalueren van de modelpassing van een redelijk alternatief model, hier het 2PL-model, minimaal 400 de norm. Uit bijlage 4 'Itemparameters IEP LVS Rekenen' (zie kolom 'Aantal afnames') blijkt dat het aantal observaties van de IEP LVS items conform de eisen is die worden beschreven in het document 'Verantwoording onderzoek werkgroep Meijer: Aanvulling COTAN Beoordelingssysteem' wat betreft het aspect normering referentieniveaus en computer adaptief toetsen van andere eindtoetsen (deel 1) (Meijer et al., 2015).
Naast aantallen afnames is de representativiteit op achtergrondgegevens van de normeringspopulatie (groep leerlingen in de eerste helft van leerjaar 6 van het basisonderwijs die aan het normeringsonderzoek heeft deelgenomen en waar de IEP LVS toets Rekenen 6a is afgenomen) ten opzichte van de doelpopulatie (alle leerlingen in leerjaar 3 t/m 8 van het regulier basisonderwijs) van belang om een oordeel te kunnen toekennen aan dit aspect. De auteurs stellen dat voor de bepaling van de kwaliteit van de items (hoofdstuk 6) en de standaardsetting (hoofdstuk 7) het niet noodzakelijk is dat de normeringspopulatie ook een normpopulatie (i.e., een representatieve steekproef) is. De reden hiervoor is dat in het onderhavige onderzoek absolute normen worden bepaald, waardoor de representativiteit van ondergeschikt belang is. Wel is het van belang vast te stellen dat de normeringspopulatie geen specifiek selecte groep is van de leerlingen in de eerste helft van leerjaar 6 van het basisonderwijs, d.w.z. dat er in ieder geval achtergrondgegevens worden gerepresenteerd in de steekproef. Voor de bepaling van de gemiddelde groeifactor van de IEP LVS populatie (hoofdstuk 9) is volgens de auteurs de representativiteit van de normeringspopulatie echter wel van belang, omdat daar een relatieve norm wordt bepaald.

Persoonlijke achtergrondgegevens van de leerlingen konden niet worden gebruikt vanwege de geldende privacyregels (zoals beschreven in de Wet bescherming persoonsgegevens) en daarom zijn in dit normeringsonderzoek alleen de schoolachtergrondgegevens denominatie, urbanisatiegraad, schoolgrootte, regio en schoolweging gebruikt om representativiteit te onderzoeken. Deze gegevens zijn openbaar beschikbaar per school bij DUO en het CBS.

Het digitale platform IEP LVS is gebruikt om de data voor het normeringsonderzoek te verzamelen. De hiervoor gebruikte toets is door de betrokken scholen tijdens het reguliere onderwijsproces in schooljaar 2020-2021 op eigen initiatief afgenomen in de eerste helft van leerjaar 6. Er is hier dus sprake van 'purposeful sampling' (een doelsteekproef), een niet-probalistische steekproeftechniek. De data in het normeringsonderzoek is door deze werkwijze dus onder gelijke afnamecondities en afnamemomenten verzameld als waaronder de IEP LVS toets Rekenen 6a ook in de komende jaren afgenomen gaat worden.

Voor de analyses wordt alleen gebruikgemaakt van responsedata van leerlingen uit leerjaar 6 die in december, januari en februari de toets Rekenen 6a hebben gemaakt. Dit geldt ook voor de responsedata van de toetsen 5b en <1F-F versie 1 (i.e., direct voorafgaande en direct volgende toets), zodat voor alle analyses is gefilterd op de passendheid van de toets bij het afnamemoment en verwacht kan worden dat de gemeten doelen uit de toets zijn aangeboden in het onderwijs.

Voor de analyses van de kwaliteit van de items en de toets als geheel en de standaardsetting (zie hoofdstuk 6 en 7) is voor de bruikbaarheid van de responsedata als selectiecriterium gebruikt dat records niet meer dan 15% niet-beantwoorde items (> 15% missing data) mogen bevatten. Hierdoor wordt o.a. onterechte inflatie van de discriminatiewaarde en deflatie van de p-waarde voorkomen. Ook wordt hiermee voorkomen dat data wordt meegenomen van leerlingen die de toets alleen maar aan het verkennen zijn ("verkenners") en de toets dus niet "serieus" maken. Voor de analyse voor het bepalen van de gemiddelde groeifactor (hoofdstuk 9) is, naast het criterium voor passendheid, gefilterd op twee criteria. Ten eerste zijn records van SBO-scholen en van scholen van Bonaire uit het databestand verwijderd. De gemiddelde groeifactor wordt bepaald op basis van reguliere BO-leerlingen en SBO-leerlingen en leerlingen uit Bonaire vallen hier niet onder. Ten tweede zijn records verwijderd waarvoor het toetsresultaat niet in een zogenaamde ontwikkelscore (i.e., scores op één en dezelfde schaal voor alle toetsen van één vaardigheid, zodat de ontwikkeling van een leerling zinvol gevolgd kan worden in de tijd; zie hoofdstuk 9) uitgedrukt kan worden, hetgeen betekent dat records met minder dan 25% goed beantwoorde items niet gebruikt zijn. De reden hiervoor is dat de gemiddelde groeifactor wordt berekend op de ontwikkelscore en de ontwikkelscore alleen aan een toetsresultaat wordt gekoppeld als er tenminste 25% van de toetsvragen correct zijn beantwoord door de leerlingen.

Tabel 4.1 laat zien hoe de normeringspopulatie zich op achtergrondgegevens verhoudt tot de doelpopulatie voor de analyses ten behoeve van de bepaling van de gemiddelde groeifactor (hoofdstuk 9). Door het filteren op andere criteria dan criteria gebruikt voor de analyses ten behoeve van de gemiddelde groeifactor (hoofdstuk 9) is de normeringspopulatie die gebruikt is voor de kwaliteitsanalyse en de standaardsetting (hoofdstuk 6 en 7) weliswaar niet identiek (iets grotere N), maar is wel vergelijkbaar qua verdeling over de diverse achtergrondvariabelen. Omdat de representativiteit voor deze relatieve normering van belang is, wordt in tabel 4.1 daarom de normeringspopulatie weergegeven zoals gebruikt in hoofdstuk 9.

Uit de verdeling die in tabel 4.1 getoond wordt, is op te maken dat de normeringspopulatie niet op alle categorieën als representatieve steekproef van de eerste helft van leerjaar 6 van het basisonderwijs beschouwd kan worden. Wel kan uit tabel 4.1 geconcludeerd worden dat alle antwoordcategorieën vertegenwoordigd zijn in de normeringspopulatie en vergelijkbaar verdeeld zijn ten opzichte van de landelijke

populatie in het gehele basisonderwijs en er dus sprake is van een representatieve steekproef. Om een indruk te krijgen van de mate waarin de steekproef afwijkt van de populatieverdeling is in tabel 4.1 voor iedere achtergrondvariabele ook nog een effectgrootte Phi (Phi = sqrt[Chi-kwadraat/N]) berekend, zodat duidelijk wordt voor welke achtergrondgegevens uit de normeringspopulatie 2020-2021 er sprake is van eventuele grote effecten (waarmee waarden van Phi ≥0.50 worden bedoeld (Cohen, 1988)). Uit de gerapporteerde waarden in tabel 4.1 kan geconcludeerd worden dat er geen sprake is van grote effecten voor de achtergrondvariabelen denominatie, urbanisatiegraad, schoolgrootte, regio en schoolweging (respectievelijk 0.18, 0.22, 0.25, 0.25 en 0.35). Dit impliceert dat voor de toekomst de gemiddelde groeifactor jaarlijks geëvalueerd wordt aan de hand van representatieve afnamedata en indien nodig wordt aangepast (zie hoofdstuk 9).

Conclusie:
De procedure voor het samenstellen van de steekproeven is onderbouwd en de omstandigheden en momenten waaronder data is verzameld, is vergelijkbaar met de afnamecondities en afnamemomenten waaronder de IEP LVS toets Rekenen 6a wordt afgenomen. Hoewel de normeringspopulatie niet op alle achtergrondcategorieën als representatieve doelsteekproef beschouwd kan worden, is deze echter wel bruikbaar voor het doel waarmee de data verzameld zijn (i.e., geen relatieve maar absolute normering van de referentieniveaus). Op aspect S1 wordt aan de IEP LVS toets Rekenen 6a het oordeel '**voldoende'** toegekend.

S2 In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:
In het schooljaar 2020-2021 is eerst de pretest IEP LVS toets Rekenen 6a afgenomen, welke uit meer items bestond dan de IEP LVS toets Rekenen 6a definitief bevatte. De selectie van de items per toets is vastgesteld op basis van de toets- en itemanalyses op de responsedata van de afnames in het normeringsonderzoek tijdens de pretestfase en het design van de definitieve toets. De afnamecondities van de pretest IEP LVS toets Rekenen 6a waren identiek aan de definitieve IEP LVS toets Rekenen 6a.

Tabel 5.1 toont het design van het normeringsonderzoek in de pretest fase, waarin per toets (toets 5b, toets 6a en toets <1F-1F versie 1) is weergegeven hoeveel items van welk niveau ieder van de drie toetsen in de pretestfase bevatte. In de laatste kolom staat het aantal observaties per toets (3505, 5220 en 4927 voor respectievelijk toets 5b, toets 6a en toets <1F-F versie 1).

Uit tabel 5.1 valt op te maken dat toets 5b (totaal 40 items in pretestfase) en toets 6a (totaal 50 items in pretestfase) in de pretestfase altijd 5 opgaven beneden en 5 opgaven boven het niveau van de betreffende toets bevatten. Zo bevat bijvoorbeeld de toets Rekenen 6a ook 5 opgaven op niveau 5b (het niveau eronder) en ook 5 opgaven op niveau <1F-F versie 1 (het niveau erboven). Dit reflecteert het onderwijskundige uitgangspunt dat toets 6a is ontwikkeld om de inhoudelijke aansluiting tussen de 5b toets en de <1F-F versie 1 toets te verbeteren. Op deze manier meten de toetsen een breed vaardigheidsgebied en geeft het een indicatie of een leerling de lesdoelen van een half jaar geleden nog steeds beheerst (i.e., op niveau 5b) en/of al misschien wel moeilijkere opgaven kan maken (i.e., op niveau <1F-1F versie 1).

De items van het niveau onder en boven het niveau van de toets hebben, naast een onderwijskundig doel, ook een psychometrisch doel. De items van het niveau onder en boven het niveau van de toets dienen als ankeritems (items die het design linken) tussen de verschillende niveautoetsen, welke het onvolledige dataverzamelingsdesign tot een verbonden ('linked design') maakt en de mogelijkheid biedt om de ontwikkeling van de leerlingen psychometrisch verantwoord te plotten en blijven volgen op één en dezelfde schaal (unidimensionaliteit). Hierdoor zijn de verschillende niveautoetsen dus onderling vergelijkbaar.

Paragraaf 6.1 beschrijft de wijze waarop van de pretest toets Rekenen 6a van 50 items de selectie van 40 items voor de definitieve samenstelling van toets Rekenen 6a tot stand is gekomen op basis van TIA-analyses (toets- en itemanalyses) en het design van de definitieve toets. Tabel 6.1 toont het ankerdesign waaraan de itemselectie voor de definitieve samenstelling van toets Rekenen 6a moest voldoen: 30 items van het niveau van de toets (i.e., toets 6a), 5 items van het niveau eronder (i.e., toets 5b) en 5 items van het niveau erboven (i.e., toets <1F-1F versie 1). In totaal bestaat de definitieve samenstelling van toets Rekenen 6a dus uit 40 items. Daarnaast is er weer sprake van enige overlap tussen de toetsen van het opvolgende niveau. Iedere toets heeft 5 items overlap met een toets van een lager niveau en 5 items overlap met een toets van een hoger niveau.

Op basis van dit definitieve design heeft de itemselectie plaatsgevonden door te kijken naar de statistieken afkomstig uit de TIA-analyse (bijlage 5) over de responsedata uit het normeringsonderzoek, waarbij de volgende uitgangspunten golden:

- In de samenstelling moet de inhoudelijke dekking van het IEP LVS Toetskader (zie hoofdstuk 2) gewaarborgd zijn;
- Items hebben een rit-waarde (item-totaalcorrelatie) van groter of gelijk aan 0.20, zodat de items als voldoende beoordeeld kunnen worden volgens het beoordelingssysteem van de COTAN (Evers, Lucassen, Meijer, & Sijtsma, 2010);
- Items van het niveau van de toets hebben een p-waarde (proportie correct) van minimaal 0.30 en maximaal 0.90;
- Items van het niveau onder de toets hebben een p-waarde van minimaal 0.30 en maximaal 0.95;
- Items van het niveau boven de toets hebben een p-waarde van minimaal 0.25 en maximaal 0.90.

Alle items in de definitieve samenstelling van de IEP LVS toets Rekenen 6a voldoen aan alle bovenstaande uitgangspunten (op het eerste uitgangspunt worden geen uitzonderingen gemaakt), waarmee er dus sprake is van zowel een goede psychometrische afspiegeling als een goede inhoudelijke dekking van het IEP LVS Toetskader (zie hoofdstuk 2) met variatie in domeinen en domeinonderwerpen (i.e., inhoudsvaliditeit). Ook blijkt uit tabel 6.2 dat de globale betrouwbaarheid (zoals geschat met Cronbach's Alpha) van de toets 6a (0.865) vergelijkbaar is met de globale betrouwbaarheden van de toetsen 5b en <1F-1F versie 1 (respectievelijk 0.868 en 0.866), waarmee toets 6a overlap heeft.

Het 1PLM (Rasch model) met de Marginal Maximum Likelihood (MML) schattingsmethode is toegepast voor de IRT-analyses van de toetsen IEP LVS toetsen en is uitgevoerd in Versie 0.4.0 van de applicatie Lexter (2021). De keuze voor een 1PLM analyse in Lexter wordt door de auteurs beargumenteerd vanuit de gedachte dat het scoremodel van een

punt per item (ongeacht het discriminatievermogen) het voor het onderwijs het meest begrijpelijke en dus bruikbare model is.

De items van de toetsen IEP LVS Rekenen <1F-1F versie 1, 6a en 5b zijn op dezelfde parameterschaal gezet, d.w.z. kalibratie van alle items op een doorlopende lijn van vaardigheidsschattingen op dezelfde onderliggende vaardigheidsschaal. De toetsen Rekenen <1F-1F versie 1 en 5b zijn bestaande toetsen die in respectievelijk 2019 en 2020 zijn verantwoord. In het schooljaar 2020-2021 is in de toets Rekenen 5b twee keer een aantal items gezaaid, welke het anker vormen (in totaal 7 items) tussen de toets Rekenen 5b en 6a. Daarom zijn er in de schaling van de itemparameters (bijlage 4) twee booklets voor de 5b toets te zien. Vanaf schooljaar 2021-2022 maken 5 van deze 7 gezaaide items deel uit van de toets 5b, welke het oude anker vervangen in de toets 5b dat de 5b toets met de <1F-1F versie 1 toets verankerde. De schalingen die in hoofdstuk 7 worden besproken, zijn gebaseerd op de nieuwe samenstelling van de 5b toets.

De vaardigheid bij de verschillende scoringspunten (inclusief het cesuurpunt) wordt bij de kalibratie van de items in Lexter berekend. In het document Itemparameters IEP LVS Rekenen (bijlage 4) zijn het aantal afnames, de itemmoeilijkheid (β), de meetfout van de moeilijkheid (SE(β)) en de iteminformatie (ook wel Fisher-informatie genoemd), bij de gemiddelde vaardigheid van de afnamegroepen van de IEP LVS toetsen Rekenen <1F-F versie 1, 6a en 5b in een tabel weergegeven voor alle parameterschattingen. Hierbij is de vaardigheidsschaal genormeerd door van de afnamegroep van de 5b toets de gemiddelde vaardigheid en de standaarddeviatie hiervan op respectievelijk 0 en 1 te zetten (de gemiddelde vaardigheid van de afnamepopulatie van de twee booklets 5b verschilde nauwelijks van elkaar met -0.004 en 0 respectievelijk).

In paragraaf 7.3 wordt een methode besproken om de nauwkeurigheid van de parameterschattingen te beoordelen (Evers et al., 2010). Deze methode bestaat eruit om de nauwkeurigheid van de parameterschattingen na te gaan aan de hand van de constante 'c', die de relatie weergeeft tussen de standaardfout van de moeilijkheidsparameter van een item (SE(β)) en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie (σ_θ). Volgens het COTAN-beoordelingssysteem (Evers et al., 2010) worden waarden van c lager of gelijk aan 0.2 als 'goed' beoordeeld en waarden tussen 0.3 en 0.4 als 'voldoende'. De nauwkeurigheid van de parameterschattingen is onderzocht door de c-waarden te berekenen voor de parameterschattingen uit de gezamenlijke kalibratie van de moeilijkheidsparameters van de items van de IEP LVS toetsen Rekenen <1F-1F versie 1, 6a en 5b. Tabel 7.2 rapporteert de standaarddeviatie van de vaardigheid van de kalibratiepopulatie en de minimum-, maximum- en gemiddelde waarde van de constante c per kaibratie. Uit Tabel 7.2 is af te lezen dat de gemiddelde waarde van de constante c (i.e., 0.036), berekend over alle items in de kalibratie, veel lager is dan de vereiste waarde van 0.2 en geen item heeft een c-waarde boven de 0.2 (maximum-waarde voor de constante c is 0.067). Op basis van deze resultaten kan de nauwkeurigheid van de parameterschattingen als goed beoordeeld worden.

Modelpassing is onderzocht, mede in het kader van constructvalidering (zie hoofdstuk 8), via een Differential Item Functioning (DIF) analyse (uitgevoerd in de applicatie Lexter) op de 12 ankeritems tussen de opeenvolgende niveautoetsen. Daarnaast is via statistische toetsen nagegaan of de Item Response Curven (IRC's) de responsies goed representeren (eveneens uitgevoerd in Lexter). De DIF van de ankeritems tussen de opeenvolgende niveautoetsen is berekend op basis van de gezamenlijke kalibratie van de IEP LVS toetsen beschreven in paragraaf 6.2, waarbij ieder van deze ankeritems in twee booklets is

opgenomen geweest en het aantal vrijheidsgraden dus 1 is (df = 2 – 1 = 1). In bijlage 7 is per item de Lagrange Multiplier statistiek (LM) weergegeven, het aantal vrijheidsgraden (df), de overschrijdingskans (Prob), het absolute verschil (Abs. Diff) en is daarnaast aangegeven in welke boekjes het item was opgenomen.

Met name de absolute verschillen zijn informatief, omdat significantie altijd gevoelig is voor de steekproefgrootte (bij grote steekproeven is een Chi-kwadraat toets bijna altijd significant). De tabel in bijlage 7 laat zien dat geen van de ankeritems een absoluut verschil van groter dan 0.10 heeft. Het gemiddelde absolute verschil van de 12 ankeritems is 0.02 (max = 0.05; min = 0.00). Op basis van deze resultaten kan gesteld worden dat, ongeacht het boekje, de observaties op de ankeritems tussen de IEP LVS boekjes Rekenen in deze schaling weinig van de verwachting verschillen. Er kan derhalve geconcludeerd worden dat met betrekking tot DIF een goede modelpassing aannemelijk is.

Bijlage 8 toont door middel van de First order Statistics optie ('Lagrange multiplier tracelines for Rasch-Type Model') uit de applicatie Lexter de mate waarin de Item Response Curven de responsies goed representeren door dit statistisch te toetsen. Er zijn 155 effectgroottes berekend, waarvan er geen groter is dan 0.1. Dit toont aan dat alle items in deze schaling in het Rasch-model blijken te passen en dat de modelpassing ook volgens deze analyse dus zeer goed is.

Op pagina 22 wordt gesproken over 12 ankeritems tussen de opeenvolgende niveautoetsen, waarvan de DIF zal worden berekend. Volgens het 'Beoordelingskader voor instrumenten binnen leerlingvolgsystemen' (zie website van de Expertgroep Toetsen PO) zou het anker bij een onvolledig maar 'verbonden' dataverzamelingsdesign echter uit tenminste 15 ankeritems moeten bestaan. Het in het onderhavige onderzoek gehanteerde design voldoet dus niet aan deze minimale eis. De toetsaanbieder heeft aangegeven dit punt op te nemen.

Conclusie:
Het onvolledige maar 'verbonden' dataverzamelingsdesign is adequaat. Op aspect S2 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende'** toegekend.

De Expertgroep merkt op dat, gezien het feit dat de gegevensverzameling al uitgevoerd is, de investering in een nieuwe dataverzameling niet opweegt tegen de winst in nauwkeurigheid door het gebruik van een groter anker. Bij de kalibratie van toekomstige LVS toetsen wordt nadrukkelijk gevraagd om een anker met een grootte boven het absolute minimum te gebruiken.

S3 In het geval van een observatie-instrument: is er sprake van een adequate steekproef van observatoren en randvoorwaarden waaronder de observatie wordt uitgevoerd?

Bevindingen:
Dit criterium is niet van toepassing (n.v.t.), omdat er hier sprake is van de IEP LVS-toets Rekenen 6a en er dus geen sprake is van een observatie-instrument.

Conclusie:
n.v.t.

S4 Er is een handleiding met duidelijke instructies voor de leerkracht over het zo objectief mogelijk uitvoeren en weergeven van de observaties door de leerkracht.

Bevindingen:
Dit criterium is niet van toepassing (n.v.t.), omdat er hier sprake is van de IEP LVS toets Rekenen 6a en er dus geen sprake is van een observatie-instrument.

Conclusie:
n.v.t.

Normering

N1.1 Is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd?

Bevindingen:
Er is een standaardsetting uitgevoerd om vast te kunnen stellen vanaf welke ruwe score de leerling het gemeten niveau op de IEP LVS toets Rekenen 6a heeft behaald (cesuur). Er is dus sprake van een criteriumgerichte interpretatie van de toetsscores, waarbij de toetsscores vergeleken worden met een absolute norm. Experts uit het betreffende vakgebied, en deel uitmakend van een geïnstalleerde standaardsettingscommissie, bepalen de cesuur bij deze methode. Er is een licht aangepaste Bookmark methode (Karantonis & Sireci, 2006) tijdens deze standaardsettingsprocedure gevolgd, welke uit drie stappen bestaat. Als eerste stap in deze methode ('Informeren van standaardsetters') worden experts uit het betreffende vakgebied zodanig geïnformeerd dat zij een beeld over de 'borderline kandidaat' krijgen, dat wil zeggen, van de leerling die het niveau van de toets net behaalt. De experts worden daarnaast ook geïnformeerd over de consequenties van de niveau-uitspraak voor de leerling en de betekenis daarvan voor het onderwijs. Als tweede stap in deze methode geven de experts individueel per item aan hoeveel procent van de borderline kandidaten het item goed zullen maken, ofwel de kans dat een borderline kandidaat het item goed zal maken. Als derde en laatste stap in deze methode ('Bereiken van consensus') wordt er een discussie gevoerd tussen de experts om tot een unanieme uitspraak te komen over de resultaten van stap 2, welke vervolgens resulteert in een cesuur op de toets.

Omdat de tweede stap van deze methode lastig is voor onervaren standaardsetters, is die stap vervangen door een deel van de Bookmark methode. De standaardsetters moesten hierbij eerst de 40 items uit de definitieve toetssamenstelling voor toets 6a (zie tabel 6.1) individueel op volgorde van makkelijk naar moeilijk leggen (OIB = Ordered Item Booklet), waarbij ze geen psychometrische informatie over de items (bijvoorbeeld p-waarden uit de KTT of β-waarden uit de IRT) en de toets (bijvoorbeeld globale betrouwbaarheid) kregen en de standaardsetters dus zelf moesten nadenken over het niveau en de moeilijkheid van de items van de verschillende niveaus. Vervolgens moesten de standaardsetters individueel een 'bookmark' (bladwijzer) plaatsen in de door henzelf eerder gemaakte volgorde van items van makkelijk naar moeilijk. De 'bookmark' betekent dat leerlingen die het niveau net beheersen (i.e., de borderline kandidaat) de items tot de 'bookmark' goed zouden moeten maken en de items na de 'bookmark' niet goed zouden hoeven te maken.

De hierboven beschreven stappen van de vier individuele standaardsetters zijn geanalyseerd, waarbij uit de lijsten van de standaardsetters met items op volgorde van makkelijk naar moeilijk als eerste de items van het niveau onder (5 items) en boven het niveau van de toets (5 items) zijn verwijderd. Alle lijsten waren hierna dus teruggebracht naar 30 items van het niveau van de toets. De 'bookmark' die de standaardsetters hebben gezet, is dus nu een 'bookmark' geworden op 30 items. Omdat de leerling ook de 5 items van het niveau onder de toets goed moet maken om het niveau te behalen, is elke bookmark (één bookmark voor elke standaardsetter) omgezet in een cesuur door het aantal items tot aan de bookmark te tellen en daar 5 bij op te tellen. Op dit moment bleek er in de standaardsettingsprocedure nog geen consensus te bestaan over de vast te stellen cesuur.

In stap 3 van de gevolgde standaardsettingsprocedure ('Bereiken van consensus') is daarom door de procesbegeleider (Bureau ICE heeft de rol van procesbegeleider in deze standaardsetting op zich genomen) aan de standaardsetters een voorstel voor de cesuur gedaan op basis van het gemiddelde van de individuele cesuren. Daarbij verstrekte de procesbegeleider feedback aan de standaardsetters over het slagingspercentage bij de voorgestelde cesuur (impact feedback) en over de cesuren van de andere standaardsetters (normatieve feedback). Dit werd gedaan om de standaardsetters te helpen het beoordelingsproces goed te doordenken. Omdat het toetskader dat ten grondslag ligt aan de IEP LVS toetsen een cumulatief karakter heeft (hoofdstuk 2) is het voorstel voor de cesuur waar nodig aangepast, zodanig dat de vaardigheid die nodig is om de cesuur op de verschillende toetsen te halen oploopt naarmate het gemeten niveau van de toets toeneemt (zie tabel 7.1, vaardigheid voor behalen cesuur voor toets 5b, 6a en <1F-1F versie 1 zijn respectievelijk -0.171; 0.672; 1.565). Aan de hand van dit voorstel is vervolgens de discussie gevoerd (Delphi-methode) en dit voorstel bediscussieerd net zo lang dat er voor elke cesuur volledige overeenstemming was bereikt op elke toets van de definitieve toetssamenstelling (zie tabel 7.1). Voor de IEP LVS toets Rekenen 6a resulteerde die discussie in een cesuur van 28 op in totaal 40 items (zie tabel 7.1), d.w.z. op 70% van de maximaal te behalen score van 40.

In tabel 7.1 worden de cesuren van alle drie de IEP LVS toetsen Rekenen 5b, 6a en <1F-1F versie 1 gerapporteerd in scorepunten op de definitieve toetssamenstelling (40 items). In tabel 7.1 wordt, naast de cesuren in scorepunten, voor iedere cesuur de bijbehorende vaardigheid theta (θ), de lokale meetfout van de vaardigheid, de lokale betrouwbaarheid (een lage lokale meetfout van de vaardigheid bij de cesuur resulteert in een hoge lokale betrouwbaarheid) en het percentage leerlingen dat het betreffende niveau ten onrechte wel of niet heeft gehaald (classificatiefouten) gegeven. In bijlage 5 ('Algemene toelichting methode') is meer uitleg te vinden voor wat betreft de berekening van de lokale betrouwbaarheden en misclassificaties.

Daarnaast worden in tabel 7.1 ook nog de 95% betrouwbaarheidsintervallen (BTIs) gerapporteerd, waaruit blijkt dat de betrouwbaarheidsintervallen van de theta's (en dus van de cesuren) elkaar (flink) overlappen. Volgens de auteurs is het overlappen van de betrouwbaarheidsintervallen niet problematisch, omdat het laat zien dat de vaardigheidsontwikkeling van de leerlingen heel gelijkmatig verloopt en dus dat de onderwijsdoelen van de drie toetsen goed op elkaar aansluiten c.q. voortborduren. Tabel 7.1 laat overigens ook duidelijk zien dat zowel de ondergrenzen als de bovengrenzen van de betrouwbaarheidsintervallen netjes oplopend zijn van toets 5b tot en met toets <1F-

1F versie 1, hetgeen volgens verwachting is op basis van het cumulatieve karakter van het toetskader dat ten grondslag ligt aan de IEP LVS toetsen (hoofdstuk 2).

Conclusie:
De standaardbepalingsmethode is gemotiveerd en op de juiste wijze uitgevoerd. Op aspect N1.1 wordt aan de toets IEP LVS Rekenen 6a het volgende oordeel toegekend: '**voldoende'.**

N1.2 Zijn de beoordelaars/vakdeskundigen/experts naar behoren geselecteerd en getraind?

Bevindingen:
Er is een standaardsettingscommissie geïnstalleerd, bestaande uit twee inhoudelijke experts en twee leerkrachten, om zodoende zowel kennis uit het vakgebied Rekenen als ervaring uit de praktijk samen te brengen. Omdat de toetsen van het leerlingvolgsysteem geen high stakes toetsen zijn, werd vier standaardsetters per standaardsetting als voldoende geacht. De inhoudelijke experts moesten voldoen aan de eis dat zij gespecialiseerd zijn in het vakgebied Rekenen en daarnaast werkzaam zijn als onderwijskundig of taalkundig adviseur in het primair onderwijs. Hoewel ervaring als leerkracht in het basisonderwijs voor de inhoudelijke experts niet vereist was, was dit wel wenselijk. De twee leerkrachten uit de standaardsettingscommissie moesten wel recentelijk ervaring hebben in het lesgeven aan leerjaar 6 en bij voorkeur op dit moment ook werkzaam zijn als leerkracht van leerjaar 6. Bureau ICE had de rol van procesbegeleider op zich genomen in deze standaardsetting.

Via een werving binnen het eigen netwerk van Bureau ICE en sociale media zijn de vier leden van de standaardsettingscommissie geworven. Op het competentieprofiel van de geïnteresseerden is een screening uitgevoerd, hetgeen heeft geresulteerd in een standaardsettingscommissie die voldoet aan bovenstaande eisen. Vooraf hebben de standaardsetters een geheimhoudingsverklaring ondertekend, waarin zij verklaren dat zij de toetsinhoud vertrouwelijk behandelen.

In stap 1 ('Informeren van standaardsetters') van de gevolgde standaardsettingsprocedure (i.e., licht aangepaste Bookmark methode) werden de standaardsetters goed geïnformeerd over de aanleiding en het doel van de standaardsetting en hebben zij voor het 6a-niveau de niveauomschrijvingen (bijlage 3) en de Toetsmatrijs (bijlage 1, hoofdstuk 3.4) goed bekeken. Ook ontwikkelden zij op basis hiervan een beeld van de borderline kandidaat, d.w.z. conceptualisering van de borderline kandidaat.

Conclusie:
De beoordelaars/vakdeskundigen/experts zijn naar behoren geselecteerd en getraind. Op aspect N1.2 wordt aan de IEP LVS-toets Rekenen 6a het oordeel '**voldoende**' toegekend.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
In stap 3 ('Bereiken van consensus') van de gehanteerde standaardsettingsprocedure (licht aangepaste Bookmark methode) is het door de procesbegeleider gedane voorstel (gebaseerd op het gemiddelde van de cesuren van de individuele standaardsetters) voor de cesuur bediscussieerd en bijgesteld net zo lang tot er voor de cesuur volledige overeenstemming was bereikt.

Conclusie:
Er is voldoende overeenstemming tussen de beoordelaars. Op aspect N1.3 wordt aan de toetsen IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
De IEP LVS toets Rekenen 6a toets is genormeerd voor de eerste helft van leerjaar 6. Uit tabel 4.1 op pag. 13 valt af te lezen dat de normgroep (normeringspopulatie) voor de toets IEP LVS Rekenen 6a gelijk is aan 5195. Deze normgroep is van voldoende grootte. Op pag. 7, 3e regel van onderen, van het document 'Scenario's voor ijking van de eindtoetsen op de referentieniveaus' van Glas, Emons en Berding-Oldersma (december 2016) wordt namelijk gesteld dat voor een betrouwbare cesuur het aantal leerlingen in de steekproef minimaal 1000 moet zijn. Dit document is te vinden op de homepage van de Expertgroep Toetsen PO onder 'Overige Informatie – Onderzoeken', uitgevoerd door de Expertgroep.

Conclusie:
De normgroep is groot genoeg. Op aspect N2.1 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend.

N.2.2 Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproef (normeringspopulatie) werd hierboven onder aspect S1 besproken en daar werd geconstateerd dat de representativiteit van de normeringspopulatie op de achtergrondvariabelen niet optimaal was om als normpopulatie voor het gehele reguliere basisonderwijs (doelpopulatie) te fungeren. Volgens de auteurs was het echter ook niet per se noodzakelijk dat de normeringspopulatie ook een normpopulatie is en zijn de representativiteitseisen aan een normpopulatie in het onderhavige onderzoek dan ook niet van toepassing. De argumentatie hiervoor is dat in dit normeringsonderzoek absolute cesuren op de toets 6a worden bepaald, waarvoor de samenstelling van de normeringspopulatie volgens de auteurs van ondergeschikt belang is.

Conclusie:
De normgroep is representatief. Op aspect N2.2 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend, waarbij wel dezelfde kanttekening wordt geplaatst als onder opmerking 1 bij aspect S1 ('Is de steekproef van leerlingen representatief?').

N2.3 Zijn de normen correct bepaald?

Bevindingen:
Hoewel in principe ook andere standaardsettingsprocedures gebruikt hadden kunnen worden (bijv. de originele Bookmark methode i.p.v. de hier gevolgde licht aangepaste Bookmark methode en dus empirische i.p.v. subjectieve moeilijkheden gebruiken om de 40 items van de IEP LVS toets Rekenen 6a op volgorde van makkelijk naar moeilijk te ordenen), past de gevolgde standaardsettingsprocedure bij de data van de normeringspopulatie en er zijn ook schattingsfouten van de cesuren berekend.

Conclusie:
De normen zijn correct bepaald. Op aspect N2.3 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In bijlage 5 (TIA's toets Rekenen 6a) wordt de globale betrouwbaarheid (Cronbach's alpha) weergegeven voor de IEP LVS-toets Rekenen 6a, welke is berekend met het programma TiaPlus. Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt. Onder gebruikmaking van het programma Lexter worden in tabel 7.1 ook de lokale betrouwbaarheden, gemeten bij de cesuurpunten op de latente vaardigheidsschaal, weergegeven voor de IEP LVS toetsen Rekenen <1F-F versie 1, 6a en 5b. In bijlage 6 ('Algemene toelichting methode') wordt gedetailleerde uitleg gegeven hoe deze (conditionele) lokale betrouwbaarheden, gegeven een vaardigheidsniveau θ, kunnen worden berekend via de (ook in tabel 7.1 weergegeven) lokale meetfout/meetnauwkeurigheid (SEM_θ) en de standaarddeviatie van de vaardigheidsverdeling (σ_θ). Deze (conditionele) lokale betrouwbaarheid vertoont qua interpretatie grote overeenkomst met de globale betrouwbaarheidscoëfficiënt Cronbach's alpha uit de Klassieke Testtheorie (KTT) en wordt in de psychometrische literatuur beschreven (Raju, Price, Oshima & Nering, 2007) en als correct aangemerkt.

Daarnaast worden in tabel 7.1 voor de drie IEP LVS-toetsen Rekenen ook nog het percentage leerlingen berekend dat het betreffende niveau ten onrechte wel of niet heeft gehaald (classificatiefouten). In bijlage 5 ('Algemene toelichting methode') wordt op een correcte manier beschreven hoe deze classificatiefouten worden berekend. Omdat voor de berekeningen gebruik is gemaakt van bekende en algemeen beschikbare software (i.e., TiaPlus en Lexter), kunnen we ervan uitgaan dat de betrouwbaarheidsgegevens correct zijn berekend.

Conclusie:
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1 wordt aan de toets IEP LVS Rekenen 6a het oordeel **'voldoende'** toegekend.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden?

Bevindingen:
Uit tabel 6.2 kan afgelezen worden dat de globale betrouwbaarheden in termen van Cronbach's alpha (interne consistentie betrouwbaarheden) voor de drie IEP LVS toetsen Rekenen <1F-1F versie 1, 6a en 5b gelijk zijn aan respectievelijk 0.866, 0.865 en 0.868. Omdat het behalen van één van de niveaus <1F-1F versie 1, 6a en 5b valt in de categorie van minder belangrijke beslissingen op individueel niveau, is met deze waarden voor de globale betrouwbaarheden ruimschoots voldaan aan de eis van het COTAN-beoordelingssysteem (Evers et al., 2010) dat de minimale betrouwbaarheidscoëfficiënt van toetsen voor minder belangrijke beslissingen tenminste 0.70 moet zijn. Tabel 7.1 laat zien dat aan deze minimale eis ook ruimschoots wordt voldaan voor de lokale betrouwbaarheden bij de cesuurpunten van de toetsen Rekenen <1F-1F versie 1, 6a en 5b, welke gelijk zijn aan respectievelijk 0.856, 0.865 en 0.876.

Verder laat tabel 7.1 nog zien dat de classificatiefouten (i.e., voor de drie IEP LVS-toetsen Rekenen het percentage leerlingen dat het betreffende niveau ten onrechte wel of niet heeft gehaald) loopt van 11% tot 13% (hoe hoger de lokale betrouwbaarheid, hoe lager de classificatiefout). Deze percentages hebben betrekking op scores dicht bij een cesuur en er geldt dan ook dat het percentage misclassificaties bij een score verder van de cesuur af per definitie lager is. Omdat de berekende classificatiefouten in de context van de IEP LVS toetsen geen summatieve toetsen betreft waarop een leerling kan zakken of slagen, heeft een misclassificatie daarmee voor de leerling geen directe grote gevolgen. In combinatie met het feit dat de IEP LVS toetsen volgtoetsen zijn waar geen belangrijke beslissingen mee worden genomen, kan er geconcludeerd worden dat de classificatiefouten als acceptabel gezien kunnen worden.

Conclusie:
De betrouwbaarheid van de IEP LVS-toets Rekenen 6a is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en ingestemd wordt met de eisen voor de betrouwbaarheid van het COTAN-beoordelingssysteem (Evers et al., 2010). Op aspect B2 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende'** toegekend.

Validiteit

V1 Inhoudsvaliditeit: Dragen de items in het instrument bij aan de validiteit van het instrument (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)?

Bevindingen:
De items in de toets toetsen de doelen die ze beogen te toetsen. Er kunnen geen misverstanden ontstaan over de juistheid van de gegeven antwoorden.

Conclusie:
'voldoende'; de items in de toets dragen bij aan de validiteit van de toets.

V2 Constructvaliditeit: Meet het instrument in zijn geheel datgene wat het beoogt te meten?

Bevindingen:
In paragraaf 7.3 ('Passing van het meetmodel en nauwkeurigheid van de parameterschattingen') was al aannemelijk gemaakt dat er sprake was van een passing van het geassumeerde meetmodel (i.e., het Rasch model) en er dus mag worden uitgegaan van unidimensionaliteit, hetgeen impliceert dat aan de noodzakelijke (maar niet voldoende) voorwaarde van constructvaliditeit wordt voldaan. In hoofdstuk 8 wordt aanvullend onderzoek verricht naar de onderstaande andere aspecten, welke kunnen worden opgevat als enkele argumenten die pleiten voor de constructvaliditeit van de IEP LVS-toetsen Rekenen: (1) (inter)correlationeel onderzoek tussen de inhoudelijke domeinen binnen de IEP LVS-toets Rekenen 6a, (2) onderzoek naar divergente validiteit, (3) itemkwaliteit (psychometrische kwaliteit van de items).

1. Uit tabel 8.1 valt af te lezen dat er voor de toets IEP LVS Rekenen 6a een middelmatig tot sterk correlationeel verband is tussen de scores op het domein Getallen (G) en het domein Meten & Meetkunde (M) met een correlatiecoëfficiënt van 0.68, tweezijdig significant is op het 1%-niveau. Deze waarde van de correlatiecoëfficiënt toont aan dat er sprake is van een middelmatig verband tussen de domeinen Getallen en Meten & Meetkunde (G/M) binnen de toets Rekenen 6a. Het 'slechts' middelmatige verband is echter geenszins onverwacht dan wel onwenselijk, omdat de verdeling in domeinen conform het Referentiekader is gebaseerd op zowel inhoudelijke differentiatie als de samenhang van begrippen en methoden (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008). De aandacht voor de domeinen in het basisonderwijs verschilt met het vervolgonderwijs en ook tussen de leerjaren van het basisonderwijs.

2. Er is op de IEP LVS toets een soortgenotenonderzoek uitgevoerd in de vorm van onderzoek naar divergente validiteit tussen de toets Rekenen 6a en verschillende vaardigheden (Lezen en Technisch Lezen) die binnen dezelfde periode (i.e., eerste helft van leerjaar 6) als de toets Rekenen 6a zijn afgenomen. Opdat er sprake is van divergente validiteit zouden scores op de toets Rekenen 6a laag moeten correleren met toetsen die een ander construct meten, zoals Lezen en Technisch Lezen. De resultaten van dit onderzoek worden gerapporteerd in tabel 8.2. Conform de verwachting is hieruit af te lezen dat de correlaties tussen de toetsen Rekenen 6a en Lezen (i.e., 0.44) en tussen Rekenen 6a en Technisch Lezen (i.e., 0.18) in dezelfde periode respectievelijk middelmatig en zwak zijn. Dit duidt erop dat de toets Rekenen 6a een andere vaardigheid meet dan bij de toetsen Lezen en Technisch Lezen in dezelfde periode en er dus inderdaad sprake is van divergente validiteit.

3. In tabel 8.3 zijn de gemiddelden en ranges van de p- en rit-waarden weergegeven (gebaseerd op de uiteindelijke selectie van items per toets). Uit deze tabel blijkt dat met name de gemiddelden (gemiddelde p-waarde = 0.68; gemiddelde rit-waarde = 0.40) voldoen als criterium voor de itemkwaliteit per toets. De grootte van de ranges in tabel 8.3 wordt sterk bepaald door enkele outliers (zie hiervoor ook bijlage 5).

Conclusie:
De gerapporteerde resultaten in Hoofdstuk 8 (Constructvaliditeit) vormen een psychometrische ondersteuning voor de constructvaliditeit van de toets IEP LVS Rekenen 6a en er wordt dus gemeten wat men beoogt te meten, namelijk rekenvaardigheid bij de eerste helft van leerjaar 6. Op aspect V2 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende'** toegekend.

Het volg-aspect

Va1 Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een correcte manier gemeten?

Bevindingen:
Alle IEP LVS toetsen Rekenen leerjaar 3 t/m 6 zijn gekalibreerd op één en dezelfde onderliggende vaardigheidsschaal theta (θ). Het gevolg hiervan is dat de vaardigheidsscores op deze toetsen onderling vergelijkbaar zijn en de vaardigheidsontwikkeling van de leerlingen gevolgd kan worden door hun scores op de verschillende opeenvolgende momenten met elkaar te vergelijken.

Omdat deze (latente) vaardigheidsschaal moeilijk te interpreteren is voor leerkrachten, is de vaardigheidsschaal getransformeerd naar een lineaire schaal, die de ontwikkelscoreschaal wordt genoemd. Dit is een zinvolle schaal waar leerkrachten en leerlingen aan gewend zijn en is ook inzichtelijk voor ouders, voogden of verzorgers. Deze schaal loopt van 0 tot maximaal 60 punten. In de ontwikkelscoreschaal van de IEP LVS toetsen Rekenen leerjaar 6 t/m 8 (gebaseerd op het referentiekader) is namelijk het referentieniveau 1F gelijk aan 60 ontwikkelscorepunten. De ontwikkelscoreschaal van de IEP LVS toetsen Rekenen leerjaar 3 t/m 6 moet hieronder blijven, omdat deze toetsen inhoudelijk voorbereidend zijn op het referentieniveau F1. Omdat de inhoud van de verschillende toetsen IEP LVS toetsen Rekenen leerjaar 3 t/m 6 stapelend is qua inhoudelijke onderwijsdoelen (i.e., hiërarchische opbouw), is ervoor gekozen het scorebereik van de verschillende toetsen op de ontwikkelscoreschaal niet overlappend te laten zijn.

Verder is ervoor gekozen om iets boven de nul te beginnen op de ontwikkelscoreschaal, waarmee aan de gebruikers (leerkrachten, ouders en leerlingen) wordt duidelijk gemaakt dat ook in de leerjaren 1 en 2 natuurlijk sprake is van vaardigheidsontwikkeling. Een ontwikkelscoreschaal onder de 0 is theoretisch mogelijk, maar niet wenselijk omdat de schaal inzichtelijk moet zijn voor leerkrachten, ouders, voogden of verzorgers. Vanuit dit uitgangspunt staan de plaatsen van de niveaus 3a t/m 6a voor de verschillende vaardigheden (Lezen, Technisch Lezen en Taalverzorging) op eenzelfde punt op de ontwikkelscoreschaal en deze punten staan daardoor dus ook op gelijke onderlinge afstand.

Voor het omzetten van de toetsscore (het aantal goed beantwoorde vragen) naar de ontwikkelscore (OS) zijn voor ieder van de IEP LVS toetsen Rekenen 3a t/m 6a drie vaste punten als uitgangspunt genomen: de bodem, de cesuur en het plafond. Tabel 9.1 (Ontwikkelscorebereiken per toets) laat zien dat in iedere toets het totale ontwikkelscorebeleid uit 8 te behalen ontwikkelscorepunten bestaat. De schaal voor Rekenen voor leerjaar 3 t/m 5 blijft hierbij gelijk aan hoe deze in 'Verantwoording IEP LVS

Rekenen 2020' verantwoord is en het bereik voor de ontwikkelscoreschaal voor de IEP LVS toets Rekenen 6a komt hier dus bovenop.

In paragraaf 7.1 is via een standaardsettingsprocedure de cesuur op de toetsscore voor de IEP LVS toets Rekenen 6a bepaald (zie tabel 7.1). Hoewel de cesuurpunten op de toetsen in het IEP LVS leerjaar 3 t/m 6 verschillen per toets, is voor de bodem en voor het plafond besloten deze voor alle toetsen op een gelijke toetsscore te leggen. Hierbij is rekening gehouden met de lokale betrouwbaarheid per scorepunt berekend in de IRT-analyses (zie paragraaf 6.2), waarbij de grens van een lokale betrouwbaarheid van 0.70 voor alle scorepunten is aangehouden als minimum (COTAN-norm). Bij de berekening van de ontwikkelscorepunten (altijd afgerond op een geheel getal) zijn de gebieden tussen de bodem en de cesuur, en tussen de cesuur en het plafond vervolgens lineair verdeeld. In tabel 9.2 is de omzetting van de scorepunten naar ontwikkelscore voor de IEP LVS toets Rekenen 6a weergegeven, waarbij tevens op de betreffende punten (bodem, cesuur en plafond) de lokale betrouwbaarheid (REL) is weergegeven. Ook voor alle tussenliggende punten is de lokale betrouwbaarheid weergegeven. Uit tabel 6.2 blijkt dat de COTAN-norm ook voor de IEP LVS toets Rekenen 6a geldt.

Conclusie:
Er is voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt. Op aspect Va1 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende'** toegekend.

<u>Va2 Wordt de betrouwbaarheid van de groei op die schaal correct weergegeven?</u>

Bevindingen:
De (conditionele) lokale betrouwbaarheid voor ieder scorepunt wordt op dezelfde manier geschat als de (conditionele) lokale betrouwbaarheid van de cesuurpunten (op de thetaschaal) voor de IEP LVS toets Rekenen 6a en is beschreven in paragraaf 7.2 (zie bijlage 6 voor uitleg over de gedetailleerde berekening). Deze (conditionele) lokale betrouwbaarheid vertoont qua interpretatie grote overeenkomst met de globale betrouwbaarheidscoëfficiënt Cronbach's alpha uit de Klassieke Testtheorie (KTT) en wordt in de psychometrische literatuur beschreven (Raju, Price, Oshima & Nering, 2007) en als correct aangemerkt.

Conclusie:
De betrouwbaarheid van de groei op de ontwikkelscoreschaal wordt correct weergegeven. Op aspect Va2 wordt aan de toets IEP LVS toets Rekenen 6a het oordeel '**voldoende'** toegekend.

Va3 Worden er voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden?

Bevindingen:
De interpretatie van de behaalde resultaten op de toetsen IEP LVS Rekenen leerjaar 3 t/m 6 zijn gevisualiseerd door de behaalde ontwikkelscores grafisch weer te geven op twee manieren. Ten eerste geeft de voortgangsgrafiek (zie figuur 9.1) de vaardigheidsontwikkeling weer in de tijd door de toetsresultaten van alle toetsen, die in de leerjaren 3, 4, 5 en 6 zijn afgenomen, op één en dezelfde schaal uit te drukken. Ten tweede wordt in de leergroeimeter (zie figuur 9.3) de leergroei van een leerling afgezet tegen de gemiddelde groeifactor van de populatie en geeft dus aan hoe snel een leerling groeit ten opzichte van de populatie.

Het verschil in toetsresultaat, uitgedrukt in ontwikkelscorepunten, tussen twee of meer opeenvolgende toetsmomenten duidt hierbij de leergroei van een leerling aan. De gemiddelde groeifactor wordt berekend om een relatieve beoordeling van de leergroei per leerling te visualiseren. Hiervoor zijn voor de toetsen 3a t/m 6a de scores van de leerlingen op scholen met een schoolweging van 27 t/m 32.99, behaald in schooljaar 2020-2021, gemiddeld. Uitgangspunt per toets is een vast afnamemoment in het jaar, welke voor de a toetsen eind januari is en voor de b toetsen juni. Vervolgens is de gemiddelde leergroei van de populatie bepaald door per toets alle gemiddelden in een grafiek af te zetten tegen de tijd. Deze trendlijn bleek een lineaire trend te hebben door de zeven punten (de vaste afnamepunten van leerjaar 3 t/m de a toets in leerjaar 6).

In figuur 9.2 (Ontwikkelscoreverloop gemiddelde leergroei leerjaar 3 t/m halverwege leerjaar 6) is de gemiddelde leergroei van de populatie weergegeven, waarbij op de x-as het aantal onderwijsmaanden staat (een schooljaar bestaat uit 10 onderwijsmaanden) en op de y-as de ontwikkelscore. De formule voor de lineaire trendlijn in figuur 9.2 is y = 1.54*DL + 1.6, waarbij de richtingscoefficient 1.54 staat voor de gemiddelde groeifactor (leergroei per onderwijsmaand) van de populatie.

In hoofdstuk 4 (Normeringspopulatie) is reeds vastgesteld dat de representativiteit van de normeringspopulatie op de achtergrondvariabelen niet optimaal is om als referentiepopulatie voor het gehele basisonderwijs te fungeren. Om de absolute leergroei (aantal ontwikkelscorepunten groei) te duiden is de representatie van de referentiepopulatie van ondergeschikt belang, maar voor de vergelijking van de leergroei met de gemiddelde groeifactor (weergegeven in de leergroeimeter, zie figuur 9.3) is de representativiteit echter wel van belang. Dit betekent dat voor de toekomst de gemiddelde groeifactor jaarlijks geëvalueerd wordt aan de hand van representatieve afnamedata en indien nodig wordt aangepast. Hiertoe zal jaarlijks op dezelfde wijze de berekening van de groeifactor worden uitgevoerd zoals eerder in paragraaf 9.2 is beschreven, waarbij de data van volgende schooljaren worden toegevoegd aan de data van vorige schooljaren om de groeifactor te berekenen.

De interpretatie van de leergroei van leerlingen wordt voor de leerkrachten, leerlingen en ouders/verzorgers ondersteund door de twee grafische weergaven, de voortgangsgrafiek en de leergroeimeter, gecombineerd te gebruiken in de leergroeimeter (zie fig. 9.3). Het voorbeeld van de leergroeimeter in fig. 9.3 geeft aan dat het kind in het voorbeeld op het IEP LVS Taalverzorging minder dan gemiddeld is gegroeid, op het IEP LVS Lezen meer dan gemiddeld en op het IEP LVS Rekenen precies gemiddeld.

Voor leerkrachten zijn de leervorderingen van een leerling ook digitaal beschikbaar in het IEP LVS, welke door de leerkrachten ook geprint kunnen worden in de vorm van de IEP LVS Talentenkaart (zie bijlage 10 voor een voorbeeld van een Talentenkaart). Daarnaast hebben leerkrachten ook per toetsresultaat inzicht in de scores (percentage goed beantwoorde items) op de verschillende inhoudelijke domeinen en kunnen de gegeven antwoorden van de leerling inzien, hetgeen het formatief gebruik en de bruikbaarheid van de IEP LVS toetsen bevordert. Voor de interpretatie van de leervorderingen worden aan keerkrachten handvatten gegeven in stap 6 in de handleiding van het IEP LVS https://handleiding.toets.nl/snel-op-weg-met-het-iep-lvs-245. Daarnaast kunnen leerkrachten op Mijn IEP-kanaal (https://www.bureau-ice.nl/basisonderwijs/mijniepkanaal/informatie-iep-lvs/) onder de kopjes 'Resultaten en analyse' en 'Altijd handig!' nog meer informatie vinden over het interpreteren van leervorderingen.

Conclusie:
Er worden voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden. Op aspect V31 wordt aan de toets IEP LVS Rekenen het oordeel '**voldoende'** toegekend.

Inzicht in leervorderingen

I1 Levert de aanbieder een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders /verzorgers/voogden/docenten begrijpelijk is?

Bevindingen:
De toetsaanbieder Bureau ICE levert speciaal voor ouders/voogden/verzorgers een Leeswijzer voor de IEP LVS Talentenkaart (zie bijlage 11), die hen handvatten geeft voor de interpretatie van de leervorderingen op de Talentenkaart (IEP = Inzicht in Eigen Profiel). Deze Leeswijzer is onder andere beschikbaar via de algemene informatiepagina voor ouders/voogden/verzorgers van het IEP LVS https://www.bureau-ice.nl/basisonderwijs/voor-ouders.

Conclusie:
De aanbieder (i.e., Bureau ICE) levert een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders/verzorgers/voogden/docenten begrijpelijk is. Op aspect I1 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend.

I2 Is er een evaluatie van de leervorderingen en worden op basis van deze evaluatie vervolgstappen geformuleerd?

Bevindingen:
De leerkracht wordt ondersteund bij de interpretatie van de vaardigheidsontwikkeling van de leerling door het gecombineerd gebruik van de twee grafische weergaven, de voortgangsgrafiek (zie fig. 9.1) en de leergroeimeter (zie fig. 9.3). De leerkracht kan hiermee evalueren in welke mate de leerling ten opzichte van zijn/haar verwachting en/of ten opzichte van de verwachte groeifactor zich ontwikkelt en kan hij/zij inschatten hoe waarschijnlijk het is dat de leerling het beoogde streefniveau zal gaan bereiken. In de

'Handreiking interpreteren toetsresultaten' (bijlage 9) worden leerkrachten geholpen bij de interpretatie van de ontwikkelscores en krijgen zij advies over het bepalen of een toets 'passend' was qua niveau voor de leerling.

Conclusie:
Er is een evaluatie van de leervorderingen en op basis van deze evaluatie worden vervolgstappen geformuleerd. Op aspect I2 wordt aan de toets IEP LVS Rekenen 6a het oordeel '**voldoende**' toegekend.

Referentieniveaus

R1 Sluit de inhoud van de toets aan op de kennis en vaardigheden zoals omschreven in de referentieniveaus van het betreffende domein (voor toetsen vanaf groep 6)?

Bevindingen:
In het IEP LVS zijn zowel contextopgaven als kale opgaven opgenomen op zowel het niveau <F als 1F. Vragen op het niveau 'op weg naar 1F' zijn vragen die eenvoudiger zijn dan het niveau 1F. Deze items zijn opgenomen, zodat gemeten kan worden in hoeverre een leerling in staat is om opgaven net onder 1F te beantwoorden.

De toets Rekenen 6a bevat vragen op de niveaus 5b, 6a, <1F en 1F en vormt daarmee de overbrugging tussen de rekentoetsen voor leerjaar 3 t/m 5, en de rekentoetsen voor leerjaar 6 t/m 8. Geadviseerd wordt om deze toets af te nemen tijdens of net na de eerste helft van leerjaar 6.

De inhoud van de toets sluit voldoende aan op de kennis en vaardigheden in de referentieniveaus van de verschillende domeinen.

Conclusie:
'voldoende'; de inhoud van de toets sluit voldoende aan op de kennis en vaardigheden in de referentieniveaus van de verschillende domeinen.

3. Verzamelstaat

Kwaliteitsaspect	**Code**	**Oordeel**
De kwaliteit van de steekproef	*S1*	**Voldoende**
	S2	**Voldoende**
	S3	**n.v.t.**
	S4	**n.v.t.**
Normering	*N1.1*	**Voldoende**
	N1.2	**Voldoende**
	N1.3	**Voldoende**
	N2.1	**Voldoende**
	N2.2	**Voldoende**
	N2.3	**Voldoende**
Betrouwbaarheid	*B1*	**Voldoende**
	B2	**Voldoende**
Validiteit	*V1*	**n.v.t.**
	V2	**Voldoende**
Volg-aspect	*Va1*	**Voldoende**
	Va2	**Voldoende**
	Va3	**Voldoende**
Inzicht in leervorderingen	*I1*	**Voldoende**
	I2	**Voldoende**
Referentieniveaus	*R1*	**n.v.t.**

4. Literatuurlijst

- Bezdan, E., Binsbergen, M., Winter, N., Helsloot, J. & Laan, J. (2021). Verantwoording IEP LVS toets Rekenen 6a. Culemborg: Bureau ICE.

- Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). *Standard setting: A bookmark approach*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Boulder, CO.