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 Taalverzorging maken onderdeel uit van dit volgsysteem. De ter beoordeling voorliggende toetsen zijn de toetsen voor leerjaar 6, 7 en 8: versie 2 van de toetsen <1F-1F en <1F-1F-1S, die in 2019 ter beoordeling bij de Expertgroep ingediend zijn.

De toetsen Taalverzorging meten de vaardigheid van leerlingen op het gebied van taalverzorging. De toetsen worden digitaal gemaakt. Het is mogelijk om de teksten uit de toets als boekje te downloaden en te printen voor de leerling. Op die manier kan de leerling de teksten op papier lezen.

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 6, 7 en 8. 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 taalverzorging, 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. Voor leerlingen met ondersteuningsbehoeften als dyslexie is er audio-ondersteuning.

Inhoudelijke theoretische inkadering:
Het Referentiekader Taal en Rekenen (Meijerinket al., 2009) ligt ten grondslag aan de toetsen Taalverzorging van leerjaar 6 tot en met 8. In het Referentiekader is beschreven welke kennis en vaardigheden van leerlingen worden verwacht. Binnen het taalonderwijs wordt onderscheid gemaakt tussen vier fundamentele niveaus (1F t/m 4F).Het Referentiekader heeft een cumulatief karakter. Zo beheerst een leerling op niveau 2F ook de inhoud van niveau 1F.

In het IEP LVS zijn ook vragen opgenomen op het niveau ‘op weg naar 1F’. Dit zijn vragen die eenvoudiger zijn dan het niveau 1F. Dit niveau is niet bij de wet vastgesteld. Vragen met het niveau ‘op weg naar 1F’ zijn in het IEP LVS opgenomen zodat gekeken kan worden in hoeverre een leerling op weg is naar 1F.

Inhoud van het toetspakket
Het toetspakket Taalverzorging voor leerjaar 6, 7 en 8 bestaat uit de volgende documenten:

- Verantwoording LVS-toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-1S versie 2, 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),

    - IJking van de referentieniveaus (hfdst. 7),
    - De constructvaliditeit (hfdst. 8),
- Toetswijzer (bijlage 1)
- Itemoverzicht (bijlage 2)
- Itemparameters (bijlage 3)
- TIA's (bijlage 4)
- Algemene toelichting methode (bijlage 5)
- Omzettingstabel (bijlage 6)
- Langrange Multiplier Tracelines (bijlage 7)
- Toelichtingen (bijlage 8)

De verantwoording van de toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 is in het verlengde uitgewerkt van de verantwoording 'IEP LVS toetsen Taalverzorging 1F-2F, <1F-1F-2F en <1F-1F' (Bezdan, Binsbergen, Haitjema, Helsloot, & Laan, 2019). De toetsen versie 2 die hier voorliggen ter beoordeling, hebben hetzelfde doel en functie als de toetsen versie 1, en vormen daarmee tezamen één integraal leerlingvolgsysteem voor de leerjaren 6 t/m 8.

2. Beoordeling van de kwaliteitsaspecten

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

De kwaliteit van de dataverzameling

S1 Is de steekproef van leerlingen representatief?

Bevindingen:
Uitgangspunt is een steekproefgrootte die resulteert in tenminste 400 observaties per item. Het aantal observaties per item zijn weergegeven in bijlage 3 (zie kolom 'Aantal afnames'). Dit is conform de eisen 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 ten opzichte van de doelpopulatie van belang. De auteurs stellen dat aangezien voor de normering van de referentieniveaus gebruikt wordt gemaakt van een extern criterium, het anker 2019 van de Eindtoets dat door de Expertgroep Toetsen PO vastgesteld is, het niet noodzakelijk is dat de normeringspopulatie ook een normpopulatie is. De representativiteitseisen aan een normpopulatie zijn hier dus niet van toepassing. Voor de bepaling van de kwaliteit van de items is het echter wel van belang vast te stellen dat de normeringspopulatie geen specifiek selecte groep leerlingen is, maar een reguliere doorsnede van de bovenbouw van het basisonderwijs, d.w.z. dat er in ieder geval achtergrondgegevens worden gerepresenteerd in de steekproef. De reden hiervoor is dat in het onderhavige onderzoek absolute normen worden bepaald, waardoor de representativiteit van ondergeschikt belang is.

Persoonlijke achtergrondgegevens van de leerlingen konden niet worden gebruikt vanwege het waarborgen van de privacy (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.

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. Er is hier dus sprake van 'purposeful sampling' (doelsteekproef), een niet-probabilistische 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 naar de item- en toetskwaliteit en de standaardsetting is voor de bruikbaarheid van de

responsedata als selectiecriterium gebruikt dat records met meer dan 15% niet-beantwoorde items uit de responsedata worden verwijderd.

Tabel 4.1 laat zien hoe de normeringspopulatie zich op achtergrondgegevens verhoudt tot de doelpopulatie voor de analyses. 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 niveaus <1F-1F versie 2 en <1F-1F-2F versie 2 beschouwd kan worden. Wel kan uit de verdeling in tabel 4.1 geconcludeerd worden dat alle antwoordcategorieën vertegenwoordigd zijn in de normeringspopulatie en min of meer vergelijkbaar verdeeld zijn ten opzichte van de landelijke populatie. In tabel 4.2 worden bovendien nog de effectgroottes Phi weergegeven, die bijna allemaal lager zijn dan .50. Een uitzondering wordt gevormd door Regio Zuid voor niveau <1F-1F, die sterk ondervertegenwoordigd is. De auteurs motiveren dat dit geen consequenties heeft gehad voor de schaling met IRT.

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 Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 het volgende oordeel toegekend: '**voldoende'**

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

Bevindingen:
Bij het vaststellen van de cesuren voor de referentieniveaus zijn de cesuren op het gezamenlijk anker van de eindtoetsen 2019 het uitgangspunt geweest. Deze cesuren zijn door de Expertgroep Toetsen PO bepaald na de afname van de eindtoetsen in 2019.

Om de cesuren van het gezamenlijk anker van de eindtoets 2019 ook over te zetten op de IEP LVS toetsen versie 2 wordt gebruik gemaakt van twee schalingen. Allereerst zijn de IEP LVS toetsen gelinkt aan het anker van de eindtoets 2019. Vervolgens zijn de IEP LVS toetsen versie 2 gelinkt aan de reeds goedgekeurde IEP LVS toetsen versie 1 in een ankerdesign. De itemparameters van de ankers tussen de IEP Eindtoetsen en de IEP LVS toetsen, die bij de eerste schaling vorig jaar geschat waren, zijn hierbij gefixeerd bij de tweede schaling (i.e., LVS toetsen). Op deze wijze zijn alle items van IEP Eindtoetsen en de IEP LVS toetsen versie 1 en de nieuwe IEP LVS toetsen versie 2 op dezelfde parameterschaal gezet.

De items uit reeds goedgekeurde IEP LVS zijn zo geselecteerd dat deze ankeritems tezamen een goede afspiegeling zijn van de IEP LVS toetsen. De cesuren van de referentieniveaus zijn thans gebaseerd op het gezamenlijk anker 2019 van de eindtoetsen. Omdat het niet vaststaat dat de referentiecesuren op de latente schaal onveranderd blijven de komende jaren, is het beleid om de normering van de IEP LVS toetsen te blijven controleren aan het meest recent afgenomen gezamenlijk anker van de eindtoetsen.

Bij de selectie van de items voor de IEP LVS toetsen versie 2 is ten eerste gekeken naar het niveau van de items en ten tweede naar de statistieken afkomstig uit de TIA-analyse. Er is rekening gehouden met de p-waarde (proportie correct) en rit-waarde (item-

totaalcorrelatie) van de items. De p- en maximale rit-waarde van de items geselecteerd voor de LVS toetsten zijn in het document TIA's IEP LVS toetsen Taalverzorging (bijlage 4) per toets weergegeven. De uitgangspunten bij deze selectie waren:

- Items van het referentieniveau <1F en 1F hebben een p-waarde onder de 0,95 en items van het referentieniveau 2F hebben een p-waarde onder de 0,90.
- 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 (open items) of een p-waarde onder de gokkans (meerkeuze items) worden alleen geselecteerd als het gaat om open items met een goede rit-waarde.
- Items hebben een rit-waarde 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).

Er is voor enkele items een uitzondering gemaakt op deze uitgangspunten om een goede inhoudelijke dekking van het Referentiekader taal en rekenen (Meijerink et al., 2009) en variatie in tekstonderwerpen te kunnen garanderen. Dit is terug te zien in de bijlagen waar naar verwezen wordt in de hoofdstukken 2 en 3. Deze uitzonderingen hebben een verwaarloosbare negatieve invloed op de globale betrouwbaarheid (Cronbach's Alpha) van de toetsen. Het behalen van een referentieniveau valt in de categorie minder belangrijke beslissingen

op individueel niveau en volgens het beoordelingssysteem van COTAN (Evers et al., 2010) is de minimale betrouwbaarheidscoëfficiënt van testen voor minder belangrijke beslissingen tenminste 0,70. Met deze selectie van items is aan deze eis bij alle toetsen voldaan met betrouwbaarheden van 0,87 (<1F-1F versie 2) en 0,83 (<1F-1F-2F versie 2) (zie bijlage 4).

Voor de IRT-analyse van de IEP LVS toetsen is een one-parameter logistisch model (1PLM, oftewel Rasch model) met de marginal maximum likelihood (MML) schattingsmethode toegepast. De IRT-analyses zijn uitgevoerd in Versie 0.0.7 van de applicatie Lexter (2019). De keuze voor een 1PLM analyse in Lexter is gemaakt om de resultaten goed vergelijkbaar te houden met de analyses die gedaan zijn voor de vaststelling van de referentiecesuren van de IEP Eindtoets 2019. De definitieve cesuren voor de IEP Eindtoets 2019 werden na de afname door de Expertgroep Toetsen PO vastgesteld. De cesuren zijn éérst bepaald op het gezamenlijk anker van alle aanbieders van eindtoetsen in 2019 en vervolgens overgezet op de IEP Eindtoets 2019. De analyses om tot de definitieve cesuurstelling te komen van de IEP Eindtoets 2019 zijn door de Expertgroep Toetsen PO uitgevoerd in Lexter.

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 (σθ). 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 beide kalibraties. De gemiddelde waarde van de constante c berekend over alle items in beide kalibraties is veel

lager dan de vereiste waarde van 0,2. De c-waarde van drie items in de kalibratie van de IEP Eindtoetsen komt boven de waarde van 0,2 uit, maar in alle drie de gevallen is deze waarde kleiner dan 0,3 en daarmee voldoende. De nauwkeurigheid van de parameterschattingen gebaseerd op deze resultaten kan als goed beoordeeld worden.

Om modelpassing verder te onderzoeken, mede in het kader van begripsvalidering, is er een Differential Item Functioning (DIF) analyse uitgevoerd in Lexter. Op grond van deze resultaten kan er gesteld worden dat de observaties op de items – inclusief de ankeritems tussen de nieuwe en oude toetsen – van de IEP LVS toetsen weinig van de verwachting verschillen, ongeacht het boekje, wat een goede modelpassing aannemelijk maakt. Tot slot is door middel van de First order Statistics optie ('Lagrange multipliers tracelines for Rasch-Type Model') de mate waarin de item response curven de responsies goed representeren statistisch getoetst. Van de 889 berekende effectgroottes zijn er 124 groter dan 0,10. Een aantal items blijkt niet in het Rasch-model te passen volgens bovenstaande analyse. De auteurs stellen dat het verantwoord is om deze items in de IEP LVS toetsen te behouden omdat de items in de toetsen volgens de KTT goed functioneren (zie bijlage 4). Bovendien gaat het om een relatief klein percentage items dat een (te) grote effectgrootte laat zien in bovenstaande analyse.

Conclusie:
Het onvolledige maar 'verbonden' dataverzamelingsdesign voor beide kalibraties is adequaat. Op aspect S2 wordt aan de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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:
n.v.t.

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:
n.v.t.

Conclusie:
n.v.t.

Normering

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

Bevindingen:
De cesuurbepaling is, net als de kalibratie, uitgevoerd in de applicatie Lexter. Hierbij zijn de cesuurpunten in hele scorepunten berekend op de scoreschalen van de verschillende toetsen door na de tweede schaling (zie 5.1 en 6.1) opnieuw te 'schalen' in Lexter met alleen de meetellende items met gefixeerde parameters. Deze cesuren zijn berekend op basis van de cesuren uitgedrukt in vaardigheid (θ) die zijn aangeleverd door de Expertgroep Toetsen PO voor het onderdeel Taalverzorging van de IEP Eindtoets 2019. De vaardigheid van de cesuur wordt in Lexter omgerekend naar een exacte scorepunt (observed cut-off scores). Dit exacte scorepunt, op tienden nauwkeurig, is vervolgens afgerond naar het bovenliggende hele scorepunt, daar de uitslag van de toetsten in hele scorepunten wordt gegeven. Hierbij is dus sprake van een criteriumgerichte interpretatie van de testscores, waarbij de testscores vergeleken worden met een absolute norm.

Conclusie:
In plaats van het uitvoeren van een standaardsettingsprocedure, zijn de cesuren van de Eindtoets 2019 met behulp van irt omgezet naar cesuren voor de IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2. Deze omzetting is correct gebeurd. Dit criterium is daarom **voldoende.**

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

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
n.v.t.

Conclusie:
n.v.t.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
De aantallen van de normgroepen worden niet in de verantwoording genoemd maar uit bijlage 4 valt af te lezen dat de normgroepen de gebruikt zijn voor de kalibratie voor de beide toetsen IEP LVS-toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 in

de normeringspopulatie respectievelijk 1188 en 1941 zijn. Deze normgroepen zijn van voldoende grootte.

Conclusie:
De normgroepen zijn groot genoeg. Op aspect N2.1 wordt aan de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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. De auteurs stellen dat aangezien voor de normering van de referentieniveaus gebruikt wordt gemaakt van een extern criterium, het anker 2019 van de Eindtoets dat door de Expertgroep Toetsen PO vastgesteld is, het niet noodzakelijk is dat de normeringspopulatie ook een normpopulatie is. De representativiteitseisen aan een normpopulatie zijn hier dus niet van toepassing.

Conclusie:
Omdat voor de normering gebruik gemaakt wordt van een extern criterium is aspect N2.2 voor de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 '**voldoende'.**

N2.3 Zijn de normen correct bepaald?

Bevindingen:
De gevolgde linking procedure is correct uitgevoerd en er zijn ook 95% betrouwbaarheidsintervallen, lokale meetfout, lokale betrouwbaarheid en % classificatiefouten van de cesuren berekend.

Conclusie:
De normen zijn correct bepaald. Op aspect N2.3 wordt aan de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 het oordeel '**voldoende**' toegekend

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In bijlage 4 (TIA's <1F-1F 2 en <1F-1F-2F 2) worden de globale betrouwbaarheden (Cronbach's alpha) weergegeven voor de beide IEP LVS-toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2, 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 7.1 ook de lokale betrouwbaarheden, gemeten bij de cesuurpunten op de latente vaardigheidsschaal,

weergegeven voor beide toetsen. 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 en de standaarddeviatie van de vaardigheidsverdeling (σθ). 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. 16 voor de twee IEP LVS-toetsen per referentieniveau ook nog het percentage leerlingen berekend dat het betreffende niveau ten onrechte wel of niet heeft gehaald (classificatiefouten). Dit percentage is maximaal 20%. 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 Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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 <1F-1F versie 2 en <1F-1F-2F versie 2) kan afgelezen worden dat de globale betrouwbaarheden in termen van Cronbach's alpha (interne consistentie betrouwbaarheden) voor de beide IEP LVS toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 gelijk zijn aan respectievelijk 0,87 en 0,83. Omdat het behalen van één van de niveaus <1F-1F versie 2 en <1F-1F-2F versie 2 valt in de categorie van minder belangrijke beslissingen op individueel niveau, is met deze waarden voor de globale betrouwbaarheden ruimschoots voldaan aan de eis van het COTAN beoordelingssysteem (Evers et al., 2010) dat de minimale betrouwbaarheidscoëfficiënt van toetsen voor minder belangrijke beslissingen tenminste 0.70 moet zijn. Tabel 7.1 laat zien dat aan deze minimale eis ook ruimschoots wordt voldaan voor de lokale betrouwbaarheden bij de cesuurpunten van de toetsen <1F-1F versie 2 en <1F-1F-2F versie 2. De geschatte lokale betrouwbaarheden van de cesuren (1F en 2F) zijn namelijk bij de beide IEP LVS-toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 gelijk aan respectievelijk 0.828 (1F) voor de <1F-1F versie 2 toets en 0.821 (1F) en 0.715 (2F) voor de <1F-1F-2F versie 2 toets.

Verder laat tabel 7.1 nog zien dat de classificatiefouten (i.e., voor de vier IEP LVS-toetsen Taalverzorging het percentage leerlingen dat het betreffende niveau ten onrechte wel of niet heeft gehaald) loopt van 12% tot 20% (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.

De 95%-betrouwbaarheidsintervallen van de 1F- en de 2F-cesuur in de toets <1F-1F-2F versie 2 overlappen elkaar. Gezien de functie van beide toetsen, het schatten van de vaardigheid in de context van leergroei en niet in de context van het nemen van belangrijke beslissingen, is dit acceptabel volgens de auteurs.

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 Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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 van versie 2 richten zich op de referentieniveaus <1F, 1F en 2F.
Versie 2 heeft 2 toetsen uitgewerkt met ieder 40 opgaven met verschillende vraagtypes die oplopen in niveau (<1F-1F, <1F-1F-2F).
Binnen het domein Taalverzorging worden de volgende aspecten onderscheiden:

- verschillende categorieën van spellingsproblemen en –regels;
- regels voor lettergreepgrenzen;
- regels voor woordgrenzen (3F);
- regels voor morfologische spelling;
- regels voor werkwoordspelling;
- overige regels (vanaf 2F);
- regels voor leestekens.

In de Toelichtingen Taalverzorging voor groep 3 tot en met 3F zijn de aspecten per referentieniveau vanaf groep 6 nader omschreven. Dit biedt inzicht wat er per aspect en doel getoetst wordt en op welk niveau. Voor elk toetsdoel is een voorbeeld opgenomen. Dit biedt (voor de leraar) extra inzicht in de opbouw in moeilijkheid.
De items zijn adequaat uitgewerkt op de doelen binnen de aspecten. De toetsen/toetsitems bieden een goede vertaling van de niveaubeschrijvingen Taalverzorging in het Referentiekader.

De toetsen voor de leerjaren 6 tot en met 8 maken gebruik van de vraagtypen meerkeuze, multikeuze, tekstmarkeervraag, open vragen en meervoudig dichotoom. Doordat de toetsen digitaal zijn kunnen ze ook door het systeem worden nagekeken.

De antwoordcategorieën bij de meervoudig dichotoom, multikeuze en keuzevragen leveren geen onduidelijkheden op.

Illustraties worden soms functioneel ingezet en soms als illustratie. Ze ogen allemaal fris en aansprekend en leiden niet af.

Geen afzonderlijke opmerkingen bij de uitgewerkte items.

Conclusie:
'voldoende'

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

Bevindingen:
In paragraaf 7.2 ('Passing van het meetmodel en nauwkeurigheid van de parameterschattingen') was al aannemelijk gemaakt dat er sprake was van een passing van het geassumeerde meetmodel (i.e., het Rasch model) en er dus mag worden uitgegaan van unidimensionaliteit, hetgeen impliceert dat aan de noodzakelijke (maar niet voldoende) voorwaarde van constructvaliditeit wordt voldaan. In hoofdstuk 8 wordt aanvullend onderzoek verricht naar argumenten die pleiten voor de constructvaliditeit van de IEP LVS-toetsen Taalverzorging.

Op de LVS toetsen is een soortgenotenonderzoek uitgevoerd in de vorm van onderzoek naar convergente validiteit tussen twee binnen één leerjaar opeenvolgende IEP LVS toetsen van dezelfde vaardigheid. Ook is er onderzoek gedaan naar divergente validiteit tussen toetsen van verschillende vaardigheden (Lezen, Taalverzorging en Rekenen) binnen één en hetzelfde leerjaar. Deze worden in tabel 8.1 getoond.
Met andere woorden, er is gebruikgemaakt van een Mult-Trait Multi-Method matrix (MMTM), waarbij scores op Taalverzorging 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 8.1. Conform de verwachting is hieruit af te lezen dat de correlatie tussen twee opeenvolgende toetsen Taalverzorging structureel hoger is dan tussen twee toetsen van verschillende vaardigheden binnen hetzelfde leerjaar.
In tabel 8.2 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 0.74 en 0.77. De grootte van de ranges in tabel 8.2 wordt sterk bepaald door enkele outliers (zie hiervoor ook bijlage 4).

Vanuit de onderwijskundige analyse wordt het volgende opgemerkt: In de toetsmatrijzen wordt inzichtelijk gemaakt hoe aspecten Taalverzorging aan bod komen in de uitgewerkte toetsen. De toetsmatrijzen maken inzichtelijk welke aspecten op welk niveau getoetst worden. Onder ieder aspect vallen meerdere toetsdoelen. Toetsdoelen kunnen op een lager niveau en hoger niveau terugkomen door bijvoorbeeld woorden op een hoger niveau

langer en complexer te maken. In de toetsitems versie 2 van de niveaus <1, 1F en 2F is dit ook duidelijk waarneembaar.
De items zijn per aspect en niveau duidelijk onderscheidend van elkaar uitgewerkt.
Er zit een evenwichtige verdeling tussen de aspecten. Bijvoorbeeld het aspect (niet)werkwoordspelling heeft substantieel meer items dan bijvoorbeeld het aspect Leestekens. Ook is een duidelijke accentverschuiving zichtbaar van het aantal te toetsen items bij aspecten die onder en op het niveau 1F worden getoetst en vanaf 1F. Bijvoorbeeld in de toetsmatrijs Taalverzorging <1-1F is nog gericht aandacht voor categorieën van spellingsregels. In de toetsmatrijs <1-1F-F zie je het aantal items afnemen en items rond morfologische spelling en werkwoordspelling toenemen wat ook passend is bij de opbouw van Taalverzorging en de taken in het taalonderwijs waar leerlingen mee te maken krijgen die steeds complexer worden naarmate ze in een hogere groep zitten.

Conclusie:
De gerapporteerde resultaten in Hoofdstuk 8 (Constructvaliditeit) van de WV vormen een psychometrische ondersteuning voor de constructvaliditeit van de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 en er wordt dus gemeten wat men beoogt te meten. Op aspect V2 wordt aan de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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 IEP LVS-toetsen Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 zijn gekalibreerd op één en dezelfde onderliggende vaardigheidsschaal. Het gevolg hiervan is dat de vaardigheidsscores op deze vier 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.

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 Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 het volgende oordeel toegekend: '**voldoende'.**

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

Bevindingen:
De (conditionele) lokale betrouwbaarheid voor ieder scorepunt wordt op dezelfde manier geschat als de (conditionele) lokale betrouwbaarheid van de cesuurpunten (op de thetaschaal) voor de vier 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. In bijlage 6 wordt bovendien per scorepunt de thetawaarde, ontwikkelscore (OS) en lokale betrouwbaarheid (REL) weergegeven.

Conclusie:
De betrouwbaarheid van de groei op de ontwikkelscoreschaal wordt correct weergegeven. Op aspect Va2 wordt aan de toetsen IEP LVS Taalverzorging <1F-1F versie 2 en <1F-1F-2F versie 2 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 vaardigheidsontwikkeling van de leerling wordt voor de leerkracht ondersteund door de twee grafische weergaven in de hoofdstukken 9 en 10. De voortgangsgrafiek en de leergroeimeter, zijn 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. 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:
Er worden voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden. Op aspect Va3 wordt aan de toetsen IEP LVS Taalverzorging voor groep 6, 7, 8 het oordeel **voldoende** toegekend.

Inzicht in leervorderingen

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

Bevindingen:
Leraren kunnen de leervorderingen van de leerlingen per domein inzien (in percentage goed-scores). Ook kunnen de leraren de antwoorden van de leerlingen inzien wat het formatief gebruik van de toetsen bevordert. Aan leerkrachten worden handvatten gegeven voor de interpretatie van de leervorderingen. Op de IEP portal is ook nog extra informatie te vinden voor leraren die zich nog verder in deze materie willen verdiepen.

Speciaal voor ouders/verzorgers is er een leesbare Leeswijzer die hen handvatten geeft voor de interpretatie van de leervorderingen die op de Talentenkaart zijn weergegeven. De Talentenkaart is zo gemaakt dat deze voor de meeste leerlingen te begrijpen zal zijn.

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 Taalverzorging voor groep 6, 7, 8 het oordeel **voldoende** toegekend.

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

Bevindingen:
De opmerkingen die onder de paragraaf Va3 werden gemaakt zijn ook hier van toepassing: alle relevante informatie dient in de verantwoording aanwezig te zijn.

Na enig zoekwerk werd het volgende vastgesteld. De leerkracht wordt ondersteund bij de interpretatie van de vaardigheidsontwikkeling van de leerling door het gecombineerd gebruik van de twee grafische weergaven, de voortgangsgrafiek en de leergroeimeter. 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' 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.

De resultaten kunnen gebruikt worden om op zowel leerlingniveau, als op groeps- en schoolniveau te evalueren en analyseren. In de bovenbouw zie je welk referentieniveau de leerling beheerst en welke ontwikkelscore daarbij hoort. Ook kun je de resultaten vergelijken met het landelijk gemiddelde zodat je weet hoe een leerling of groep ervoor staat. Ook is het mogelijk om te sorteren op categorie of domein, zodat je kunt zien voor welke leerlingen nog extra vervolgstappen noodzakelijk zijn.

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 Taalverzorging voor groep 6, 7, 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:
De leerdoelen uit de toetsen zijn gebaseerd op de referentieniveaus.
In de rapportages worden vanaf leerjaar 6 uitspraken gedaan over de beheersing van de referentieniveaus.
Er zijn nu 5 toetsen in het LVS gericht op taalverzorging: <1F-1F 2 versies ; <1F-1F-2F 2 versies; 1F-2F.
In de toetsen komen de volgende categorieën aan de orde: categorieën van spellingsregels, lettergreepgrenzen, morfologische spelling, werkwoordspelling, overige regels (op 2F) en leestekens aan de orde. Er is 1 op 1 koppeling met het Referentiekader.

Conclusie:
'voldoende'

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	**n.v.t.**
	N1.3	**n.v.t.**
	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*	**Voldoende**

4. Literatuurlijst

- Bezdan, E., Binsbergen, M., Haitjema, T, Helsloot, J. & Laan, J. (2020). Verantwoording IEP LVS-toetsen Lezen. 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.