



Regeling van het College voor toetsen en examens van 3 februari 2025, nummer CvTE-25.00471, houdende wijziging van de Regeling beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen PO in verband met een aantal technische verbeteringen en nader toelichting van belangrijke criteria

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 Staatssecretaris van Onderwijs, Cultuur en Wetenschap, gegeven op 19 februari 2025, nummer 50613368,

Besluit:

Artikel 1. Beoordelingskader

De regeling beoordelingskader voor de toetsen behorende tot leerling- en onderwijsvolgsystemen PO wordt als volgt gewijzigd:

De bijlage wordt vervangen door de bijlage bij deze regeling.

Artikel 2. Inwerkingtreding

Deze regeling treedt in werking met ingang van de dag na de datum van uitgifte van de Staatscourant waarin de regeling wordt geplaatst.

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> <i>Absolute normen</i>	Een indiener van een leerlingvolgsysteem. Standaarden waaraan de vaardigheidsniveaus van leerlingen gerelateerd kunnen worden. Deze standaarden worden vastgesteld op basis van inhoudelijke en/of toetsspecifieke criteria.
<i>Algoritme (beslisregel)</i>	Systeem van regels, logische of wiskundige handelingen, die als stappenplan gelden en kunnen worden uitgevoerd om tot een bepaald resultaat te komen.
<i>Beoordelaarsovereenstemming</i>	De mate waarin verschillende beoordelaars tot gelijke beoordelingen komen van (bijvoorbeeld) het werk van leerlingen. Twee veel gebruikte maten voor beoordelaarsovereenstemming zijn het percentage van overeenstemming tussen twee beoordelaars en Cohen's Kappa.
<i>Betrouwbaarheid</i> <i>CAT</i>	De mate waarin toetsscores vrij zijn van toevallige meetfouten. Een computergestuurde adaptieve toets (CAT) op itemniveau, waarbij een item wordt gekozen die qua moeilijkheid het beste aansluit op de vaardigheid zoals ingeschat op basis van de antwoorden op de eerdere opgaven.
<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.
<i>CvTE</i> <i>DIF</i>	College voor toetsen en examens Differentieel item functioneren of Differential item functioning (DIF) is een verschijnsel dat items verschillend functioneren voor bepaalde (sub)groepen leerlingen, zelfs wanneer deze (sub)groepen leerlingen een vergelijkbaar prestatieniveau hebben. Het gaat hierbij over de vraagonzuiverheid, dat wil zeggen wanneer de verschillen in de prestaties op het item tussen groepen niet geheel verklaard kunnen worden door verschillen in de (latente) vaardigheid die aan de opgave ten grondslag ligt.
<i>Discriminerend vermogen</i>	De mate waarin op grond van de scores op een item hoog- en laagscorende leerlingen (op de hele toets) onderscheiden kunnen worden.
<i>Domein</i>	Verzameling van onderling samenhangende kennis en vaardigheden waarover getoetst wordt.
<i>Effectgrootte</i>	Statistische maat om de grootte van een effect weer te geven. Er wordt een onderscheid gemaakt tussen niet gestandaardiseerde (ruwe) effectgroottes en gestandaardiseerde effectgroottes. De bekendste gestandaardiseerde effectmaat is Cohen's <i>d</i> . Deze maat geeft het verschil tussen twee gemiddelden in het aantal standaarddeviaties. Ook de correlatie kan worden gezien als een gestandaardiseerde maat voor de effectgrootte.
<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 (KTT) 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. 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>Leerdoel</i>	Het beoogde resultaat van onderwijsactiviteiten. Als het beoogde leerresultaat wordt geformuleerd in termen van uiterlijke condities en waarneembaar leerlinggedrag spreekt men van 'concrete leerdoelen' of 'gedragsdoelen'. Als het beoogde leerresultaat wordt geformuleerd in termen van cognitief gedrag of van cognitieve structuren, spreekt men van cognitieve leerdoelen.
<i>LVS</i>	Het leerling(- en onderwijs)volgsysteem (LVS) betreft een systeem met onderliggende data waarmee de ontwikkeling van de leerling dient te 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>Lokale betrouwbaarheid</i>	De mate waarin de toetsscores op een specifiek deel van de meetschaal vrij zijn van toevallige meetfouten.
<i>Module</i>	Set van items in een multistage toets (MST), vaak van vergelijkbare moeilijkheid.
<i>Moeilijkheidsgraad</i>	In hoeverre leerlingen uit de doelpopulatie in staat zijn om de vraag of vragen goed te beantwoorden. De moeilijkheid wordt meestal uitgedrukt met een p-waarde.
<i>MST</i>	Een multistage toets (MST) is een adaptieve toets op modulenniveau, die via een algoritme het niveau van de leerling bepaalt op basis van diens antwoorden op een set opgaven (module). Daarna krijgt de leerling een nieuwe set opgaven (module), afgestemd op het eerder bepaalde niveau.
<i>Normeren</i>	Het toekennen van een waardering aan een behaalde score. Onderdeel van het normeren is het vaststellen van de cesuren.
<i>Relatief beoordelen</i>	Het rapporteren op basis van relatieve normen.
<i>Relatieve normen</i>	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>Populatie</i>	Een populatie is een verzameling van vooraf gedefinieerde eenheden waarover men met behulp van statistisch onderzoek uitspraken wil doen.
<i>Pretesten</i>	Het afnemen van een aantal opgaven bij een groep leerlingen, meestal met het doel de eigenschappen van het LVS-instrument te onderzoeken. Deze groep leerlingen is zoveel mogelijk vergelijkbaar met de populatie waarvoor het LVS-instrument oorspronkelijk is ontwikkeld.
<i>P-waarde</i>	Is een getal tussen 0 en 1 waarmee de moeilijkheidsgraad van een opgave kan worden weergegeven. De p-waarde wordt berekend door de gemiddelde score op een opgave te delen door de maximaal haalbare score op die opgave. De p-waarde voor een meerkeuzevraag is gelijk aan het deel van de leerlingen die het juiste antwoord heeft gekozen.
<i>Referentiekader</i>	Het referentiekader is een beschrijving van wat leerlingen op verschillende momenten in hun schoolloopbaan op het gebied van Nederlandse taal en Rekenen moeten kennen en kunnen (zie ook: Besluit referentieniveaus Nederlandse taal en Rekenen).
<i>Referentieniveaus</i>	Beschrijvingen van beheersingsdoelen voor Nederlandse taal en Rekenen die aangeven wat leerlingen moeten kennen en kunnen op verschillende momenten tijdens een schoolloopbaan. De referentieniveaus zijn wettelijk verankerd in de Wet referentieniveaus Nederlandse taal en Rekenen.
<i>Schaal</i>	Een reeks getallen die volgens een bepaald voorschrift gekoppeld worden aan waarnemingen.
<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>Steekproef</i>	Verzameling leerlingen met specifieke kenmerken die op aselecte wijze uit de te onderzoeken populatie zijn getrokken.
<i>Terrein</i>	Een te toetsen onderdeel (bijvoorbeeld Rekenen ofwel Nederlandse taal).
<i>Toetsmatrijs</i>	Een adequate, schematische representatie van het meetdoel.
<i>Toetsreglement</i>	Document waarin de wijze van afnemen van een toets wordt geregeld.
<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. Wanneer aanbieders ervoor kiezen om een afzonderlijk onderdeel van een LVS-instrument ter erkenning of herbeoordeling in te dienen bij het CvTE, dienen zij in hun verantwoording te onderbouwen hoe dit onderdeel past binnen het volledige LVS-instrument. 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.

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 alleen leerlingvolgsystemen die worden gebruikt in het onderwijs. Daar vallen dus geen screeningsinstrumenten onder die worden afgenomen bij individuele leerlingen om diagnoses te kunnen stellen door een zorgaanbieder.

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:

- Een ingevulde leeswijzer, te vinden op de website van het CvTE, met daarin uiteengezet in welk aangeleverd document en hoofdstuk een kwaliteitscriterium uit het beoordelingskader wordt verantwoord;
- 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;
- het toetsreglement;
- Indien van toepassing: Cd's, dvd's en/of USB-sticks en log-ins en wachtwoorden bij online instrumenten.

Let op: bovenstaande opsomming is niet uitputtend en een indicatie voor de aan te leveren documentatie. Met het invullen van de leeswijzer maakt een aanbieder helder in welke documentatie de kwaliteitscriteria precies worden verantwoord. Ook wanneer een aanbieder een kwaliteitscriterium niet van toepassing acht, moet in de verantwoordingsdocumenten duidelijk worden waarom een aanbieder dit kwaliteitscriterium niet van toepassing acht en moet de leeswijzer hiernaar verwijzen.

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: een ingevulde leeswijzer, te vinden op de website van het CvTE, met daarin uiteengezet in welk aangeleverd document en hoofdstuk een kwaliteitscriterium uit het beoordelingskader wordt verantwoord;
- het instrument / de toets met bijbehorende ondersteunende materialen voor de leerkracht;
- antwoordformulieren;
- scoringssleutels;
- 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 groepsverzichten / schoolrapportages;
- cd's, dvd's en/of USB-sticks bij geautomatiseerde instrumenten en log-ins en wachtwoorden bij online-instrumenten;
- het toetsreglement.

Aanvullend dient voor een LVS-instrument van het type (III) het aangeleverde informatiepakket eveneens te bestaan uit:

- in geval van normatieve toetsing, het vertrouwelijk delen van de normtabellen, en;
- indien van toepassing: documentatie over geautomatiseerde scoring.

Let op: bovenstaande opsomming is niet uitputtend en een indicatie voor de aan te leveren documentatie. Met het invullen van de leeswijzer maakt een aanbieder helder in welke documentatie de kwaliteitscriteria precies worden verantwoord. Ook wanneer een aanbieder een kwaliteitscriterium niet van toepassing acht, moet in de verantwoordingsdocumenten duidelijk worden waarom een aanbieder dit kwaliteitscriterium niet van toepassing acht en moet de leeswijzer hiernaar verwijzen.

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.5 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.7 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 <i>inhoudsvaliditeit</i> .			
	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 <i>signaleringsfunctie</i> en/of <i>volgaspect</i> .			
	Type I	Type II	Type III	Type IV
	O.II.1	•	•	
	O.II.2	○	○	
O.II.3	○	○		



O.II.4	○	○		
O.II.5	•	•		
<i>Kwaliteitseis O.III Paragraaf 2.3.1</i>	Toetsconstructie			
O.III.1			•	•
O.III.2			•	•
O.III.3			•	•
O.III.4			•	○
O.III.5			•	•
<i>Kwaliteitseis O.IV Paragraaf 2.3.2</i>	Inhoudsvaliditeit			
O.IV.1			•	•
O.IV.2			•	•
O.IV.3			•	•
<i>Kwaliteitseis O.V Paragraaf 2.3.3</i>	Referentieniveaus			
O.V.1a			•	
O.V.1b			•	
O.V.2a			○	
O.V.2b			○	
<i>Kwaliteitseis O.VI Paragraaf 2.3.4</i>	Inzicht in de leervorderingen			
O.VI.1			•	•
O.VI.2			•	•
O.VI.3			•	•
O.VI.4			•	•
O.VI.5			•	•

Psychometrische aspecten

<i>Kwaliteitseis P.I Paragraaf 3.2.1</i>	Dataverzameling, met specifieke aandacht voor de kwaliteitseis <i>steekproef van leerlingen</i> .			
	Type I	Type II	Type III	Type IV
P.I.1	○	○		
<i>Kwaliteitseis P.II Paragraaf 3.2.2</i>	IJking, met specifieke aandacht voor de kwaliteitseis <i>niet relatieve normering</i> .			
P.II.1	○	○		
P.II.2	○	○		
P.II.3	○	○		
<i>Kwaliteitseis P.III Paragraaf 3.2.3</i>	Interbeoordelaarsbetrouwbaarheid, met specifieke aandacht voor de kwaliteitseis <i>betrouwbaarheid van de observaties</i> .			
P.III.1	○	○		
P.III.2	○	○		
<i>Kwaliteitseis P.IV Paragraaf 3.3.1</i>	De kwaliteit van de dataverzameling			
P.IV.1		○	•	•



P.IV.2		○	○	○
<i>Kwaliteitseis P.V Paragraaf 3.3.2</i>	Normering			
P.V.1		○	○	○
P.V.2		○	○	○
P.V.3		○	○	○
P.V.4		○	○	○
P.V.5		○	○	○
P.V.6		○	○	○
P.V.7			○	
P.V.8		○	○	○
<i>Kwaliteitseis P.VI Paragraaf 3.3.3</i>	Betrouwbaarheid			
P.VI.1		○	•	•
P.VI.2		○	•	•
P.VI.3		○	•	•
<i>Kwaliteitseis P.VII Paragraaf 3.3.4</i>	Constructvaliditeit			
P.VII.1		○	○	○
P.VII.2		○	○	○
P.VII.3		○	•	•
P.VII.4		○	○	○
P.VII.5		○	•	•
<i>Kwaliteitseis P.VIII Paragraaf 3.3.5</i>	Volgaspect of signaleringsfunctie			
P.VIII.1		○	•	•
P.VIII.2		○	•	•
P.VIII.3		○	•	•

Organisatorische aspecten

<i>Kwaliteitseis B.I Paragraaf 4.1</i>	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	○	○	○	○
B.I.7	○	○	○	○

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 niet voldoende zijn.

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

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 checks plaats. De aanbieder is gehouden de informatie aan het CvTE te verstrekken die het nodig heeft om een tussentijdse check uit te voeren.

Een tussentijdse check kan resulteren in het oordeel 'voldoende' of 'onvoldoende'. Bij een oordeel 'onvoldoende' volgt intrekking van de erkenning.

De volgende drie scenario's omschrijven de noodzakelijkheid voor een tussentijdse check:

a) Geen wijzigingen

Voor een instrument dat in tien jaar niet is gewijzigd, vindt vijf jaar na het afgeven van een kwaliteitsoordeel een tussentijdse check plaats. De aanbieder wordt hierbij gevraagd om aan te tonen dat het instrument nog steeds aan alle kwaliteitseisen uit het kader voldoet. Dit houdt in dat het instrument aansluit bij de actuele eisen en standaarden, en dat de items, observatiecategorieën en normering nog steeds actueel en passend zijn.

De volgende drie vragen dienen als leidraad voor de verantwoording van de beoordelingscriteria tijdens de tussentijdse check:

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)?

Dit kan onderzocht worden door middel van een inhoudsanalyse van de opgaven.

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?
- Meet het instrument nog steeds op een betrouwbare en valide manier de beoogde vaardigheden van leerlingen?

3. Sluiten de beveiligingsmaatregelen aan bij de actuele richtlijnen?

Indien nodig wordt voor een tussentijdse check zonder wijzigingen in de toelichtingen van bepaalde eisen aanvullende informatie gegeven.

b) 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 checks. Dit zou bijvoorbeeld bij jaarlijkse verversing van items in een itembank het geval kunnen zijn.

c) 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 check plaats of dient de aanbieder de instrumenten (deels) opnieuw in.

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.I.1	Is er een bij de inhoud van het instrument passende, gedegen en complete verantwoording, inclusief leeswijzer waarin is onderbouwd hoe de visie van de aanbieder op de ontwikkeling van het kind is geborgd in de actuele pedagogische en/of psychologische wetenschappelijke literatuur?	ja/nee
O.I.2	Is er een actuele inhoudelijke verantwoording waarin gedegen en compleet is onderbouwd hoe de visie van de aanbieder op de ontwikkeling van het kind is geoperationaliseerd in te observeren ontwikkelingsdoelen?	ja/nee
O.I.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.I.4	Voltoet 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.I.5	Is de doelgroep van het instrument gedegen en compleet gedefinieerd?	ja/nee
O.I.6	Is het gebruiksdoel van het instrument gedegen en compleet gedefinieerd?	ja/nee

Toelichting bij O.I.1:

De aanbieder beschrijft ten minste op passende, gedegen en complete wijze zijn visie op de ontwikkeling van het 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.I.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 kind.

Toelichting bij O.I.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.I.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.

Toelichting bij O.I.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.I.6:

Er is in de verantwoording aangegeven dat het gebruiksdoel het observeren, registreren en, indien van toepassing, signaleren van de individuele ontwikkeling van het 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 leerling een rapport samen met daarin een gedegen en compleet overzicht van de in ieder geval vaker dan één keer door de leerkracht geregistreerde observaties?	ja/nee
O.II.2	Bevat het individuele rapport per 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 geregistreerde indicatoren geïnterpreteerd dienen te worden, zodat leerkrachten de ontwikkeling van de leerling volledig en duidelijk kunnen begrijpen?	ja/nee/n.v.t.
O.II.4	Biedt de aanbieder een gedegen en complete toelichting over hoe de geregistreerde indicatoren geïnterpreteerd dienen te worden, zodat leerkrachten de ontwikkeling van de leerling op een begrijpelijke en toegankelijke manier kunnen uitleggen aan ouders, verzorgers en voogden?	ja/nee/n.v.t.
O.II.5	Biedt de informatie in het individuele rapport een eenduidige en gedegen basis voor de leerkracht om de ontwikkeling van de leerling te evalueren en, waar nodig op basis daarvan passende 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 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 leerling. Het rapport maakt het op deze manier voor de leerkracht mogelijk om observaties van een 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 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 geregistreerde 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 leerling door de leerkrachten.

Toelichting O.II.4:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, dient er een toelichting te worden geboden waarin begrijpelijke handvatten worden gegeven voor de interpretatie van deze leervorderingen ten behoeve van ouders, verzorgers en voogden. De aanbieder is vrij om te kiezen hoe deze informatie beschikbaar wordt gesteld; dit kan bijvoorbeeld via een leerlingrapportage, een ouderbrochure, een website of andere geschikte communicatiemiddelen. Het is belangrijk dat de gekozen manier leerkrachten in staat stelt om ouders goed en volledig te informeren, in lijn met de informatieplicht.

Toelichting O.II.5:

De informatie in het individuele rapport dient dusdanig concreet en gedetailleerd te zijn, dat het duidelijk is welk verloop in de ontwikkeling van de leerling aandacht verdient. Het ondersteunt de leerkracht in het zelfstandig en professioneel 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? Of is er expliciet benoemd voor welke doelgroep(en) het instrument niet geschikt is?	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. De aanbieder onderbouwt het meetdoel op gedegen wijze en

verwijst naar actuele en relevante brondocumenten ter onderbouwing van het meetdoel.

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 toetsmatrijs een adequate, schematische representatie van het meetdoel?	ja/nee
O.IV.2	Voldoen alle toetsopgaven in het instrument aan de kwaliteitscriteria voor toetsopgaven: relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit?	ja/nee
O.IV.3	Voldoen de toetsopgaven aan de constructievoorschriften passend bij het vraagtype?	ja/nee

Toelichting O.IV.1:

De toetsmatrijs 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;
- een aanbieder aantoont in welke mate het instrument type (III) vanaf groep 6 aansluit bij de referentieniveaus voor de onderdelen Nederlandse taal en Rekenen;
- de toetsmatrijs 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, inclusief een verantwoording of toelichting die inzicht biedt in de gemaakte keuzes bij deze verdeling;
 5. de toegestane hulpmiddelen;
 6. en in het geval van een niet adaptieve toets: de toetsduur.

Toelichting O.IV.2:

De vijf kwaliteitscriteria relevantie, objectiviteit, efficiëntie, specificiteit en neutraliteit staan beschreven in de Checklist, deze wordt gebruikt voor het beoordelen van de kwaliteit van observatie-categorieën en toetsopgaven. Alle toetsopgaven dienen hieraan te voldoen.

Toelichting bij O.IV.3:

De constructievoorschriften voor toetsvragen staan beschreven in de Checklist voor het beoordelen van de kwaliteit van observatie-categorieën en toetsopgaven. Alle toetsopgaven dienen hieraan te voldoen. Aanvullend beschrijft de aanbieder het constructieproces van de toetsopgaven.

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

Deze paragraaf beschrijft de kwaliteitseisen voor referentieniveaus die van toepassing zijn op instrumenten van het type (III).

In het Referentiekader Taal en Rekenen (SLO, 2011) is vastgelegd wat leerlingen aan het einde van groep 8 van de basisschool moeten beheersen op het gebied van Nederlandse taal en Rekenen. Deze basisvaardigheden zijn verdeeld over een aantal referentieniveaus. De niveaus zijn beschreven in twee kwaliteiten: fundamenteel (F) en streef (S). De niveaus zijn cumulatief: beheersing van een hoger niveau impliceert beheersing van de vaardigheden van de lagere niveaus. De beschrijving van de niveaus verschilt tussen Nederlandse taal en Rekenen vanwege de verschillen tussen de vakken.

Voor het primair onderwijs zijn de volgende referentieniveaus wettelijk vastgelegd:

- voor Nederlandse taal: 1F (fundamenteel) en 2F (streef);
- voor Rekenen: 1F (fundamenteel) en 1S (streef).

LVS-instrumenten van het type (III) maken deel uit van de doorlopende, methodevrije leerlijn richting

de doorstroomtoets in groep 8. Omdat de doorstroomtoetsen zijn gekoppeld aan referentieniveaus, is het van belang dat LVS-instrumenten vanaf (eind) groep 6 aansluiten bij deze niveaus. Ongeacht de toetsvorm moeten de instrumenten in staat zijn om vanaf (eind) groep 6 per leerling uitspraken te doen over de beheersing van het niveau 1F en/of 2F (voor Taal) en 1F en/of 1S (voor Rekenen). Dit geeft inzicht in de voortgang van de leerling richting het eindniveau van groep 8.

Code	Vragen	Mogelijke antwoorden
O.V.1a	Worden de wettelijk verplichte inhouden van de terreinen Nederlandse taal en/of Rekenen, en indien van toepassing de optionele inhouden van Taal, evenwichtig getoetst?	ja/nee
O.V.1b	Zijn de opgaven van de toets voor de wettelijk verplichte inhouden van Nederlandse taal en/of Rekenen, en indien van toepassing de optionele inhouden van Taal, evenwichtig samengesteld qua referentieniveaus 1F en/of 2F/1S?	ja/nee
O.V.2a	Indien er sprake is van een CAT (adaptieve toets op itemniveau), voldoet de itembank van de verzameling van LVS-instrumenten van het type (III) van (eind) groep 6 tot en met 8 aan de kwaliteitseis O.V.1a en O.V.1b?	ja/nee/n.v.t.
O.V.2b	Indien er sprake is van een MST (adaptieve toets op moduleniveau), voldoet het design van de verzameling van LVS-instrumenten van het type (III) van (eind) groep 6 tot en met 8 aan de kwaliteitseis O.V.1a en O.V.1b?	ja/nee/n.v.t.

Toelichting O.V.1a:

Dit aspect moet uit de toetsmatrijs naar voren komen. Evenwichtig wil zeggen dat de verdeling van de vragen over de domeinen, subdomeinen, onderdelen en kenmerken van de taakuitvoering evenredig verdeeld is.

Toelichting O.V.1b:

Dit aspect moet uit de toetsmatrijs naar voren komen. Dat wil zeggen dat de verhouding 1F en/of 2F/1S vragen in de toetsmatrijs is gespecificeerd en dat de verhouding van de vragen erin wordt benoemd. Evenwichtig wil zeggen dat het aantal en de verhouding 1F en/of 2F/1S vragen gezamenlijk inhoudelijk de referentieniveaus 1F en/of 2F/1S van de terreinen Nederlandse taal en Rekenen representeren.

Toelichting O.V.2a:

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

Toelichting O.V.2b:

Indien er sprake is van een MST, dient elke routing binnen het MST design voor elke individuele leerling voor de terreinen Taal en Rekenen in de periode groep 6-8 een uitspraak over het behaalde niveau 1F en/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	Biedt de aanbieder een gedegen en begrijpelijke toelichting bij het gemeten niveau en de leervorderingen van de leerling, zodat leerkrachten deze adequaat kunnen interpreteren en toepassen?	ja/nee
O.VI.4	Stelt de aanbieder begrijpelijk en toegankelijk informatie aan leerkrachten beschikbaar over het gemeten niveau en de leervorderingen van de leerling, zodat zij ouders, verzorgers, en voogden voldoende en volledig kunnen informeren?	ja/nee
O.VI.5	Is de informatie in het leerlingrapport een voldoende 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, die de leerkracht duidelijke en begrijpelijke

handvatten biedt voor de interpretatie van de leervorderingen van de leerling. Deze toelichting moet de leerkracht in staat stellen de voortgang van de leerling goed te begrijpen en toe te passen in de begeleiding.

Toelichting O.VI.4:

Er dient een geschreven toelichting te worden geboden waarin begrijpelijke handvatten worden gegeven voor de interpretatie van de leervorderingen van de leerling ten behoeve van ouders, verzorgers en voogden. De aanbieder is vrij om te kiezen hoe deze informatie beschikbaar wordt gesteld; dit kan bijvoorbeeld via een leerlingrapportage, een ouderbrochure, een website of andere geschikte communicatiemiddelen. Het is belangrijk dat de gekozen manier leerkrachten in staat stelt om ouders goed en volledig te informeren, in lijn met de informatieplicht.

Toelichting O.VI.5:

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 of gewogen naar betekenisvolle achtergrondvariabelen, representatief is voor de indicatoren van observeerbare (gedrags-)kenmerken, en is verzameld onder omstandigheden die redelijk vergelijkbaar zijn met de omstandigheden waarin de gegevens in een operationele setting zouden worden verzameld. Indien van toepassing, wordt de aanbieder gevraagd om effectgroottes aan te leveren ten behoeve van de gemaakte vergelijking.

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.

Het volgen van de ontwikkeling van de leerling met betrekking tot de verschillende ontwikkelingsdoelen en eventuele tussentijdse mijlpalen heeft nadrukkelijk niet als doel om waardeoordeel over het niveau van de 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 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 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, onderbouwt de aanbieder de selectieprocedure op gedegen wijze, 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. De aanbieder beargumenteert waarom het aantal experts voor de gekozen methode als voldoende wordt beschouwd.

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

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 overeenstemming waarbij de proportie overeenstemming representatief is voor een instrument voor minder belangrijke beslissingen (een zogenaamd medium stakes instrument) op individueel niveau. De toetsaanbieder dient de gehanteerde methode en gekozen grenswaarden te verklaren en deze te onderbouwen aan de hand van relevante literatuur.

3.2.3 Interbeoordelaarsbetrouwbaarheid

Deze paragraaf beschrijft de kwaliteitseisen voor de interbeoordelaarsbetrouwbaarheid. 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 interpretatie van de observatie / resultaten. De toetsaanbieder dient de gehanteerde methode en gekozen grenswaarden te verklaren en deze te onderbouwen aan de hand van relevante literatuur.

Toelichting P.III.2:

Indien de ontwikkelingsdoelen in het instrument zijn verbijzonderd in inhoudelijk bijpassende indicatoren van observeerbare (gedrags-)kenmerken, verduidelijkt de aanbieder dat de observatoren / leerkrachten die zijn ingezet om de interbeoordelaarsovereenstemming te schatten, op dezelfde wijze zijn geïnstrueerd en/of getraind als beoordelaars in de praktijk (bijv. d.m.v. de handleiding voor leerkrachten).

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).

Fasen psychometrische analyse	Toelichting	Kwaliteitseisen
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 volgens onderstaande kwaliteitseisen.

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.

Voor een tussentijdse check dient de aanbieder te onderzoeken of er vaardigheidsverschillen zijn tussen de huidige steekproef van het instrument en de steekproef waarop het instrument oorspronkelijk is gebaseerd. Daarnaast moet de aanbieder aantonen dat er geen sprake is van differential item functioning (DIF) tussen de initiële afname en de huidige afname van het instrument. Als er geen vaardigheidsverschillen zijn en geen DIF is tussen de afnames, kan worden aangenomen dat de psychometrische kwaliteit van het instrument in de tussentijd niet is veranderd en dat het instrument aan de gestelde eisen voldoet. Indien er wel vaardigheidsverschillen worden gevonden en/of DIF is tussen de afnames, moet de aanbieder op basis van de onderstaande psychometrische eisen aantonen dat het instrument alsnog voldoet aan de gestelde eisen.

Code	Vragen	Mogelijke antwoorden
P.IV.1	Is de steekproef van leerlingen groot genoeg voor het gekozen model en representatief voor de doelpopulatie?	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 dient groot genoeg te 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 itemrespons theorie (1PLM of 2PLM).

Voor alle genoemde pretestmogelijkheden en toetsvormen geldt in ieder geval dat de steekproef representatief is voor de doelgroep (c.q. de landelijke populatie leerlingen). De steekproef dient beschreven te worden op basis van de variabelen en niveaus zoals beschreven in de onderstaande tabel en indien nodig te worden gestratificeerd/gewogen.

Variabele	#N	Welke niveaus (NB: #N = aantal niveaus)
regio	4	noord, zuid, oost, west
stedelijkheid	2	verstedelijkt, landelijk ¹
schoolgrootte	3	aantal leerlingen: <100, 100-300, >300 leerlingen
schoolweging	5	gewichtscategorieën: <23 / 23-26,99 / 27-32,99 / 33-36,99 / ≥ 37

¹ De mate van verstedelijking is terug te voeren op de bij het CBS gebruikelijke indeling naar vijf niveaus, namelijk zeer sterk, sterk, matig, weinig en niet verstedelijkt. De vijfdeling is hier teruggebracht naar een tweedeling in enerzijds verstedelijkt (zeer sterk en sterk) en anderzijds landelijk (matig, weinig en niet verstedelijkt) die in de praktijk goed lijkt te volstaan (c. f. van Boxtel en Hemker, 2009). De toetsaanbieder kan ervoor kiezen om van deze indeling af te wijken, mits deze keuze voldoende is onderbouwd.

Tevens dient er informatie geleverd te worden over hoe de gerealiseerde steekproef zich verhoudt tot de landelijke populatiewaarden met betrekking tot regio, stedelijkheid, schoolgrootte en schoolweging. Indien van toepassing, wordt de aanbieder gevraagd om effectgroottes aan te leveren ten behoeve van de gemaakte vergelijking.

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

Bij het pretesten van nieuwe items in een proeftoets of in een operationele toets bij gebruik van KTT is een steekproefgrootte van minimaal 200 observaties van elk pretest item vereist. Voor het pretesten van nieuwe items in een proeftoets of in een operationele toets bij gebruik van een 1PL is een steekproefgrootte van minimaal 300 observaties van elk pretest item vereist bij gebruik van een 2PL model en minimaal 600 observaties.

Wanneer de LVS-instrumenten worden ontwikkeld voor een specifiek vaardigheidsschaal, dan geldt het minimumaantal observaties per pretest item voor die specifieke schaal.

Aanbieders mogen van de boven genoemde aantallen afwijken, mits dit gedegen en volledig wordt onderbouwd.

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 verbonden 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 verbonden 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 verbonden 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 als volgt hoe de passing van het IRT model is geëvalueerd:
 - de aanbieder toont aan dat de itemparameters tussen verschillende afnamen of boekjes niet veranderen, en;
 - de aanbieder voert een DIF analyse op itemniveau uit op minimaal de achtergrondvariabel schoolweging, 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 literatuur. De aanbieder wordt verzocht ook effectgroottes te vermelden.
 - De aanbieder vermeldt effectgroottes waarop de modelpassing is geëvalueerd (berekend als het gemiddelde absolute verschil tussen de geobserveerde en verwachte proporties juiste

antwoorden, gegeven het IRT-model). De aanbieder toont aan dat dit verschil op itemniveau $\leq 10\%$. De toetsaanbieder mag hier met de juiste motivering van afwijken.

- de toetsaanbieder onderbouwt de modelpassing aan de hand van een grafische weergave van de modelfit en de geobserveerde scores van alle items.
- De aanbieder wordt verzocht om informatie over de scoreverdeling aan te leveren, zodat inzichtelijk is welke informatiewaarde (en omgekeerd de standard error) correspondeert met een klassieke betrouwbaarheidscoëfficiënt van 0,70.
- 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 de aanbieder duidelijk verantwoorden 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 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
Bepalen van absolute normen		
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.
Vaststellen van relatieve normen		
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.
Linken met referentieniveaus		
P.V.7	Is de toets op gedegen manier verbonden aan passende referentieniveaus?	ja/nee/n.v.t.
Andere methodes		

Code	Vragen	Mogelijke antwoorden
P.V.8	Als naast/in plaats van de bovengenoemde methoden (eisen P.V.1 t/m P.V.7) ook andere normeringsmethoden worden gebruikt, zijn deze methodes goed omschreven, voldoende onderbouwd en correct uitgevoerd?	ja/nee/n.v.t.

Toelichting P.V.1:

De toetsaanbieder verantwoordt de kwaliteit van de gekozen standaard- of cesuurbepalingsmethode op gedegen en complete wijze en onderbouwt dat de methode correct is uitgevoerd door de erbij betrokken experts met vakdeskundigheid.

Toelichting P.V.2:

De toetsaanbieder onderbouwt de selectieprocedure op gedegen en complete wijze; er zijn voldoende experts met aantoonbaar inhoudelijke vakdeskundigheid geselecteerd, zij hebben een gedocumenteerde training gevolgd in de betreffende standaardbepalingsmethode.

Toelichting P.V.3:

De toetsaanbieder toont aan dat de mate van overeenstemming tussen de experts resulteert in een coëfficiënt voor overeenstemming 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 toetsaanbieder toont aan dat 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 regio, stedelijkheid, schoolgrootte en schoolweging is beschikbaar en er is gebruik gemaakt van een aselekt steekproefmodel. Als er geen gebruik is gemaakt van een aselekt steekproefmodel dient de aanbieder dit op juiste en volledige wijze te onderbouwen.

Toelichting P.V.6:

De toetsaanbieder onderbouwt op gedegen en complete wijze de betekenis en beperkingen van de normschaal. Ook toont de aanbieder aan dat het type normschaal in overeenstemming is met het doel van het LVS-instrument van het type (III) of (IV).

Toelichting P.V.7:

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. De toetsaanbieder verantwoordt op gedegen en complete wijze dat de cesuren correct zijn overgebracht.

Toelichting P.V.8:

Het beoordelingskader LVS laat ruimte voor alternatieve vormen van normeren, zoals bijvoorbeeld continue normeren waarbij leerlingen uit een lagere groep bij de normering ook een 'moeilijkere' toets maken voor een hoger leerjaar. Alternatieve vormen van normeren dienen inhoudelijk en psychometrisch te worden onderbouwd door de toetsaanbieder.

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.

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).

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 (III) of (IV)?	ja/nee

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 misclassificaties in verantwoordingen van andere LVS-toetsen en doorstroomtoetsen voor het primair onderwijs (zie ook P.VI.3).

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) en dient deze te duiden door middel van bijvoorbeeld de plus-minus 1 niveau-index (Pilliner 1969) of de marginal classification accuracy. De aanbieder legt uit waarom voor een bepaalde statistische methode en de gehanteerde grenzen is gekozen.

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/n.v.t.
P.VII.2	Is aangetoond dat de itemkwaliteit op orde is?	ja/nee/n.v.t.

Code	Vragen	Mogelijke antwoorden
P.VII.3	Is de convergente en divergente validiteit van het instrument door onderzoek aangetoond?	ja/nee
P.VII.4	Is aangetoond dat er geen sprake is van item bias voor relevante subgroepen/categorieën?	ja/nee/n.v.t.
P.VII.5	Zijn verschillen in prestaties van relevante subgroepen voldoende onderbouwd?	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 bijvoorbeeld latente correlaties tussen verschillende inhoudelijke subvaardigheden of een confirmatieve factoranalyse gekoppeld aan multidimensionale betrouwbaarheidsindices². Het gebruik van IRT biedt daarnaast de mogelijkheid om met multidimensionele 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. De aanbieder onderbouwt de gekozen grenswaarden voor de p- en rit-waarden op een juiste en gedegen wijze.

Toelichting P.VII.3:

De aanbieder laat op basis van onderzoek zien dat het te meten onderdeel hoog correleert met andere toetsen die hetzelfde construct meten (convergente validiteit) en laag correleert met andere toetsen die een ander construct meten (divergente validiteit). Indien het aantonen van convergente validiteit niet mogelijk is, bijv. als er geen vergelijkbare instrumenten beschikbaar zijn of als 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 om de convergente validiteit te onderzoeken. 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.

Toelichting P.VII.4:

In het kader van itembias is er onderzoek uitgevoerd naar differentieel item functioneren (DIF) voor relevante subgroepen en/of categorieën, zoals bijvoorbeeld afnamevormen (papier/digitaal).

Als er sprake is van een tussentijdse check dient de aanbieder onderzoek naar DIF uit te voeren tussen de afnames van de initiële indiening en de tussentijdse check om aan te tonen dat de items nog hetzelfde functioneren.

Toelichting P.VII.5:

De prestaties van relevante subgroepen is onderzocht door de gemiddelde score en standaarddeviatie per subgroep te bepalen. Tevens dienen de effectgroottes te worden toegevoegd, waarbij geldt dat een klein effect < 0.2 voldoende is. Op gedegen en complete wijze onderbouwt de aanbieder de keuze van de relevante subgroepen en verantwoordt eventuele verschillen op basis van theoretische inzichten en/of eerder onderzoek.

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

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

(IV) tot drie criteria met betrekking tot de schaal waarop groei wordt uitgedrukt: (1) de opbouw van de schaal, (2) de betrouwbaarheid van de metingen, 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 groei van een leerling nauwkeurig en betrouwbaar gemeten?	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/of 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 alle 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 aanbieder de schaal waarop de groei wordt weergegeven grondig en compleet onderbouwt.

De aanbieder dient ook aan te geven of het instrument deel uitmaakt van een overkoepelend LVS (bijvoorbeeld, als alleen de toetsen voor groepen 3 tot en met 5 zijn ingediend voor een leerlingvolgsysteem dat bedoeld is voor groepen 3 tot en met 8). Als dit het geval is, dient de aanbieder duidelijk en volledig te onderbouwen hoe de schalen van de verschillende toetsen binnen het overkoepelend LVS aan elkaar zijn gekoppeld.

Toelichting P.VIII.2:

Voor een LVS-instrument van het type (III) of (IV) is het van belang om een indicatie van de betrouwbaarheid van die gevolgtrekking weer te geven. Hiervoor is het belangrijk om te bepalen of de focus ligt op individuele verandering of op verschillen tussen personen op groepsniveau.³ Wanneer het instrument bedoeld is voor het nemen van beslissingen op individueel niveau, dient de aanbieder de meetprecisie van de verschillende meetmomenten in kaart te brengen en de betrouwbaarheid van de verschillen (reliable change index ³ en⁴) te berekenen en de gekozen methode te onderbouwen. Als het instrument wordt gebruikt om de groei van leerlingen relatief te positioneren, bijvoorbeeld ten opzichte van een gemiddelde groeiscore, moet de aanbieder ook de betrouwbaarheid van de groeimetingen berekenen en deze zorgvuldig onderbouwen.⁵

Bij gebruik van percentielscores moet de betrouwbaarheid van de verandering van de percentielscores geschat zijn. Binnen de CTT-methodologie zijn percentielscores gebaseerd op een somscore die een meetfout bevat. Het verschil in percentielscores kan vertaald worden in het verschil in onderliggende scores en hun betrouwbaarheid. Bij IRT zijn de percentielscores gebaseerd op vaardigheidsscores die geschat zijn met een zekere mate van betrouwbaarheid.⁶ Omdat de transformatie van ruwe

³ Zie bijvoorbeeld: Gu, Z., Emons, W.H.M. & Sijsma, K. (2018). Review of Issues About Classical Change Scores: A Multilevel Modeling Perspective on Some Enduring Beliefs. *Psychometrika* 83, 674–695. <https://doi.org/10.1007/s11336-018-9611-3>

⁴ Zie bijvoorbeeld: Jacobson, N.S., & Truax, P. (1991) Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19. <https://psycnet.apa.org/record/1992-97624-042>

⁵ Zie bijvoorbeeld: Trafimow, D. (2015) A defense against the alleged unreliability of difference scores, *Cogent Mathematics*, 2:1, 1064626, DOI: 10.1080/23311835.2015.1064626

⁶ Zie bijvoorbeeld: May, K., & Nicewander, W. A. (1994). *Reliability and information functions for percentile ranks*. *Journal of Educational Measurement*, 31(4), 313-325.

scores of theta scores naar percentielscores niet lineair is, kan dit wel resulteren in een zekere bias⁷.

Toelichting P.VIII.3:

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/of 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 meetprecisie van de resultaten en de meetpretentie van het instrument als geheel.

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).

Indien de aanbieder ISO 27001 is gecertificeerd, volstaat het om dit certificaat inclusief een verklaring van toepasbaarheid (c.q. de scope van het certificaat) aan te leveren. Indien voorgaande wordt aangeleverd, zijn onderstaande kwaliteitseisen voldoende verantwoord.

Code	Vragen	Mogelijke antwoorden
B.I.1	Heeft de aanbieder een risico-analyse omtrent informatiebeveiliging en privacy uitgevoerd bij het gebruik van het LVS-instrument?	ja/nee/n.v.t.
B.I.2	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/n.v.t.
B.I.3	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/n.v.t.
B.I.4	Waarborgt de aanbieder op gedegen en complete wijze de privacy van leerkracht, leerling en ouder conform de Algemene Verordening Gegevensbescherming?	ja/nee/n.v.t.
B.I.5	Hanteert de aanbieder op gedegen en complete wijze incidentmanagement inclusief een protocol datalekken, zoals voorgeschreven door de Autoriteit Persoonsgegevens?	ja/nee/n.v.t.
B.I.6	Heeft de aanbieder duidelijke richtlijnen (in het toetsreglement/ verantwoordingsdocumenten) opgesteld voor bij de afname betrokkene personen (zoals de schoolleider en leerkracht) over de veilige opslag en het veilige gebruik van het instrument en de daarmee verzamelde gegevens?	ja/nee/n.v.t.
B.I.7	Heeft de aanbieder een protocol (in het toetsreglement / verantwoordingsdocumenten) opgesteld voor een regelmatige controle of de inhoud van het LVS-instrument of andere niet-openbare informatie over het LVS-instrument uitgelekt/openbaar geworden is, bijv. door regelmatige controle van het internet en andere media?	ja/nee/n.v.t.

Toelichting B.I.1 t/m B.I.7:

Indien een aanbieder niet ISO 27001-gecertificeerd is, kan hij voldoen aan de kwaliteitseisen door bijvoorbeeld te verwijzen naar een verwerkersovereenkomst. Deze overeenkomst moet ingaan op de eisen B.I.1 tot en met B.I.7. De aanbieder moet daarbij bewijsmateriaal aanleveren voor de gevraagde documenten, contracten en protocollen bij B.I.2, B.I.3, B.I.4, B.I.5 en B.I.6. Dit bewijsmateriaal kan bestaan uit afzonderlijke documenten of specifieke hoofdstukken in het beleid die duidelijk maken hoe aan de eisen wordt voldaan.

Voor B.I.6 moet de aanbieder een beschrijving leveren van het gebruik van het instrument in verschillende afname-omstandigheden (zoals papier, digitaal en observatie). Dit helpt de schoolleider en leerkracht om het instrument correct toe te passen en te voldoen aan de eisen voor beveiliging en privacy.

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.

⁷ 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.



TOELICHTING

Op 7 november 2022 heeft het College voor toetsen en examens het beoordelingskader voor de toetsen en beoordelingsinstrumenten die onderdeel uitmaken van leerling- en onderwijsvolgsystemen in het primair onderwijs vastgesteld.

De ervaringen die opgedaan zijn met dit kader leidden tot de volgende conclusie dat het beoordelingskader op een aantal punten aangepast moet worden. Deze aanpassingen zien met name op verheldering van reeds bestaande criteria en begrippen. Ook is de structuur van het kader verbeterd en is er een aantal criteria toegevoegd voor de toetsconstructie- en validiteit. Tevens zijn psychometrische criteria voor betrouwbaarheid en normering van toetsen aangescherpt. Daarnaast introduceert het kader ook specifieke richtlijnen voor de periodieke evaluatie en bijstelling van toetsen.

De omvang van de aanpassingen is zodanig dat het beoordelingskader in zijn geheel is vervangen.

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