1. Algemene informatie

Algemeen en meetpretentie
Het IEP LVS is een volgsysteem waarin leerlingen vanaf leerjaar 3 tot aan de eindtoets gevolgd kunnen worden in hun ontwikkeling. De rekentoetsen maken onderdeel uit van dit volgsysteem. De rekentoetsen meten de rekenvaardigheid van leerlingen.

Doelgroep
Het IEP LVS Rekenen voor groep 6, 7 en 8 bevat toetsen voor leerlingen in jaar 6 t/m 8. Het IEP LVS is zo ingericht dat alle toetsen voor elke leerling toegankelijk zijn. Het IEP LVS is hierdoor ook geschikt voor leerlingen uit het SBO.

Inhoudelijke theoretische inkadering:
Het Referentiekader Taal en Rekenen (Meijerink et al., 2009) ligt ten grondslag aan de rekentoetsen van leerjaar 6 t/m 8. In het Referentiekader voor rekenen is per referentieniveau opgedeeld in vier domeinen: Getallen, Verhoudingen, Meten en meetkunde en Verbanden. De verhouding waarin de vier domeinen worden getoetst is in lijn met de beheersing van de rekenvaardigheid van leerlingen aan het einde van de basisschool. Het IEP LVS legt daarom meer nadruk op de domeinen Getallen en Verhoudingen dan op de domeinen Meten en meetkunde en Verbanden, aansluitend bij de lesstof van het primair onderwijs. De LVS-toetsen bevatten 30% Getallen, 30% Verhoudingen, 20% Meten en Meetkunde en 20% Verbanden opgaven. In de Toelichting Rekenen (bijlage 8) worden de verschillende domeinen per referentieniveau nader omschreven.

Inhoud van het toetspakket
Het toetspakket Rekenen gr. 6 t/m 8 bestaat uit de volgende documenten:

- Inlog instructie
- Verantwoording, deze bevat informatie over:
    - De uitgangspunten van de toetsen (hfdst. 2),
    - De inhoud van de toetsen (hdfst. 3),
    - De steekproef (hfdst 4),
    - Het design van de dataverzameling (hfdst 5),
    - De kalibratie en kwaliteit van de items (hfdst 6),
    - De ijking van de referentieniveaus (hfdst 7),
    - Het volgaspect (hfdst 8)
    - Inzicht in leervorderingen (hfdst 9)
- 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)
- Toelichting (bijlage 8)

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor (reeksen van) toetsen uit 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 mevrouw Jennifer Roubiës MSc (secretaris).

De kwaliteit van de dataverzameling

S1 Is de steekproef representatief?

Bevindingen:

De representativiteit op achtergrondgegevens van de normeringspopulatie (groep leerlingen die aan het normeringsonderzoek heeft deelgenomen) ten opzichte van de doelpopulatie (alle leerlingen in leerjaar 6, 7 en 8 van het regulier basisonderwijs) is naast aantallen afnames uiteraard ook van belang om een oordeel te kunnen toekennen aan dit aspect. Omdat voor de normering van de referentieniveaus (volgens het Referentiekader Taal en Rekenen van de commissie Meijerink, 2009) een extern criterium wordt gebruikt, het anker 2019 van de Eindtoets dat door de Expertgroep Toetsen PO is vastgesteld, wordt door de auteurs van de Wetenschappelijke Verantwoording (WV) gesteld dat het niet noodzakelijk is dat de normeringspopulatie ook een normpopulatie is en dientengevolge de representativiteitseisen aan een normpopulatie hier dus niet van toepassing zijn. Het hoofddoel van onderhavig onderzoek is volgens de auteurs namelijk het vaststellen van absolute cesuren op de referentieniveaus, waarvoor de samenstelling van de normeringspopulatie volgens de auteurs van ondergeschikt belang is. Anderzijds stellen de auteurs echter ook vast dat het voor de bepaling van de kwaliteit van de items wel van belang is om na te gaan dat de normeringspopulatie geen specifiek selecte groep leerlingen is, maar een reguliere doorsnede van de leerlingen in de leerjaren 6, 7 en 8 van het basisonderwijs.

Voor bovengenoemd doel worden de volgende schoolachtergrondgegevens gebruikt: denominatie (openbaar en bijzonder), urbanisatiegraad (G4, G5-37 en >G37), schoolgrootte (<100, 100-300 en >300), regio (Noord, Oost, West en Zuid) en schoolgewicht (percentage gewichtenleerlingen, t.w. <1%, 1%-10% en >10%), welke per school bij DUO openbaar beschikbaar zijn. Er worden geen persoonlijke achtergrondgegevens gebruikt omwille van de privacy. De IEP Advieswijzer, het digitale platform van de voorloper van het IEP LVS om scholen handvatten te geven bij het onderbouwen van hun schooladvies, is gebruikt om de data voor het normeringsonderzoek te verzamelen. De betrokken scholen hebben de toetsen <1F-1F (aangeraden om in het tweede deel van leerjaar 6 af te nemen), <1F-1F-1S (aangeraden om in leerjaar 7 af te nemen) en 1F-1S (aangeraden om in leerjaar 8 af te nemen voordat het definitieve schooladvies wordt vastgesteld) op eigen initiatief tijdens het reguliere onderwijsproces in schooljaar 2018-2019 afgenomen in respectievelijk de leerjaren 6, 7 en 8. Hierdoor is gegarandeerd dat de data in het normeringsonderzoek is verzameld onder gelijke afnamecondities en afnamemomenten als waaronder de IEP LVS-toetsen afgenomen gaan worden. Of responsdata al dan niet bruikbaar is voor het normeringsonderzoek wordt besloten op basis van het percentage niet-beantwoorde items per leerling. Als een leerling meer dan 15% van de items niet heeft beantwoord, wordt zijn of haar data niet meegenomen in de analyses.

Hoe de normeringspopulatie zich op achtergrondgegevens verhoudt tot de doelpopulatie na de selectie van de data op bruikbaarheid is weergegeven in tabel 4.1 op pag. 10. Hierbij was de steekproeflengte voor de (in moeilijkheid opklimmende) toetsen <1F-1F, <1F-1F-1S en 1F-1S respectievelijk 731, 4912 en 3053. Tabel 4.1 laat zien dat niet alle categorieën van de achtergrondgegevens van de normeringspopulatie als een representatieve steekproef van de leerjaren 6, 7 en 8 van het basisonderwijs beschouwd kunnen worden. Zoals reeds eerder gememoreerd is de samenstelling van de normeringspopulatie volgens de auteurs echter van ondergeschikt belang voor het onderhavige onderzoek (dus niet noodzakelijkerwijs ook een normpopulatie), maar moet wel een reguliere doorsnede van de leerlingen in de leerjaren 6, 7 en 8 van het basisonderwijs zijn. De auteurs concluderen derhalve dat de gebruikte normeringspopulatie voor het onderhavige onderzoek bruikbaar is, omdat alle schoolachtergrondgegevens vertegenwoordigd zijn en er weinig grote effectgroottes phi (phi = sqrt[Chi-kwadraat/N]) zijn op basis van de berekende Chi-kwadraat waarden (zie tabel 4.2 op pag. 10). Met grote effectgroottes worden waarden van ≥ 0.50 bedoeld (Cohen, 1988).

Het hoofddoel van het onderhavige onderzoek is het bepalen van absolute cesuren op de referentieniveaus, waarvoor de samenstelling van de normeringspopulatie van ondergeschikt belang is. Voor de normering van de referentieniveaus wordt gebruik gemaakt van een extern criterium, het anker 2019 van de Eindtoets dat door de Expertgroep Toetsen PO is vastgesteld.

Er is sprake van een afgeleide normering. Omdat de ankeritems van de Eindtoetsen met behulp van een genormeerde steekproef berekend zijn en de LVS toetsen via deze items uiteindelijk gelinkt zijn aan het gezamenlijke anker, zijn daarmee de cesuren geborgd en is de concrete samenstelling van de normeringssteekproef minder belangrijk. Om deze redenering te onderbouwen is met behulp van statistische toetsen aangetoond dat één-en-hetzelfde IRT model van toepassing is op alle steekproeven die in de schaling gebruikt zijn.

Conclusie:

De steekproeven zijn adequaat gestratificeerd naar denominatie, urbanisatiegraad, schoolgrootte, regio en schoolgewicht en geven informatie over hoe de steekproeven zich verhouden tot de populatiewaarden. De procedure voor het samenstellen van de steekproeven is onderbouwd en de omstandigheden waaronder data is verzameld, is redelijk vergelijkbaar met de omstandigheden waaronder de toets wordt afgenomen. De steekproef wijkt weliswaar af, maar is wel bruikbaar voor het doel waarmee de dat verzameld zijn. Daarmee wordt aan aspect S1.1. het oordeel '**voldoende**' toegekend.

S1.2. In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:
In paragraaf 5.1 ('Normeringsonderzoek voor de IEP LVS-toetsen') wordt beschreven welk design is gebruikt om de data te verzamelen ten behoeve van het vaststellen van de huidige cesuren op de referentieniveaus. Zoals reeds eerder gememoreerd onder aspect S1 zijn de IEP LVS-toetsen qua inhoudelijke samenstelling identiek aan de IEP Advieswijzer-toetsen (operationele afname in het schooljaar 2018-2019) en zijn de afnamecondities van de IEP LVS-toetsen nagenoeg identiek aan de IEP Advieswijzer-

toetsen. Uitgangspunt bij het vaststellen van de cesuren voor de referentieniveaus zijn de cesuren op het gezamenlijk anker van de Eindtoetsen 2019 geweest, een extern criterium, welke door de Expertgroep Toetsen PO zijn bepaald na de afname van de Eindtoetsen in 2019. Deze cesuren zijn middels twee schalingen in het normeringsonderzoek voor het IEP LVS overgebracht naar de IEP LVS-toetsen (zie figuur 5.1 op pag. 11), waarbij gebruik werd gemaakt van de applicatie Lexter. De reeds bestaande linking tussen het gezamenlijk anker Eindtoetsen 2019 en de IEP Eindtoetsen 2017, 2018 (en 2019) is gebruikt in de eerste schaling. Voor de linking tussen de IEP Eindtoetsen 2017, 2018 (en 2019) en het IEP LVS is in de tweede schaling gebruikgemaakt van een ankerdesign (zie tabel 5.1 op pag. 12 voor een visuele weergave van het gebruikte ankerdesign).

Bij de samenstelling van de IEP Advieswijzer-toetsen zijn unieke items gebruikt en items die eerder in een IEP Eindtoets waren opgenomen. Om cesuren op de referentieniveaus te kunnen vaststellen (i.e., de normering van de referentieniveaus) zijn aan deze toetsen extra ankeritems (i.e., items die het design linken) uit de IEP Eindtoetsen 2017 en 2018 in blokjes sequentieel gezaaid, zodat de link tussen de IEP LVS-toetsen en de IEP Eindtoetsen uit (tenminste) 15 items per vaardigheid per niveau bestaat. Hiermee wordt dus voldaan aan de eis op pag. 5 uit het (op de homepage van de Expertgroep Toetsen PO te vinden) document 'Beoordelingskader voor (reeksen van) toetsen uit leerlingvolgsystemen (LVS)' dat het anker uit minstens 15 items moet bestaan. Verder zijn de verschillende zaaiblokjes herhaaldelijk verwisseld (na tenminste 400 afnames) om te garanderen dat van alle gezaaide ankeritems voldoende observaties verzameld konden worden. Op deze manier zijn in alle drie de toetsen alle ankeritems per niveau opgenomen geweest, al dan niet als meetellend item of als gezaaid niet-meetellend item.

Belangrijk is om na te gaan of de geselecteerde ankeritems uit de IEP Eindtoetsen 2017 en 2018 die zijn toegevoegd aan de IEP Advieswijzer-toetsen in het schooljaar 2018-2019 tezamen een goede afspiegeling zijn van de IEP LVS-toetsen (en dus representatieve ankeritems zijn), zowel inhoudelijk als psychometrisch. De auteurs concluderen dat de ankeritems inhoudelijk een goede afspiegeling zijn, omdat alle referentieniveaus en inhoudelijke toetsdomeinen van Rekenen (Getallen (30%), Verhoudingen (30%), Meten en meetkunde (20%) en Verbanden (20%)) vertegenwoordigd zijn (tabel 5.1 en Referentiekader Taal en Rekenen (Meijerink et al., 2009)). De auteurs concluderen dat de ankeritems psychometrisch ook representatief zijn, omdat er sprake is van een goede spreiding van de b-parameters (moeilijkheidsparameters in het Rasch-model als gehanteerd IRT-model) van de ankeritems (zie document 'Itemparameters IEP LVS Rekenen' in bijlage 3).

Het Rasch-model (1PLM) met de Marginal Maximum Likelihood (MML) schattingsmethode is toegepast voor de IRT-analyse (d.w.z. kalibratie van alle itemparameters op dezelfde schaal) van de IEP LVS-toetsen Rekenen <1F-1F, <1F-1F-1S en 1F-1S zijn uitgevoerd in Versie 0.0.7 van de applicatie Lexter (2019), waarmee door de Expertgroep Toetsen PO ook de analyses gedaan zijn voor de vaststelling van de referentiecesuren van de IEP Eindtoets 2019. Zodoende zijn de resultaten goed vergelijkbaar.

De itemparameters van het gezamenlijke anker van alle aangeboden Eindtoetsen 2019 zijn gefixeerd in de IRT-analyse meegenomen bij de eerste schaling (i.e., IEP Eindtoetsen). De Expertgroep Toetsen PO leverde de parameterwaarden van de gezamenlijke ankeritems aan. De itemparameters van de ankers tussen de IEP Eindtoetsen 2017 en 2018, die dus bij de eerste schaling geschat waren, zijn vervolgens gefixeerd bij de tweede

schaling (i.e., LVS-toetsen). Alle items van de IEP Eindtoets en LVS-toetsversies zijn zodoende op dezelfde onderliggende schaal gezet (i.e., gekalibreerd op de veronderstelde zelfde onderliggende vaardigheidsschaal). In bijlage 3 ('Itemparameters IEP LVS Rekenen') zijn het aantal afnames, de moeilijkheid (β), de meetfout van de moeilijkheid (SE(β)) en de iteminformatie (i.e., Fisher-informatie) bij de gemiddelde vaardigheid van de afnamegroepen voor de IEP Eindtoets 2019 en de drie LVS-toetsen <1F-1F, <1F-1F-1S en 1F-1S in een tabel weergegeven, waarbij de vaardigheidsschaal is genormeerd door het gemiddelde en de standaarddeviatie van de IEP Eindtoets 2019 op respectievelijk 0 en 1 te zetten.

In paragraaf 7.2 wordt een methode toegepast om de modelpassing van het toegepaste IRT-model (i.e., het Rasch-model) te verantwoorden, welke besproken wordt in het COTAN-beoordelingssysteem (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 worden waarden van c lager of gelijk aan 0.2 als 'goed' beoordeeld en waarden tussen 0.3 en 0.4 als 'voldoende'. Tabel 7.2 op pag. 19 laat zien dat de gemiddelde waarde van constante c berekend over alle items in beide kalibraties (i.e., de kalibratie van de items in de drie IEP Eindtoetsen 2017, 2018 en 2019 en de kalibratie van de items in de IEP LVS-toetsen <1F-1F, <1F-1F-1S en 1F-1S) veel lager is dan de vereiste 0.2 en geen item heeft een c-waarde boven de 0.2 (maximum-waarde voor de drie IEP Eindtoetsen is 0.091 en maximum-waarde voor de drie IEP LVS-toetsen is 0.102). Op basis van deze resultaten concluderen de auteurs dat de nauwkeurigheid van de parameterschattingen als goed kan worden beschouwd, waardoor een goede passing van het meetmodel aannemelijk is.

In paragraaf 6.3 ('Uitgangspunten selectie items IEP LVS-toetsen Rekenen') worden de volgende uitgangspunten besproken voor de selectie van items voor de IEP LVS-toetsen Rekenen gebaseerd ten eerste op het niveau van de items en ten tweede op de statistieken afkomstig uit de TIA-analyse (zie bijlage 4 voor de TIA-statistieken van het normeringsonderzoek):

(i) Items van de twee laagste referentieniveaus <1F en 1F hebben een p-waarde onder de 0.95 en items van het hoogste referentieniveau 1S hebben een p-waarde onder de 0.90;
(ii) Items hebben een p-waarde boven de 0.30 (open items) of een p-waarde boven de gokkans (meerkeuze items). Items met een p-waarde onder 0.30 worden alleen geselecteerd als het gaat om open items met een goede rit-waarde;
(iii) Items hebben een rit-waarde van groter of gelijk aan 0.20, zodat items als voldoende beoordeeld kunnen worden volgens het COTAN-beoordelingssysteem (Evers et al., 2010).

Voor een aantal items is een uitzondering gemaakt op deze uitgangspunten teneinde een goede inhoudelijke dekking van het Referentiekader Taal en Rekenen en variatie in inhoudelijke domeinen te kunnen garanderen.

Met behulp van statistische toetsen is aangetoond dat het design adequaat is.

Conclusie:

Het onvolledige maar 'verbonden' design van de proefonderzoeken is adequaat. Het volledige design van de toets <1F-1F, <1F-1F-2F en 1F-2F is eveneens adequaat. De normen worden jaarlijks geëvalueerd en eventueel aangepast. Een modelfit analyse is toegevoegd om de geschiktheid van het design te onderbouwen. Aan aspect S1.2. wordt het oordeel '**voldoende**' toegekend.

Normering

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

Bevindingen:
Zoals reeds beschreven bij het gebruikte dataverzamelingsdesign onder aspect S2 zijn de cesuren op het gezamenlijk anker van de Eindtoetsen 2019 (bepaald door de Expertgroep Toetsen PO na de afname van de Eindtoetsen in 2019) het uitgangspunt geweest bij het vaststellen van de absolute cesuren voor de referentieniveaus IEP LVS. Door middel van twee schalingen/kalibraties (kalibratie van de items in de drie IEP Eindtoetsen en kalibratie van de items in de IEP LVS-toetsen) zijn deze cesuren in het normeringsonderzoek voor het IEP LVS overgebracht naar de IEP LVS-toetsen (zie figuur 5.1 op pag. 11). Evenals de kalibratie is de cesuurbepaling uitgevoerd in de applicatie Lexter. De cesuurpunten zijn berekend in hele scorepunten op de scoreschalen van de verschillende toetsen. Er is dus sprake van een criteriumgerichte interpretatie van de toetsscores, waarbij de toetsscores vergeleken worden met een absolute norm (i.e., het gezamenlijk anker van de Eindtoetsen 2019 als extern criterium).

In het Referentiekader Taal en Rekenen (Meijerink et al., 2009) wordt een onderscheid gemaakt tussen drie (cumulatieve) fundamentele niveaus (1F, 2F en 3F) en drie (cumulatieve) streefniveaus (1S, 2S en 3S). In het IEP LVS Rekenen zijn ook vragen opgenomen die eenvoudiger zijn dan het niveau 1F, welke worden aangeduid met niveau <1F ('op weg naar 1F'). Dit niveau is niet bij wet vastgesteld. Vragen met het niveau <1F zijn in het IEP LVS opgenomen, zodat gemeten kan worden in hoeverre een leerling in staat is om rekenopgaven net onder niveau 1F te beantwoorden. In het IEP LVS Rekenen worden dus drie referentieniveaus onderscheiden, t.w. <1F, 1F en 1S. Om die reden zijn er drie toetsen samengesteld die oplopend zijn in niveau. Gekozen is voor één toets met items van de niveaus <1F en 1F (<1F-1F), één toets met items van de niveaus <1F, 1F en 1S (<1F-1F-1S) en één toets met items van de niveaus 1F en 1S (1F-1S). In paragraaf 6.2 ('Referentieniveaus binnen toetsen') wordt de keuze van referentieniveaus binnen de drie verschillende toetsen nader toegelicht met o.a. in tabel 6.1 een weergave van de gemiddelde vaardigheid en informatie per itemniveau (<1F, 1F en 1S) en afnamegroep (IEP Eindtoets 2019, <1F-1F, <1F-1F-1S en 1F-1S). De gemiddelde vaardigheid is hierbij genormeerd door de gemiddelde vaardigheid van de afnamegroep van de IEP Eindtoets 2019 op 0 te zetten. Geconcludeerd wordt dat de gegevens in tabel 6.1 de keuze voor de niveaus in de drie verschillende toetsen ondersteunen.

In tabel 7.1 op pag. 18 worden in scorepunten de equivalente cesuren 1F en 1S van de IEP Eindtoets 2019 en van de drie (in moeilijkheid opklimmende) LVS-toetsen <1F-1F, <1F-1F-1S en 1F-1S gerapporteerd. De drie toetsen voor leerjaar 6 t/m 8 (Rekenen <1F-1F, Rekenen <1F-1F-1S en Rekenen 1F-1S) bestaan uit 40 vragen met een maximale

score van 40. Daarnaast kunnen de toetsen een paar extra vragen bevatten die niet meetellen voor de uitslag, maar gebruikt worden om de kwaliteit van en de vergelijkbaarheid tussen toetsen te waarborgen. In tabel 7.1 wordt, naast de cesuren in scorepunten, voor iedere cesuur de bijbehorende vaardigheid (de vaardigheid van het naar boven afgeronde scorepunt van de cesuur), de lokale meetfout van de vaardigheid, de lokale betrouwbaarheid en de lokale informatiewaarde van die cesuur gerapporteerd voor de IEP Eindtoets 2019 en de drie IEP LVS-toetsen. Voor de drie IEP LVS-toetsen is per referentieniveau ook nog het percentage leerlingen berekend dat onterecht gezakt of geslaagd is (misclassificaties). In bijlage 5 ('Algemene toelichting methode') is meer uitleg te vinden voor wat betreft de berekening van de lokale informatiewaarden, lokale betrouwbaarheid en misclassificaties. Hierbij dient nog opgemerkt te worden dat de in tabel 7.1 gerapporteerde cesuren van de toets 1F-1S zijn vastgesteld op basis van de vaardigheidsschaal berekend op basis van 10.000 virtuele afnames, omdat ten tijde van de hier gerapporteerde analyses er nog geen afnamedata beschikbaar waren van drie items van de 1F-1S toets in de definitieve samenstelling van deze toets.

Conclusie:
De standaardbepalingsmethode is gemotiveerd en op de juiste wijze uitgevoerd. Op aspect N1.1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het volgende oordeel toegekend: '**voldoende'**.

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

Bevindingen:
In de WV wordt niets vermeld over wie de beoordelaars/vakdeskundigen/experts zijn die de absolute cesuren op de drie IEP LVS-toetsen Rekenen hebben bepaald. Maar zoals hierboven onder aspect N1.1. is geconstateerd, is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd met gebruikmaking van de applicatie Lexter. Ook kan in dit verband nog worden opgemerkt dat in het normeringsonderzoek voor de IEP LVS-toetsen de linking tussen het gezamenlijke anker Eindtoetsen 2019 en de IEP Eindtoets 2019 is verantwoord in de 'Verantwoording afname 2019 en normeringsonderzoek 2020' van de IEP Eindtoets 2019 die Bureau ICE op 1 augustus 2019 bij de Expertgroep Toetsen PO heeft ingediend. Verder is de linking tussen de IEP Eindtoetsen 2017, 2018 (en 2019) en het IEP LVS beschreven in de paragraaf 'Linking IEP LVS-toetsen met IEP Eindtoetsen' op pag. 12 (zie ook tabel 5.1 op pag. 12 voor een visuele weergave van het gebruikte ankerdesign). De beoordelaars mogen er dus vanuit gaan dat de beoordelaars/vakdeskundigen/experts naar behoren zijn geselecteerd en getraind.

Conclusie:
De beoordelaars/vakdeskundigen/experts zijn naar behoren geselecteerd en getraind. Op aspect N1.2. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel '**voldoende'** toegekend.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
In de WV wordt niets vermeld over of er voldoende overeenstemming tussen de beoordelaars is voor wat betreft de gevolgde normeringsprocedure (i.e., hoe de vaststelling van de absolute cesuren op de referentieniveaus tot stand is gekomen). Maar zoals hierboven onder aspect N1.1 is geconstateerd, is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd met gebruikmaking van de applicatie Lexter. De beoordelaars mogen er dus vanuit gaan dat er voldoende overeenstemming is tussen de beoordelaars.

Conclusie:
Er is voldoende overeenstemming tussen de beoordelaars. Op aspect N1.3. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel '**voldoende'** toegekend.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
Uit tabel 4.1 op pag. 10 valt af te lezen dat de normgroepen voor de drie IEP LVS-toetsen Re <1F-1F, Re <1F-1F-1S en Re 1F-1S in de normeringspopulatie respectievelijk 731, 4912 en 3053 zijn, welke aantallen groot genoeg zijn om schattingen te kunnen maken voor de doelpopulatie. Er moet hierbij echter wel de kanttekening worden geplaatst dat een steekproeflengte van N = 731 voor de toets Re <1F-1F eigenlijk aan de kleine kant is. 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 normgroepen zijn groot genoeg. Echter wel met de kanttekening dat de steekproef van N = 731 voor de toets RE <1F-1F volgens het document 'Scenario's voor ijking van de eindtoetsen op de referentieniveaus' van Glas, Emons en Berding-Oldersma (december 2016) eigenlijk aan de kleine kant is, omdat hier als eis een minimale toetslengte van 1000 wordt geformuleerd voor een betrouwbare cesuur. Op aspect N2.1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 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 basisonderwijs te fungeren. Volgens de auteurs was het echter ook niet perse 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 het hoofddoel van dit onderzoek is het vaststellen van absolute cesuren op de referentieniveaus, waarvoor de samenstelling van de normeringspopulatie volgens de auteurs van ondergeschikt belang is. Het is volgens de auteurs voor de bepaling van de kwaliteit van de items echter wel van belang vast te stellen dat de normeringspopulatie geen specifiek selecte groep leerlingen is, maar een reguliere doorsnede van de leerlingen in de leerjaren 6, 7 en 8 van het basisonderwijs. Tabel 4.1 op pag. 10 laat zien dat de normeringspopulatie inderdaad niet op alle categorieën als een representatieve random steekproef van de leerjaren 6, 7 en 8 van het basisonderwijs beschouwd kan worden. Omdat er echter maar weinig grote effectgroottes zijn (zie tabel 4.2 op pag. 10) concluderen de auteurs dat de gebruikte normeringspopulatie toch bruikbaar is voor dit onderzoek (i.e., de normeringspopulatie is geen specifiek selecte groep leerlingen, maar kan beschouwd worden als 'een reguliere doorsnede van de leerlingen in de leerjaren 6, 7 en 8 van het basisonderwijs').
Bureau ICE heeft er bewust voor gekozen om (anders dan wat eerder is gedaan door andere toetsaanbieders) de normen, zoals beschreven in de hoofdstukken 6, 7 en 8, 1 jaar geldig te laten zijn (i.e., tot en met het einde van het schooljaar 2019-2020). De reden hiervoor is dat de absolute cesuren van de referentieniveaus van de IEP LVS-toetsen thans zijn gebaseerd op het gezamenlijk anker 2019 van de Eindtoetsen zoals bepaald door de Expertgroep Toetsen PO na de afname van de Eindtoetsen in 2019. Omdat jaarlijks de normering van de Eindtoetsen wordt gedaan (en, indien nodig, de cesuren op de Eindtoets jaarlijks worden herzien door de Expertgroep Toetsen PO), zal daarom ook jaarlijks de normering van de IEP LVS-toetsen, indien nodig, worden herzien op basis van het meest recent afgenomen gezamenlijk anker van de Eindtoetsen. Zodoende kan worden gegarandeerd dat de cesuren 'up to date' blijven en daarmee de resultaten op de IEP LVS-toetsen ook het best vergelijkbaar zijn met de resultaten op de Eindtoets.
Op gelijke wijze als in dit normeringsonderzoek, worden er elk nieuw schooljaar zaaiblokken met ankeritems toegevoegd aan de IEP LVS-toetsen. Deze zaaiblokken worden elk schooljaar ververst met minimaal 15 ankeritems per referentieniveau uit de laatste IEP Eindtoets, waarbij bij de selectie van de ankeritems rekening zal worden gehouden met de representativiteit van deze items op gelijke wijze zoals beschreven in paragraaf 5.1 ('Normeringsonderzoek voor de IEP LVS-toetsen'). Door gebruik te maken van de jaarlijkse verversing van de ankeritems in de IEP LVS-toetsen is het mogelijk om na elk schooljaar de IEP LVS-toetsen opnieuw te normeren op dezelfde wijze zoals is beschreven in paragraaf 5.1.
Op pag. 7 wordt door de auteurs nog opgemerkt dat m.i.v. het schooljaar 2020/2021 de Inspectie van het Onderwijs het oordeel over de basisvaardigheden baseert op de referentieniveaus voor Taal en Rekenen behaald op de Eindtoets in leerjaar 8. Ook kijkt men naar de vorderingen van de leerlingen op de volgtoetsen in het leerlingvolgsysteem, voor de leerjaren 6, 7 en 8 naar het behaalde referentieniveau. Dit sluit volgens de auteurs dan mooi aan op de manier van rapporteren van het IEP LVS.

Conclusie:
Op aspect N2.2. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het volgende oordeel toegekend: '**voldoende'**.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In bijlage 4 (TIA's) worden de globale betrouwbaarheden (Cronbach's alpha) weergegeven voor de drie IEP LVS-toetsen (Re <1F-1F, Re <1F-1F-1S en Re 1F-1S), welke zijn berekend met het programma TiaPlus. Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt. In tabel 7.1 op pag. 18 worden ook de lokale betrouwbaarheden, gemeten bij de cesuurpunten op de latente vaardigheidsschaal, weergegeven voor zowel de IEP Eindtoets 2019 als voor de drie IEP LVS-toetsen (Re <1F-1F, Re <1F-1F-1S en Re 1F-1S). 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 7.1 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 op pag. 18 voor de drie IEP LVS-toetsen per referentieniveau ook nog het percentage leerlingen berekend dat onterecht gezakt of geslaagd is (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 software, kunnen de beoordelaars ervan uitgaan dat de betrouwbaarheidsgegevens correct zijn berekend.

Conclusie:
De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel **'voldoende'** toegekend.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen:
Op pag. 16 wordt vermeld dat de globale betrouwbaarheden in termen van Cronbach's alpha voor de drie IEP LVS toetsen Rekenen <1F-1F, <1F-1F-1S en 1F-1S gelijk zijn aan respectievelijk 0.90, 0.87 en 0.89 (zie ook bijlage 4). Omdat het behalen van een referentieniveau in de categorie valt 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 op pag. 18 laat zien dat aan deze minimale eis ook wordt voldaan voor de lokale betrouwbaarheden bij de cesuurpunten. De geschatte lokale betrouwbaarheden van de referentiecesuren zijn namelijk bij alle drie LVS-toetsen hoger dan 0.800 en voor de 1F cesuur bij de 1F-1S toets zelfs hoger dan 0.900 (i.e., 0.909). Verder laat tabel 7.1 op pag. 19 nog zien dat de classificatiefouten (i.e., voor de drie LVS-toetsen per referentieniveau het percentage leerlingen dat onterecht gezakt of geslaagd is) loopt van 7% tot 12% (hoe hoger de lokale betrouwbaarheid, hoe lager de classificatiefout).

Conclusie:
Op basis van het voorgaande wordt op aspect B2. aan het IEP LVS Rekenen groep 6 t/m 8 het volgende oordeel toegekend: '**voldoende**'.

Validiteit

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

Bevindingen:
Er zijn zorgvuldige keuzes gemaakt om een zo valide mogelijke toets te creëren:

- Men heeft zich goed rekenschap gegeven van het referentiekader taal en rekenen. Dat is o.a. zichtbaar in de invulling van 'functioneel rekenen', wat tot uitdrukking komt in een redelijk goed gebruik van contexten (zie opm. bij 'gebruikte items' over items waar dat minder gelukt is).
- De inschatting van het niveau (<1F e.d.) is door de bank genomen goed uitgevoerd. Bij 1S is het soms nog een zoektocht (kan bijv. leiden tot probleem-oplospuzzeltjes die er minder toe doen), maar verder levert e.e.a. een goede spreiding op.
- De verhouding tussen kale sommen en contextsommen is uitgevoerd volgens het kader, maar het oogt ook als een uitgebalanceerde keuze om leerlingen bij deze vaardigheden goed te ondersteunen.

Conclusie:
Op aspect V1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel '**voldoende'** toegekend.

V2 Constructvaliditeit: Meet de toets in zijn geheel datgene wat hij beoogt te meten?

Bevindingen:
In de wetenschappelijke verantwoording wordt de constructvaliditeit onderbouwd door aan te tonen dat een unidimensioneel IRT model op de data past.

Conclusie:

Op aspect V.2. wordt 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 adequate manier gemeten?

Bevindingen:
Uit het kalibratieonderzoek in paragraaf 7.2 ('Passing van het meetmodel en nauwkeurigheid van de parameterschattingen') blijkt dat de items van het IEP LVS

Rekenen groep 6 t/m 8 op één en dezelfde vaardigheidsschaal (thetaschaal) afgebeeld kunnen worden en dat aan de hand van de door de leerling behaalde vaardigheidsscores op de drie IEP LVS toetsen (<1F-1F-1S, <1F-1F en 1F-1S) diens groei adequaat gemeten kan worden. In het IEP LVS wordt dit o.a. gerealiseerd doordat alle toetsen een uitspraak doen op referentieniveau, welke zich beperkt tot drie niveaus: onder 1F (<1F), 1F en 1S. Om ook de ontwikkeling binnen de niveaus te kunnen duiden is een scoreschaal ontwikkeld die, helder en eenduidig voor leerkrachten en leerlingen, verfijnd weergeeft welk vaardigheidsniveau een leerling heeft.

Deze scoreschaal wordt aangeduid met de term ontwikkelscoreschaal en is een lineaire transformatie van de thetaschaal (de getalswaarden van deze schaal zijn voor leerkrachten en leerlingen betekenisloos) door de referentieniveaus 1F en 1S te koppelen aan respectievelijk de waarden 60-79 en 80-99 en de 1F- en 1S-cesuurpunten op de thetaschaal dus overeenkomen met de ontwikkelscores 60 en 80. Voor Rekenen geldt daarbij dat de 1F-cesuur met theta -1.292 overeenkomt met 60 op de ontwikkelscoreschaal en dat de 1S-cesuur van 0.265 overeenkomt met 80, hetgeen resulteert in de volgende lineaire omzettingsformule: Ontwikkelscore Rekenen = θ * 12.8 + 76.6. Als de berekende ontwikkelscore op basis van het inhoudelijk domein boven een benoemde maximale ontwikkelscore voor een IEP LVS-toets uitkomt (toets <1F-1F heeft een maximale ontwikkelscore van 79 en de toetsen <1F-1F-1S en 1F-1S hebben een maximale ontwikkelscore van 99), wordt deze gecorrigeerd en krijgt een leerling een ontwikkelscore gelijk aan de hoogst haalbare begrensde ontwikkelscore voor de betreffende IEP LVS-toets. Dit levert een voor leerkrachten en leerlingen bekende en betekenisvolle schaal (bestaande uit alleen gehele positieve getallen) op, omdat zij wel bekend zijn met de ontwikkelscoreschaal die Bureau ICE al jaren gebruikt in het PO, VO en MBO.

Met bovenstaande lineaire omzettingsformule en beschreven begrenzing is per IEP LVS-toets (<1F-1F, <1F-1F-1S en 1F-1S) de ontwikkelscore voor alle ruwe scores berekend. In tabel 8.1 op pag. 21 is de omzettingstabel voor de LVS-toetsen Rekenen weergegeven, waarin per (ruwe) scorepunt (lopend van 0 t/m 40) de thetawaarde (θ), ontwikkelscore (OS) en lokale betrouwbaarheid (REL) zijn opgenomen. Indien een leerling een scorepunt onderaan de scoreschaal haalt met een lokale betrouwbaarheid onder de 0.70 (de minimale betrouwbaarheid volgens het COTAN-beoordelingssysteem voor minder belangrijke beslissingen op individueel niveau), zal de uitslag "geen resultaat" worden gegeven. Aan de scorepunten aan de bovenkant van de scoreschaal met een lokale betrouwbaarheid onder de 0.70 wordt de ontwikkelscore van het hoogste scorepunt met een lokale betrouwbaarheid van > 0.70 toegekend.

Conclusie:
Op aspect Va1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel '**voldoende**' toegekend.

Va2 Wordt de betrouwbaarheid van de groei op die schaal adequaat 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 drie IEP LVS-toetsen en is beschreven in paragraaf 7.1 (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. De lokale betrouwbaarheden van de thetaschaal en de ontwikkelscoreschaal zijn gelijk, omdat de ontwikkelscoreschaal een lineaire transformatie is van de thetaschaal.

Opm.:
Pag. 22, 1e en 2e regel van boven: Lokale betrouwbaarheid i.p.v. globale betrouwbaarheid.

Conclusie:
Op aspect Va2. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het volgende oordeel toegekend '**voldoende**'.

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

Bevindingen:
In paragraaf 8.2 ('Leergroei') worden twee grafieken gegenereerd in het IEP LVS om de interpretatie van de behaalde resultaten te ondersteunen, namelijk de voortgangsgrafiek (zie fig. 8.1 op pag. 23) en de leergroeimeter (zie fig. 8.3 op pag. 25). De voortgangsgrafiek geeft de vaardigheidsontwikkeling weer in de tijd (door de resultaten van alle toetsen, die in de leerjaren 6, 7 en 8 zijn afgenomen, op één en dezelfde schaal uit te drukken) en in de leergroeimeter wordt de leergroeifactor van een leerling afgezet tegen de gemiddelde groeifactor van de normeringspopulatie. Het verschil in toetsresultaat, uitgedrukt in ontwikkelscorepunten, tussen twee of meer opeenvolgende toetsmomenten duidt hierbij de leergroei van een leerling aan en in paragraaf 8.2 ('Leergroei') op pag. 23 wordt gedetailleerd beschreven hoe de gemiddelde groeifactor van de normeringspopulatie wordt berekend. Het voorbeeld van de leergroeimeter in fig. 8.3 op pag. 25 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.

De interpretatie van de vaardigheidsontwikkeling van de leerling wordt voor de leerkracht ondersteund door de twee grafische weergaven, de voortgangsgrafiek en de leergroeimeter, gecombineerd te gebruiken. 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 'IEP LVS Talentenkaart' voor een voorbeeld van een Talentenkaart). Voor de interpretatie van de leervorderingen worden aan keerkrachten handvatten gegeven in de handleiding van het IEP LVS https://handleiding.toets.nl/snel-op-weg-met-het-iep-lvs-245.

Conclusie:
Op aspect Va3. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het volgende oordeel toegekend: '**voldoende**'.

Inzicht in leervorderingen

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

Bevindingen:
Zie ook de opmerkingen die gemaakt zijn door de beoordelaars bij LVS Lezen, o.a. over het gebruik van de talentenkaart.

Wat sterk is uitgevoerd (in de toelichting rekenen) is dat men voor elk domein van rekenen en voor elk niveau (<1F, 1F en 1S) enkele goed beschreven inhouden heeft (bolletjes-lijst) en dat men voorbeelden heeft gebruikt ter toelichting van deze inhouden. Dit is een goede hulp voor leerkrachten, ouders (en wellicht ook kinderen).

NB: Het zou goed zijn de gebruikte voorbeelden door te lopen en te kijken of dit de meest gelukkige voorbeelden zijn. De indruk van de beoordelaar is van niet. Juist enkele items waar een kritische kanttekening bij geplaatst wordt (zie bijlage met de bespreking van de items) worden hier gebruikt in de voorbeelden, en dat zal dus tot een 'vertekend' (en minder goed) beeld leiden als het gaat om de beoogde inhoud.

Conclusie:
Op aspect I1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 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:
Men heeft goede notie genomen van de inhouden van het referentiekader rekenen, zowel wat betreft de inhouden van de domeinen, als de niveau-aanduidingen waar mee gewerkt wordt (<1F, 1F, 1S), hoewel de beoordelaars over dat laatste willen meegeven dat hier op item-niveau nog wel eens twijfel is over de inschatting (zie wederom de item-analyse).

Conclusie:
Op aspect R1. wordt aan het IEP LVS Rekenen groep 6 t/m 8 het oordeel '**voldoende'** toegekend.

3. Verzamelstaat

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

4. Literatuurlijst

- Bezdan, E., Binsbergen, M., Haitjema, T, Helsloot, J. & Laan, J. (2019). Verantwoording IEP LVS-toetsen Rekenen. Culemborg: Bureau ICE.