# Verantwoording onderzoek werkgroep Meijer

Meijer, Rob R.; Egberink, Iris J.L.; Albers, Casper J.; Tendeiro, Jorge N.; Niessen, A. Susan M.

# Verantwoording onderzoek werkgroep Meijer

Meijer, Rob R.; Egberink, Iris; Albers, Casper; Tendeiro, Jorge; Niessen, Anna

rijksuniversiteit groningen
faculteit gedrags- en maatschappijwetenschappen
psychometrie & statistiek

# Verantwoording onderzoek werkgroep Meijer: Aanvullingen COTAN Beoordelingssysteem wat betreft volg-aspect van leerling- en onderwijsvolgsystemen (deel 2)

prof. dr. Rob R. Meijer
dr. Iris J. L. Egberink
dr. Casper J. Albers
dr. Jorge N. Tendeiro
A. Susan M. Niessen, MSc.

## Voorwoord

Voor u ligt de verantwoording van het onderzoek dat is uitgevoerd voor het schrijven van aanvullingen op het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer, & Sijtsma, 2010) in het kader van medewerking van de COTAN aan de werkzaamheden van de Expertgroep Toetsen PO. Specifiek gaat het om aanvullingen die het mogelijk maken om (andere) eindtoetsen en/of leerling- en onderwijsvolgsystemen (lovs) te kunnen beoordelen op de toepassing van de referentieniveaus, computer adaptief toetsen en het volg-aspect.

Een eerste uitgangspunt bij het opstellen van deze aanvullingen was om zo dicht mogelijk te blijven bij de inhoud van bestaande documenten en notities. Enerzijds omdat deze documenten geschreven zijn door experts uit de psychometrie en onderwijsveld en anderzijds omdat – niet onbelangrijk – een aantal documenten dienden als leidraad voor de toetsconstructeurs. Verder wordt bij de aanvullingen uitgegaan van het principe dat de bewijslast bij de toetsconstructeur ligt. Dit is immers ook een van de uitgangspunten bij het huidige COTAN Beoordelingssysteem. Dit klinkt wat zwaar, maar het blijkt in veel gevallen niet eenvoudig om voor elke situatie geldende en passende vuistregels op te stellen vanwege de diversiteit aan mogelijkheden en beslissingen die genomen moeten worden bij de toetsconstructie en bij de uit te voeren (psychometrische) analyses. Het gaat er daarom bij dit uitgangspunt om dat de toetsconstructeur aannemelijk moet maken dat alle beslissingen die genomen zijn tijdens het ontwikkelingsproces uitgebreid beschreven en verantwoord worden middels argumentatie ondersteund door de psychometrie. Er is getracht waar mogelijk te werken met vuistregels, richtlijnen en/of uitgebreide informatie over de meest gangbare mogelijkheden en/of analysetechnieken. Aan de andere kant maakt het ook duidelijk dat naast ‘kunde’, toetsconstructie ook een ‘kunst’ is waarbij het gaat om overtuigd te worden - met valide argumenten - door de toetsconstructeur. Om met Abelson (1995) te spreken: "Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations" (p. 2).

Tenslotte, ervaring zal moeten uitwijzen in hoeverre deze aanvullingen een werkzaam geheel vormen. Waar nodig zijn in de toekomst - op basis van opgedane ervaringen en argumenten - wellicht aanpassingen nodig.

prof. dr. Rob R. Meijer
dr. Iris J. L. Egberink
dr. Casper J. Albers
dr. Jorge N. Tendeiro
A. Susan M. Niessen, MSc.

Groningen, juli 2015

**Inhoudsopgave**

# Hoofdstuk 1 Inleiding en doel onderzoek

## 1.1 Inleiding

De ‘Wet centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs’, hierna ‘Wet Eindtoetsing PO’ (Stb. 2014, 13), verplicht basisscholen om vanaf het schooljaar 2014/2015 bij alle leerlingen in groep 8 een eindtoets voor Nederlandse taal en rekenen af te nemen. Hierbij gaat het om het bepalen van het eindniveau van leerlingen ten opzichte van de referentieniveaus (zie Wet Referentieniveaus Nederlandse taal en rekenen). Voor Nederlandse taal betreft het ten minste de twee domeinen Lezen en Begrippenlijst en taalverzorging, voor rekenen de vier domeinen Getallen, Verhoudingen, Meten en meetkunde, en Verbanden. De wet voorziet tevens in keuzevrijheid; scholen zouden naast de centrale eindtoets ook moeten kunnen kiezen voor een andere eindtoets. De centrale eindtoets wordt door de overheid beschikbaar gesteld. De diverse toetsaanbieders wordt de mogelijkheid geboden om andere eindtoetsen (hierna ‘eindtoetsen’ genoemd) te ontwikkelen. Tevens is door het ministerie aangegeven dat men toe wil naar adaptieve eindtoetsing.

Naast een verplichte eindtoetsafname schrijft de ‘Wet Eindtoetsing PO’ (Stb. 2014, 13) het gebruik van een leerling- en onderwijsvolgsysteem (lovs) voor elke leerling voor. Het gaat hierbij om het systematisch in kaart brengen van de leervorderingen, ook wel groei van kennis en vaardigheden van leerlingen. Scholen zijn vrij om te bepalen welke kennis en vaardigheden zij in kaart brengen.

Een voorwaarde die geldt voor zowel de eindtoetsen als het lovs die een basisschool gaan gebruiken, is dat deze van goede kwaliteit moeten zijn. Daarom is door de minister een onafhankelijke commissie ingesteld die enerzijds de minister adviseert over toelating van eindtoetsen en anderzijds een geschiktheidsoordeel uitspreekt over de toetsen van lovs-en, zodat basisscholen een goede, onderbouwde keuze kunnen maken uit het aanbod eindtoetsen en lovs-en. Deze commissie is de Expertgroep Toetsen PO.

## 1.2 Doel van het huidige onderzoek

Om de Expertgroep Toetsen PO in staat te stellen een goede afweging te maken over de psychometrische kwaliteit van de eindtoetsen en lovs toetsen, wordt bijgedragen door de Commissie Testaangelegenheden Nederland (afgekort tot COTAN) in de vorm van het beoordelen van dergelijke toetsen aan de hand van het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer, & Sijtsma, 2010). Hoewel het COTAN Beoordelingssysteem toegepast kan worden op zowel psychologische als onderwijskundige tests/toetsen en vragenlijsten, is er een lacune wat betreft het beoordelen van de rapportage en toepassing van de referentieniveaus en het gebruik van computer adaptief toetsen (CAT) van andere eindtoetsen alsmede het beoordelen van het ‘volg-aspect’ van lovs toetsen. Deze aspecten zijn van cruciaal belang bij het doen van een uitspraak over de psychometrische kwaliteit van dergelijke toetsen en/of toetssystemen. Het is complexe materie die enerzijds hoogwaardige psychometrische kennis vereist en waarvoor anderzijds kennis van het (complexe) veld nodig is.

De COTAN is een bestuurscommissie van het Nederlands Instituut van Psychologen (NIP). Vanuit die ‘relatie’ is het NIP opdrachtgever voor dit onderzoek. De opdracht voor de werkgroep is tweeledig, enerzijds beoordelingsvragen opstellen voor de aspecten ‘normering referentieniveaus’ en ‘computer adaptief toetsen’ van eindtoetsen en anderzijds beoordelingsvragen opstellen voor het ‘volg-aspect’ van lovs toetsen. Het is de bedoeling dat de COTAN deze beoordelingsvragen kan gebruiken als aanvulling op het huidige COTAN Beoordelingssysteem.
Dit document (deel 2) betreft de beoordelingsvragen aangaande het ‘volg-aspect’ van lovs toetsen.

### 1.3 Werkwijze

Bij het formuleren van de aanvullende beoordelingsvragen is zoveel mogelijk gebruik gemaakt van de documenten die ook zijn geformuleerd voor de toetsaanbieders. Deze documenten zijn uitgebreid bestudeerd. Verder zijn interviews gehouden met diverse experts en stakeholders om een goed overzicht te krijgen van wat mogelijk is en waar op gelet dient te worden bij het beoordelen van dergelijke toetsen. Al deze informatie is samengenomen, zaken zijn toegevoegd en/of gespecificeerd, waarbij rekening is gehouden om het in lijn te houden met het huidige COTAN Beoordelingssysteem. Dit eerste basisstuk is besproken met een aantal COTAN leden voor een eerste feedback ronde. De hierna aangepaste versie is inclusief een aantal specifieke vragen voorgelegd aan Anton Béguin en Cees Glas, als auteurs van een aantal belangrijke documenten omtrent deze onderwerpen. Vervolgens is de werkwijze toegelicht in een bijeenkomst met toetsaanbieders op het ministerie van OCW en is de toetsaanbieders de gelegenheid geboden om onduidelijkheden aan te geven. De aangegeven onduidelijkheden en vragen zijn zo veel en zo goed als mogelijk verwerkt in een versie die op 2 juli 2015 is voorgelegd en besproken tijdens de COTAN vergadering. De zaken die daar besproken zijn en de feedback van COTAN leden die per mail in de week erna is ontvangen, zijn verwerkt tot deze voor de werkgroep definitieve versie van aanvullende beoordelingsvragen wat betreft normering referentieniveaus, computer adaptief toetsen en volg-aspect van andere eindtoetsen en/of leerling- en onderwijsvolgsystemen.

### 1.4 Relatie ten opzichte van het huidige COTAN Beoordelingssysteem

Hoewel in eerste instantie getracht is de aanvullende beoordelingsvragen zoveel mogelijk als verduidelijking van en aanvulling op bestaande beoordelingsvragen op te nemen in het huidige COTAN Beoordelingssysteem, is gebleken dat dit een te grote aanpassing van het systeem vergt. De aanvullende beoordelingsvragen zijn daarom als losse aanvullingen op het COTAN Beoordelingssysteem geschreven. Dit betekent dat alleen voor de groep toetsen die in het kader van de werkzaamheden van de Expertgroep Toetsen PO worden beoordeeld een extra beoordeling zal worden gegeven op de betreffende aanvullende aspecten, die de Expertgroep Toetsen PO kan helpen bij het uitbrengen van een advies aan de minister van OCW of een eigenstandig geschiktheidsoordeel.

Het is denkbaar dat de aanvullende beoordelingsvragen in de toekomst wel verweven zullen worden in een nieuwe editie van het COTAN Beoordelingssysteem, maar daarover zullen betrokken partijen tijdig worden geïnformeerd.
Overigens zijn wij ons ervan bewust dat het ontwikkelen van een toets een ingewikkeld en langdurig proces is en dat de toetsaanbieders tijdens dat proces nog niet op de hoogte waren van de aanvullende beoordelingsvragen zoals die in dit document zijn geformuleerd. Dit is voor ons een van de redenen geweest om zoveel mogelijk aan te sluiten bij informatie die voor de toetsaanbieders beschikbaar was.

### 1.5 Bepaling beoordeling per aspect

Net als bij het huidige COTAN Beoordelingssysteem blijft het belangrijk dat de verschillende keuzes in het ontwikkelingstraject uitgebreid beschreven en verantwoord worden, zodat de redenering gevolgd kan worden en de correctheid kan worden beoordeeld. Echter, onder andere vanwege de complexiteit van de materie op een groot aantal punten en het niet altijd even duidelijk te maken onderscheid tussen ‘voldoende’ en ‘goed’ aan de hand van concrete richtlijnen is ervoor gekozen om de aanvullende aspecten te beoordelen als ‘onvoldoende/voldoende’.
Als uitgangspunt voor het bepalen van het eindoordeel per aanvullend aspect geldt dat alle van toepassing zijnde beoordelingsvragen als ‘voldoende’ moeten zijn beoordeeld.

Het is verder goed om te benadrukken dat een COTAN beoordeling als geheel (d.w.z. inclusief beoordeling op de aanvullende aspecten) dient als advies voor de psychometrische beoordeling door de Expertgroep Toetsen PO en dat tevens via de Expertgroep Toetsen PO een onderwijskundige beoordeling wordt uitgevoerd.

### 1.6 Opbouw document

Hierna is voor het volg-aspect een hoofdstuk opgenomen, waarin na een korte uitleg een beschrijving zal worden gegeven van de gebruikte termen, gevolgd door een tabel met aanvullende beoordelingsvragen voor het betreffende aspect. Onder de tabel volgt per beoordelingsvraag een toelichting.

# Hoofdstuk 2 Het volgen van de groei van leerlingen

## 2.1 Introductie

Het volgende staat beschreven in het Toetsbesluit PO over een lovs:
“Een leerling- en onderwijsvolgsysteem bestaat uit twee of meer toetsen die leervorderingen op diverse kennis en vaardigheden meten. Het kan ook bestaan uit één toets die op verschillende meetmomenten wordt afgenomen, of meerdere versies van een toets. Veel leerling- en onderwijsvolgsystemen toetsen kennis en vaardigheden van Nederlandse taal en rekenen en wiskunde. De toetsen worden verspreid over een bepaalde periode afgenomen. Bijvoorbeeld gedurende de onderbouw, de bovenbouw of de hele basisschoolperiode (groep 1/2 tot en met groep 8). De leerlingscores op verschillende losse toetsen uit het leerling- en onderwijsvolgsysteem kunnen in samenhang worden bezien. Op deze wijze brengen scholen de leervorderingen oftewel de groei van kennis en vaardigheden van leerlingen in kaart. Hiermee vormt het leerling- en onderwijsvolgsysteem een waardevol instrument voor scholen bij het vormgeven van het onderwijs. De gegevens uit het leerling- en onderwijsvolgsysteem zijn ook bruikbaar bij het evalueren van de leerresultaten van groepen leerlingen gedurende meerdere jaren.” (p.22)

“De Expertgroep toetsen PO beoordeelt de psychometrische kwaliteit van losse toetsen binnen het leerling- en onderwijsvolgsysteem: de validiteit van de toets, de betrouwbaarheid van de toets en de deugdelijkheid van de normering. Deze drie aspecten zijn toegelicht in paragraaf 2.2.1 van dit besluit. Naast de psychometrische kwaliteit beoordeelt de Expertgroep ook enkele onderwijskundige aspecten van het leerling- en onderwijsvolgsysteem. De toetsen uit het leerling- en onderwijsvolgsystemen meten in ieder geval de kennis en vaardigheden van leerlingen op het gebied van Nederlandse taal en rekenen. De Expertgroep zal beschrijven op welke wijze de referentieniveaus taal en rekenen zijn verwerkt in de leerling- en onderwijsvolgsystemen. Scholen kunnen naast taal en rekenen ook kiezen voor leerling- en onderwijsvolgsystemen die de kennis en vaardigheden van leerlingen in andere vakken toetsen. De gezamenlijke toetsen of toetsversies van een leerling- en onderwijsvolgsysteem moeten op systematische en valide wijze de leervorderingen van leerlingen in kaart brengen. Bij de meting van de leervorderingen op verschillende meetmomenten wordt in principe gebruik gemaakt van verschillende toetsopgaven, dus twee of meer toetsen of toetsversies. In sommige specifieke gevallen bestaat een leerling- en onderwijsvolgsysteem uit één toets, waarbij dezelfde toetsopgaven dus vaker worden gebruikt. Dit kan alleen als deze werkwijze niet strijdig is met een valide meting van leervorderingen. De toetsontwikkelaar die een dergelijk leerling- en onderwijsvolgsysteem aanbiedt, moet aannemelijk maken dat een eventuele groei in kennis en vaardigheid van een leerling niet het gevolg is van bekendheid met de opgaven.
Om leervorderingen te kunnen meten, moeten de toetsscores van de leerling op een schaal te plaatsen zijn die de ontwikkeling van leerlingen zichtbaar maakt. Vaak omvat een dergelijke schaal het conceptuele bereik van een bepaalde kennis of vaardigheid, zoals de woordenschat van leerlingen. De schaal is cumulatief opgebouwd: van lage scores naar hoge scores. De

losse toetsen van het leerling- en onderwijsvolgsysteem meten steeds een deel van de schaal om zo de vooruitgang van de leerling gedurende een bepaalde tijd vast te stellen." (p.23-24)

Het uitgangspunt bij onderstaande beoordelingsvragen is dat wanneer gesteld wordt dat de groei van leerlingen over jaren heen in kaart kan worden gebracht en dat afwijkingen in de groei kunnen worden opgespoord, de interpretatie van die groei uitgelegd en onderbouwd moet worden. Waarbij tevens aangegeven dient te worden wanneer er sprake is van afwijkende en/of normale groei en/of wanneer voorzichtigheid geboden is bij de interpretatie.

**2.2 Terminologie**

Leervorderingen worden op systematische wijze gemeten wanneer:

(1) individuele toetsen psychometrisch in orde zijn
(2) de toetsen gelinkt zijn, dat wil zeggen dat aannemelijk is gemaakt dat toetsen schaalbaar zijn op dezelfde schaal over de jaren/onderwijsniveaus heen.

Verder zijn twee typen groei te onderscheiden, namelijk relatieve groei (d.w.z. ten opzichte van de normpopulatie, hoe goed doet de leerling het ten opzichte van de medeleerlingen) en absolute of inhoudelijke groei (in de zin van toename van vaardigheden/kennis en ten opzichte van jezelf over de jaren heen).
Onderstaande beoordelingsvragen zijn van toepassing op beide typen groei. Voor beide typen groei zal een beoordeling gegeven moeten worden. Indien over een van de twee typen groei geen informatie wordt gegeven, zal daarbij worden aangegeven 'geen onderzoek'.

Tevens dient hier vermeld te worden dat de COTAN het gebruik van het Didactisch Leeftijds Equivalent (DLE), vanwege vele praktische en theoretische bezwaren, afkeurt. Een uitgebreide bespreking is te vinden in Evers en Resing (2007). Wanneer toetsen uitsluitend groei in termen van DLE's rapporteren, wordt de beoordeling op het volg-aspect 'onvoldoende'.

## 2.3 Beoordelingsvragen m.b.t. het volgen van de groei van leerlingen

Tabel 2.1. Beoordelingsvragen m.b.t. het volgen van de groei van leerlingen

| | | onv. | vold. |
|---|---|---|---|
| Vraag 1 | Worden er gegevens verstrekt over de volgtijdelijkheid van de toets? | 1 | 2 |
| Vraag 2 | Worden er gegevens verstrekt over de uitgevoerde linking procedure? | | |
| | a. Is de gekozen dataverzameling adequaat? | 1 | 2 |
| | b. Is de kwaliteit van het design adequaat? | 1 | 2 |
| | c. Is de gekozen psychometrische methode adequaat? | 1 | 2 |
| Vraag 3 | Groei | | |
| | a. Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden? | 1 | 2 |
| | b. Wordt groei op een adequate manier gemeten? | 1 | 2 |

## 2.4. Toelichting beoordelingsvragen

*Aanwijzingen bij vraag 1: Worden er gegevens verstrekt over de volgtijdelijkheid van de toets?*

In de handleiding dient een verantwoording te zijn gegeven voor wat betreft de volgtijdelijkheid van de toets. Dat wil zeggen dat er zowel een theoretische als empirische onderbouwing moet zijn van de wijze waarop leervorderingen van leerlingen in kaart worden gebracht. Dit kan door zowel dezelfde toets meerdere malen aan te bieden als, hetgeen vaker zal voorkomen, scores van dezelfde leerlingen op verschillende toetsen te vergelijken. Wanneer dezelfde toets wordt aangeboden dient aannemelijk te worden gemaakt dat hogere scores niet worden veroorzaakt door kennis van dezelfde vragen.

Vaak zal het gaan om een schaal die het conceptueel bereik van een bepaald kennis of vaardigheid meet, bijvoorbeeld woordenschat of rekenvaardigheid. De schaal is opgebouwd van lage scores naar hoge scores. De afzonderlijk toetsen meten dan een deel van de schaal (aanvullende criteria document, 2014).

Een belangrijk punt is dat het PO besluit stelt dat “Om leervorderingen te kunnen meten, moeten de toetsscores van de leerling op een schaal te plaatsen zijn die de ontwikkeling van leerlingen zichtbaar maakt. Vaak omvat een dergelijke schaal het conceptuele bereik van een bepaalde kennis of vaardigheid, zoals de woordenschat van leerlingen.”

Kanttekening hierbij: Hoewel wij de COTAN criteria in overeenstemming hebben geformuleerd met deze “eis”, is deze “eis” niet per se nodig om leervorderingen van leerlingen te meten. Bijvoorbeeld per toets zou de score van een leerling vergeleken kunnen

worden met de scores van andere relevante leerlingen. Dan kan het zo zijn dat leerling A tot de 20% beste leerlingen behoort in groep 6, maar in groep 7 “slechts” tot de 40% beste leerlingen. Deze informatie geeft in ieder geval weer dat de groei van leerling A minder snel is dan zijn groepsgenoten, hetgeen waardevolle informatie kan opleveren. Hierbij dient opgemerkt te worden dat dergelijke metingen wel over hetzelfde inhoudsdomein moeten gaan en bij voorkeur op dezelfde schaal gescoord zijn.

*Aanwijzingen bij vraag 2: Worden er gegevens verstrekt over de uitgevoerde linking procedure?*

Aangetoond moet worden dat de scores over de verschillende toetsen binnen een bepaald domein (bijvoorbeeld Begrijpend Lezen) op één schaal liggen. De handleiding moet beschrijven hoe men dit gedaan heeft.

Zowel KTT methoden (zoals verticaal linken, zie bijvoorbeeld Kolen & Brennan, 2004) als IRT methoden kunnen hiervoor worden gebruikt. Het is in het algemeen belangrijk dat de gevolgde procedure uitgebreid en correct wordt verantwoord.

*Aanwijzingen bij vraag 2a: is de gekozen data verzameling adequaat?*

In de notitie “Aanvullende criteria COTAN voor andere eindtoetsen en leerlingvolgsystemen. Versie 13-5-2014” wordt gesteld dat dezelfde criteria voor kwaliteitshandhaving gelden als bij de normering van de referentieniveaus(o.a. wat betreft representativiteit en grootte van de steekproeven) maar met de nuancering dat “sommige opgaven in verschillende leerjaren zich relatief anders gedragen. Waar prestaties op een groep opgaven toenemen in opeenvolgende leerjaren zal de prestatie op een andere groep opgaven stagneren of zelfs afnemen. Het gevolg hiervan is dat de groei in vaardigheid tussen leerjaren afhangt van de gekozen ankeropgaven.” Het is dus voor de toetsconstructie van belang goed te kijken naar deze ankeropgaven en hier aandacht aan te besteden in de handleiding.

Ook zal het zo zijn dat in verschillende toetsen niet alle deelvaardigheden aan bod komen. Bijvoorbeeld bij leesvaardigheid kan op toets 0 het onderdeel reflecteren niet aan bod komen, terwijl dit wel aan bod komt in latere toetsen. De inhoud van Leesvaardigheid verandert dus met de jaren.

Qua grootte van de steekproeven wordt als richtlijn meegegeven dat wanneer een simpel IRT model wordt gebruikt elk item bij minimaal 200 personen moet zijn afgenomen en wanneer het 2-parameter logistisch model wordt gebruikt elk item bij minimaal 400 personen moet zijn afgenomen. Voor de volledigheid wordt hierbij vermeld dat niet elke persoon alle items (uit de toets) hoeft te hebben ingevuld. Het gebruikte design moet daarvoor wel uitgebreid zijn beschreven.

Bij gebruik van de KTT methoden moet elk item door minimaal 300 personen zijn ingevuld.

*Aanwijzingen bij vraag 2b: is de kwaliteit van het design adequaat?*

Er verschillende manieren om toetsen met elkaar te linken (zie bijvoorbeeld Engelen & Eggen, 1993; Kolen & Brennan, 2004). Deze dienen op een adequate manier te zijn beschreven en te worden verantwoord. Vaak zal gebruik worden gemaakt van een kalibratie

onderzoek met een onvolledig design. Dat wil zeggen dat niet iedere leerling alle opgaven krijgt. Opgaven worden verdeeld over groepen van items en aan leerlingen worden “boekjes” gegeven (een of meer groepen items). De verschillende boekjes hebben gemeenschappelijke opgaven waardoor er een gelinked design ontstaat. Wanneer bepaalde opgaven uit twee opeenvolgende toetsen in hetzelfde boekje worden opgenomen kunnen itemparameters van alle opgaven met elkaar worden vergeleken.

*Aanwijzingen bij vraag 2c: Is de gekozen psychometrische methode adequaat?*
Voor zowel KTT als IRT methoden geldt dat de aannames waaronder de toetsen zijn ‘gekalibreerd/gelinked’ beschreven moeten worden en dat de houdbaarheid c.q. passing ervan voldoende aannemelijk gemaakt moet worden.

*Aanwijzingen bij vraag 3: Groei*
*Aanwijzingen bij vraag 3a: Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden?*
De handleiding moet een beschrijving bevatten van hoe de gebruiker de gegevens met betrekking tot de groei (en/of stagnatie) van een leerling dient te interpreteren en/of hoe hiermee omgegaan dient te worden.

*Aanwijzingen bij vraag 3b: Wordt groei op een adequate manier geïnterpreteerd gemeten?*
Er dient aangetoond te worden dat de toetsen voldoende betrouwbaar zijn om verschillen in groei (en/of stagnatie) te kunnen meten. Wanneer gebruik wordt gemaakt van IRT kan bijvoorbeeld de standaardschattingsfout worden gerapporteerd conditioneel op het vaardigheidsniveau. Op deze manier kan inzichtelijk worden gemaakt hoe betrouwbaar scores zijn op een bepaald vaardigheidsniveau.

Een opmerking hierbij is dat of de testscores betrouwbaar genoeg zijn afhangt van de scoring. Bij absolute normering zou gekeken kunnen worden naar de meetnauwkeurigheid op individueel niveau. Bij relatieve normering zou gekeken kunnen worden naar inter-individuele verschillen in veranderingen, waarbij men afhankelijk is van de betrouwbaarheid van verschilscores en die is over het algemeen niet erg hoog.

## Literatuur


Abelson, R. P. (1995). *Statistics as a principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie, mei 2009; gewijzigde herdruk mei 2010)*. Amsterdam: NIP.

Evers, A., & Resing, W. C. M. (2007). Het drijfzand van didactische leeftijdsequivalenten. *De Psycholoog, 42*, 466-472.

Engelen, R. J. H., & Eggen, T. J. H. M. (1993). Equivaleren. In T. J. H. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 309-348). Arnhem: Cito.

Geen auteur. *Aanvullende criteria COTAN voor andere eindtoetsen en leerlingvolgsystemen.* Versie 13-5-2014.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking.* New York, NY: Springer.

Toetsbesluit PO. *Staatsblad 2014,* 000. 3 juni 2014.

Wet centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs. *Staatsblad* 2014, 13. 16 januari 2014.