



Regeling van het College voor toetsen en examens van 7 november 2022, nummer CvTE-22.00963, houdende vaststelling van het beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen in het primair onderwijs (Regeling beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen PO)

Het College voor toetsen en examens,

Gelet op artikel 3a, eerste lid, onderdeel g, van de Wet College voor toetsen en examens;

Gezien de goedkeuring van de Minister voor Primair en Voortgezet Onderwijs, gegeven op 4 november 2022, nummer 1301072,

Besluit:

Artikel 1. Beoordelingskader

Het beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen als bedoeld in artikel 3a, eerste lid, onderdeel g van de Wet College voor toetsen en examens wordt vastgesteld als opgenomen in de bijlage van deze regeling.

Artikel 2. Inwerkingtreding

Deze regeling treedt in werking op het tijdstip waarop artikel VIII, onderdeel B, van de Wet van 9 februari 2022 tot wijziging van een aantal onderwijswetten in verband met aanpassingen op het gebied van de doorstroom van het basisonderwijs naar het voortgezet onderwijs en wijziging van de stelselinrichting van doorstroomtoetsen en toetsen verbonden aan leerling- en onderwijsvolgsystemen in het basisonderwijs (Stb. 2022, 135) in werking treedt.

Artikel 3. Citeertitel

Deze regeling wordt aangehaald als: Regeling beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen po.

Deze regeling zal met de toelichting in de Staatscourant worden geplaatst.

*Het College voor toetsen en examens,
de voorzitter,
J.H. van der Vegt*



BIJLAGE 1 BEOORDELINGSKADER VOOR DE TOETSEN BEHORENDE TOT LEERLING- EN ONDERWIJS VOLGSYSTEMEN

Bijlage behorende bij artikel 1 van de Regeling beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen po

1. Inleiding

1.1 Begrippen en definities

<i>Aanbieder</i>	Een indiener van een leerlingvolgsysteem.
<i>Absolute normen</i>	Standaarden waaraan de vaardigheidsniveaus van leerlingen gerelateerd kunnen worden. Deze standaarden worden vastgesteld op basis van inhoudelijke en/of toetsspecifieke criteria.
<i>Betrouwbaarheid</i>	De betrouwbaarheid is de mate waarin de toetsscores vrij zijn van toevallige meetfouten.
<i>CAT</i>	Een Computergestuurde adaptieve toets op itemniveau. Een CAT is een toets waarvan de opgaven tijdens de afname geselecteerd worden . Hierdoor krijgt elke leerling opgaven aangeboden die aansluiten bij diens niveau. Bij een goed antwoord krijgt de leerling een moeilijkere opgave, bij een fout antwoord een makkelijkere opgave.
<i>(Referentie)Cesuur</i>	De minimale prestatie (gerepresenteerd als vaardigheid of toetsscore) die net indicatief genoeg is voor het halen van een (referentie)niveau.
<i>Constructvaliditeit</i>	De eigenschap die een LVS-instrument heeft als kan worden aangetoond dat het instrument het door de constructeur beoogde kenmerk van de leerling (onderliggende trek, vaardigheid) meet. Deze wordt per kenmerk binnen een LVS bepaald.
<i>CvTE</i>	College voor toetsen en examens
<i>DIF</i>	Differentieel item functioneren is een verschijnsel dat items verschillend functioneren voor bepaalde (sub)groepen leerlingen, zelfs wanneer deze (sub)groepen leerlingen een vergelijkbaar prestatieniveau hebben.
<i>Domein</i>	Verzameling van onderling samenhangende kennis en vaardigheden waarover getoetst wordt.
<i>Gebruiksdoel</i>	Omschrijving van de beoogd te meten construct(en), doelgroep(en), oftewel een populatie, en functie(s) van het volgsysteem.
<i>IJken</i>	Bij een observatie- of registratie-instrument worden de indicatoren van observeerbare (gedrags-)kenmerken van de ontwikkelingsdoelen en eventuele tussentijdse mijlpalen in het instrument voorzien van ijkpunten (c.q. referentiepunten). IJken is het volgens bepaalde regels omzetten van de periodieke registraties van geobserveerde gedragingen in een signaal waarmee de gebruiker kan vaststellen in welke mate een individuele leerling zich op een zeker moment een specifiek ontwikkelingsdoel eigen heeft gemaakt.
<i>Inhoudvaliditeit</i>	De eigenschap dat de operationalisering van de inhoud van het instrument een representatieve weergave is van de te observeren en te volgen ontwikkeling.
<i>IRT</i>	Itemresponstheorie (IRT) is een theorie binnen de testleer waarbij op het niveau van items op een wiskundige manier een relatie wordt gelegd tussen de antwoorden die leerlingen geven en een onderliggende latente vaardigheid.
<i>KTT</i>	De klassieke testtheorie is gebaseerd op een model waarbij men er vanuit gaat dat de waargenomen toetsscore (bijvoorbeeld het aantal correct) uit twee componenten bestaat: (1) een ware score en (2) een toevallige meetfout.
<i>LVS</i>	De meetfout is geheel toevallig en hangt daarom met geen enkele andere variabelen samen. Dit model wordt ook wel het klassieke testmodel genoemd. Alle eigenschappen die volgen uit dit model vormen gezamenlijk de klassieke testtheorie.
<i>Lokale betrouwbaarheid</i>	Het (leerling)volgsysteem (LVS) betreft een systeem met onderliggende data waarmee de ontwikkeling van de leerling kan worden gevolgd over de tijd heen en waarmee kan worden geïnterpreteerd in hoeverre deze ontwikkeling in lijn is met de leerdoelen van de verschillende leerjaren en, vanaf groep 3, met de ontwikkeling van de andere leerlingen uit hetzelfde leerjaar. Er is sprake van een volgaspect wanneer een aantal (twee of meer) verschillende observatiemomenten op hetzelfde construct plaatsvinden; de afnamemomenten kunnen daarbij vast of variabel zijn.
<i>MST</i>	De mate waarin de toetsscores op een specifiek deel van de meetschaal vrij zijn van toevallige meetfouten.
<i>Normeren</i>	Bij een multistage toets bepaalt een algoritme het niveau van de leerling op basis van diens antwoorden op een set opgaven. Daarna krijgt de leerling een nieuwe set opgaven, afgestemd op het eerder bepaalde niveau.
<i>Relatief beoordelen</i>	Het toekennen van een waardering aan een behaalde score. Onderdeel van het normeren is het vaststellen van de cesuren.
<i>Relatieve normen</i>	Het rapporteren op basis van relatieve normen. Relatieve normen plaatsen de leerling op een schaal die betrekking heeft op de relatieve positie van een leerling in een specifieke, goed gedefinieerde populatie (zie gebruiksdoel). Een relatieve norm is derhalve gebaseerd op een onderlinge vergelijking van de toetsprestaties van de kandidaten binnen die populatie.
<i>Signaleringsfunctie</i>	De functie van een toets die de leerkracht helpt aan te zetten tot handelen om de leerling te helpen (beoogde) vorderingen te maken (op de gemeten trek).
<i>Terrein</i>	Een te toetsen onderdeel bijvoorbeeld rekenen ofwel Nederlandse taal.
<i>Toetsmatrijs</i>	Een adequate, schematische representatie van het meetdoel.
<i>Unidimensionaliteit</i>	Eigenschap van een schaal wanneer de items in een toets dezelfde vaardigheid meten. Met elke willekeurige subset van items uit de gekalibreerde itembank kan dezelfde onderliggende vaardigheid worden vastgesteld.
<i>Wet CvTE</i>	Wet College voor toetsen en examens
<i>WEC</i>	Wet op de expertisecentra
<i>WPO</i>	Wet primair onderwijs



1.2 Reikwijdte

Scholen zijn ingevolge artikel 45b WPO en artikel 48c, eerste en tweede lid WEC respectievelijk artikel 51a en 51b van de Wet primair onderwijs BES verplicht om in ieder geval gebruik te maken van door het CvTE erkende toetsen verbonden aan het LVS voor de terreinen Nederlandse taal en rekenen behoudens voor de eerste twee schooljaren. Dit beoordelingskader is daarnaast ook toepasbaar op LVS-instrumenten voor het derde tot en met het achtste leerjaar voor de niet wettelijk verplichte terreinen. Scholen mogen ook LVS-instrumenten voor het derde tot en met het achtste leerjaar inzetten die niet door het CvTE goedgekeurd zijn voor de wettelijk verplichte terreinen Nederlandse taal en rekenen, mits zij daarnaast ook door het CvTE erkende LVS-instrumenten voor het derde tot en met het achtste leerjaar gebruiken. Verder staat het scholen vrij om LVS-instrumenten van verschillende toetsaanbieders te combineren. Toetsaanbieders kunnen ook onderdelen van een LVS door het CvTE laten beoordelen. Het onderhavige beoordelingskader biedt criteria voor de kwaliteit van de toetsen behorende tot het LVS. Aanbieders van een LVS voor (ook) de eerste twee leerjaren kunnen hun instrument eveneens door het CvTE laten beoordelen.

Het CvTE kan ook onderdelen van het LVS beoordelen. In dit beoordelingskader wordt een onderscheid gemaakt in vier typen instrumenten die kunnen worden ingezet om de leerling te volgen:

- i. Observatie- en registratie-instrumenten voor de eerste twee leerjaren.
- ii. Observatie- en registratie-instrumenten voor het derde tot en met het achtste leerjaar.
- iii. LVS-instrumenten voor het derde tot en met het achtste leerjaar voor de voorgeschreven terreinen Nederlandse taal en rekenen.
- iv. LVS-instrumenten voor het derde tot en met het achtste leerjaar voor niet voorgeschreven terreinen, zoals genoemd in artikel 8, tweede lid WPO en artikel 11, derde lid, WEC. De instrumenten richten zich bijvoorbeeld op het meten van de emotionele en de verstandelijke ontwikkeling, op de ontwikkeling van creativiteit, en/of op de verwerving van noodzakelijke kennis van sociale, culturele en lichamelijke vaardigheden.

Het CvTE beoordeelt of de toets voor een periode van tien jaar kan worden erkend. De beoordeling wordt uit verschillende delen opgebouwd. De onderdelen van de beoordeling worden na elkaar uitgevoerd. Als op één van de onderdelen de toets niet voldoet aan het beoordelingskader, kan het daarop volgende onderdeel onbesproken blijven.

1.3 Is sprake van een leerlingvolgsysteem?

Een LVS bestaat uit observatie- en registratie instrumenten die in vier typen worden onderscheiden.

1.3.1 Observatie- en registratie-instrumenten - type (I) en (II)

De aanvraag voor erkenning van observatie- en registratie-instrument type (I) en (II) bevat de volgende kenmerken en bescheiden:

- Er is sprake van een observatie- en registratie-instrument als het als doel heeft de ontwikkeling van leerlingen op basis van een onderbouwde theorie (c.q. inhoud) op een objectieve manier in kaart te brengen en te volgen. Aan deze (kwalitatieve) beschrijvingen geeft ofwel de gebruiker ofwel het instrument vervolgens betekenis.
- Kenmerkend voor het volgen van de ontwikkeling van de leerling met het instrument is dat:
 - er sprake is van een aantal (twee of meer) verschillende observatiemomenten op hetzelfde construct; de afnamemomenten kunnen daarbij vast of variabel zijn;
 - er tenminste vergelijkingen met zichzelf (t.o.v. het vorige observatiemoment) worden gemaakt, en vanaf groep 3 eventueel ook met andere leerlingen.

Alvorens het CvTE een aanvraag in behandeling neemt, moet de aanvrager de navolgende gegevens en bescheiden verstrekken:

- Het instrument met bijbehorende ondersteunende materialen voor de gebruiker (c.q. leerkracht) van het instrument.
- Een handleiding voor de leerkracht met informatie over de minimaal vereiste expertise van de leerkracht, en met duidelijke instructies voor de leerkracht over het zo objectief mogelijk en professioneel uitvoeren en periodiek registreren van de observaties van de individuele leerling(en), inclusief richtlijnen voor het veilig bewaren van en omgaan met het materiaal zodra dit op de scholen aanwezig is. Een en ander aangevuld met een overzicht van veel gestelde vragen.
- De (digitale) registratieformulieren waarin de leerkracht periodiek de verzamelde observaties over een langere periode invult.
- Een voorbeeld van het (observatie)rapport waarin de observaties op overzichtelijke wijze zijn weergegeven voor de individuele leerling. Daar waar van toepassing aangevuld met voorbeelden van de (observatie)rapporten per groep en per school.
- Indien van toepassing: Cd's, dvd's en/of usbsticks en log-ins en wachtwoorden bij online instrumenten.



Als de aanvraag niet compleet is, wordt de aanvrager in gelegenheid gesteld zijn aanvraag aan te vullen. Als de aanvrager niet alsnog de informatie verstrekt die noodzakelijk is voor het behandelen van diens aanvraag, kan het CvTE besluiten de aanvraag buiten behandeling te laten.

1.3.1.1 Aanvullende voorwaarden voor observatie- en registratie-instrumenten groep 1-2 - type (I)

Verder gelden als aanvullende voorwaarden voor observatie- en registratie-instrumenten voor leerlingen in de eerste twee leerjaren (type I):

- Een toets die wordt afgenomen bij leerlingen in het eerste of tweede leerjaar, beschrijft uitsluitend de leervorderingen door de leerling te observeren. ‘Schoolse’ LVS-Toetsen voor de eerste twee leerjaren zullen daarom niet door het CvTE worden erkend. Onder een schoolse toets wordt verstaan het in een voor de kleuter onnatuurlijke setting (c.q. klassikaal aan een tafel in stilte) afnemen van een (papieren) toets uit een toetsboekje. Met behulp van concrete gedragsbeschrijvingen observeert en evalueert de leerkracht gedurende de kleuterjaren de leerling in dagelijkse situaties, met als doel het beschreven gedrag meermaals en in verschillende situaties daadwerkelijk waar te nemen en periodiek te registreren in het observatie- en registratie-instrument.
- De rapportage van observatie- en registratie-instrumenten voor leerlingen in ^{de} eerste twee leerjaren mag niet bestaan uit relatieve scores of interpretaties.
- Op basis van de rapportage en inhoudelijke gronden wordt onderbouwd dat individuele leerlingen gedurende de eerste twee leerjaren ten opzichte van zichzelf een verandering laten zien, rekening houdend met al dan niet tijdelijke ontwikkelingsgaten en -sprongen. In die zin hebben observatie- en registratie-instrumenten een signaleringsfunctie, doordat de leerkracht op de langere termijn aan de hand van de kwalitatieve beschrijvingen zicht krijgt op de voortgang van de individuele leerling. Het interpreteren van de geobserveerde gedragscriteria mag niet op basis van een relatieve normering plaatsvinden.

1.3.2 Kenmerken LVS-instrumenten - type (III) en (IV)

De aanvraag voor erkenning van LVS-instrument type (III) of (IV) bevat de volgende gegevens en bescheiden:

- Het aangeleverde informatiepakket bevat de door het CvTE voorgeschreven informatie / documentatie zijnde in ieder geval:
 - het instrument / de toets met bijbehorende ondersteunende materialen voor de leerkracht;
 - antwoordformulieren;
 - scoringsleutels;
 - een praktische handleiding voor de leerkracht inclusief richtlijnen voor het veilig bewaren van en omgaan met het materiaal zodra dit op de scholen aanwezig is;
 - een wetenschappelijk inhoudelijke verantwoording, inclusief leeswijzer;
 - voorbeeld leerlingrapport en groepsoverzichten / schoolrapportages;
 - cd's, dvd's en/of usb-sticks bij geautomatiseerde instrumenten en log-ins en wachtwoorden bij online-instrumenten.
- Aanvullend dient voor een LVS-instrument van het type (III) het aangeleverde informatiepakket eveneens te bestaan uit:
 - het toetsreglement;
 - in geval van normatieve toetsing, het vertrouwelijk delen van de normtabellen, en;
 - indien van toepassing: documentatie over geautomatiseerde scoring.

Als de aanvraag niet compleet is, wordt de aanvrager in gelegenheid gesteld zijn aanvraag aan te vullen. Als de aanvrager niet alsnog de informatie verstrekt die noodzakelijk is voor het behandelen van diens aanvraag, kan het CvTE besluiten de aanvraag buiten behandeling te laten.

1.4 Algemene criteria

Op basis van artikel 11, eerste lid, onderdeel b Toetsbesluit PO geeft het CvTE een oordeel over de volgende algemene criteria van observatie- en registratie-instrumenten en toetsinstrumenten: inhoudelijke validiteit, betrouwbaarheid en deugdelijke normering. In paragraaf 1.4.1 en 1.4.2 worden deze algemene criteria voor observatie- en registratie-instrumenten en LVS-instrumenten toegelicht.

1.4.1 Oordeel observatie- en registratie-instrumenten - type (I) en (II)

Conform het Toetsbesluit PO worden bij het kwaliteitsoordeel over observatie- en registratie-instrumenten (type I en II) in ieder geval betrokken:

- a) de wijze waarop de ontwikkeling van leerlingen op cognitief of niet-cognitief gebied met behulp van het instrument systematisch wordt geobserveerd en in het instrument wordt geregistreerd op



- één of meerdere van de onder b genoemde gebieden;
- b) de mate waarin de instrumenten de ontwikkelingsmomenten van de leerling(en) registreren op de gebieden, genoemd in artikel 8, tweede lid WPO of artikel 11, derde lid WEC: de emotionele en de verstandelijke ontwikkeling, het ontwikkelen van creativiteit, het verwerven van kennis en sociale, culturele en lichamelijke vaardigheden;
- c) de wijze waarop de ontwikkeling van leerlingen voor de ouders, voogden of verzorgers en leerkrachten inzichtelijk worden gemaakt.
- In paragraaf 1.5 wordt uiteengezet hoe deze criteria worden geoperationaliseerd.

1.4.2 Oordeel LVS-instrumenten - type (III) en (IV)

Bij het kwaliteitsoordeel over LVS-instrumenten (type III en IV) worden in ieder geval betrokken:

- a) de wijze waarop de vorderingen van leerlingen op cognitief of niet-cognitief gebied systematisch worden gemeten op de onder b genoemde gebieden;
- b) de mate waarin de instrumenten de kennis en/of vaardigheden van de leerling(en) meten op de gebieden, genoemd in artikel 8, tweede lid WPO of artikel 11, derde lid WEC: de emotionele en de verstandelijke ontwikkeling, het ontwikkelen van creativiteit, het verwerven van kennis en van sociale, culturele en lichamelijke vaardigheden;
- c) de wijze waarop de leervorderingen van leerlingen voor de ouders, voogden of verzorgers en docenten inzichtelijk worden gemaakt.

In de paragraaf 1.5 wordt uiteengezet hoe deze criteria worden geoperationaliseerd.

1.5 Operationalisering in onderwijskundige, psychometrische en organisatorische aspecten

De voorgaande algemene criteria uit paragraaf 1.4 zijn in dit beoordelingskader in de onderstaande tabellen per type instrument gespecificeerd naar de voor dat type instrument geldende kwaliteitseisen voor onderwijskundige aspecten (kwaliteitseisen O.I.1 t/m O.VI.4 in hoofdstuk 2), psychometrische aspecten (kwaliteitseisen P.I.1 t/m P.VIII.3 in hoofdstuk 3) en organisatorische aspecten (kwaliteitseisen B.I.1 t/m B.I.6 in hoofdstuk 4).

De onderwijskundige aspecten richten zich op de inhoudsvaliditeit en de eerder genoemde punten b) en c) in de paragrafen 1.4.1 en 1.4.2. De psychometrische aspecten richten zich op de in paragraaf 3.1 genoemde betrouwbaarheid en deugdelijke normering en het eerder genoemde punt a) in de paragrafen 1.4.1 en 1.4.2. De organisatorische aspecten richten zich op de criteria voor veilig toetsen.

Per kwaliteitseis is per type instrument met zwarte balletjes aangegeven wanneer de eis altijd beantwoord en beoordeeld moet worden. Doorzichtige balletjes betekenen dat de kwaliteitseis ook niet van toepassing kan zijn. Wanneer dat geldt, wordt beschreven bij de kwaliteitseisen in de betreffende paragrafen. Voor wat betreft de psychometrische kwaliteitseisen voor instrument type II heeft de aanbieder de keuzemogelijkheid om de psychometrische kwaliteitseisen voor instrument type I te volgen, de psychometrische kwaliteitseisen voor instrument type III en IV te volgen, of om een combinatie van de psychometrische kwaliteitseisen voor instrument type I en voor instrument type III en IV te volgen. Welke combinaties mogelijk zijn, wordt eveneens in onderstaande tabellen aangegeven.

Onderwijskundige aspecten

Kwaliteitseis O.I Paragraaf 2.2.1	Uitgangspunten van de constructie van het observatie- en registratie-instrument, met specifieke aandacht voor de kwaliteitseis inhoudsvaliditeit.			
	Type I	Type II	Type III	Type IV
O.I.1	•	•		
O.I.2	•	•		
O.I.3	○	○		
O.I.4	○	○		
O.I.5	•	•		
O.I.6	•	•		
Kwaliteitseis O.II Paragraaf 2.2.2	Inzicht in de individuele ontwikkeling, met specifieke aandacht voor de kwaliteitseisen signaleringsfunctie en/of volgaspect.			
O.II.1	•	•		
O.II.2	○	○		
O.II.3	○	○		
O.II.4	•	•		
Kwaliteitseis O.III Paragraaf 2.3.1	Toetsconstructie			



Kwaliteitseis O.I Paragraaf 2.2.1	Uitgangspunten van de constructie van het observatie- en registratie-instrument, met specifieke aandacht voor de kwaliteitseis inhoudvaliditeit.			
	Type I	Type II	Type III	Type IV
O.III.1			•	•
O.III.2			•	•
O.III.3			•	•
O.III.4			•	○
O.III.5			•	•
Kwaliteitseis O.IV Paragraaf 2.3.2	Inhoudvaliditeit			
O.IV.1			•	•
O.IV.2			•	•
O.IV.3			•	•
Kwaliteitseis O.V Paragraaf 2.3.3	Referentieniveaus			
O.V.1			•	
O.V.1a			○	
O.V.1b			○	
Kwaliteitseis O.VI Paragraaf 2.3.4	Inzicht in de leervorderingen			
O.VI.1			•	•
O.VI.2			•	•
O.VI.3			•	•
O.VI.4			•	•

Psychometrische aspecten

Kwaliteitseis P.I Paragraaf 3.2.1	Dataverzameling, met specifieke aandacht voor de kwaliteitseis steekproef van leerlingen.			
	Type I	Type II	Type III	Type IV
P.I.1	○	○ en/of P.IV.1 t/m P.IV.2		
Kwaliteitseis P.II Paragraaf 3.2.2	IJking, met specifieke aandacht voor de kwaliteitseis niet relatieve normering.			
P.II.1	○	○ en/of P.V.1 of P.V.4		
P.II.2	○	○ en/of P.V.2 of P.V.5		
P.II.3	○	○ en/of P.V.3 of P.V.6		
Kwaliteitseis P.III Paragraaf 3.2.3	Interbeoordelaarsbetrouwbaarheid, met specifieke aandacht voor de kwaliteitseis betrouwbaarheid van de observaties.			
P.III.1	○	○ en/of P.VI.1 t/m P.VI.3		
P.III.2	○	○		
Kwaliteitseis P.IV Paragraaf 3.3.1	De kwaliteit van de dataverzameling			
P.IV.1		○ en/of P.I.1	•	•
P.IV.2		○ en/of P.I.1	○	○
Kwaliteitseis P.V Paragraaf 3.3.2	Normering			
P.V.1		○ en/of P.II.1	○	○
P.V.2		○ en/of P.II.2	○	○
P.V.3		○ en/of P.II.3	○	○
P.V.4		○ en/of P.II.1	○	○
P.V.5		○ en/of P.II.2	○	○
P.V.6		○ en/of P.II.3	○	○
Kwaliteitseis P.VI Paragraaf 3.3.3	Betrouwbaarheid			
P.VI.1		○ en/of P.II.1 t/m P.II.2	•	•
P.VI.2		○ en/of P.II.1 t/m P.II.2	•	•
P.VI.3		○ en/of P.II.1 t/m P.II.2	•	•
Kwaliteitseis P.VII Paragraaf 3.3.4	Constructvaliditeit			
P.VII.1		○	○	○
P.VII.2		○	○	○
P.VII.3		○	•	•



Kwaliteitseis P.I Paragraaf 3.2.1	Dataverzameling, met specifieke aandacht voor de kwaliteitseis steekproef van leerlingen.			
	Type I	Type II	Type III	Type IV
P.VII.4		○	○	○
P.VII.5		○	•	•
Kwaliteitseis P.V.III Paragraaf 3.3.5	Volgaspect of signaleringsfunctie			
P.VIII.1		○	•	•
P.VIII.2		○	•	•
P.VIII.3		○	•	•

Organisatorische aspecten

Kwaliteitseis B.I Paragraaf 4.1	Afname en organisatorische aspecten			
	Type I	Type II	Type III	Type IV
B.I.1	•	•	•	•
B.I.2	•	•	•	•
B.I.3	•	•	•	•
B.I.4	•	•	•	•
B.I.5	•	•	•	•
B.I.6			•	•

1.6 Wijze van beoordelen

De aanbieder kan het observatie- en registratie-instrument voor het volgen van de leerling indienen bij het CvTE. De adviseur heeft de taak om het CvTE te adviseren over de onderwijskundige, psychometrische en organisatorische aspecten van de LVS-instrumenten en de observatie- en registratie-instrumenten. De procedure voor het beoordelen van de instrumenten en het bijbehorende tijdspad zijn gepubliceerd op de website van het CvTE.

1.6.1 Beslisregel om te komen tot een kwaliteitsoordeel

De onderwijskundige aspecten, psychometrische aspecten en organisatorische aspecten uit paragraaf 1.5 zijn alle drie geconcretiseerd in kwaliteitseisen in vraagvorm inclusief bijbehorende codes. De vragen van de separate kwaliteitseisen zijn te beantwoorden met 'ja', 'nee' of soms met 'niet van toepassing'. De beslisregel is als volgt: Om een aspect met een voldoende te kunnen afsluiten, dienen alle vragen van de kwaliteitseisen van het betreffende aspect met 'ja' of 'niet van toepassing' te worden beantwoord.

Op basis van de beslisregels worden de drie aspecten uit paragraaf 1.5 separaat gescoord met:

- voldoende;
- onvoldoende.

Bij 'onvoldoende' geeft het CvTE aan welke kwaliteitseisen per aspect nog niet voldoende zijn.

Volgens artikel 3a, derde lid, Wet CvTE besluit het College binnen 15 weken.

1.7 Initiële beoordeling en tussentijdse check

1.7.1 Initiële beoordeling

De initiële beoordeling door het CvTE is een uitgebreide beoordeling van de onderwijskundige, psychometrische en organisatorische aspecten van het instrument. De erkenning van de toetsen is vanaf de datum van afgifte tien jaar geldig.

1.7.2 tussentijdse check

De erkenning door het CvTE van een LVS-instrument is tien jaar geldig. In tien jaar kunnen er echter veel ontwikkelingen plaatsvinden die de kwaliteit van een LVS-instrument beïnvloeden. Daarom vinden er één of meerdere tussentijdse beoordelingen plaats. De aanbieder is gehouden de informatie aan het Cvte te verstrekken die het nodig heeft om een tussentijdse beoordeling uit te voeren.

1. Geen wijzigingen

Voor een instrument dat in tien jaar niet verandert, vindt er vijf jaar na afgifte van een kwaliteitsoordeel een tussentijdse check plaats.



2. **Periodieke wijzigingen**
Als de inhoud van de toets periodiek wezenlijk verandert en dit in de verantwoording beschreven staat, maakt het CvTE bij de initiële beoordeling met de aanbieder afspraken over de frequentie en omvang van tussentijdse beoordelingen. Dit zou bijvoorbeeld bij jaarlijkse verversing van items in een itembank het geval kunnen zijn.
3. **Incidentele wijzigingen**
Als er incidenteel fundamentele wijzigingen worden voorzien, zoals het vervangen van enkele items, het wijzigen van de afnamevorm (digitaal in plaats van papier) of het wijzigen van de normering (bijvoorbeeld het type normering of een wijziging van de schaal) meldt de aanbieder dit bij het CvTE. Over de relevantie en impact van de veranderingen en de noodzaak van een tussentijdse check, vindt overleg plaats tussen het CvTE en de aanbieder. Afhankelijk van de wijzigingen vindt er al dan niet een extra tussentijdse beoordeling plaats of dient de aanbieder de instrumenten (deels) opnieuw in.

De beoordeling richt zich vooral op de items en observatiecategorieën. De LVS-ontwikkelaars dienen inzicht te geven in de volgende twee vragen:

1. **Zijn de items of de observatie-categorieën nog steeds actueel?**
 - Is het taalgebruik in de opgaven verouderd of niet langer passend?
 - Zijn er veranderingen in het onderwijsaanbod geweest (bijv. wijzigingen in het curriculum en dus in methodes, toetswijzer en referentiekader)?
 - Zijn er maatschappelijke ontwikkelingen geweest (bijv. verandering munteenheid of nieuw ontstane gevoeligheden)?
 2. **Is de normering nog steeds passend en actueel?**
 - Functioneren de items of observatiecategorieën nog steeds hetzelfde als bij eerdere indiening? Dat wil zeggen, zijn de psychometrische eigenschappen (p-waarden of IRT itemparameters, en de relatie tussen items of observatiecategorieën en het totale instrument, zoals IRT discriminatie waarden, en/of item/testcorrelaties) nog vergelijkbaar?
 - Zijn items niet te zeer bekend geworden of hebben de onder punt 1 aangehaalde problemen de psychometrische eigenschappen van de items of observatiecategorieën veranderd?
- Een tussentijdse check kan resulteren in het oordeel 'voldoende of 'onvoldoende. Bij een oordeel 'onvoldoende' volgt intrekking van de erkenning.

1.8 Leeswijzer

De hoofdstukken in dit kader behandelen respectievelijk de onderwijskundige aspecten (Hoofdstuk 2), de psychometrische aspecten (Hoofdstuk 3) en de organisatorische aspecten (Hoofdstuk 4). Zoals beschreven in paragraaf 1.5 zijn de aspecten geoperationaliseerd in verplichte of optionele kwaliteitseisen per type instrument. Daarnaast kunnen, afhankelijk van het type instrument, kwaliteitseisen ook niet van toepassing zijn. Een en ander is gevisualiseerd in de tabellen in paragraaf 1.5.

2. Onderwijskundige aspecten

2.1. Inleiding

De beoordeling van de onderwijskundige aspecten van een LVS-instrument richt zich op de in paragraaf 1.4 genoemde inhoudsvaliditeit.

De kwaliteitseisen onderwijskundige aspecten voor observatie- en registratie-instrumenten type (I) en (II) zijn te vinden in de paragrafen 2.2.1 en 2.2.2. De kwaliteitseisen onderwijskundige aspecten voor LVS-instrumenten type (III) en (IV) zijn te vinden in de paragrafen 2.3.1 tot en met 2.3.4.

2.2 Kwaliteitseisen onderwijskundige aspecten observatie- en registratie-instrumenten - type (I) en (II)

2.2.1 Uitgangspunten van de constructie

Deze paragraaf beschrijft de kwaliteitseisen voor de uitgangspunten van de constructie van het observatie- en registratie-instrument, met specifieke aandacht voor de kwaliteitseis *inhoudsvaliditeit*.

Code	Vragen	Mogelijke antwoorden
O.1.1	Is er een bij de inhoud van het instrument passende, gedegen en complete verantwoording, indien nodig inclusief leeswijzer waarin is onderbouwd hoe de visie van de aanbieder op de ontwikkeling van het (jonge) kind is geborgd in de actuele pedagogische en/of psychologische wetenschappelijke literatuur?	ja/nee



Code	Vragen	Mogelijke antwoorden
O.1.2	Is er een actuele inhoudelijke verantwoording waarin gedegen en compleet is onderbouwd hoe de visie van de aanbieder op de ontwikkeling van het (jonge) kind is geoperationaliseerd in te observeren ontwikkelingsdoelen?	ja/nee
O.1.3	Is er een actuele gedegen en gedetailleerde onderbouwing van het ontwikkelingsverloop, verbijzonderd in bijpassende indicatoren van observeerbare (gedrags)kenmerken en met een verwijzing naar relevante brondocumenten?	ja/nee/n.v.t.
O.1.4	Voldoet de formulering van de indicatoren van observeerbare (gedrags)kenmerken in het instrument aan de kwaliteitscriteria voor observatie-categorieën: relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit?	ja/nee/n.v.t.
O.1.5	Is de doelgroep van het instrument gedegen en compleet gedefinieerd?	ja/nee
O.1.6	Is het gebruiksdoel van het instrument gedegen en compleet gedefinieerd?	ja/nee

Toelichting bij O.1.1:

De aanbieder beschrijft ten minste op gedegen en complete wijze zijn visie op de ontwikkeling van het (jonge) kind en geeft hiervoor een deugdelijke onderbouwing, indien nodig inclusief leeswijzer, bijvoorbeeld door middel van referenties naar de actuele pedagogische en/of psychologische wetenschappelijke literatuur.

Toelichting bij O.1.2:

Met de actuele inhoudelijke verantwoording is er sprake van een (kwalitatieve) verantwoording van de representativiteit van de ontwikkelingsdoelen voor de onderbouwde theorie (c.q. inhoud) van het instrument. Daarmee wordt voldaan aan de eis voor inhoudsvaliditeit, zijnde de eigenschap dat de operationalisering van de inhoud van het instrument een representatieve weergave is van de te observeren en te volgen ontwikkeling van het (jonge) kind.

Toelichting bij O.1.3:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren (c.q. observatie-categorieën) van observeerbare (gedrags)kenmerken, wordt er een actuele gedegen en gedetailleerde onderbouwing van het ontwikkelingsverloop van de ontwikkelingsdoelen en eventuele tussentijdse mijlpalen gegeven.

Als voorbeeld: er is een actuele gedegen en gedetailleerde beschrijving van het ontwikkelingsverloop van taal, met een verbijzondering in het domein Mondelinge taalvaardigheid. De inhoud van dit domein wordt gedegen en in detail beschreven in bijvoorbeeld de mijlpalen Gesprekken voeren en Geletterdheid. De mijlpalen worden daarbij geoperationaliseerd in indicatoren van observeerbare (gedrags)kenmerken (bijvoorbeeld: de leerling begrijpt dat lezen van links naar rechts gaat) op basis waarvan de leerkracht zijn of haar observaties over een zelf gekozen dan wel over een door het instrument voorgeschreven periode kan registreren in het instrument. Ter inspiratie kan bijvoorbeeld gekeken worden naar de inhoudskaarten met aanbodsdoelen voor het jonge kind van Expertisecentrum SLO, zoals opgesteld voor de verschillende domeinen uit de WPO.

Toelichting bij O.1.4:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags)kenmerken, voldoet de formulering van deze indicatoren aan de kwaliteitscriteria voor observatie-categorieën: relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit, een en ander zoals gespecificeerd in de Checklist voor het beoordelen van de kwaliteit van observatie-categorieën en toetsopgaven, versie 2022.1006cl.

Toelichting bij O.1.5:

De definitie van de doelgroep van het instrument betreft het noemen van de leerjaren waarin het instrument ingezet kan worden. Indien van toepassing dient de aanbieder in aanvulling hierop aan te geven in welke bijzondere situaties(s) en/of bij welke individuele ondersteuningsbehoeften van een leerling het instrument niet geschikt is.

Toelichting O.1.6:

Er is in de verantwoording aangegeven dat het gebruiksdoel het observeren, registreren en, indien van toepassing, signaleren van de individuele ontwikkeling van het (jonge) kind betreft, zoals gedefinieerd in de doelgroep en een en ander conform de in hoofdstuk 1 genoemde bepalingen.

2.2.2 Inzicht in de individuele ontwikkeling

Deze paragraaf beschrijft de kwaliteitseisen betreffende het inzicht in de individuele ontwikkeling, met specifieke aandacht voor de kwaliteitseisen *signaleringsfunctie* en/of *volgaspect*

Code	Vragen	Mogelijke antwoorden
O.II.1	Stelt het instrument voor iedere (jonge) leerling een rapport samen met daarin een gedegen en compleet overzicht van de in ieder geval vaker dan één keer door de leerkracht geregistreeerde observaties?	ja/nee
O.II.2	Bevat het individuele rapport per (jonge) leerling per registratiemoment de bij diens door de leerkracht geobserveerde indicatoren?	ja/nee/n.v.t.
O.II.3	Bevat het individuele rapport een gedegen en complete toelichting over hoe de geregistreeerde indicatoren geïnterpreteerd dienen te worden, zodat de ontwikkeling van de (jonge) leerling (ook) voor ouders, verzorgers, voogden en leerkrachten begrijpelijk is?	ja/nee/n.v.t.
O.II.4	Biedt de informatie in het individuele rapport een eenduidige en gedegen basis voor de leerkracht om de ontwikkeling te evalueren en, waar nodig, door zichzelf gekozen vervolgstappen te formuleren?	ja/nee

Toelichting O.II.1:

Er is bij een volgaspect altijd sprake van een aantal (twee of meer) verschillende observatiemomenten op hetzelfde construct; de afnamemomenten kunnen daarbij vast of variabel zijn.

Toelichting O.II.2:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, bevat het rapport per (jonge) leerling per registratiemoment de wel (en nog niet of niet langer) bij hem of haar door de leerkracht geobserveerde indicatoren van de verbijzonderde ontwikkelingsdoelen. Het betreft hier een neutrale registratie van de observatie, zonder waardeoordeel. Zo geeft het rapport een overzichtelijk beeld van de individuele ontwikkeling in de tijd zowel op de verschillende ontwikkelingsdoelen afzonderlijk als in samenhang met elkaar voor wat betreft de ontwikkelingsdoelen per construct, waarbij geen sprake is van een relatieve normering van de individuele (jonge) leerling. Het rapport maakt het op deze manier voor de leerkracht mogelijk om observaties van een (jonge) leerling op verschillende momenten en in verschillende situaties met elkaar te kunnen vergelijken. Daarmee heeft het instrument tevens een signaleringsfunctie in de zin dat het instrument de eigen ontwikkeling van de individuele (jonge) leerling weergeeft.

Toelichting O.II.3:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags)kenmerken, bevat het individuele rapport een toelichting over hoe de geregistreeerde indicatoren geïnterpreteerd dienen te worden, waarmee het instrument begrijpelijke handvatten geeft voor de interpretatie van het verloop van de ontwikkeling (groei en/of stagnatie) van de individuele (jonge) leerling door de diverse betrokkenen, zoals ouders, verzorgers, voogden en leerkrachten.

Toelichting O.II.4

De informatie in het individuele rapport dient dusdanig concreet en gedetailleerd te zijn, dat het duidelijk is welk verloop in de ontwikkeling van de (jonge) leerling aandacht verdient. Het ondersteunt de leerkracht in het zelfstandig en op basis van de eigen professionaliteit formuleren van passende vervolgstappen in lijn met de door de aanbieder van het instrument geoperationaliseerde en bij de visie passende ontwikkelingsdoelen en eventuele tussentijdse mijlpalen.

2.3 Kwaliteitseisen onderwijskundige aspecten LVS-instrumenten - type (III) en (IV)

2.3.1 Uitgangspunten van de toetsconstructie

Deze paragraaf beschrijft de kwaliteitseisen voor de uitgangspunten van de toetsconstructie.

Code	Vragen	Mogelijke antwoorden
O.III.1	Is de afnamevorm van het instrument gedegen en compleet gedefinieerd?	ja/nee
O.III.2	Is (zijn) de doelgroep(en) van het instrument gedegen en compleet gedefinieerd?	ja/nee
O.III.3	Is het meetdoel van het instrument gedegen en compleet gedefinieerd?	ja/nee
O.III.4	Is er een actuele gedegen en gedetailleerde inhoudelijke theoretische inkadering van het instrument?	ja/nee/n.v.t.
O.III.5	Is het gebruiksdoel van het instrument gedegen en compleet gedefinieerd?	ja/nee

Toelichting O.III.1:

In de handleiding van het instrument is op gedegen en complete wijze aangegeven voor welk afnamevorm(en), papier of digitaal, het instrument geschikt is.



Toelichting O.III.2:

De definitie van de doelgroep(en) van het instrument betreft het noemen van de leerjaren waarin het instrument kan worden ingezet. Daarnaast is er ook relevante informatie gegeven over het vereiste niveau van de leerlingen en is er vermeld in welke situatie(s) en/of bij welke individuele ondersteuningsbehoeften het instrument niet geschikt is. In plaats van genoemde toelichting mag de toetsaanbieder er ook voor kiezen om expliciet te benoemen voor welke doelgroep(en) het instrument *niet* geschikt is.

Toelichting O.III.3:

De definitie van het meetdoel betreft het benoemen van wat de leerling op het moment van inzetten van het instrument dient te beheersen. Hieraan is toegevoegd een gedetailleerde beschrijving van domeinen en onderdelen, een toetsmatris en/of een verwijzing naar actuele en relevante en brondocumenten.

Toelichting O.III.4:

Er is, indien van toepassing op het type instrument, aangegeven hoe de inhoud van het instrument aansluit bij het theoretische kader. En er is aangegeven op welke (analyse)methoden de indeling van de vaardigheden is gebaseerd.

Toelichting O.III.5:

De definitie van het gebruiksdoel bestaat uit een gedegen en complete beschrijving van de combinatie van niveaubepaling en inzicht in de leervorderingen.

2.3.2 Inhoudsvaliditeit

Deze paragraaf beschrijft de kwaliteitseisen voor de inhoudsvaliditeit/ representativiteit van het LVS-instrument.

Code	Vragen	Mogelijke antwoorden
O.IV.1	Is de toetsmatris een adequate, schematische representatie van het meetdoel?	ja/nee
O.IV.2	Voldoen de toetsopgaven in het instrument aan de kwaliteitscriteria voor toetsopgaven: relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit?	ja/nee
O.IV.3	Voldoen de toetsopgaven in het instrument aan de constructievoorschriften voor gesloten vragen, open vragen en/of nieuwe vraagtypes?	ja/nee

Toelichting O.IV.1:

De toetsmatris is een adequate, schematische representatie van het meetdoel. Er is sprake van een adequate, schematische representatie wanneer de eind- en toetstermen het meetdoel representeren. Dit blijkt uit het gegeven dat:

- de toetstermen concreet en eenduidig zijn omschreven en, voor wat betreft instrument type (III) vanaf groep 6, aansluiten bij de referentieniveaus voor de onderdelen Nederlandse taal en rekenen;
- de toetsmatris in ieder geval bevat:
 1. het aantal vragen met bijbehorende scorepunten;
 2. de toetsvorm en/of het type vragen;
 3. de verdeling van de verschillende vraag- en teksttypes in de toets;
 4. de verdeling van de toetsopgaven over de onderdelen en domeinen in de toets;
 5. de toegestane hulpmiddelen;
 6. en in het geval van een niet adaptieve toets: (7) de toetsduur.

Toelichting O.IV.2:

De adviseur evalueert in opdracht van het CvTE de relevantie van de inhoud van de toets en de inhoudsvaliditeit van alle papieren toetsopgaven dan wel alle digitale items uit de itembank op de vijf kwaliteitscriteria relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit, zoals beschreven in de Checklist voor het beoordelen van de kwaliteit van observatie-categorieën en toetsopgaven, versie 2022.1006cl.

Toelichting bij O.IV.3:

De auditor van de adviseur evalueert in opdracht van het CvTE of de toetsopgaven in de toets voldoen aan de constructievoorschriften voor toetsvragen, zoals beschreven in de Checklist voor het beoordelen van de kwaliteit van observatie-categorieën en toetsopgaven, versie 2022.1006cl.



2.3.3 Referentieniveaus (deze paragraaf is alleen van toepassing op type (III) instrumenten)

Deze paragraaf beschrijft ten behoeve van instrumenten van het type (III) de kwaliteitseisen voor de referentieniveaus.

LVS-instrumenten van het type (III) maken onderdeel uit van de doorlopende, methodevrije leerlijn richting de doorstroomtoets in groep 8. Omdat de doorstroomtoetsen zijn gekoppeld aan referentieniveaus, is het belangrijk dat ook de LVS-instrumenten van het type (III) vanaf groep 6 hierbij aansluiten. Ongeacht de toetsvorm, moet het instrument het mogelijk maken om voor elke leerling vanaf groep 6 uitspraken te doen over de beheersing van het 1F niveau en het 2F/1S niveau. Er is dan inzicht in de vorderingen van de leerling richting het eindniveau.

Code	Vragen	Mogelijke antwoorden
O.V.1	Bevat de verzameling van LVS-instrumenten van het type (III) van groep 6-8 voor wat betreft de wettelijk verplichte voorgeschreven terreinen Nederlandse taal en rekenen én de optionele domeinen van taal, de inhouden van 1F en 2F/1S?	ja/nee
O.V.1a	Indien er sprake is van een CAT (adaptieve toets op itemniveau), voldoet de gehele itembank van de verzameling van LVS-instrumenten van het type (III) van groep 6-8 voor wat betreft de bij O.V.1 genoemde domeinen aan de kwaliteitseis O.V.1?	ja/nee/n.v.t.
O.V.1b	Indien er sprake is van een MST (adaptieve toets op moduleniveau), voldoet het gehele design van de verzameling van LVS-instrumenten van het type (III) van groep 6 tot en met 8 voor wat betreft de bij O.V.1 genoemde domeinen aan de kwaliteitseis O.V.1?	ja/nee/n.v.t.

Toelichting O.V.1:

Dit aspect moet uit de toetsmatrijs naar voren komen. Dat wil zeggen dat de verhouding 1F en 2F/1S vragen in de toetsmatrijs is gespecificeerd en dat de verhouding van de vragen erin wordt benoemd.

Toelichting O.V.1a:

Indien er sprake is van een CAT, dient voor elke individuele leerling voor de onderdelen taal en rekenen in de periode groep 6-8 een uitspraak over het behaalde niveau 1F of 2F/1S te worden gedaan.

Toelichting O.V.1b

Indien er sprake is van een MST, dient elke routing binnen het MST design voor elke individuele leerling voor de onderdelen taal en rekenen in de periode groep 6-8 een uitspraak over het behaalde niveau 1F of 2F/1S te doen.

2.3.4 Inzicht in de leervorderingen

Deze paragraaf beschrijft de kwaliteitseisen betreffende het inzicht in de leervorderingen.

Code	Vragen	Mogelijke antwoorden
O.VI.1	Stelt de toetsaanbieder voor elke individuele leerling een leerlingrapport op?	ja/nee
O.VI.2	Bevat het leerlingrapport het gemeten niveau van de leerling en geeft het rapport de leervordering van de leerling weer?	ja/nee
O.VI.3	Bevat het leerlingrapport een toelichting bij het gemeten niveau en de leervorderingen van de leerling die (ook) voor ouders, verzorgers, voogden en docenten begrijpelijk is?	ja/nee
O.VI.4	Is de informatie in het leerlingrapport een goede basis voor de leerkracht om de leervorderingen te evalueren en vervolgstappen te formuleren?	ja/nee

Toelichting O.VI.1:

Deze kwaliteitseis behoeft geen nadere toelichting.

Toelichting O.VI.2:

In het leerlingrapport van een LVS-instrument van het type (III) vanaf groep 6 hoeft het referentieniveau niet verplicht te worden vermeld. Wel kan de toetsaanbieder ervoor kiezen om een vaardigheidschaal te construeren aan de hand van de LVS-instrumenten, op basis waarvan er een uitspraak kan worden afgeleid over de weg richting de referentieniveaus.

Toelichting O.VI.3:

Er dient een geschreven toelichting te worden geboden, waarin begrijpelijke handvatten gegeven worden voor de interpretatie van de leervorderingen van de leerling door diverse betrokkenen.

Toelichting O.VI.4:

De informatie moet dusdanig concreet en gedetailleerd zijn dat het duidelijk is welke lacunes in de



ontwikkeling van de leerling in het onderwijs aandacht verdienen.

3. Psychometrische aspecten

3.1 Inleiding

De beoordeling van de psychometrische aspecten van een LVS-instrument richt zich op de in paragraaf 1.4 genoemde betrouwbaarheid en deugdelijke normering, de (construct)validiteit en op het in paragraaf 1.4.1 en 1.4.2 genoemde punt a) betreffende de wijze waarop de vorderingen van leerlingen op cognitief of niet-cognitief gebied systematisch worden gemeten.

De (optionele) kwaliteitseisen psychometrische aspecten voor observatie- en registratie-instrumenten type (I) zijn te vinden in de paragrafen 3.2.1 en 3.2.2. De kwaliteitseisen psychometrische aspecten voor LVS-instrumenten type (III) en (IV) zijn te vinden in de paragrafen 3.3.1 tot en met 3.3.5.

Voor een observatie- en registratie-instrument van het type (II) heeft de aanbieder de keuzemogelijkheid om de psychometrische kwaliteitseisen voor instrument type I te volgen, de psychometrische kwaliteitseisen voor instrument type III en IV te volgen, of om een combinatie van de psychometrische kwaliteitseisen voor instrument type I en voor instrument type III en IV te volgen. Welke combinaties mogelijk zijn, is in de tabellen in paragraaf 1.5 aangegeven.

3.2 Kwaliteitseisen psychometrische aspecten observatie- en registratie-instrumenten - type (I) en (II)

De beoordeling van de psychometrische aspecten van een observatie- en registratie-instrument type (I) richt zich op de dataverzameling (paragraaf 3.2.1), de onderbouwing van de kwaliteit van de resultaten / uitspraken van het instrument (paragraaf 3.2.2) en op de interbeoordelaarsbetrouwbaarheid (paragraaf 3.2.3).

3.2.1 Dataverzameling

Deze paragraaf beschrijft de kwaliteitseisen voor de dataverzameling, met specifieke aandacht voor de kwaliteitseis *steekproef van leerlingen*.

Deze paragraaf 3.2.1 en de hierin opgenomen kwaliteitseis P.I.1 is van toepassing op een observatie- en registratie-instrument type (I), indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, zoals verantwoord in de kwaliteitseisen O.I.3 en O.I.4 in paragraaf 2.2.1.

Code	Vragen	Mogelijke antwoorden
P.I.1	Beschrijft de aanbieder een representatieve steekproef van de indicatoren van observeerbare (gedrags-)kenmerken (c.q. de observatie-categorieën) op basis waarvan de steekproefresultaten kunnen worden gegeneraliseerd naar de doelgroep van het instrument, zoals gedefinieerd bij kwaliteitseis O.I.5?	ja/nee/n.v.t.

Toelichting P.I.1:

Indien van toepassing dient de aanbieder te verantwoorden dat de steekproef groot genoeg is, adequaat is gestratificeerd naar betekenisvolle achtergrondvariabelen, en is verzameld onder omstandigheden die redelijk vergelijkbaar zijn met de omstandigheden waarin de gegevens in een operationele setting zouden worden verzameld.

3.2.2 IJking van het instrument¹

Deze paragraaf beschrijft de kwaliteitseisen voor de ijking van het instrument, met specifieke aandacht voor de kwaliteitseis *niet relatieve normering*.

Deze paragraaf 3.2.2 en de hierin opgenomen kwaliteitseisen P.II.1 tot en met P.II.3 zijn van toepassing op een observatie- en registratie-instrument type (I), indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, zoals verantwoord in de kwaliteitseisen O.I.3 en O.I.4 in paragraaf 2.2.1.

¹ In het geval van een instrument van het type (II) of (III) wordt gesproken over ijken. In het geval van een instrument van het type (III) of (IV) wordt in dit beoordelingskader de term normeren gehanteerd.



Het volgen van de ontwikkeling van de (jonge) leerling met betrekking tot de verschillende ontwikkelingsdoelen en eventuele tussentijdse mijlpalen heeft nadrukkelijk niet als doel om waardeoordeel over het niveau van de (jonge) leerling in relatie tot het niveau van andere leerlingen uit te spreken. Het doel is om de leerkracht van informatie te voorzien waarmee hij of zij zelfstandig en op basis van de eigen professionaliteit de specifieke ontwikkelbehoefte van de individuele (jonge) leerling kan signaleren en desgewenst een voor die specifieke leerling passende interventie kan doen.

Om de gebruiker te kunnen ondersteunen in het bepalen van de mate waarin een individuele (jonge) leerling zich op zeker moment een specifiek ontwikkelingsdoel eigen heeft gemaakt, dienen de indicatoren van observeerbare (gedrags-)kenmerken van de ontwikkelingsdoelen en eventuele tussentijdse mijlpalen in het instrument te worden voorzien van ijkpunten (c.q. referentiepunten). Dit noemt men ook wel het normeren van het instrument.

De leerkracht kan met behulp van het instrument op basis van observaties vaststellen en volgen hoe individuele leerlingen gedurende de kleuterjaren zich ontwikkelen in het aantal en soort bij hem of haar geobserveerde indicatoren van (gedrags-)kenmerken. De ijkpunten van de indicatoren geven daarbij een objectiever beeld van de individuele ontwikkeling, zonder dat de leerkracht een waardeoordeel over het niveau van de leerling in relatie tot het niveau van andere leerlingen hoeft uit te spreken.

De ijkpunten worden bij de initiële ontwikkeling van het instrument door experts bepaald en vervolgens periodiek geëvalueerd. Het iken vindt plaats op grond van de inhoud van het instrument, zoals verantwoord bij de kwaliteitseisen O.I.1 tot en met O.I.4 uit paragraaf 2.2.1 en derhalve niet op grond van de relatieve positie van het ene kind ten opzichte van andere kinderen. Er zijn verschillende methodes om een instrument te iken, bijvoorbeeld de in de literatuur beschreven standaardbepalingsmethoden. Deze methoden hebben gemeenschappelijk dat externe experts worden geraadpleegd en dat deze experts op inhoudelijke en/of op kwantitatieve gronden tot voldoende overeenstemming moeten komen over de te bepalen ijkpunten van de indicatoren van observeerbare (gedrags-)kenmerken van de ontwikkelingsdoelen en eventuele tussentijdse mijlpalen in het instrument.

Code	Vragen	Mogelijke antwoorden
P.II.1	Is de keuze voor de ijkingsmethode gedegen gemotiveerd en op de juiste wijze uitgevoerd?	ja/nee/n.v.t.
P.II.2	Zijn de experts met inhoudelijke vakdeskundigheid die de ijkpunten van de indicatoren van observeerbare (gedrags-)kenmerken van de ontwikkelingsdoelen en eventuele tussentijdse mijlpalen vaststellen naar behoren geselecteerd en getraind?	ja/nee/n.v.t.
P.II.3	Is er voldoende overeenstemming tussen de experts?	ja/nee/n.v.t.

Toelichting P.II.1:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, blijkt uit de verstrekte informatie dat er gebruik is gemaakt van een beproefde ijkingsmethode.

In dat geval blijkt uit de verstrekte informatie ook waarom de gebruikte methode past bij het doel.

Toelichting P.II.2:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, is er een onderbouwing van de selectieprocedure, heeft er een selectie van voldoende experts voor de gekozen methode met aantoonbaar inhoudelijke vakdeskundigheid plaatsgevonden en zijn deze experts getraind in de betreffende ijkingsmethode.

Toelichting P.II.3:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, toont de aanbieder aan dat de mate van overeenstemming tussen de experts resulteert in een coëfficiënt voor beoordelaarsovereenstemming waarbij de proportie overeenstemming representatief is voor een instrument voor minder belangrijke beslissingen (een zogenaamd medium stakes instrument) op individueel niveau.

3.2.3 Interbeoordelaarsbetrouwbaarheid

Deze paragraaf beschrijft de kwaliteitseisen voor de interbeoordelaarsbetrouwbaarheid, met specifieke aandacht voor de kwaliteitseis *betrouwbaarheid van de observaties*.

Deze paragraaf 3.2.3 en de hierin opgenomen kwaliteitseisen P.III.1 en P.III.2 zijn van toepassing op een observatie- en registratie-instrument type (I), indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags)kenmerken, zoals

verantwoord in de kwaliteitseisen O.I.3 en O.I.4 in paragraaf 2.2.1.

Bij observatie- en registratie-instrumenten van het type (I) die van in het instrument getrainde observatoren / leerkrachten (holistische) oordelen van de gedragingen van jonge kinderen in de dagelijkse praktijk verlangen, dient te worden aangetoond dat de betreffende observatoren / leerkrachten deze gedragingen eenduidig kunnen interpreteren en correct in het instrument kunnen registreren. De mate waarin op deze manier betrouwbaar met het instrument wordt gewerkt, dient de aanbieder aan te tonen, zowel na de initiële ontwikkeling van het instrument als periodiek tijdens de tussentijdse check (zie paragraaf 1.7.2) van het instrument.

Code	Vragen	Mogelijke antwoorden
P.III.1	Wordt er met behulp van stelselmatig onderzoek aangetoond dat de in het instrument getrainde observatoren / leerkrachten de indicatoren van observeerbare (gedrags-)kenmerken (c.q. de observatie-categorieën) eenduidig kunnen interpreteren en registreren?	ja/nee/n.v.t.
P.III.2	Zijn de observatoren / leerkrachten die zijn ingezet om de interbeoordelaarsovereenstemming te schatten, naar behoren getraind?	ja/nee/n.v.t.

Toelichting P.III.1:

De aanbieder van het observatie- en registratie-instrument dient bewijslast in te leveren, waaruit blijkt dat de in het instrument getrainde observatoren / leerkrachten betrouwbaar werken met het instrument.

De bewijslast bevat een methode die strookt met de (absolute dan wel relatieve bij de LVS 3-8) interpretatie van de observatie / resultaten.

Toelichting P.III.2:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, zijn de observatoren / leerkrachten die door de aanbieder zijn ingezet om de interbeoordelaarsovereenstemming te schatten, getraind op een niveau dat gelijk is aan het niveau van de training of instructie voor beoordelaars in de praktijk.

3.3 Kwaliteitseisen psychometrische aspecten LVS-instrumenten - type (III) en (IV)

De verantwoording en de beoordeling van de kwaliteit van de psychometrische aspecten van een LVS-instrument type (III) of type (IV) bestaat uit de volgende vijf fasen.

Fasen psychometrische analyse	Toelichting	Kwaliteitseisen
Fase 1. Data verzameling in een pretestprocedure	De toetsaanbieder pretest nieuwe items, voordat deze onderdeel worden van (de itembank van) het LVS-instrument (type III en IV). Pretesten kan plaatsvinden in een proeftoets op vrijwillig deelnemende scholen of in een operationele setting.	De kwaliteit van de dataverzameling wordt beoordeeld met de kwaliteitseisen voor het steekproefkader en de samenstelling van de steekproef (zie § 3.3.1)
Fase 2. Normeren van het LVS-instrument	Het LVS-instrument (type III en IV) wordt genormeerd.	De kwaliteit van de normering wordt beoordeeld met de kwaliteitseisen voor kalibreren en voor normeren (zie § 3.3.2).
Fase 3. Betrouwbaarheid van het LVS-instrument	Naar aanleiding van het instrument kunnen leerlingen worden gecategoriseerd. De toetsaanbieder schat de betrouwbaarheid van het volledige LVS-instrument en berekent het percentage misclassificaties van de verschillende categorieën (type III en IV).	De kwaliteit van de betrouwbaarheid wordt beoordeeld met de kwaliteitseisen voor betrouwbaarheid (zie § 3.3.3).
Fase 4. Constructvaliditeit van het LVS-instrument	De toetsaanbieder toont de constructvaliditeit van het LVS-instrument kwantitatief en/of kwalitatief aan.	De kwaliteit van de constructvaliditeit van het LVS-instrument wordt beoordeeld met de kwaliteitseisen voor constructvaliditeit (zie § 3.3.4).
Fase 5. Volgaspect of signaleringsfunctie	De toetsaanbieder toont de kwaliteit aan van: de opbouw; de betrouwbaarheid (type III en IV), en; het gebruik van de schaal, waarop de ontwikkeling (groei) van de leerling zichtbaar wordt gemaakt, aan.	De kwaliteit van het volgaspect van het LVS-instrument wordt beoordeeld met de kwaliteitseisen voor (1) de opbouw van de schaal, (2) de betrouwbaarheid van de schaal, en (3) het gebruik van de schaal (zie § 3.3.5).

De toetsaanbieder verantwoordt de psychometrische analyses in de wetenschappelijke handleiding.

3.3.1 Kwaliteit van de dataverzameling

Deze paragraaf beschrijft de kwaliteitseisen voor Fase 1 van de psychometrische verantwoording van



een LVS-instrument van het type (III) en (IV), zijnde de kwaliteit van de dataverzameling in een pretestprocedure.

De kwaliteit van de normering van een LVS-instrument type (III) en (IV) en van de betrouwbaarheid van de beslissingen die op basis van de genormeerde scores worden genomen, hangt grotendeels af van de kwaliteit van de dataverzameling. Hierop zijn de normering en de betrouwbaarheidsgegevens gebaseerd. Daarom wordt de kwaliteit van de dataverzameling eerst behandeld.

De kwaliteit van de dataverzameling wordt beoordeeld met de kwaliteitseisen voor het steekproefkader en de samenstelling van de steekproef.

De dataverzameling kan plaatsvinden door nieuwe items aan te bieden aan scholen via een vrijwillige proeftoets, of door nieuwe items direct toe te voegen aan (de itembank van) het bestaande LVS-instrument. Wanneer, in dit tweede scenario, blijkt dat de nieuw geconstrueerde items naar behoren functioneren, kan de toetsaanbieder ervoor kiezen om deze direct mee te laten tellen. Voor een leerling mag het niet herkenbaar zijn of een vraag wel of niet meetelt.

Voor beide scenario's geldt dat de toetsaanbieder er voor kan kiezen om (een deel van) de nieuwe items die worden gepretest eerst door het CvTE te laten beoordelen op de kwaliteitseisen voor inhoudsvaliditeit O.IV.2 en O.IV.3 uit paragraaf 2.3.2.

Code	Vragen	Mogelijke antwoorden
P.IV.1	Is de steekproef van leerlingen groot genoeg en representatief voor de landelijke populatie?	ja/nee
P.IV.2	Indien er sprake is van een onvolledig dataverzamelingsdesign: is het dataverzamelingsdesign adequaat?	ja/nee/n.v.t.

Toelichting P.IV.1:

- De steekproef moet groot genoeg zijn voor het schatten van de moeilijkheidsgraad, het discriminerend vermogen en de kwaliteit van de afleiders van de items.
- Voor het pretesten van nieuwe items in een proeftoets omgeving, volstaan zowel het gebruik van de klassieke toets theorie als de itemresponstheorie (1PLM of 2PLM IRT model).
- Voor het pretesten van nieuwe items in een operationele setting (het zaaien van nieuwe items) binnen een papieren of digitaal lineaire toets volstaan zowel het gebruik van de klassieke toets theorie als de itemresponstheorie.
- Voor het pretesten van nieuwe items in een operationele setting binnen een adaptieve toets op itemniveau (CAT) of een adaptieve toets op moduleniveau (MST) volstaat het gebruik van een IRT model.

Voor alle genoemde pretestmogelijkheden en toetsvormen geldt in ieder geval:

- De steekproef moet representatief zijn voor de doelgroep (c.q. de landelijke populatie leerlingen) in termen van het onderwijsniveau. Dit betekent dat de steekproef adequaat moet zijn gestratificeerd naar de volgende achtergrondvariabelen:
 - regio (Noord, Zuid, Oost, West)
 - urbanisatiegraad (G4: de vier grootste gemeenten; G5-G37: de 5 tot 37 grootste steden; >G37: de overige steden)
 - schoolgrootte (<100; 100-300; >300)
 - schoolweging (<23; 23-26,99; 27-32,99; 33-36,99; ≥ 37)

Tevens moet er informatie geleverd worden over hoe de gerealiseerde steekproef zich verhoudt tot de landelijke populatiewaarden met betrekking tot geslacht, regio, urbanisatiegraad, schoolgrootte en schoolweging.

- Bij de steekproef is in een steekproefkader een beschrijving gegeven van de procedure waarmee de steekproef tot stand is gekomen.
- De omstandigheden waaronder de data verzameld zijn, moeten redelijk vergelijkbaar zijn met de omstandigheden waaronder de operationele toets wordt afgenomen.
- De steekproefgegevens over SBO en SO scholen worden, indien van toepassing, buiten beschouwing gelaten en aanvullend separaat gerapporteerd en verantwoord.

Pretesten van nieuwe items in een proeftoets of in een operationele toets bij gebruik van een papieren of digitaal lineaire toets:

- Een steekproefgrootte van minimaal 300 observaties van elk pretest-item.
- Wanneer de LVS-instrumenten worden ontwikkeld voor een specifiek schooljaar (bv. groep 3 of eind groep 4), dan geldt het minimum aantal observaties per pretest-item voor de leerlingen uit dat specifieke leerjaar.
- Wanneer het uitgangspunt is een vaardigheidsschaal met overlap in vaardigheid tussen de verschillende leerjaren, dan geldt als voorwaarde dat er gebruik wordt gemaakt van een IRT model. In dat geval geldt het minimum aantal observaties per vaardigheidsniveau (categorieën op basis van theta, bij gebruik van een IRT model).

Pretesten van nieuwe items in een proeftoets of in een operationele toets bij gebruik van een



adaptieve toets op itemniveau (CAT) of bij gebruik van een adaptieve toets op moduleniveau (MST):

- Een steekproefgrootte van minimaal 1.000 observaties van elk pretest item, verzameld via een proeftoets of via het zaaien in de itembank.
- Wanneer de LVS-instrumenten worden ontwikkeld voor een specifiek schooljaar (bv. groep 3 of eind groep 4), dan geldt het minimum aantal observaties per pretest-item voor de leerlingen uit dat specifieke leerjaar.
- Wanneer het uitgangspunt is een vaardigheidsschaal met overlap in vaardigheid tussen de verschillende leerjaren, dan geldt als voorwaarde dat er gebruik wordt gemaakt van een IRT model. In dat geval geldt het minimum aantal observaties per vaardigheidsniveau (categorieën op basis van theta, bij gebruik van een IRT model).
- Met behulp van het itemresponsemodel maakt de toetsaanbieder de vertaling van itemparameters naar de bandbreedtes van vaardigheidsniveau (theta) die horen bij de categorieën van het LVS-instrument van het type (III) of (IV).

Toelichting P.IV.2:

Wanneer een LVS-instrument van het type (III) of (IV) of eventueel de itembank van het betreffende instrument uit afzonderlijke items bestaat, worden data vaak verzameld in een onvolledig 'gelinked' design, waarbij niet alle leerlingen alle items maken. Men spreekt vaak van een boekjesdesign. Het staat de toetsaanbieder vrij om zelf een best passend dataverzamelingsdesign te kiezen, mits de keuze wordt onderbouwd.

Als gebruik wordt gemaakt van een gelinked design, dan dient deze te bestaan uit een overzicht of beschrijving van:

- het totaal aantal boekjes ('booklets');
- het aantal booklets per deelpopulatie óf het aantal beantwoorde items per kandidaat bij een adaptieve toets;
- het aantal items per domein én per niveau.

Verder gelden de volgende aanvullende eisen:

- Met het oog op de kalibratie moeten de boekjes voldoende 'gelinked' zijn. Dat betekent dat er voldoende overlap in observaties tussen de verschillende items en boekjes moet zijn.
- Bij gebruik van een IRT model moet er voldoende evidentie voor de passing van dat IRT model worden gepresenteerd, zodat kan worden nagegaan of de dataverzameling adequaat is en of er aan de eisen van steekproefgrootten voor het schatten van (item)parameters is voldaan. De toetsaanbieder beschrijft hiertoe:
 - hoe de passing van het IRT model is geëvalueerd en voldoet. Dit wordt gedaan door:
 - aan te tonen dat de itemparameters tussen verschillende afnamen of boekjes niet veranderen, en;
 - een DIF analyse (differential item functioning) op itemniveau uit te voeren op minimaal de achtergrondvariabele geslacht, waarbij de itemparameterschattingen uit de verschillende boekjes in de pretest worden vergeleken. Wanneer er sprake is van DIF op itemniveau, dient de toetsaanbieder aan te tonen dat dit in overeenstemming is met de verwachting op basis van de relevante literatuur².
 - Aanvullend worden de effectgroottes vermeld, waarop de passing van het model is geëvalueerd. Hierbij geldt als uitgangspunt dat, het verschil tussen het geobserveerd aantal correcte antwoorden en het verwacht aantal correcte antwoorden op basis van de globale modeltoets R1c (Glas, 1988) en/of de DIF toets op itemniveau kleiner is dan 10%. De toetsaanbieder mag hier met de juiste motivering van afwijken.
 - Wat betreft de schattingsfout (standard error) van de nieuwe items dient de toetsaanbieder informatie over de scoreverdeling aan te leveren, zodat inzichtelijk is welke informatiewaarde (en omgekeerd de standard error) correspondeert met een klassieke betrouwbaarheidscoëfficiënt van 0,80.
- Andere kalibratiemethoden, zoals bijvoorbeeld kernel-equating van Von Davier en Holland (2004), zijn ook toegestaan, zolang de betrouwbaarheid analoog is aan de voor IRT geformuleerde eisen.
- Er is beargumenteerd op welke wijze er rekening is gehouden met vermoeidheids- en/of volgorde effecten.
- Wanneer de pretestdata zijn verzameld met een CAT of MST is het niet zinvol om het hele onvolledige design weer te geven, omdat iedere leerling dan in principe een unieke verzameling items maakt. Ook dan moet wel duidelijk zijn op welke gegevens de linking en normering gebaseerd zijn.

3.3.2 Normering

Deze paragraaf beschrijft de kwaliteitseisen voor Fase 2 van de psychometrische verantwoording van

² Zie bijvoorbeeld: Bechger Exploring Differential Items Functioning with Dexter, 2022.



een LVS-instrument van het type (III) en (IV), zijnde de kwaliteit van het normeren van het LVS-instrument.

Na het verzamelen van de observaties voert de toetsaanbieder een kalibratie uit met de verzamelde pretestdata. In de kalibratie schat de toetsaanbieder de itemparameters, zodat de kwaliteit van de in de proeftoets en/of in de operationele versie van de toets nieuw gepreteste items kan worden vastgesteld. Vervolgens kan de toetsaanbieder het LVS-instrument normeren. Er is onderscheid te maken tussen absolute en relatieve normen.

Absolute normen plaatsen een leerling op een schaal die betrekking heeft op een analyse van een inhoudelijk domein. Een absolute norm is gebaseerd op een minimaal acceptabel beheersingsniveau van dat inhoudelijke domein. Absolute normen kan men bepalen via een standaardsettingsmethode waarbij inhoudelijke experts de minimaal acceptabele beheersingsniveaus vaststellen.

Relatieve normen plaatsen de leerling op een schaal die betrekking heeft op de relatieve positie van een leerling in een populatie. Een relatieve norm is derhalve gebaseerd op een onderlinge vergelijking van de toetsprestaties van de kandidaten in een adequaat omschreven populatie. Om deze normen vast te stellen is een pretest met een zuivere steekproef nodig, waarna de normen met een psychometrische methode worden vastgesteld.

Het staat de toetsaanbieder vrij om, binnen de eisen van het beoordelingskader, hiervoor zelf een best passende wijze van normeren te kiezen. De toetsaanbieder beargumenteert dat de eigen gemaakte keuze deugdelijk is. Dit betekent dat er naast de methode van standaardsetting met experts of het definiëren van relatieve normen via pretesten ook gebruik gemaakt mag worden van andere manieren. Bijvoorbeeld door de normen af te leiden van de referentieniveaus. Dat kan op verschillende manieren worden gerealiseerd. Bijvoorbeeld door een toets voor de leerjaren 6-8 te pretesten samen met items uit beschikbare referentiesets of door LVS items via een pretest te kalibreren op dezelfde schaal als de ankeritems van de doorstroomtoets.

Code	Vragen	Mogelijke antwoorden
Absoluut normeren		
P.V.1	Is de keuze voor de standaardbepalingsmethode gedegen gemotiveerd en op de juiste wijze uitgevoerd?	ja/nee/n.v.t.
P.V.2	Zijn de experts met inhoudelijke vakdeskundigheid die de standaarden (cesuren) vaststellen naar behoren geselecteerd en getraind?	ja/nee/n.v.t.
P.V.3	Is er voldoende overeenstemming tussen de experts?	ja/nee/n.v.t.
Relatief normeren		
P.V.4	Als de normeringssteekproef afwijkt van de kalibratiesteekproef, zijn de normgroepen groot genoeg?	ja/nee/n.v.t.
P.V.5	Als de normeringssteekproef afwijkt van de kalibratiesteekproef, zijn de normgroepen representatief?	ja/nee/n.v.t.
P.V.6	Zijn de normen correct bepaald?	ja/nee/n.v.t.

Toelichting P.V.1:

- Uit de verstrekte informatie blijkt dat de kwaliteit van de gekozen standaard- of cesuurbepalingsmethode hoog is en dat de methode correct is uitgevoerd en is onderbouwd door de erbij betrokken experts met vakdeskundigheid.
- Uit de verstrekte informatie blijkt waarom de gebruikte methode past bij het doel.
- Bij LVS-instrumenten van het type (III), waarbij de standaard / cesuur al min of meer vastligt in de onderliggende referentiesets dient de cesuur te worden overgebracht middels linking en niet door een afzonderlijke standaardbepaling.
- Het beoordelingskader LVS laat tevens ruimte voor alternatieve vormen van normeren, zoals bijvoorbeeld continue normeren waarbij leerlingen uit een lagere groep bij de normering ook een 'moeilijker' toets maken voor een hoger leerjaar. Alternatieve vormen van normeren dienen inhoudelijk en psychometrisch te worden beargumenteerd door de toetsaanbieder.

Toelichting P.V.2:

- Er is een onderbouwing van de selectieprocedure, er heeft een selectie van voldoende experts voor de gekozen methode met aantoonbaar inhoudelijke vakdeskundigheid plaatsgevonden en deze experts hebben een gedocumenteerde training gevolgd in de betreffende standaardbepalingsmethode.

Toelichting P.V.3:

- De aanbieder toont aan dat de mate van overeenstemming tussen de experts resulteert in een coëfficiënt voor beoordelaarsovereenstemming waarbij de proportie overeenstemming representatief is voor een instrument voor minder belangrijke beslissingen (een zogenaamd medium stakes instrument) op individueel niveau.



Toelichting P.V.4:

- De normgroep bestaat uit het bij kwaliteitseis P.IV.1 genoemde minimum aantal observaties per item.

Toelichting P.V.5:

- Een beschrijving van de samenstelling van de normgroep voor in ieder geval de achtergrondvariabelen sekse, regio, urbanisatiegraad, schoolgrootte en schoolweging is beschikbaar en er is gebruik gemaakt van een aselekt steekproefmodel.

Toelichting P.V.6:

- Er is een heldere beschrijving van de betekenis en beperkingen van de normschaal en het type normschaal is in overeenstemming met het doel van het LVS-instrument van het type (III) of (IV).

3.3.3 Betrouwbaarheid

Deze paragraaf beschrijft de kwaliteitseisen voor Fase 3 van de psychometrische verantwoording van een LVS-instrument van het type (III) en (IV), zijnde de kwaliteit van de betrouwbaarheid van het LVS-instrument.

Code	Vragen	Mogelijke antwoorden
P.VI.1	Zijn of worden de betrouwbaarheidsgegevens van de metingen correct berekend?	ja/nee
P.VI.2	Zijn de betrouwbaarheidsgegevens voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden?	ja/nee
P.VI.3	Is er een indicatie gegeven van het percentage misclassificaties van de verschillende categorieën van het LVS-instrument van het type (I) of (II)?	ja/nee

Leerlingen worden na afname van het LVS-instrument gecategoriseerd. De toetsaanbieder schat de betrouwbaarheid van het volledige LVS-instrument van het type (III) of (IV) en berekent het percentage misclassificaties van de verschillende categorieën van het LVS-instrument van het type (III) of (IV).

Voor het schatten van de betrouwbaarheid van niveaubepaling en groei mag alleen gebruik gemaakt worden van items die onderwijskundig en psychometrisch goed functioneren. Wanneer de toetsaanbieder er toch voor kiest om, om andere redenen, minder goed functionerende items mee te nemen, dan dient de toetsaanbieder dit te beargumenteren.

Toelichting P.VI.1:

Bij een LVS-instrument van het type (III) of (IV) is er normaliter sprake van meerdere categorieën of van een relatieve categorisatie ten opzichte van een referentiepopulatie. De betrouwbaarheid van het LVS-instrument van het type (III) of (IV) hangt samen met percentage verwachte misclassificaties. De betrouwbaarheid moet adequaat worden aangetoond. In de psychometrie maakt men een onderscheid tussen globale en lokale betrouwbaarheid:

Globale betrouwbaarheid heeft betrekking op de mate waarin men twee willekeurig getrokken leerlingen uit een adequaat gedefinieerde populatie kan onderscheiden. Een maat voor de globale betrouwbaarheid is een variantieratio: de verhouding tussen de relevante variatie in de te meten vaardigheid (de ware variantie) en de totale variantie, die de som is van de ware variantie en irrelevante variatie (de ruis). Deze ratio is zowel met CTT als met een IRT model te berekenen. Het is hierbij belangrijk om alle variantie-componenten in de schatting van de betrouwbaarheid mee te nemen. Dus ook variantie tussen beoordelaars, of de effecten van clustering van leerlingen in schoolklassen op de variantieschatting. Zoiets is bijvoorbeeld te realiseren met generaliseerbaarheids-theorie en multilevelanalyse. Vaak kan men de globale betrouwbaarheid correct schatten met Cronbach's Alpha, maar bij hiërarchische steekproeven en beoordelingen door beoordelaars, zijn de genoemde meer geavanceerde technieken nodig.

Lokale betrouwbaarheid heeft betrekking op de meetprecisie bij specifieke punten op de schaal. Bijvoorbeeld bij categoriegrenzen. Voor het schatten van de lokale betrouwbaarheid is IRT het meest geëigend. Meestal gebruikt men Fishers informatie bij bepaalde punten op de latente vaardigheids-schaal.

Toelichting P.VI.2:

Bij een LVS-instrument van het type (III) of (IV) is de coëfficiënt voor globale betrouwbaarheid berekend over minimaal 300 kandidaten. Bij hiërarchische steekproeven moet bij het bepalen van steekproefgrootten ook rekening gehouden worden met de effecten van clustering van leerlingen. De betrouwbaarheidscoëfficiënt moet voldoende zijn voor een medium stakes instrument, wat neerkomt op een betrouwbaarheid $\geq 0,70$.

Lokale betrouwbaarheid kan gebruikt worden om een schatting te maken van het percentage leerlingen dat foutief wordt geclassificeerd. Wat in dit nog acceptabel is, hangt af van de meetprentie van de toets. Goed vergelijkingsmateriaal is te vinden in de schattingen van percentages misclassi-



ficaties in verantwoordingen van andere LVS-toetsen en doorstroomtoetsen voor het primair onderwijs.

Toelichting P.VI.3:

De toetsaanbieder geeft een indicatie van het percentage misclassificaties van de verschillende categorieën van het LVS-instrument van het type (III) of (IV). Voor een voldoende op kwaliteitseis B3 dient het percentage misclassificaties < 20% te zijn.³

3.3.4 Constructvaliditeit

Deze paragraaf beschrijft de kwaliteitseisen voor Fase 4 van de psychometrische verantwoording van een LVS-instrument van het type (III) en (IV), zijnde de kwaliteit van de constructvaliditeit van het LVS-instrument.

De construct- of begripsvaliditeit is de eigenschap die het LVS-instrument heeft als kan worden aangetoond dat het instrument het door de constructeur beoogde kenmerk van de leerling (onderliggende trek, vaardigheid) meet. De toetsaanbieder kan de constructvaliditeit kwantitatief en/of kwalitatief aantonen.

Code	Vragen	Mogelijke antwoorden
P.VII.1	Is aangetoond dat er sprake is van unidimensionaliteit?	ja/nee
P.VII.2	Is aangetoond dat de itemkwaliteit op orde is?	ja/nee
P.VII.3	Is er onderzoek uitgevoerd naar de convergente en divergente validiteit?	ja/nee
P.VII.4	Is aangetoond dat er geen sprake is van item bias?	ja/nee
P.VII.5	Is onderzoek uitgevoerd naar de prestaties van relevante subgroepen?	ja/nee

Toelichting P.VII.1:

De resultaten van de uitgevoerde kalibratie maken het aannemelijk dat er bij het LVS-instrument van het type (III) of (IV) sprake is van unidimensionaliteit. Dit betekent dat met elke willekeurige subset van items uit de gekalibreerde itembank dezelfde onderliggende vaardigheid kan worden vastgesteld. De unidimensionaliteit kan op klassieke wijze worden aangetoond met een confirmatieve factoranalyse gekoppeld aan multidimensionale betrouwbaarheidsindices⁴. Het gebruik van IRT biedt daarnaast de mogelijkheid om met multidimensionale IRT modellen, testlet modellen of bi-factor modellen een complexere dimensiestructuur te modelleren.

Toelichting P.VII.2:

De itemkwaliteit wordt aangetoond met een passende moeilijkheidsparameter en met voldoende discriminerend vermogen.

Toelichting P.VII.3:

Naar mate de complexiteit van het te meten onderdeel toeneemt (bv. het meer complexe onderdeel sociaal-emotionele ontwikkeling versus het meer overzichtelijk terrein rekenen) wegen de inhoudelijke argumenten zwaarder dan de kwantitatieve argumenten. Met bijvoorbeeld een Multi Trait Multi Method matrix kan worden onderbouwd hoe een instrument zich gedraagt. Voor de meer complexe onderdelen is er vaak beperkt of geen extern vergelijkingsmateriaal. In die situatie biedt een meer argument based approach (Kane, 2004) of bijvoorbeeld een operationalisering van de inhoudsvaliditeit uitkomst. Dit betreft een meer beschrijvende manier op basis van de blueprint/toetsmatrijs, waarbij gelet wordt op de representativiteit en evenwichtigheid. Bovendien kan een DIF analyse worden toegevoegd als onderdeel van een validity argument.⁵

Toelichting P.VII.4:

In het kader van itembias is er onderzoek uitgevoerd naar differentieel item functioneren (DIF) met betrekking tot in ieder geval de achtergrondvariabele sekse.

Toelichting P.VII.5:

De prestaties van relevante subgroepen is onderzocht door de gemiddelde score en standaarddeviatie per leeftijdsgroep en per sekse te bepalen. Tevens dienen de effectgroottes te worden toegevoegd,

³ Zie bijvoorbeeld: Eggen & Sanders (1993), *Psychometrie in de Praktijk*, Tabel 3.13. Arnhem: Cito.

⁴ Zie bijvoorbeeld: Widhiarso, W., & Ravand, H. (2014). *Estimating reliability coefficient for multidimensional measures: A pedagogical illustration*. *Review of psychology*, 21(2), 111-121.

⁵ Zie bijvoorbeeld: Benito, J. G., Sireci, S., García, J. L. P., Montesinos, M. D. H., & Baena, I. B. (2018) *Differential item functioning: Beyond validity evidence based on internal structure*. *Psicothema*, 30(1), 104-109.



waarbij geldt dat een klein effect < 0.2 voldoende is.

3.3.5 Volgaspect

Deze paragraaf beschrijft de kwaliteitseisen voor Fase 5 van de psychometrische verantwoording van een LVS-instrument van het type (III) en (IV), zijnde de kwaliteit van het volgaspect of de signaleringsfunctie.

Om leervorderingen te kunnen meten, moeten de scores van de leerling op een schaal te plaatsen zijn die de ontwikkeling van leerlingen zichtbaar maakt. Dit leidt voor LVS-instrumenten van het type (III) of (IV) tot drie criteria met betrekking tot de schaal waarop groei wordt uitgedrukt: (1) de opbouw van de schaal, (2) de betrouwbaarheid van de schaal, en (3) het gebruik van de schaal.

Code	Vragen	Mogelijke antwoorden
P.VIII.1	Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een adequate manier gemeten?	ja/nee
P.VIII.2	Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?	ja/nee
P.VIII.3	Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden?	ja/nee

Toelichting P.VIII.1:

- Er dient duidelijk invulling te worden gegeven aan het begrip ‘volgsysteem’. Het betreft een systeem met onderliggende data waarmee de ontwikkeling van de leerling kan worden gevolgd over de tijd heen en waarmee kan worden geïnterpreteerd in hoeverre deze ontwikkeling in lijn is met de leerdoelen van de verschillende leerjaren en met de ontwikkeling van de andere leerlingen uit hetzelfde leerjaar. Het volgsysteem moet hierom bestaan uit werkelijke gegevens om scores van verschillende afnames met elkaar te kunnen vergelijken en/of om duiding te kunnen geven in termen van bijvoorbeeld groei. Tevens dienen er duidelijke handvatten voor testgebruikers te komen hoe zij de toetsresultaten kunnen gebruiken als volgsysteem. Bij het gebruik van bijvoorbeeld IRT modellen wordt de vergelijkbaarheid al iets makkelijker (uitgaande van een juiste linking tussen allen items/toetsversies). Omdat veel docenten en ouders niet weten hoe IRT werkt en hoe dit moet worden toegepast, moeten de toetsaanbieders uitleggen hoe zij de resultaten moeten lezen en interpreteren.
- Alle overwegingen over steekproeftrekking, betrouwbaarheid en normering gelden eveneens voor een reeks van twee of meer opeenvolgende meetmomenten. Dus beide steekproeven moeten representatief zijn en het design, dat in dit geval bijna per definitie onvolledig is, moet adequaat zijn in de termen die hierboven zijn gedefinieerd.
- Verder moet er empirische informatie zijn over de schaalbaarheid van opeenvolgende meetmomenten. Het hoeft niet het geval te zijn dat de schaal strikt uni-dimensioneel is in de zin van een uni-dimensioneel IRT model (hoewel dit wel de meest voor de hand liggende schaal is). Wel dient er in ieder geval betekenisvolle informatie gegeven te worden over de samenhang tussen de twee (of meer) meetmomenten. Essentieel is dat de schaal waarop de groei wordt weergegeven grondig is onderbouwd.

Toelichting P.VIII.2:

- Daarbij is het voor LVS-instrumenten van het type (III) of (IV) ook van belang om een indicatie van de betrouwbaarheid van die gevolgtrekking weer te geven. Vertaald naar een uni-dimensioneel IRT model betekent dit dat de schattingsfout van het verschil van de vaardigheid op twee tijdstippen geschat moet zijn. Voor percentielscores betekent dit dat de betrouwbaarheid van de verandering van de percentielscores geschat moet zijn.
- Bij gebruik van CTT zijn de percentielscores gebaseerd op een somscore met een meetfout. Het verschil in percentielscores kan vertaald worden in het verschil in onderliggende scores en hun betrouwbaarheid⁶. Omdat de transformatie van ruwe scores of theta scores naar percentielscores niet lineair is, kan dit wel resulteren in een zekere bias. Deze kan de toetsaanbieder onderbouwen in de verantwoording.
- Bij gebruik van IRT zijn de percentielscores gebaseerd op vaardigheidsscores die geschat zijn met een zekere mate van betrouwbaarheid.⁷

Toelichting P.VIII.3:

⁶ Zie bijvoorbeeld: Rogosa, D. R. (2000). *Accuracy of individual scores expressed in percentile ranks: Classical test theory calculations*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

⁷ Zie bijvoorbeeld: May, K., & Nicewander, W. A. (1994). *Reliability and information functions for percentile ranks*. *Journal of Educational Measurement*, 31(4), 313-325.



De handleiding moet een beschrijving bevatten van hoe de gebruiker (zoals docenten en ouders) de gegevens met betrekking tot de groei (en/of stagnatie) van een leerling inhoudelijk en relatief ten opzichte van een referentiepopulatie dient te interpreteren. De geschreven toelichting moet consistent zijn met de resultaten uit het betrouwbaarheids-, validiteits- en normeringsonderzoek, dat wil zeggen dat de gebruiker een goed beeld moet krijgen van de (relatieve) onderwijskundige waarde en meetpretentie van de resultaten.

4. Organisatorische aspecten

4.1 Afname

De beoordeling van de organisatorische aspecten van een LVS-instrument richt zich op de criteria voor veilig toetsen en observeren / registreren. Onderstaande kwaliteitseisen zijn van toepassing op zowel observatie- en registratie-instrumenten type (I) en (II) als op LVS-instrumenten type (III) en (IV).

Code	Vragen	Mogelijke antwoorden
B.I.1	Heeft de aanbieder de functies en verantwoordelijkheden van de functionarissen die betrokken zijn bij de inhoud en de beveiliging van het LVS-instrument volledig en juist beschreven?	ja/nee
B.I.2	Hebben de functionarissen van de aanbieder die betrokken zijn bij de inhoud en de beveiliging van het LVS-instrument aantoonbaar een geheimhoudingsverklaring ondertekend?	ja/nee
B.I.3	Waarborgt de aanbieder op gedegen en complete wijze de privacy van de leerkracht, (jonge) leerling en ouder conform de Algemene Verordening Gegevensbescherming)?	ja/nee
B.I.4	Hanteert de aanbieder op gedegen en complete wijze incidentmanagement inclusief een protocol datalekken, zoals voorgeschreven door de Autoriteit Persoonsgegevens?	ja/nee
B.I.5	Bevat de handleiding voor de schoolleider en leerkracht op gedegen en complete wijze informatie over de wijze waarop het instrument en de ermee verzamelde informatie op een veilige wijze gebruikt en opgeslagen dient te worden?	ja/nee
B.I.6	Controleert de aanbieder regelmatig het internet en andere media op het bekend raken van (delen) van de inhoud van het LVS-instrument of andere niet-openbare informatie over het LVS-instrument?	ja/nee/n.v.t.

Toelichting B.I.1 t/m B.I.6:

Indien de aanbieder ISO 27001 is gecertificeerd, volstaat het om naar dit certificaat en het achterliggende Information Security Management System (ISMS) te verwijzen.

In alle andere situaties kan de aanbieder bijvoorbeeld verwijzen naar een verwerkersovereenkomst, waar op de kwaliteitseisen B.I.1 t/m B.I.6 wordt ingegaan. Daarbij dient bewijslast te worden overlegd van de bij B.I.2, B.I.3 en B.I.4 gevraagde documenten, contracten en protocollen.

Ter verduidelijking wordt vermeld dat de verantwoordelijkheid voor het goed gebruik van het instrument en de ermee verzamelde informatie in de dagelijkse praktijk bij de school / de eindgebruiker zelf ligt.



TOELICHTING

Sinds de invoering van de Wet centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs (Stb. 2014, 13) is het bevoegd gezag in het primair onderwijs verplicht tot het volgen van de ontwikkeling van kennis en vaardigheden van al hun leerlingen door gebruik te maken van een leerling- en onderwijsvolgsysteem (LVS). Hierbij dient het in ieder geval gebruik te maken van toetsen die voldoen aan de op grond van de Wet op het Primair Onderwijs onderscheidenlijk de Wet op de expertisecentra, en het Toetsbesluit PO gestelde voorschriften voor de inhoud, de validiteit, de betrouwbaarheid en de deugdelijke normering.

Op 1 januari 2023 treedt de Wet van 9 februari 2022 tot wijziging van een aantal onderwijswetten in verband met aanpassingen op het gebied van de doorstroom van het basisonderwijs naar het voortgezet onderwijs en wijziging van de stelselinrichting van doorstroomtoetsen en toetsen verbonden aan leerling- en onderwijsvolgsystemen in het basisonderwijs (Stb. 2022, 135) in werking. Met de inwerkingtreding van deze wet krijgt het College voor toetsen en examens de taak om de toetsen verbonden aan het leerling- en onderwijsvolgsysteem te erkennen. Deze erkenning behoudt haar geldigheid voor een periode van 10 jaar. Om tot erkenning te komen, beoordeelt het CvTE of de toetsen voldoen aan de in de Wet op het Primair Onderwijs en het Toetsbesluit PO vastgelegde voorschriften. Op grond van artikel 3a, eerste lid, onderdeel g, van de Wet College voor toetsen en examens is het College vanaf 1 januari 2023 bevoegd om bij regeling een beoordelingskader vast te stellen voor de erkenning van toetsen verbonden aan het LVS. Dit beoordelingskader wordt met deze regeling vastgelegd.

Het bevoegd gezag is verplicht om in ieder geval gebruik te maken van door het College erkende toetsen van het LVS voor de onderdelen Nederlandse taal en rekenen voor het derde tot en met het achtste leerjaar. Deze onderdelen moeten op grond van artikel 45b, tweede lid, van de Wet op het primair onderwijs, artikel 48c, tweede lid, van de Wet op de expertisecentra en artikel 51a tweede lid, van de Wet primair onderwijs BES, verplicht getoetst worden. Het College maakt daarnaast gebruik van het beoordelingskader bij de beoordeling van toetsen voor leerlingen uit deze groepen waarmee de niet wettelijk verplichte onderdelen worden getoetst, zoals de sociaal-emotionele ontwikkeling. Het bevoegd gezag mag ook niet door het College erkende toetsen voor de wettelijk verplichte onderdelen Nederlandse taal en Rekenen gebruiken, mits zij daarnaast door het College erkende toetsen gebruiken. Verder staat het een bevoegd gezag vrij om toetsen van verschillende toetsaanbieders te combineren.

*Het College voor toetsen en examens,
de voorzitter,
J.H. van der Vegt*