Expertgroep Toetsen PO

Beoordelingskader voor instrumenten binnen leerlingvolgsystemen (LVS)

Ontwikkeld door:	*Psychometrisch experts,* Hans Vos Arnold Brouwer Bernard Veldkamp Piet Sanders *SLO* *Expertgroep Toetsen PO*, Cees van der Vleuten Cees Glas Desirée Joosten-Ten Brinke
Vastgesteld op:	01-09-2020
Referentienummer:	20.02

Inhoud

1. Inleiding

De Expertgroep Toetsen PO (hierna: Expertgroep) heeft op grond van artikel 8, lid 6 van de Wet op het primair onderwijs en artikel 2, lid 2b van haar Instellingsbesluit onder meer tot taak een eigenstandig kwaliteitsoordeel te geven over de inhoudelijke validiteit, betrouwbaarheid en deugdelijke normering van een tussentijdse toets of reeks van tussentijdse toetsen. Toetsen dienen, volgens artikel 11 van het Instellingsbesluit, uiterlijk binnen 13 weken te zijn beoordeeld door de Expertgroep.

Algemene criteria waaraan LVS-instrumenten (toetsen) dienen te voldoen

Volgens het Toetsbesluit PO geeft de Expertgroep een kwaliteitsoordeel over de inhoudelijke validiteit, betrouwbaarheid en deugdelijke normering van de toetsen uit een leerling- en onderwijsvolgsysteem (LVS). Onder het begrip 'toetsen' wordt voor leerlingvolgsystemen ook het begrip 'instrumenten' in het algemeen verstaan. Daarom wordt in dit beoordelingskader hierna het begrip 'LVS-instrumenten' gehanteerd.
Bij het kwaliteitsoordeel over LVS-instrumenten 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 vaardigheden van de leerlingen meten op de gebieden, genoemd in artikel 8, tweede lid van de WPO of artikel 11, derde lid van de WEC: de emotionele en de verstandelijke ontwikkeling, het ontwikkelen van creativiteit, het verwerven van noodzakelijke kennis en/of 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 twee volgende paragrafen wordt uiteen gezet hoe deze criteria worden geoperationaliseerd.

Beoordeling van de onderwijskundige en psychometrische aspecten van LVS-instrumenten

De hierboven genoemde kwaliteitsaspecten worden opgedeeld in onderwijskundige en psychometrische kwaliteitsaspecten.
De beoordeling van de onderwijskundige aspecten van de LVS-instrumenten richt zich op de inhoudelijke validiteit en de hierboven genoemde punten b) en c).
De beoordeling van de psychometrische aspecten van de LVS-instrumenten richt zich op de onderbouwing van de betrouwbaarheid, de deugdelijkheid van de normering, psychometrische onderbouwing van de inhoudelijke validiteit en het hierboven benoemde punt a).

De Expertgroep laat zich voor de beoordeling van de onderwijskundige en psychometrische aspecten bijstaan door een adviesgroep van externe onafhankelijke beoordelaars, die geen banden hebben met de aanbieders. De namen van de externe beoordelaars zijn openbaar, zonder dat ze gekoppeld kunnen worden aan concreet beoordeelde instrumenten.

De procedure, inclusief het tijdpad, is conform de notitie "Procesbeschrijving LVS-instrumenten" zoals die op de website van de Expertgroep (http://expertgroeptoetsenpo.nl/) gepubliceerd is.

Beoordelingskader

Het voorliggende document behelst een beoordelingskader voor de LVS-instrumenten. De criteria in het kader zijn algemeen verwoord. Ze zijn mede gebaseerd op de beoordelingssystemen van het COTAN en het RCEC en het document "Aanvullingen COTAN Beoordelingssysteem wat betreft volg-aspect van leerling- en onderwijsvolgsystemen". Voor zover een LVS-instrument een adaptief aspect heeft, is ook "Aanvullingen COTAN Beoordelingssysteem wat betreft normering referentieniveaus en computer adaptief toetsen van andere eindtoetsen" van belang. Voor achtergrondinformatie met betrekking tot de criteria voor LVS-instrumenten worden de aanbieders verwezen naar genoemde documenten, zij het alleen voor aspecten die betrekking hebben op LVS-instrumenten (links naar deze documenten onderaan dit document). Verdere relevante informatie is te vinden in de vakliteratuur (bijvoorbeeld het boek Educational Measurement, Brennan, 2006) of door raadplegen van reeds door de Expertgroep goedgekeurde LVS-instrumenten.

Het beoordelingskader richt zich op de aspecten normering, betrouwbaarheid, validiteit, volg-aspect, inzicht in de leervordering en referentieniveaus. Naast deze wettelijk vastgelegde aspecten is het aspect 'Kwaliteit van de dataverzameling' toegevoegd. De kwaliteit van de dataverzameling is onderliggend aan de kwaliteitsaspecten normering, betrouwbaarheid en het volg-aspect.
Voor de beoordeling van deze aspecten zijn enkele vragen opgesteld op basis waarvan de aspecten gescoord kunnen worden met:

- voldoende;
- voldoende, mits;
- onvoldoende.

Bij 'voldoende, mits' en 'onvoldoende' worden door de beoordelaars aanwijzingen gegeven om op het betreffende onderdeel of de betreffende onderdelen een voldoende te scoren.

De Expertgroep gebruikt de beoordeling om te komen tot een kwaliteitsoordeel over het voorliggende instrument. Voor een positief kwaliteitsoordeel dienen alle criteria met een 'voldoende' of een 'voldoende, mits' beoordeeld te zijn. In het geval van een 'voldoende, mits' worden aan de aanvullende voorwaarden voorgelegd. Indien niet aan deze voorwaarden wordt voldaan, wordt het oordeel 'onvoldoende'.

Geldigheidsduur kwaliteitsoordelen Expertgroep

Een kwaliteitsoordeel dat op basis van een beoordeling over een instrument wordt afgegeven, is vanaf de datum van afgifte tien jaar geldig. Binnen de termijn van tien jaar vindt, afhankelijk van de aard van het instrument, tenminste één tussentijdse beoordeling plaats. Meer informatie over de tussentijdse check(s) vindt u in hoofdstuk 9 van dit beoordelingskader.
Na de termijn van tien jaar is voor gecontinueerde inzet van het betreffende instrument een nieuwe beoordeling, volgens het dan geldende beoordelingskader, en een nieuw kwaliteitsoordeel nodig.

Overige bepalingen m.b.t. LVS-instrumenten bij kleuters

- Voor kleuters wordt idealiter geen gebruik gemaakt van schoolse toetsen, maar kan wel gebruik gemaakt worden van observatie-instrumenten.
- Taken die aan leerlingen worden voorgelegd, worden aangeboden in een dagelijkse schoolpraktijk, zonder tijdsdruk, waarbij er voor de leerling geen merkbaar onderscheid is tussen leren, spelen en geobserveerd worden.
- Observatie-instrumenten dienen genormeerd te zijn en inhoudelijke, diagnostische informatie op te leveren over de ontwikkeling van een kleuter. Zij dienen goedgekeurd te zijn door de Expertgroep.
- In de nieuwe wet PO (met daarin het nieuwe Toetsbesluit), die waarschijnlijk vanaf 1 augustus 2022 van kracht wordt, is bepaald dat bij kleuters in het kader van LVS geen schoolse toetsen mogen worden afgenomen. Vóór die datum mogen schoolse LVS-toetsen nog wel bij kleuters worden afgenomen.

2. De kwaliteit van de dataverzameling

Code	*Vraag*
S1	Is de steekproef van leerlingen representatief?
S2	In het geval van een onvolledig dataverzamelingsdesign: is het design adequaat?
S3	In het geval van een observatie-instrument: is er sprake van een adequate steekproef van observatoren en randvoorwaarden waaronder de observatie wordt uitgevoerd.
S4	Er is een handleiding met duidelijke instructies voor de leerkracht over het zo objectief mogelijk uitvoeren en weergeven van de observaties door de leerkracht.

Toelichting

De kwaliteit van de normering van een instrument en de betrouwbaarheid van de beslissingen die op basis van de genormeerde scores genomen worden, hangt vooral af van de kwaliteit van de dataverzameling waarop de normering en de betrouwbaarheidsgegevens gebaseerd zijn. Daarom wordt de kwaliteit van de dataverzameling eerst behandeld.

Bij S1: De steekproef moet representatief zijn voor de doelgroep (c.q. de populatie leerlingen) in termen van het onderwijsniveau. De steekproef moet adequaat gestratificeerd zijn naar (tenminste) geslacht, regio en urbanisatiegraad, en er moet informatie geleverd worden over hoe de gerealiseerde steekproef zich verhoudt tot de populatie waarden m.b.t. geslacht, regio (minimaal de indeling Noord, Oost, West en Zuid), urbanisatiegraad. De procedure voor het samenstellen van de steekproef moet onderbouwd zijn en de omstandigheden waaronder de data verzameld zijn moeten redelijk vergelijkbaar zijn met de omstandigheden waaronder het instrument wordt afgenomen.

Bij S2: Als een toets of eventueel een itembank uit afzonderlijke items bestaat, worden data vaak verzameld in een onvolledig design, waarbij niet alle leerlingen alle items maken. Men spreekt vaak van een boekjesdesign. Vervolgens worden normering en betrouwbaarheid berekend met een Item Respons Theorie (IRT) model. Naast de vigerende eisen voor de betrouwbaarheid van schatting van IRT parameters, zijn er eisen aan het design. De boekjes in het kalibratiedesign moeten voldoende ‘gelinked’ zijn, dat wil zeggen dat er voldoende overlap in observaties is tussen de verschillende items en

boekjes. Het is niet doenlijk hier alle mogelijke designs te behandelen. Voor een ankeritem design geldt de eis dat het anker tenminste uit 15 ankeritems (items die het design linken) moet bestaan en dat de IRT parameters van de items representatief zijn voor alle items in het design. Voor andere designs geldt dat het design zodanig moet zijn dat de schattingsfouten vergelijkbaar zijn aan een ankeritem design met tenminste 15 items. Verder moeten de ankeritems voldoende representatief zijn voor het te meten domein. Tenslotte moet er evidentie voor de passing van het IRT model worden gepresenteerd. Overigens zijn andere kalibratie-methoden (bijvoorbeeld kernel-equating van von Davier en Holland) ook toegestaan zolang de betrouwbaarheid analoog is aan de voor IRT geformuleerde eisen.

Bij S3: Niet ieder LVS-instrument bestaat noodzakelijkerwijs uit items. Zo kunnen observatie-instrumenten bestaan uit een verzameling observatierubrieken. Behalve dat de dataverzameling voor de normering en bepaling van de betrouwbaarheid moet plaatsvinden bij een representatieve steekproef leerlingen, moeten ook de gebruikte observatoren een goede afspiegeling zijn van de observatoren die het instrument in de praktijk gaan gebruiken, en ook de randvoorwaarden waaronder de observaties worden gemaakt moet een goede afspiegeling zijn van de praktijk waarin het instrument gebruikt gaat worden. Een proefonderzoek met uiterst getrainde onderwijskundigen in een laboratoriumsituatie geldt bijvoorbeeld niet als zodanig. Verder zijn bij het proefonderzoek meestal meerdere observatoren betrokken. Daardoor is ook hier weer sprake van een onvolledig design voor de dataverzameling. De eisen die men kan stellen aan het design zijn vergelijkbaar met de onder S2 genoemde eisen. Zo moet de samenhang in het design zo zijn dat alle relevante parameters die nodig zijn voor de normering en het evalueren van de betrouwbaarheid voldoende betrouwbaar geschat kunnen worden.

Bij S4: De leerkrachten voeren de observaties zelfstandig uit. Om ervoor te zorgen dat alle leerkrachten de observaties op dezelfde wijze uitvoeren, dient een eenduidige handleiding aanwezig te zijn.

3. Normering

Er is onderscheid te maken tussen absoluut en relatief normeren. Kenmerkend voor absoluut normeren is dat de normen van het instrument voor de afname worden bepaald. De absolute norm is gebaseerd op een minimaal acceptabel beheersingsniveau van een nauwkeurig omschreven leerstofdomein. Bij relatief normeren wordt de norm aan de hand van afnames van het instrument bepaald. De relatieve norm is gebaseerd op een onderlinge vergelijking van de prestaties van de kandidaten met de prestaties van een nauwkeurig omschreven referentiepopulatie.

Code	*Vraag*
N1	**ABSOLUTE NORMEN**
N1.1	Is de standaardbepalingsmethode gemotiveerd en op de juiste wijze uitgevoerd?
N1.2	Zijn de beoordelaars/vakdeskundigen/experts naar behoren geselecteerd en getraind?
N1.3	Is er voldoende overeenstemming tussen de beoordelaars?
N2	**RELATIEVE NORMEN**
N2.1	Zijn de normgroepen groot genoeg?
N2.2	Zijn de normgroepen representatief?
N2.3	Zijn de normen correct bepaald?

Toelichting

De eisen voor relatieve normering zijn in feite een combinatie van de eisen voor de dataverzameling en de betrouwbaarheid. Zo moet de steekproef representatief zijn en groot genoeg zijn om een betrouwbare schatting van de betrouwbaarheid te maken.
Bij N.2.3 gaat het om het model dat gebruikt is om de normen te bepalen, de mate waarin dit model bij de data past, en schattingsfout van de normering.
N.B. De aanbieder dient duidelijk aan te geven tot wanneer (i.e. tot welk jaartal) de normen geldig zijn.

4. Betrouwbaarheid

Code	*Vraag*
B1	Zijn of worden de betrouwbaarheidsgegevens correct berekend?
B2	Zijn de betrouwbaarheidsgegevens voldoende gezien de conclusies en eventuele beslissingen die met het instrument genomen worden?

Toelichting

Leerlingen worden naar aanleiding van een instrument gecategoriseerd. Bij een examen gaat het meestal om de categorieën gezakt en geslaagd, bij een LVS is er normaliter sprake van meer categorieën of van een relatieve categorisatie t.o.v. een referentiepopulatie. De betrouwbaarheid van een instrument hangt samen met het percentage verwachte misclassificaties. Bij de te verstrekken betrouwbaarheidsgegevens hoort ook het percentage verwachte misclassificaties. Concreet betekent dit dat als de leerlingen bijvoorbeeld gecategoriseerd worden in categorieën A, B, C en D aangegeven moet worden wat de kans is dat ze ook daadwerkelijk in de genoemde categorie zitten en niet in een andere categorie. Hetzelfde geldt uiteraard ook voor ieder ander classificatiesysteem.

De betrouwbaarheid moet adequaat worden aangetoond. Dat betekent dat alle aspecten die van invloed zijn op de betrouwbaarheid moeten worden meegenomen. Voor observatie-instrumenten zijn dus ook de beoordelaarsovereenstemming en –betrouwbaarheid van belang. Bij observatie-instrumenten wordt elk kind meestal maar door een leerkracht geobserveerd. Daarbij moeten ongewilde aspecten, zoals bijvoorbeeld onbewuste verwachtingspatronen, en halo- en horn-effecten, zo veel mogelijk beperkt worden. Om de beoordelaarsbetrouwbaarheid te onderzoeken zijn in het proefonderzoek meerdere beoordelaars nodig. Ook de betrouwbaarheid van het volgaspect dient te worden onderzocht. De betrouwbaarheid in een situatie met een of meer observatoren en een volgaspect is bijvoorbeeld te berekenen met een generaliseerbaarheidstheoriemodel (zie bijvoorbeeld Brennan, 2001, Generalizability Theory).

5. Validiteit

Code	*Vraag*
V1	Inhoudsvaliditeit: Dragen de items in het instrument bij aan de validiteit van het instrument (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)
V2	Constructvaliditeit: Meet het instrument in zijn geheel datgene wat het beoogt te meten?

Toelichting

Bij V1: De relevantie van een item heeft betrekking op de mate waarin een item meet wat het beoogt te meten, met andere woorden op de mate waarin het item een directe relatie heeft met het leerdoel en of de items een representatieve afspiegeling zijn van de te meten doelen, vastgelegd in een blueprint/toetsmatrijs. Een item is objectief indien de score van het item onafhankelijk is van degene die de scoring uitvoert. Dit impliceert dat de procedure voor het beoordelen van het antwoord op een item zo is geregeld dat geen beslissingen aan het oordeel van de beoordelaar worden overgelaten en het antwoordmodel geen ruimte voor interpretatie open laat en/of dat deskundigen het eens zijn over het juiste antwoord of over de criteria waaraan een juist antwoord moet voldoen. Efficiëntie betreft de formulering van de items en heeft betrekking op de afname en de correctie.

Bij V2: De constructvaliditeit kan op twee manieren bepaald worden. De eerste manier is statistisch, door het schatten van de correlatie met een andere verwante of juist helemaal niet verwant instrument. De tweede manier is meer beschrijvend, op basis van een blueprint/toetsmatrijs, waarbij gelet wordt op de representativiteit en evenwichtigheid.

6. Het volg-aspect

De aspecten 'De kwaliteit van de steekproef', 'betrouwbaarheid', 'normering' en 'validiteit' hebben betrekking op de beoordeling van een eigenstandig instrument. Echter, het PO besluit stelt "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 tot drie criteria met betrekking tot de schaal waarop groei wordt uitgedrukt: de bouw van de schaal, de betrouwbaarheid van de schaal en het gebruik van de schaal.
Instrumenten voor kleuters zijn vooral bedoeld om inhoudelijke, diagnostische informatie op te leveren over de ontwikkeling van een kleuter. In die zin heeft een instrument voor kleuters een signaleringsfunctie met betrekking tot de ontwikkeling van de kleuter. De groei van een kleuter gaat echter meestal sprongsgewijs en ook de inzet van het instrument is niet gebonden aan vaste afnamemomenten. Daarom is het niet mogelijk om met nauwkeurig gedefinieerde en geobjectiveerde groeiscores te werken, en zijn de hieronder gedefinieerde eisen aan de betrouwbaarheid van groeiscores voor die instrumenten nog niet aan de orde.

Code	*Vraag*
Va1	Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een correcte manier gemeten?
Va2	Wordt de betrouwbaarheid van de groei op die schaal correct weergegeven?
Va3	Worden er voldoende gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd moet worden?

Toelichting
Bij Va1: Alle overwegingen m.b.t. steekproeftrekking, betrouwbaarheid en normering gelden a-fortiori ook voor een reeks van twee of meer opeenvolgende instrumenten. Dus beide steekproeven moeten representatief zijn en het design, dat in dit geval bijna per definitie onvolledig zal zijn, moet adequaat zijn in de termen die hierboven zijn gedefinieerd. Verder moet er empirische informatie zijn over de schaalbaarheid van opeenvolgende instrumenten. 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). Maar er dient in ieder geval betekenisvolle informatie gegeven te worden over de samenhang tussen de twee (of meer) meetmomenten. Essentieel is dat de schaal waarop de groei wordt weergegeven grondig is onderbouwd.

Bij Va2: Daarbij is het ook van belang om een indicatie van de betrouwbaarheid van die gevolgtrekking weer te geven. Vertaald naar een uni-dimensioneel IRT model betekent dit dat de schattingsfout van het verschil van de vaardigheid op twee tijdstippen geschat moet zijn. Voor percentielscores betekent dit dat de betrouwbaarheid van de verandering van de percentielscores geschat moet zijn.
Bij Va3: De handleiding moet een beschrijving bevatten van hoe de gebruiker (docenten, ouders, etc.) de gegevens met betrekking tot de groei (en/of stagnatie) van een leerling inhoudelijk en relatief t.o.v. een referentiepopulatie dient te interpreteren.
De geschreven toelichting moet consistent zijn met de resultaten uit het betrouwbaarheids-, validiteits- en normeringsonderzoek, dat wil zeggen dat de gebruiker een goed beeld moet krijgen van de (relatieve) onderwijskundige waarde en meetpretentie van de resultaten.

7. Inzicht in de leervorderingen

Code	*Vraag*
I1	Levert de aanbieder een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders, verzorgers, voogden en docenten begrijpelijk is?
I2	Is er een evaluatie van de leervorderingen en worden op basis van deze evaluatie vervolgstappen geformuleerd?

Toelichting

Bij I1: Er dient een geschreven toelichting te worden geboden, waarin begrijpelijke handvatten gegeven worden voor de interpretatie van de leervorderingen van de leerling door diverse betrokkenen.

Bij I2: De informatie moet dusdanig concreet en gedetailleerd zijn dat het voor de docenten duidelijk is welke lacunes in de ontwikkeling van de leerling in het onderwijs aandacht verdienen.

8. Referentieniveaus

Code	*Vraag*
R1	Sluit de inhoud van het instrument aan op de kennis en vaardigheden zoals omschreven in de referentieniveaus van het betreffende domein (voor instrumenten vanaf groep 6)?

Toelichting

LVS-instrumenten maken onderdeel uit van de doorlopende, methodevrije leerlijn richting de eindtoets die in groep 8 afgenomen wordt. Vanwege het feit dat de eindtoetsen zijn gekoppeld aan referentieniveaus, is het belangrijk dat ook de LVS-instrumenten hierbij aansluiten. Er is dan inzicht in de vorderingen van de leerling richting het eindniveau.

N.B. Het betreft hier enkel de inhoud van het instrument. In het leerlingrapport hoeft het referentieniveau niet per se gerapporteerd te worden.

9. De Tussentijdse Check

Zoals in de inleiding aangegeven is een kwaliteitsoordeel van een LVS-instrument tien jaar geldig. In tien jaar kunnen er echter veel ontwikkelingen plaatsvinden die de kwaliteit van een LVS-instrument beïnvloeden. Daarom vinden er tussentijdse beoordelingen plaats. Voor een papieren toets die in 10 jaar niet verandert, gebeurt dat na 5 jaar. Voor digitaal afgenomen toetsen gebeurt dat vaker, bijvoorbeeld eenmaal 3 jaar en eenmaal 7 jaar na afgifte van een kwaliteitsoordeel. Over de exacte frequentie en termijnen vindt overleg plaats tussen de Expertgroep en de aanbieder. Afspraken hierover worden al bij de initiële beoordeling door de Expertgroep met de aanbieder gemaakt.
Als de inhoud van de toets wezenlijk verandert, bijvoorbeeld als een itembank wordt ververst, vindt er ook een tussentijdse beoordeling plaats. Over de belangrijkheid van de veranderingen en de noodzaak van een beoordeling, vindt overleg plaats tussen de Expertgroep en de aanbieder. De beoordeling richt zich vooral op de items en observatie-categorieën. De LVS-ontwikkelaars dienen analyses op te leveren die inzicht geven in de volgende twee vragen:

1. Zijn de items of de observatie-categorieën nog steeds up-to-date? Het kan bijvoorbeeld zijn dat het taalgebruik van de opgaven veranderd is, of dat er maatschappelijke ontwikkelingen zijn waardoor de verwoording niet meer aansluit bij modern taalgebruik, of bij nieuwe maatschappelijke gevoeligheden. Dit kan onderzocht worden door een inhoudsanalyse van de opgaven.
2. Functioneren de items of observatie-categorieën nog steeds hetzelfde als in de proeftoets, dat wil zeggen, zijn hun psychometrische eigenschappen (p-waarden of IRT itemparameters, en de relatie tussen items en het totale instrument, zoals IRT discriminatie waarden, en/of item/testcorrelaties) nog vergelijkbaar? Deze vraag is bijvoorbeeld belangrijk om na te gaan of items niet te zeer bekend zijn geworden en veel worden geoefend, of dat de onder het vorige punt aangehaalde problemen de psychometrische eigenschappen van de items of observatie-categorieën veranderd hebben.

Geconstateerde problemen kunnen worden opgelost door items aan te passen of te verwijderen. Het is daarbij wel van belang dat de eigenschappen van het LVS als geheel niet sterk veranderen. Als er een of twee items uit een toets of itembank verwijderd worden, zal het effect daarvan beperkt zijn; als het aantal problematische items erg groot is, wordt de reparatie van het LVS uiteraard ingrijpender.
Een negatief oordeel in een tussentijdse check leidt niet automatisch tot het intrekken van het oorspronkelijk afgegeven kwaliteitsoordeel. Bij de volgende tussentijdse check wordt wel nagegaan of de aanwijzingen van de Expertgroep met betrekking tot het

oplossen van geconstateerde problemen adequaat zijn opgevolgd. Als dat niet het geval is, kan dat alsnog tot intrekking van het oorspronkelijke kwaliteitsoordeel leiden.

10. Achtergrondinformatie

- COTAN Beoordelingssysteem (https://www.psynip.nl/uw-beroep/cotan/)
- RCEC Beoordelingssysteem (http://www.rcec.nl/beoordelingssysteem/)
- 'Aanvullingen COTAN Beoordelingssysteem wat betreft volg-aspect van leerling- en onderwijsvolgsysteem' (opvraagbaar bij de COTAN via cotan@psynip.nl)
- 'Aanvullingen COTAN Beoordelingssysteem wat betreft normering referentieniveaus en computer adaptief toetsen van andere eindtoetsen' (opvraagbaar bij de COTAN via cotan@psynip.nl)
- Brennan, R. L. (2001), *Generalizability Theory*. Springer.
- Brennan, R. L. (2006, Ed.). *Educational Measurement* (4th Edition). ACE, NCME.