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. De toetsen Rekenen maken onderdeel uit van dit volgsysteem. De ter beoordeling voorliggende toetsen zijn de toetsen voor leerjaar 3, 4 en 5:

- Rekenen 3a
- Rekenen 3b
- Rekenen 4a
- Rekenen 4b
- Rekenen 5a
- Rekenen 5b

De toetsen Rekenen zijn digitaal en meten de rekenvaardigheid van leerlingen. De toetsen Rekenen meten de volgende domeinen: Getallen, Verhoudingen, Meten en meetkunde, Verbanden (de rekentoetsen voor leerjaar 3 t/m 5 bevatten enkel de domeinen Getallen en Meten en meetkunde).

Doelgroep

Het IEP LVS bevat toetsen voor leerlingen in leerjaar 3 tot en met 8. De ter beoordeling voorliggende toetsen zijn bedoeld voor leerlingen in leerjaar 3, 4 en 5. Het IEP LVS is echter zo ingericht dat alle toetsen voor elke leerling toegankelijk zijn. Zo kan een leerling in leerjaar 5 die moeite heeft met Rekenen, ook de toets op het niveau van leerjaar 4 maken. Maar het is bijvoorbeeld ook mogelijk om een leerling in leerjaar 6 die relatief goed is in Rekenen al een toets op 2F-niveau te geven. Het voorgaande maakt dat het IEP LVS ook geschikt is voor leerlingen uit het SBO.

Inhoudelijke theoretische inkadering:

Het IEP Toetskader Rekenen ligt ten grondslag aan de toetsen Rekenen van leerjaar 3, 4 en 5. In het toetskader is beschreven welke kennis en vaardigheden van leerlingen op een bepaald moment worden verwacht. In het toetskader wordt onderscheid gemaakt tussen de niveaus: 3a, 3b, 4a, 4b, 5a en 5b. 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, 2007)
- 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)

Daarnaast is onderzocht hoe de (tussen)doelen uit deze bronnen terugkomen in de meest gebruikte reken-wiskundemethodes voor het primair onderwijs. Aan de hand hiervan is

bepaald wanneer bepaalde doelen beheerst zouden moeten worden door de leerling. Er is onderscheid gemaakt tussen doelen voor de a-en b-toetsen. In de a-toetsen ligt de nadruk op de lesdoelen die in de meeste methodes in de eerste helft van het schooljaar behandeld worden. In de b-toetsen ligt de nadruk op de lesdoelen die in de meeste methodes in de tweede helft van het schooljaar behandeld worden.

Inhoud van het toetspakket

Het toetspakket Rekenen voor leerjaar 3, 4 en 5 bestaat uit de volgende documenten:

- Verantwoording LVS-toetsen Rekenen 3a, 3b, 4a, 4b, 5a, 5b, deze bevat informatie over:
    - De uitgangspunten van de toetsen (hfdst. 2),
    - De inhoud van de toetsen (hfdst. 3),
    - De normeringspopulatie (hfdst. 4),
    - Het design van de dataverzameling (hfdst. 5),
    - De kalibratie en kwaliteit van de items (hfdst. 6),
    - De bepaling van de cesuren (hfdst. 7),
    - De constructvaliditeit (hfdst. 8),
    - Het volgaspect (hfdst. 9),
    - Inzicht in leervorderingen (hfdst. 10)
- Toetswijzer (bijlage 1)
- Itemoverzicht (bijlage 2)
- Itemparameters (bijlage 3)
- TIA's (bijlage 4)
- Algemene toelichting methode (bijlage 5)
- IEP LVS Talentenkaart (bijlage 6)
- Leeswijzer voor de IEP LVS Talentenkaart (bijlage 7)
- Toelichtingen (bijlage 8)
- Langrange Multiplier Tracelines (bijlage 9)
- Handreiking interpreteren toetsresultaten (bijlage 10)

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 Jennifer Roubiës MSc (secretaris).
Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording veelal letterlijk vermeld.

De kwaliteit van de dataverzameling

S1 Is de steekproef van leerlingen representatief?

Bevindingen:

Uitgangspunt 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 3 (zie kolom 'Aantal afnames') blijkt dat het aantal observaties van de IEP LVS items conform de eisen is die worden beschreven in het document 'Aanvulling COTAN Beoordelingssysteem' m.b.t. het aspect normering referentieniveaus (d.d. 16-06-2016). Naast aantallen afnames is de representativiteit op achtergrondgegevens van de normeringspopulatie (groep leerlingen in leerjaar 3, 4 en 5 die aan het normeringsonderzoek heeft deelgenomen) ten opzichte van de doelpopulatie (alle leerlingen in leerjaar 3, 4 en 5 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, omdat 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 leerjaren 3, 4 en 5 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 privacy regels (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 schoolachtergrondgegevens 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 toetsen zijn door de betrokken scholen tijdens het reguliere onderwijsproces in schooljaar 2019-2020 op eigen initiatief afgenomen in respectievelijk de leerjaren 3, 4 en 5. Er is hier dus sprake van 'purposeful sampling' (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 toetsen ook in de komende jaren afgenomen gaan worden. 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 met meer dan 15% niet-beantwoorde items mogen bevatten. Voor de analyse voor het bepalen van de gemiddelde groeifactor (hoofdstuk 9) is gefilterd op twee criteria. De representativiteit van de normeringspopulatie is hier namelijk wel van belang vanwege het bepalen van een relatieve norm. Ten eerste zijn records van SBO-scholen en van scholen van Bonaire uit het databestand verwijderd. 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.

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 random steekproef van de leerjaren 3, 4 en 5 van het basisonderwijs beschouwd kan worden. Wel kan uit de verdeling in 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. 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 toetsen worden afgenomen. Op aspect S1 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: 'voldoende'.

<u>S2 In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?</u>

Bevindingen:
In het schooljaar 2019-2020 is eerst de pretest IEP LVS leerjaar 3 t/m 5 toetsen afgenomen, welke uit meer items bestonden dan de IEP LVS leerjaar 3 t/m 5 toetsen definitief bevatten. 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 toetsen. De afnamecondities van de pretest IEP LVS leerjaar 3 t/m 5 toetsen waren identiek aan de definitieve IEP LVS leerjaar 3 t/m 5 toetsen.

Tabel 5.1 toont het design van het normeringsonderzoek, waarin per toets van 50 items (toets 3a, toets 3b, toets 4a, toets 4b, toets 5a en toets 5b) is weergegeven hoeveel items van welk niveau (niveau <3a, niveau 3a, niveau 3b, niveau 4a, niveau 4b, niveau 5a, niveau 5b, niveau <1F en niveau 1F) de toets in de pretestfase bevatte. In de laatste kolom staat het aantal observaties per toets (1794, 1811, 1855, 1733, 1871 en 1951 voor respectievelijk toets 3a, toets 3b, toets 4a, toets 4b, toets 5a en toets 5b).

Uit tabel 5.1 valt op te maken dat iedere toets in de pretestfase altijd 7 opgaven beneden en 7 opgaven boven het niveau van de betreffende toets bevatten en dus 36 opgaven (50 – 7 – 7 = 36) bevat van het niveau van de betreffende toets. Zo bevat bijvoorbeeld de toets Rekenen 5a ook 7 opgaven op niveau 4b en ook 7 opgaven op niveau 5b. 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 4b) en/of al misschien wel moeilijkere opgaven kan maken (i.e., op niveau 5b). Verder valt uit tabel 5.1 op te maken dat er sprake is van enige overlap tussen de toetsen van het opvolgende niveau. Iedere toets heeft namelijk 8 items overlap met een toets van een lager niveau en 8 items overlap met een toets van een hoger niveau. 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 blijven volgen op één en dezelfde schaal (unidimensionaliteit).

Tabel 6.1 toont het ankerdesign waaraan de itemselectie voor de definitieve samenstelling van de toetsen moest voldoen: Iedere toets bestaat uit 30 items van het niveau van de toets, 5 items van het niveau eronder en 5 items van het niveau erboven (in totaal bestaat iedere toets dus uit 40 items). Daarnaast is er weer sprake van enige overlap tussen de toetsen van het opvolgende niveau. Iedere toets heeft 6 items overlap met een toets van een lager niveau en 6 items overlap met een toets van een hoger niveau. In totaal was er sprake van 30 ankeritems (zie ook bijlage 3), waarmee voldaan wordt aan de eis dat het anker uit teminste 15 ankeritems moet bestaan (zie 'Beoordelingskader voor instrumenten binnen leerlingvolgsystemen' op de website van de Expertgroep Toetsen PO). De ankeritems zijn afgenomen bij alle 6 de toetsen, of als meetellend item of als niet-meetellend item (gebruikt om de kwaliteit en de vergelijkbaarheid tussen de toetsen te waarborgen).

Op basis van dit definitieve design heeft de itemselectie plaatsgevonden door te kijken naar de statistieken afkomstig uit de TIA-analyse (bijlage 4) over de responsedata uit het normeringsonderzoek tijdens de pretestfase, waarbij de volgende uitgangspunten golden:

- 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.

Voor een aantal items is een uitzondering gemaakt op deze uitgangspunten teneinde, naast een goede psychometrische afspiegeling, een goede inhoudelijke dekking van het IEP LVS Toetskader (zie hoofdstuk 2) met variatie in domeinen en domeinonderwerpen te kunnen garanderen (i.e., inhoudsvaliditeit). Deze uitzonderingen hadden volgens de auteurs een verwaarloosbare negatieve invloed op de globale betrouwbaarheid (zoals geschat met Cronbach's Alpha) van de toetsen.

Het 1PL-model (Rasch model) met de Marginal Maximum Likelihood (MML) schattingsmethode is toegepast voor de IRT-analyse (d.w.z. de itemparameters van alle items worden verondersteld dezelfde onderliggende vaardigheid te meten, welke gezamenlijk zijn gekalibreerd op dezelfde schaal) van de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b en zijn uitgevoerd in Versie 0.07 van de applicatie Lexter (2019). In de schaling in Lexter zaten zes booklets, ieder booklet komt overeen met één toets. In bijlage 3 zijn het aantal afnames, de itemmoeilijkheid (β), de meetfout van de moeilijkheid (SE(β)) en de iteminformatie (i.e., Fisher-informatie), bij de gemiddelde vaardigheid van de afnamegroepen van de IEP LVS toetsen Rekenen in een tabel weergegeven, waarbij de vaardigheidsschaal is genormeerd door van de afnamegroep van de 5b toets de gemiddelde vaardigheid en de standaarddeviatie hiervan op respectievelijk 0 en 1 te zetten.

In paragraaf 7.2 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 3a, 3b, 4a, 4b, 5a en 5b. Tabel 7.3 laat zien dat de gemiddelde waarde van de constante c (i.e., 0.041), 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 0.086). Op basis van deze resultaten kan de nauwkeurigheid van de parameterschattingen als goed beoordeeld worden.

Modelpassing is onderzocht via een Differential Item Functioning (DIF) analyse (uitgevoerd in de applicatie Lexter) op de 30 ankeritems en 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 kalibratie van de IEP LVS toetsen, waarbij ieder van deze ankeritems in 2 booklets is opgenomen geweest en het aantal vrijheidsgraden dus 1 is (df = 2 – 1 = 1). In bijlage 3 is per item de Lagrange Multiplier statistiek (LM) weergegeven, het aantal vrijheidsgraden (df), de overschrijdingskans (Prob), het absolute verschil (Abs. Diff) en is aangegeven in welke booklets 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). Eén van de ankeritems heeft een absoluut verschil van groter dan 0.10, namelijk een absoluut verschil van 0.12, en dit item heeft tevens de hoogste LM-waarde van 205.37 bij een overschrijdingskans van 0.00. Het gemiddelde absolute verschil van

de 30 ankeritems is 0.04 (max = 0.12; min = 0.00). Op basis van deze resultaten kan geconcludeerd worden dat met betrekking tot DIF een goede modelpassing aannemelijk is.

Bijlage 9 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 statistisch goed representeren. Er zijn 240 effectgroottes (6 * 40) berekend met 3 waardes groter dan 0.10 (alle 3 met een effectgrootte van 0.12), hetgeen aantoont dat de modelpassing ook in dit opzicht zeer goed is. Volgens bovenstaande analyse blijken 3 items niet in het Rasch-model te passen vanwege een effectgrootte groter dan 0.10. De auteurs vinden het echter toch verantwoord om deze items in de IEP LVS toetsen te behouden, omdat het om een relatief klein percentage gaat van 1.25% ((3/240) * 100)) en deze items volgens de KTT (Klassieke Testtheorie) bovendien goed functioneren (zie bijlage 4).

Conclusie:
Het onvolledige maar 'verbonden' dataverzamelingsdesign is adequaat. Op aspect S2 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende'**.

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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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 op elke IEP LVS toets de leerling het gemeten niveau (3a, 3b, 4a, 4b, 5a en 5b) heeft behaald, waarbij elk niveau wordt gemeten in één toets (niveau 3b wordt bijvoorbeeld gemeten in toets 3b). Er is dus sprake van een criteriumgerichte interpretatie van de

toetsscores, waarbij de toetsscores vergeleken worden met een absolute norm. Volgens de auteurs is een combinatie van de uit drie stappen bestaande Angoff methode (Assessment Strategies inc., Canada's Testing Company, 2014) en de Bookmark methode (Karantonis & Sireci, 2006) tijdens deze standaardsettingsprocedure gevolgd. Als eerste stap in de Angoff methode worden experts uit het betreffende vakgebied (i.e., Rekenen leerjaar 3 t/m 5) zodanig geïnformeerd dat zij een beeld over de 'borderline kandidaat' (grensleerling) krijgen, 'dat wil zeggen, van de leerling die het niveau van de toets net behaalt'. Als tweede stap in de Angoff 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 de Angoff methode 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 de Angoff methode lastig is voor onervaren standaardsetters, is die stap vervangen door een deel van de Bookmark methode. De standaardsetters moesten hierbij in stap 1 van de Angoff methode de 50 items uit de pretestfase (zie tabel 5.1) 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 als tweede stap van de Angoff methode individueel een 'bookmark' (bladwijzer) plaatsen in de door henzelf gemaakte volgorde van items van makkelijk naar moeilijk uit stap 1. De 'bookmark' betekent dat leerlingen die het niveau net beheersen (i.e., de grensleerlingen) 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 1 en 2 van de individuele standaardsetters zijn geanalyseerd, zodat de 'bookmark' die de standaardsetters in stap 2 hebben gezet op de lijst met 50 items op volgorde uit de pretestfase nu een 'bookmark' wordt op 30 items van het niveau van de definitieve samenstelling van iedere toets (zie tabel 6.1). De items van het niveau onder (7 items) en boven het niveau van de toets (7 items) zijn als eerste verwijderd uit de lijsten van de toets met 50 items van individuele standaardsetters met items op volgorde van makkelijk naar moeilijk (stap 1). De items van het niveau van de toets die niet in de definitieve zijn opgenomen (6 items) zijn vervolgens ook uit deze lijsten verwijderd, waardoor alle lijsten met 50 items dus nu zijn teruggebracht naar 30 items (i.e., 30 = 50 – 7 – 7 – 6) van het niveau van de toets (zie tabel 6.1). Omdat de leerling ook de 5 items uit de definitieve van het niveau onder de toets goed moet maken om het niveau te behalen, is elke bookmark (één bookmark voor elke standaardsetter voor elke toets) omgezet in een cesuur door het aantal items tot aan de bookmark te tellen en daar 5 bij op te tellen.

In stap 3 is door de procesbegeleider (Bureau ICE heeft de rol van procesbegeleider in deze standaardsetting) aan de standaardsetters een voorstel voor de cesuur (voor elke toets) gedaan op basis van het gemiddelde van de individuele cesuren, waarbij de standaardsetters bij dit voorstel ook de definitieve te zien kregen. 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 7.2, vaardigheid voor behalen cesuur voor toets 3a t/m 5b is respectievelijk -0.548; 0.181; 1.276; 1.734; 2.168; 2.689). 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).

In tabel 7.2 worden de cesuren van de 5b, 5a, 4b, 4a, 3b en 3a niveaus van de IEP LVS toetsen Rekenen 5b, 5a, 4b, 4a, 3b en 3a gerapporteerd in scorepunten op de definitieve toetssamenstelling (40 items). In tabel 7.2 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.

Conclusie:
De standaardbepalingsmethode is gemotiveerd en op de juiste wijze uitgevoerd. Op aspect N1.1 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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 3, 4 en 5 en bij voorkeur op dit moment ook werkzaam zijn als leerkracht van leerjaar 3, 4 en 5. 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 van de gevolgde standaardsettingsprocedure werden de standaardsetters goed geïnformeerd over de borderline kandidaat en ontwikkelden zij hierover een beeld. Ook werden zij in stap 1 geïnformeerd over de consequenties van de niveau-uitspraak voor de leerling en de betekenis daarvan voor het onderwijs.

Conclusie:
De beoordelaars/vakdeskundigen/experts zijn naar behoren geselecteerd en getraind. Op aspect N1.2 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende.**

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

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

Conclusie:
Er is voldoende overeenstemming tussen de beoordelaars. Op aspect N1.3 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende**'.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
De toetsen zijn genormeerd voor de afnamemomenten a (eerste helft van het schooljaar) en b (tweede helft van het schooljaar) in de leerjaren 3, 4 en 5. Uit tabel 4.1 op pag. 12 valt af te lezen dat de normgroepen voor de zes toetsen IEP LVS-toetsen Rekenen 3a, 3b, 4a, 4b, 5a en 5b in de normeringspopulatie respectievelijk 1649, 1669, 1627, 1550, 1629 en 1729 zijn. Deze normgroepen zijn van voldoende grootte. Op pag. 7, 3^{e} 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 normgroepen zijn groot genoeg. Op aspect N2.1 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het oordeel '**voldoende**' toegekend.

N2.2 Zijn de normgroepen representatief?

Bevindingen:
De representativiteit van de steekproeven 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 toetsen werden vastgesteld, waarvoor de samenstelling van de normeringspopulatie volgens de auteurs van ondergeschikt belang is. Wel is het volgens de auteurs voor de bepaling van de kwaliteit van de items (hoofdstuk 6) en de standaardsetting (hoofdstuk 7) van belang vast te stellen dat de normeringspopulatie geen specifiek selecte groep is van de leerlingen in de leerjaren 3, 4 en 5 van het reguliere basisonderwijs. Dit kon volgens de auteurs gegarandeerd worden door records met meer dan 15% niet-beantwoorde items uit de responsedata te verwijderen. Het is volgens de auteurs voor de bepaling van de gemiddelde groeifactor (zie hoofdstuk 9) echter wel van belang vast te stellen dat de normeringspopulatie een representatieve steekproef is, omdat daar een relatieve norm wordt bepaald. Voor het bepalen van de gemiddelde groeifactor kon volgens de auteurs representativiteit gegarandeerd worden door de data te filteren op twee criteria, namelijk ten eerste door records van SBO-scholen en van scholen van Bonaire uit het databestand te verwijderen en ten tweede door records met minder dan 25% goed beantwoorde items niet te gebruiken.

Conclusie:
De normgroepen zijn groot genoeg. Op aspect N2.2 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het oordeel '**voldoende**' toegekend.

N2.3 Zijn de normen correct bepaald?

Bevindingen:
Hoewel ook andere standaardsettingsprocedures gebruikt hadden kunnen worden (bijv. de originele Bookmark methode i.p.v. de gevolgde aangepaste Bookmark methode en dus empirische i.p.v. subjectieve moeilijkheden gebruiken om de 50 items per toets op volgorde van makkelijk naar moeilijk te leggen), 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het oordeel '**voldoende**' toegekend.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In bijlage 4 (TIA's 3a t/m 5b) worden de globale betrouwbaarheden (Cronbach's alpha) weergegeven voor de zes IEP LVS-toetsen Rekenen (3a, 3b, 4a, 4b, 5a en 5b), welke zijn 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 7.2 ook de lokale betrouwbaarheden, gemeten bij de cesuurpunten op de latente vaardigheidsschaal, weergegeven voor de toetsen 3a t/m 5b. In bijlage 5 ('Algemene toelichting methode') wordt gedetailleerde uitleg gegeven hoe

deze (conditionele) lokale betrouwbaarheden, gegeven een vaardigheidsniveau θ, kunnen worden berekend via de (ook in tabel 7.2 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 op pag. 18 voor de drie IEP LVS-toetsen per referentieniveau 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het oordeel '**voldoende**' toegekend.

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

Bevindingen:
In bijlage 4 (TIA's 3a t/m 5b) kan afgelezen worden dat de globale betrouwbaarheden in termen van Cronbach's alpha (interne consistentie betrouwbaarheden) voor de zes IEP LVS toetsen Rekenen 3a, 3b, 4a, 4b, 5a en 5b gelijk zijn aan respectievelijk 0.89, 0.90, 0.91, 0.92, 0.92 en 0.91. Omdat het behalen van één van de niveaus 3a, 3b, 4a, 4b, 5a 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 7.2 laat zien dat aan deze minimale eis ook ruimschoots wordt voldaan voor de lokale betrouwbaarheden bij de cesuurpunten van de toetsen 3a t/m 5b. De geschatte lokale betrouwbaarheden van de cesuren zijn namelijk bij de zes IEP LVS-toetsen Rekenen 3a, 3b, 4a, 4b, 5a en 5b gelijk aan respectievelijk 0.891, 0.900, 0.908, 0.907, 0.905 en 0.906.

Verder laat tabel 7.2 nog zien dat de classificatiefouten (i.e., voor de zes IEP LVS-toetsen Rekenen het percentage leerlingen dat het betreffende niveau ten onrechte wel of niet heeft gehaald) loopt van 9% tot 10% (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 betrouwbaarheidsgegevens zijn voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden. Op aspect B2 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende'**.

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 toetsen de leerdoelen die ze beogen te toetsen. Er kunnen geen misverstanden ontstaan over de juistheid van de gegeven antwoorden.

Conclusie:
De items van de toets dragen bij aan de validiteit van de toets. Op aspect V1 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende**'.

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 veronderstelde 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) correlationeel onderzoek tussen de inhoudelijke domeinen binnen de IEP LVS-toetsen Rekenen leerjaar 3 t/m 5, (2) onderzoek naar convergente en divergente validiteit, (3) itemkwaliteit (psychometrische kwaliteit van de items).

1. Uit tabel 8.1 valt af te lezen dat er voor de toetsen IEP LVS Rekenen 3a t/m 5b een middelmatig tot sterk correlationeel verband is tussen de scores op het domein Getallen (G) en het domein Meten & Meetkunde (M) met correlatiecoëfficiënten tussen de 0.66 en 0.74 (alle gerapporteerde correlaties zijn tweezijdig significant op het 1%-niveau).
2. Er is op de IEP LVS toetsen een soortgenotenonderzoek uitgevoerd in de vorm van onderzoek naar convergente validiteit tussen twee binnen één leerjaar opeenvolgende IEP LVS toetsen van dezelfde vaardigheid en tevens is er onderzoek gedaan naar divergente validiteit tussen toetsen van verschillende vaardigheden (Lezen en Rekenen) binnen één en hetzelfde leerjaar. Met andere woorden, er is gebruikgemaakt van een Mult-Trait Multi-Method matrix (MMTM), waarbij scores

op rekentoetsen 3a t/m 5b hoog zouden moeten correleren met toetsen die hetzelfde construct meten en laag met toetsen die een ander construct meten. De resultaten van deze beide onderzoeken worden gerapporteerd in tabel 8.2. Conform de verwachting is hieruit af te lezen dat de correlatie tussen twee opeenvolgende toetsen Rekenen structureel hoger is dan tussen twee toetsen van verschillende vaardigheden binnen hetzelfde leerjaar.

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 voldoen als criterium voor de itemkwaliteit per toets met p-waarden tussen de 0.57 en 0.75. De grootte van de ranges in tabel 8.3 wordt sterk bepaald door enkele outliers (zie hiervoor ook bijlage 4).

Conclusie:
De gerapporteerde resultaten in Hoofdstuk 8 (Constructvaliditeit) van de WV vormen een psychometrische ondersteuning voor de constructvaliditeit van de toetsen IEP LVS Rekenen 3a t/m 5b en er wordt dus gemeten wat men beoogt te meten, namelijk rekenvaardigheid bij de leerjaren 3, 4 en 5. Op aspect V2 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende**'.

Het volg-aspect

<u>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?</u>

Bevindingen:
De zes IEP LVS-toetsen Rekenen 3a, 3b, 4a, 4b, 5a en 5b zijn gekalibreerd op één en dezelfde onderliggende vaardigheidsschaal theta (θ). Het gevolg hiervan is dat de vaardigheidsscores op deze zes 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 loopt van 0 tot maximaal 60 punten. In de ontwikkelscoreschaal van de IEP LVS Rekenen leerjaar 6 t/m 8 is namelijk het referentieniveau 1F gelijk aan 60 ontwikkelscorepunten en de ontwikkelscoreschaal van de IEP LVS Rekenen leerjaar 3 t/m 5 moet hieronder blijven. Omdat de inhoud van de verschillende toetsen IEP LVS leerjaar 3 t/m 5 stapelend is qua inhoudelijke onderwijsdoelen, 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.

Voor het omzetten van de toetsscore (het aantal goed beantwoorde vragen) naar de ontwikkelscore (OS) zijn voor ieder van de zes toetsen 3a t/m 5b 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. In paragraaf 7.1 is via een standaardsettingsprocedure per vaardigheid per toets de cesuur op de toetsscore bepaald, welke verschillen per toets (zie tabel 7.1). Voor de bodem en het plafond is besloten deze voor alle toetsen op een gelijke toetsscore te leggen, waarbij rekening is gehouden met de lokale betrouwbaarheid per scorepunt berekend in de IRT-analyses (zie paragraaf 6.2). Hierbij is de grens van een lokale betrouwbaarheid van 0.70 aangehouden als minimum, die COTAN hanteert. 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 per toets weergegeven, waarbij tevens per scorepunt de lokale betrouwbaarheid (REL) is weergegeven.

Conclusie:
Er is voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt. Op aspect Va1 wordt aan de toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende**'.

Va2 Wordt de betrouwbaarheid van de groei op die schaal correct weergegeven?

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 zes IEP LVS-toetsen en is beschreven in paragraaf 7.2 (zie bijlage 5 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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 3a t/m 5b wordt 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 en 5 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 normeringspopulatie en geeft dus aan hoe snel een leerling groeit ten opzichte van de normeringspopulatie. 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. In paragraaf 9.2 ('Leergroei') wordt gedetailleerd beschreven hoe de gemiddelde groeifactor van de normeringspopulatie wordt berekend. 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 normeringspopulatie bepaald door per toets alle gemiddelde ontwikkelscores van de leerlingen in een grafiek af te zetten tegen de tijd. Deze trendlijn bleek bij alle drie de vaardigheden een lineaire trend te hebben door de zes afpunten (de vaste afnamepunten per leerjaar). In figuur 9.2 (Ontwikkelscoreverloop) is de gemiddelde leergroei van de normeringspopulatie 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 6.2 is y = 1.5579x + 1.3889, waarbij de richtingscoefficient 1.5579 staat voor de gemiddelde groeifactor (leergroei per onderwijsmaand) van de normeringspopulatie.

In hoofdstuk 4 (Normeringspopulatie) is reeds vastgesteld dat de representativiteit van de normeringspopulatie niet optimaal is om per definitie als referentiepopulatie voor het gehele reguliere basisonderwijs te fungeren. Dit probleem zou kunnen worden verholpen door het trekken van een aantal steekproeven conform de gewenste samenstelling (gelijk de landelijke verdeling) uit de normeringspopulatie. Dit bleek bij de normeringspopulatie van de leerjaren 3, 4 en 5 echter nog niet mogelijk. Om de absolute leergroei 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 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.

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 Lezen minder dan gemiddeld is gegroeid, op het IEP LVS Technisch 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 6 ' Talentenkaart' 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 'Info voor jou' en 'Handleiding & FAQ' 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b het volgende oordeel toegekend: '**voldoende**'.

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 7 'Leeswijzer talentenkaart'), die het 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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 10) 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 toetsen IEP LVS Rekenen 3a, 3b, 4a, 4b, 5a en 5b 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:
Dit criterium is niet van toepassing (n.v.t.).

Conclusie:
n.v.t.

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*	**Voldoende**
	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., Haitjema, T, Helsloot, J. & Laan, J. (2020). Verantwoording IEP LVS-toetsen Rekenen. Culemborg: Bureau ICE.

- Brennan , R.L., & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*, 219-240.

- Fitzpatrick, A.R. (1984, April). *Social influences in standard setting: The effect of group interaction on individuals' judgments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

- 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.