

1. Algemene informatie

Algemeen en meetpretentie

Kleuter in beeld – Taal is een methodeonafhankelijk volgsysteem. Het gaat om een observatieinstrument, dat bestaat uit een leerkrachtroute en een kindroute. Het belangrijkste doel van het instrument is de vaardigheid op het gebied van taal van kleuters op een objectieve manier in beeld te brengen.

Doelgroep

De doelgroep van het instrument zijn kinderen in groep 1 en 2 van het primair en speciaal onderwijs. Het is daardoor ook in te zetten bij kinderen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften.

Gebruiksdoel en functie

Kleuter in beeld – Taal is bedoeld om de taalvaardigheid van kleuters op een objectieve manier in beeld te brengen. Er worden drie verschillende doelen onderscheiden:

- Niveaubepaling: het bepalen van het niveau van de leerling of de groep voor de acht subdomeinen binnen Taal.
- Progressiebepaling: het bepalen van vooruitgang, achteruitgang of stabilisering in de ontwikkeling.
- Afstemmen onderwijsaanbod: de rapportage geeft handvaten om het onderwijs zo goed mogelijk op de kinderen af te stemmen.

Inhoudelijke theoretische inkadering:

Er is voor gekozen om Taal (hoofddomein) op te delen in Mondelinge taal (domein) en Beginnende geletterdheid (domein). Deze verdeling wordt beargumenteerd in de wetenschappelijke verantwoording. Zowel Mondelinge taal als Beginnende geletterdheid is onderverdeeld in vier subdomeinen. Er is gekozen voor de meest onderscheidende subdomeinen en onderdelen, gebaseerd op onder andere de inhoudskaart taal van SLO. De definitieve selectie van subdomeinen, onderdelen en observatiepunten is gebaseerd op literatuur. Ook is de selectie van de verschillende inhoudsvoorwerpen voorgelegd aan leerkrachten en intern begeleiders. Hun feedback is meegenomen bij de selectie.

Keuzes bij de verdeling in subdomeinen worden verantwoord in de wetenschappelijke verantwoording.

Inhoud van het toetspakket

Het toetspakket Kleuter in beeld – Taal bestaat uit de volgende documenten:

- Wetenschappelijke verantwoording, deze bevat informatie over
 - Uitgangspunten van de ontwikkeling
 - Beschrijving meetinstrument
 - Dataverzameling
 - Kalibratie kindroute
 - Standaardbepaling voor de kindroute
 - Beschrijving algoritmes
 - Betrouwbaarheid en meetnauwkeurigheid
 - Validiteit
- Handleiding Kleuter in beeld online (via Basispoort)
- Leerkrachtmap Kleuter in beeld – Taal
- Papieren opdrachtenboekjes

- Digitale opdrachten (via Basispoort)

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor instrumenten binnen leerlingvolgsystemen (LVS)', zoals opgesteld door de Expertgroep Toetsen PO. De Expertgroep Toetsen PO wordt gevormd door Prof. Dr. Cees Van der Vleuten (voorzitter), Prof. dr. Cees Glas (psychometrisch expert), Dr. Desiree Joosten-Ten Brinke (onderwijskundig expert) en Jennifer Roubiës MSc (secretaris).

Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording en handleiding veelal letterlijk vermeld.

De kwaliteit van de dataverzameling

S1 Is de steekproef representatief?

Bevindingen:

Bij Kleuter in beeld - Taal staan (directe en indirecte) observaties centraal. De leerkrachtroute is volledig gebaseerd op indirecte observaties, waarbij de leerkracht het niveau van de kleuter inschat op basis van zijn eigen indirecte observaties over een langere periode vanwege de sprongsgewijze ontwikkeling van kleuters. De kindroute wordt door de leerkracht gekozen bij twijfels bij een of meer kleuters over een of meer subdomeinen en bestaat uit opdrachten (papier of digitaal) en activiteiten. Bij de activiteiten staan directe observaties centraal, waarbij de leerkracht samen met de kleuter een activiteit uitvoert en observatiepunten invult op basis van wat de leerkracht op dat moment ziet. Om de kwaliteit te waarborgen, zijn er diverse onderzoeken uitgevoerd om de observaties zo gestandaardiseerd en objectief mogelijk te maken. De leerkracht plant de observaties, opdrachten en activiteiten op een moment waarop de leerkracht de informatie nodig heeft en is dus flexibel. Omdat het niveau waar kleuters mee binnenkomen zeer divers is, past de inhoud van het observatieinstrument dan ook bij uiteenlopende vaardigheidsniveaus van de kleuters. Voor elk van de activiteiten is er één versie die geschikt is voor zowel kinderen in groep 1 als groep 2. Bij de opdrachten zijn er drie verschillende versies voor verschillende niveaus: * = niveau groep 1, ** = niveau midden groep 2 en *** = niveau eind groep 2.

Er zijn twee onderzoeken uitgevoerd, een proefonderzoek in september 2018 en een kwaliteitsonderzoek in februari 2019.

Een belangrijk doel van het proefonderzoek in 2018 was om de kwaliteit en de moeilijkheidsgraad van de opdrachten te bepalen. Een ander doel was om op basis van de eerste ervaringen en feedback van leerkrachten aanpassingen te doen, zodat het uiteindelijke instrument zo goed mogelijk aansluit bij de wensen uit het veld. In het steekproefkader voor het proefonderzoek zaten ruim 6000 scholen, waaruit een aselechte steekproef van 840 scholen is getrokken en waarbij rekening is gehouden met het percentage gewichtenleerlingen en regio. Deze scholen hebben per post een uitnodiging ontvangen om te participeren in het onderzoek, waarvoor zich 41 scholen hebben aangemeld. Uiteindelijk hebben 37 scholen en 922 kinderen uit groep 1 en 2 daadwerkelijk meegedaan. Op basis van psychometrische gegevens (p- en rit-waarden) en feedback van de leerkrachten zijn de beste opdrachten geselecteerd voor het kwaliteitsonderzoek.

In februari 2019 is een grootschalig kwaliteitsonderzoek georganiseerd om de kwaliteit van het instrument te onderzoeken. Hierbij zijn de leerkrachtroute, de activiteiten uit de kindroute en papieren en digitale opdrachten afgenomen. Evenals in het steekproefkader

voor het proefonderzoek zaten in het steekproefkader voor het kwaliteitsonderzoek ruim 6000 scholen. In eerste instantie is een aselecte steekproef van 361 scholen getrokken voor de digitale afname van de opdrachten en 344 scholen voor de papieren afname van de opdrachten (uitgaande van een response rate van 0,04%) in groep 1 en 2, waarbij wederom rekening is gehouden met het percentage gewichtenleerlingen en regio. De tweede steekproef voor de papieren en digitale afname bestond uit respectievelijk 856 en 571 scholen, welke beide tweemaal zijn aangeschreven. Uiteindelijk zijn alle scholen aangeschreven, omdat het aantal scholen dat zich aanmeldde te gering was. Door middel van een advertentie op Facebook en te flyereren op beurzen is er daarnaast nog extra geworven. Verder is een beloning (VVV-bon van 25 euro) gegeven aan deelnemende scholen om de deelnamebereidheid te vergroten. Uiteindelijk hebben in totaal 82 scholen en 2252 kinderen uit groep 1, 2 en 3 meegedaan, waarbij de scholen een voorkeur hadden voor de papieren afname van de opdrachten (zie Tabel 4.1). Vijf scholen deden met zowel de papieren als de digitale versie mee en meer dan de helft van de scholen (53) deden met twee of zelfs drie leerjaren mee (zie Tabel 4.2). Negen scholen deden daarnaast een afname bij kinderen uit groep 3, waarvan de gegevens alleen zijn gebruikt om te controleren of de opdrachten/observatiepunten voldoende onderscheid maakten tussen leerjaar 2 en leerjaar 3.

De representativiteit van de steekproef uit het kwaliteitsonderzoek ten behoeve van de normering is onderzocht door te kijken naar de vier variabelen regio, urbanisatiegraad, percentage achterstandsleerlingen en geslacht, waarbij de eerste drie variabelen op schoolniveau zijn gedefinieerd en de laatste variabele op leerlingniveau. Vier scholen konden niet meegenomen worden in de representativiteitsanalyse, omdat voor deze scholen de regio, de urbanisatiegraad en het percentage achterstandsleerlingen onbekend waren. Dit betekent dat de representativiteit bij het kwaliteitsonderzoek heeft plaatsgevonden bij in totaal 78 scholen (82 – 4) en 1889 kinderen uit groep 1, 2 en 3 (zie Tabel 4.4). Bij regio is uitgegaan van de vier landsdelen / regio's van de CBS-indeling (noord, oost, west, zuid). Bij urbanisatiegraad is er voor gekozen om de indeling naar vijf niveaus, die gebruikelijk is bij het CBS, te reduceren tot een tweedeling in enerzijds niet tot matig verstedelijkt (platteland) en anderzijds sterk tot zeer sterk verstedelijkt (stad). Een dergelijke tweedeling blijkt in de praktijk goed te volstaan (cf. Van Boxtel en Hemker, 2009). Bij percentage achterstandsleerlingen is uitgegaan van de formatiegewichten van de leerlingen binnen een school volgens de meest recente regeling van OCW. In navolging van OCW worden op basis van het opleidingsniveau van de ouders drie opleidingsniveaus onderscheiden: 0.0 = één van de ouders of beide ouders heeft of hebben een opleiding gehad uit categorie 3; 0.3 = beide ouders of de ouder die belast is met de dagelijkse verzorging hebben of heeft een opleiding uit categorie 2 gehad; 1.2 = één van de ouders heeft een opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2. In deze indeling wordt verwezen naar de volgende categorieën in het opleidingsniveau van de ouders: categorie 1 = maximaal basisonderwijs of (V)SO-ZMLK, categorie 2 = maximaal LBO/VBO, praktijkonderwijs of VMBO basis- of kaderberoepsgerichte leerweg en categorie 3 = overig VO en hoger. Leerlingen met een formatiegewicht van 0.3 of 1.2 zijn te definiëren als achterstandsleerlingen. Scores zijn ingedeeld naar het percentage achterstandsleerlingen volgens een indeling in twee typen: (1) percentage achterstandsleerlingen [0 tot 0,10], (2) percentage achterstandsleerlingen [0,10 tot 1]. Bij geslacht is een tweedeling gemaakt naar jongens en meisjes. Voor het bepalen van de populatiegegevens is gebruikgemaakt van gegevens van DUO en CBS.

Uit de resultaten van de representativiteitsanalyse bleek dat de verdeling van het percentage scholen met meer dan 10% achterstandskinderen niet significant afweek van de hele populatie scholen in Nederland en ook de verdeling van jongens en meisjes in de steekproef niet significant verschillend was van de verhouding in de populatie (zie Tabel 4.3 en Tabel 4.4 met een statistische toetsing van de representativiteit). Tabel 4.3 laat ook zien dat de verdeling over regio's en de urbanisatiegraad in de steekproef afwijkt van de verdeling in heel Nederland. Scholen uit het westen zijn ondervertegenwoordigd (29,5% in de steekproef versus 47,9% van de populatie), scholen uit het noorden en oosten zijn oververtegenwoordigd en scholen uit het zuiden zijn wel representatief. Verder is uit Tabel 4.3 af te lezen dat scholen uit niet tot matig verstedelijkte gebieden (platteland) zijn oververtegenwoordigd (70,5% in de steekproef versus 54,8% in heel Nederland). Omdat in Kleuter in beeld – Taal geen relatieve normering wordt gebruikt, maar gewerkt wordt met absolute standaarden die verbonden zijn aan de inhoud (kleuters ontwikkelen zich immers sprongsgewijs en zijn daarom op basis van resultaten niet meteen in een hokje te plaatsen) heeft de representativiteit van de steekproef geen impact op de standaarden. Dit impliceert dat bijvoorbeeld een weging naar een proportionele afspiegeling van de populatie achterwege kan blijven.

Conclusie:

De steekproeven zijn geschikt voor het beoogde doel en de steekproeftrekking is adequaat beschreven. De steekproeven zijn representatief naar percentage achterstandskinderen en geslacht. De verdeling over regio's en urbanisatiegraad daarentegen wijkt echter significant af van de verdeling in heel Nederland. Dit heeft echter geen impact op de standaarden, omdat geen relatieve maar absolute normering wordt gebruikt. Op aspect S1 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

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

Bevindingen:

Op grond van inhoudelijke criteria en psychometrische gegevens (Klassieke Test Theorie en Item Respons Theorie) zijn de definitieve selecties van items (opdrachten en activiteiten/observatiepunten) gemaakt voor het observatieinstrument Kleuter in beeld – Taal (zie Tabel 4.8). Er zijn 56 tot 60 opdrachten ontwikkeld om uit te testen in het kwaliteitsonderzoek voor de drie subdomeinen Woordenschat, Luisteren en Fonologisch bewustzijn. Er zijn voor elk van deze drie subdomeinen (toets)boekjes ('booklets') gemaakt van 15 opdrachten om voor te leggen aan de kinderen. Voor elk subdomein zijn 12 opdrachtenboekjes gemaakt van 5 opdrachten, die als makkelijk, gemiddeld of moeilijk ingeschat werden. Daarmee zijn vervolgens 11 verschillende boekjes gemaakt met overlap en de opdrachten die de link zijn tussen de boekjes komen in meer boekjes voor (zie Tabel 4.5). Deze boekjes konden digitaal gemaakt worden of op papier en voor elke leerling vulde de leerkracht het (indirecte) observatieformulier van de leerkrachtroute in. De kinderen deden daarnaast vier subdomeinen uit de kindroute: drie subdomeinen met opdrachten (Luisteren, Woordenschat en Fonologisch bewustzijn) en één activiteit (Spreken, Gesprekken, Oriëntatie op geschreven taal) of de drie subdomeinen met opdrachten, aangevuld met opdrachten Receptieve letterkennis (zie Tabel 4.6). De (indirecte) observaties in de leerkrachtroute en de (directe) observaties van de activiteiten werden door de leerkracht ingevuld op een papieren observatieformulier. De designs in Tabel 4.5-4.6 laten zien dat de boekjes stevig

verankerd zijn. Er kan dus geconcludeerd worden dat er sprake is van een onvolledig maar 'verbonden' dataverzamelingsdesign, waartoe een longitudinale opzet werd gebruikt. Uit Tabel 4.7 valt verder nog af te lezen dat bij het subdomein Gesprekken de meeste (directe) observaties zijn gedaan, hetgeen te verklaren is doordat deze activiteit bij twee of drie kinderen moest worden afgenomen terwijl de overige activiteiten individueel werden afgenomen.

Uit het kalibratieonderzoek voor de kindroute blijkt dat het instrument Kleuter in beeld – Taal past bij het gehanteerde itemresponstheoriemodel OPLM (met dit statistische meetmodel zijn de moeilijkheidsparameters en discriminatie-indices van de items geschat). De items passen in het OPLM, zoals blijkt uit de (niet-significante) S-toetsen; zie de grafische weergave van de ICC's in Figuur 5.4 en de uniforme verdeling van overschrijdingskansen in Tabel 5.1 voor de acht subdomeinen uit de kindroute. Ook het instrument als geheel past in het OPLM, zoals blijkt uit de R1c-toetsen. Met uitzondering van het subdomein Gesprekken geldt voor alle acht subdomeinen dat de R1c-waarde niet groter is dan ongeveer anderhalf maal het aantal vrijheidsgraden van de bijbehorende chi-kwadraat verdeling en daarnaast zijn de subdomeinen Spreken en (Receptieve) Letterkennis niet significant (zie Tabel 5.2). Ook de zogenoemde constante 'c' uit het COTAN beoordelingsstelsel (deze constante geeft de relatie weer tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie) geeft volgens de auteurs een verantwoording van de modelpassing. De gemiddelde waarde van de constante is goed ($c \leq 0.2$) voor de opdrachten en observatiepunten in alle niveauversies. Bij de grote meerderheid van de itemparameters (86%) is de nauwkeurigheid volgens de criteria die de COTAN hanteert eveneens goed. Op basis van de hierboven beschreven psychometrische resultaten kan geconcludeerd worden dat de kalibratie (het hele proces van het schatten van de itemparameters en verifiëren of de modelveronderstellingen houdbaar zijn voor de data) voor de kindroute van Kleuter in beeld – Taal geslaagd is. Dit betekent dat er sprake is van een ééndimensionale vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden. Een geslaagde kalibratie op een unidimensioneel construct (hier: taalvaardigheid voor kleuters) kan worden beschouwd als een noodzakelijke (doch niet voldoende) voorwaarde voor constructvaliditeit (komen de gemeten concepten inderdaad overeen met het begrip-zoals-bedoeld?). In hoofdstuk 9 van de Wetenschappelijke Verantwoording (WV) wordt de constructvaliditeit nader onderzocht.

Conclusie:

Het onvolledige maar 'verbonden' dataverzamelingsdesign is adequaat. Op aspect S2 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: '**voldoende**'.

S3 In het geval van een observatie-instrument: is er sprake van een adequate steekproef van observatoren en randvoorwaarden waaronder de observatie wordt uitgevoerd.

Bevindingen:

Onder aspect S1 is vastgesteld dat de 83 scholen uit de steekproef die uiteindelijk hebben deelgenomen aan het kwaliteitsonderzoek (zie Tabel 4.2) representatief zijn naar percentage achterstandskinderen en geslacht, maar niet naar regio's en urbanisatiegraad.

Dit heeft echter geen impact op de standaarden, omdat geen relatieve maar absolute normering wordt gebruikt. In de WV wordt niet expliciet vermeld of de gebruikte observatoren van deze 83 scholen uit de steekproef een goede afspiegeling zijn van de observatoren die het observatie-instrument Kleuter in beeld – Taal in de praktijk gaan gebruiken, en ook of de randvoorwaarden waaronder de observaties worden gemaakt een goede afspiegeling zijn van de praktijk waarin het instrument gebruikt gaat worden.

Conclusie:

Op aspect S3 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

S4 Er is een handleiding met duidelijke instructies voor de leerkracht over het zo objectief mogelijk uitvoeren en weergeven van de observaties door de leerkracht.

Bevindingen:

De leerkrachtmap Kleuter in beeld – Taal bevat een handleiding met de volgende onderdelen: (i) Doel, doelgroep en opzet, (ii) Stappenplan kleuter in beeld, (iii) Interpreteren en analyseren, (iv) Aanvullende signalering bij problemen, (v) Van signaleren naar handelen en (vi) Veel gestelde vragen over Kleuter in beeld. Verder bevat de leerkrachtmap nog een Instructie Leerkrachtroute en Materiaal Kindroute in de vorm van instructies, nakijkkaarten en gestandaardiseerde observatieformulieren voor de drie onderscheiden versies. Al deze materialen samen zorgen ervoor dat de leerkracht 'in principe' (als er althans sprake is van een adequate steekproef van observatoren, zie in dit verband de gemaakte opmerking onder aspect S3) zo objectief mogelijk observaties kan uitvoeren en weergeven.

Conclusie:

Op aspect S4 wordt aan het instrument Kleuter in beeld – Taal het oordeel **'voldoende'** toegekend.

Normering

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

Bevindingen:

Het doel van de standaardbepaling in het onderhavige observatieinstrument is om met leerkrachten en onderwijsprofessionals na te gaan wat qua taal doorgaans gezien wordt bij kleuters eind groep 1 (E1) en wat doorgaans gezien wordt bij kleuters eind groep 2 (E2), waardoor vroegtijdig gesignaleerd kan worden of een kind mogelijk behoefte heeft aan extra aandacht of juist aan extra uitdaging. Voor de twee onderscheiden domeinen van Taal (hoofddomein), Mondelinge taal en Beginnende geletterdheid, dienen tijdens de standaardbepaling de grenzen bepaald te worden tussen vijf opeenvolgende functioneringsniveaus (<E1, E1, M2, E2 en >E2), waarbij <E1, M2 en >E2 staan voor respectievelijk onder niveau E1, midden groep 2 en boven niveau E2. Per domein worden tijdens de standaardbepaling dus uiteindelijk vier standaarden bepaald. Bij elk(e) observatiepunt/opdracht is alleen bij de niveaus E1 en E2 een beschrijving opgenomen, welke aangeeft wat de meeste kinderen eind groep 1 en eind groep 2 geacht worden te

beheersen (zie Figuur 7.1 voor voorbeeld vijfpuntsschaal om het functioneringsniveau te beoordelen in de leerkrachtroute).

Voor de standaardbepaling is de 3DC-methode (Data Driven Direct Consensus) gebruikt (Feskens, Keuning, Van Til, & Verheyen, 2014, Keuning, Straat & Feskens, 2017), welke al succesvol is toegepast voor onder meer het bepalen van de standaarden voor het Europees Referentie Kader (Feskens, Keuning, Van Til, & Verheyen, 2014) en voor de Diagnostische Tussentijdse Toets (Schouwstra, 2018). Deze methode is een uitbreiding van de uit de standaardbepalingsliteratuur bekende DC-methode (Direct Consensus), ondersteund met empirische informatie uit het kwaliteitsonderzoek. Bij de DC-methode (in feite een speciaal geval van de uit de standaardbepalingsliteratuur bekende Ja/Nee methode, een variant op de Angoff methode) bekijken experts clusters van opdrachten of observatiepunten en zetten per cluster een grens. De methode veronderstelt dus dat een instrument bestaat uit meerdere opdrachten of observatiepunten die in te delen zijn in een aantal clusters. In dit geval is er sprake van in totaal 8 clusters, namelijk de vier clusters van het domein Mondelinge taal (Luisteren, Woordenschat, Spreken en Gesprekken) en de vier clusters van het domein Beginnende geletterdheid (Fonologisch bewustzijn, Receptieve letterkennis, Oriëntatie op boek en verhaal en Oriëntatie op geschreven taal). Ieder subdomein wordt dus opgevat als een cluster. Tijdens de standaardbepalingssessie krijgen de experts voor elk subdomein een cluster met opdrachten of observatiepunten te zien en vormen zij zich een oordeel over hoeveel opdrachten/observatiepunten het kind naar verwachting correct zou hebben op dat cluster als zijn/haar vaardigheid zich precies op de grens van twee functioneringsniveaus (onder E1/E1, E1/boven E1, onder E2/E2 en E2/boven E2) bevindt. Voor elk subdomein moet voor een grensleerling (risicoleerling) dus door de experts een oordeel worden gegeven over de onder- en bovengrens eind leerjaar 1 (E1) en onder- en bovengrens voor eind leerjaar 2 (E2).

Als aanvulling op deze DC-methode kregen de experts bij de 3DC-methode empirische informatie te zien in de vorm van het verwachte aantal correcte opdrachten/observatiepunten op een subdomein (cluster) voor elk mogelijk aantal correcte opdrachten/observatiepunten op het volledige domein van Mondelinge taal of Beginnende geletterdheid, welke bepaald werden met OPLM en simulaties (zie het standaardbepalingsformulier in Figuur 6.1). Op het standaardbepalingsformulier zien de experts dus de relatieve moeilijkheid van een zeker cluster in verhouding tot de andere clusters van het domein Mondelinge taal of Beginnende geletterdheid. Enkele experts werden uitgenodigd hun oordeel toe te lichten nadat alle experts hun oordelen hebben gegeven over het verwachte aantal correcte observatiepunten/opdrachten van de grensleerling per cluster en volgde er een discussie. De experts konden hun definitieve oordeel op het formulier aangeven na deze discussie. Indien daar een discussie over ontstond kregen de experts soms ook feedback over de moeilijkheid van bepaalde opdrachten (i.e., normatieve feedback) of over de impact van hun oordelen op het percentage leerlingen dat in een bepaald functioneringsniveau zou vallen als de standaarden op een bepaalde waarde zouden worden gezet (i.e., impact feedback).

Conclusie:

De standaardbepalingsmethode is gemotiveerd en op de juiste wijze uitgevoerd. Op aspect N1.1 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

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

Bevindingen:

De beoordelaars/vakdeskundigen/experts bestonden uit leerkrachten en andere onderwijsprofessionals die als volgt zijn geselecteerd en getraind:

In maart 2019 startte de werving voor deelnemers aan de standaardbepaling op twee verschillende dagen (in juni 2019 en in september 2019). Hiertoe is er per mail een uitnodiging gestuurd aan de volgende leerkrachten en intern begeleiders: Contactpersonen van de deelnemende scholen aan het proef- en kwaliteitsonderzoek, aan de mensen die in focusgroepen mee hebben gedacht over de concepten voor Kleuter in beeld – Taal en aan degenen die screeningswerkzaamheden hebben uitgevoerd voor dit instrument. Daarnaast is een mail gestuurd aan experts die hebben deelgenomen aan de expertbijeenkomsten voor Kleuter in beeld – Taal en aan experts die bekend zijn uit de netwerken van Cito. Op social media is tot slot een oproep geplaatst om onderwijsprofessionals te werven voor de standaardbepaling. In totaal hebben 31 onderwijsprofessionals zich aangemeld, waarvan er uiteindelijk 20 ook mee hebben gedaan aan de twee standaardbepalingsdagen: 10 kleuterleerkrachten, 2 intern begeleiders, 3 kleuterleerkrachten die tevens intern begeleider zijn en 5 experts op het gebied van taal (onderwijsadviseurs/- ontwikkelaars/logopedist).

De deelnemers kregen voorafgaand aan de standaardbepalingsdag informatie over leerlijnen en elke standaardbepalingsdag begon met een uitleg over het ontwikkelingsproces van Kleuter in beeld – Taal en een beschrijving van de inhoud van het instrument. Het doel van de standaardbepaling en de methode (i.e., de 3DC-methode) werd vervolgens uitgelegd, waarna de groep in tweeën werd gesplitst. De ene helft van de deelnemers zette standaarden voor Mondelinge taal en de andere helft voor Beginnende geletterdheid. De onderwijsprofessionals werd gevraagd tijdens de standaardbepalings sessie de grensleerling voor ogen te houden.

Conclusie:

De beoordelaars/vakdeskundigen/experts zijn naar behoren geselecteerd en getraind. Op aspect N1.2 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:

De oordelen van de experts per cluster (subdomein) van de domeinen Mondelinge taal en Beginnende geletterdheid zijn te zien in respectievelijk Tabel 6.4 en Tabel 6.5. De grens voor het hele domein per expert, Mondelinge taal of Beginnende geletterdheid, is gelijk aan de som van de expert-oordelen per cluster (zie Tabel 6.4 en Tabel 6.5). De vier definitieve standaarden voor de domeinen Mondelinge taal en Beginnende geletterdheid zijn berekend door het gemiddelde van de expert-domeingrenzen te nemen.

Bij het analyseren van de resultaten van de standaardbepalings sessies is naar de impact van elke individuele expert op de grens van het hele domein gekeken, welke bepaald werd door het absolute verschil te berekenen tussen de grens zoals bepaald en de grens zoals die zou zijn als de betreffende expert buiten beschouwing werd gelaten. Daarnaast werd

er gekeken naar de samenhang tussen oordelen van de individuele expert en de oordelen van de overige experts door middel van de Ranking Similarity Index (RSI), de gemiddelde correlatie tussen de oordelen van een expert met de oordelen van de overige experts. Ten slotte is de overeenstemming tussen de expert-oordelen geëvalueerd met behulp van de gemiddelde interbeoordelaarscorrelatie (*GIR*) en de Finn coëfficiënt. Voor de analyse zijn de grenzen op de clusters omgezet naar de schaal van het hele domein (Mondelinge taal of Beginnende geletterdheid), omdat de clusters niet evenveel opdrachten/observatiepunten bevatten.

Aan de RSI is uit Tabel 6.4 te zien dat voor elke expert de oordelen redelijk tot goed samenhangen met de oordelen van de andere experts voor het domein Mondelinge taal ($RSI = 0,79 - 0,93$). Verder laat Tabel 6.4 zien dat voor Mondelinge taal de impact van de oordelen van de individuele experts op de domeingrenzen gering is. Een domeingrens verandert slechts in één geval met 1 schaalpunt als een individuele expert buiten beschouwing wordt gelaten, in alle andere gevallen met minder dan een punt (op een schaal van 0 tot 96). Tenslotte laten de gemiddelde interbeoordelaarscorrelatie ($GIR = 0,94$) en de Finn coëfficiënt ($Finn = 0,95$) zien dat voor Mondelinge taal de overeenstemming tussen de experts goed is.

Aan de RSI is uit Tabel 6.5 te zien dat voor elke expert de oordelen redelijk tot goed samenhangen met de oordelen van de andere experts voor het domein Beginnende geletterdheid ($RSI = 0,85 - 0,94$). Verder laat Tabel 6.5 zien dat voor Beginnende geletterdheid de impact van de oordelen van de individuele experts op de domeingrenzen gering is. Een domeingrens verandert in alle gevallen minder dan een punt (op een schaal van 0 tot 87). Tenslotte laten de gemiddelde interbeoordelaarscorrelatie ($GIR = 0,93$) en de Finn coëfficiënt ($Finn = 0,94$) zien dat voor Beginnende geletterdheid de overeenstemming tussen de experts goed is.

Conclusie:

Er is voldoende overeenstemming tussen de beoordelaars. Op aspect N1.3 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:

Dit criterium heeft betrekking op relatief normeren en is dus niet van toepassing (n.v.t.) op het instrument Kleuter in beeld - Taal, omdat we hier te maken hebben met absoluut normeren.

Conclusie:

n.v.t.

N2.2 Zijn de normgroepen representatief?

Bevindingen:

Dit criterium heeft betrekking op relatief normeren en is dus niet van toepassing (n.v.t.) op het instrument Kleuter in beeld - Taal, omdat we hier te maken hebben met absoluut normeren.

Conclusie:

n.v.t.

N2.3 Zijn de normen correct bepaald?

Bevindingen:

Dit criterium heeft betrekking op relatief normeren en is dus niet van toepassing (n.v.t.) op het instrument Kleuter in beeld - Taal, omdat we hier te maken hebben met absoluut normeren.

Conclusie:

n.v.t.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:

De interbeoordelaarsbetrouwbaarheid is bepaald om de betrouwbaarheid van de vier directe observaties (Oriëntatie op boek en verhaal, Oriëntatie op geschreven taal, Spreken, Gesprekken) bij de activiteiten in de kindroute te onderzoeken. Hiertoe is scholen gevraagd of een van de negen Cito-experts met ervaring voor de klas, mocht langskomen terwijl de leerkracht een activiteit met een kind uitvoerde. Zowel de leerkracht als de Cito-expert vulden steeds hetzelfde observatieformulier in. In totaal hebben 16 scholen meegedaan aan dit betrouwbaarheidsonderzoek. De interbeoordelaarsbetrouwbaarheid is berekend op basis van de correlatie tussen het leerkrachtoordeel en het oordeel van de Cito-expert, waarvoor per activiteit de som van de oordelen van de observatiepunten uit de uitgave is berekend. De eerder vastgestelde vier standaarden zijn vervolgens toegepast op de leerkrachtoordelen en de expertoordelen om het functioneringsniveau (<E1, E1, M2, E2 en >E2) te bepalen. Hierna is de overeenstemming tussen het functioneringsniveau volgens het leerkrachtoordeel en volgens het Cito-expertoordeel bepaald. Hiertoe zijn de absolute overeenstemming (percentage leerlingen dat volgens de oordelen van de leerkracht en de expert in hetzelfde niveau valt) en de gewogen Cohen's kappa, waarbij rekening wordt gehouden met toevallige overeenstemming en met partiële overeenstemming, uitgerekend.

Door het onvolledige (maar verbonden) design kunnen bij de subdomeinen geen traditionele maten als Cronbach's alpha worden uitgerekend als maat voor betrouwbaarheid. De betrouwbaarheden van de subdomeinen zijn echter te schatten door gebruik te maken van het feit dat alle opdrachten en observatiepunten, die zijn opgenomen in de kindroute, OPLM-geschaald zijn. Met behulp van het programma OPLAT uit OPLM is een coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie (KTT): de MAcc 'Accuracy of Measurement' (zie Box 1). Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt.

Conclusie:

De betrouwbaarheidsgegevens worden correct berekend. Op aspect B1 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

B2 Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen:

Tabel 8.2 laat zien dat de interbeoordelaarsbetrouwbaarheid in de vorm van de correlatie tussen de observaties van de leerkracht en de observaties van een Cito-expert uitstekend zijn bij de activiteiten van de subdomeinen voor Beginnende geletterdheid (Oriëntatie boek en verhaal = 0,88; Oriëntatie geschreven taal = 0,93). Tabel 8.2 toont ook dat bij Mondelinge taal de interbeoordelaarsbetrouwbaarheid bij het subdomein Spreken wederom uitstekend is (correlatie = 0,95) en bij Gesprekken redelijk (correlatie = 0,68). Een mogelijke verklaring voor de relatief lage correlatie bij Gesprekken is dat bij de activiteit Gesprekken kinderen in twee- of drietallen werken en geobserveerd worden, waardoor het denkbaar is dat waarnemingen (kinderen) verwisseld zijn. De absolute overeenstemming tussen functioneringsniveaus voor de vier activiteiten volgens het leerkrachtoordeel en volgens het Cito-expertoordeel is met waarden tussen 0,70 en 0,80 goed te noemen, terwijl deze overeenstemming bij de gewogen Cohen's kappa met waarden tussen 0,63 en 0,81 als goed tot uitstekend kan worden beschouwd (zie Tabel 8.6). Voor de interpretatie van deze overeenstemmingsmaten is uitgegaan van de vuistregel volgens Landis en Koch (1977) dat waarden van 0 – 0,2 als onvoldoende worden beschouwd; 0,21 – 0,40 als matig; 0,41 – 0,60 als redelijk; 0,61 – 0,80 als goed en 0,81 – 1,00 als uitstekend.

Tabel 8.7 en Tabel 8.8 tonen in de laatste kolom de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de niveauversies van respectievelijk de subdomeinen van Beginnende geletterdheid met waarden tussen 0,76 en 0,88 en Mondelinge taal met waarden tussen 0,70 en 0,96 in de kindroute. Op grond van het COTAN beoordelingssysteem voor meetinstrumenten waaraan geen zware consequenties voor leerlingen zijn verbonden, zoals bij Kleuter in beeld - Taal, kan derhalve de betrouwbaarheid van de niveauversies van Kleuter in beeld – Taal voldoende tot goed worden genoemd.

De geschatte betrouwbaarheidscoëfficiënt (MAcc) heeft alleen betrekking op de globale meetnauwkeurigheid van de subdomeinen en geeft geen beeld van de lokale meetnauwkeurigheid van de subdomeinen. Om inzicht te krijgen in de lokale meetnauwkeurigheid van de subdomeinen is de grootte van de meetfout op de vaardigheidsschaal afgebeeld. In Figuur 8.1 is ter illustratie de grootte van de meetfout op de vaardigheidsschaal afgebeeld van het subdomein Oriëntatie op boek en verhaal, waarbij tevens de verdeling van de vaardigheid van kleuters in groep 1 en in groep 2 is weergegeven. Figuur 8.1 laat zien dat volgens verwachting de meetfout kleiner is (en dus de meetnauwkeurigheid groter is) in het gemiddelde vaardigheidsgebied. Voor de andere subdomeinen kan op dezelfde wijze de lokale meetnauwkeurigheid afgebeeld worden en hieruit blijkt ook dat de meetfout kleiner is in het gemiddelde vaardigheidsgebied.

Betrouwbaarheidstabellen (ook wel verwarringsmatrices genoemd) laten de betekenis van de (lokale) meetnauwkeurigheid zien voor de functieniveaus die met het meetinstrument

gerapporteerd worden. Met behulp van simulaties is voor elk subdomein en niveauversie (* = niveau groep 1, ** = niveau midden groep 2, *** = niveau eind groep 2) nagegaan hoe vaak het functieniveau dat volgt uit de resultaten gelijk is aan het ware (gesimuleerde) functieniveau op het betreffende domein. In Tabel 8.9 - 8.12 is voor alle subdomeinen en niveauversies te zien hoe vaak het werkelijke functieniveau overeenkomt met het waargenomen functieniveau. Tabel 8.9 laat bijvoorbeeld zien dat 86% van de kleuters die volgens de resultaten in functieniveau 1 (<E1) vallen op het subdomein Oriëntatie op boek en verhaal, ook werkelijk in functieniveau 1 zitten. En ook dat bijvoorbeeld 14% van de kleuters die volgens de resultaten in functieniveau 2 (E1) vallen, in werkelijkheid het functieniveau 3 (M2) hebben voor Oriëntatie op boek en verhaal.

In de psychometrische literatuur zijn verschillende indices voorgesteld die de nauwkeurigheid (accuraatheid) van de classificaties in een betrouwbaarheidstabel samenvatten (zie onder andere Lee, Hanson, & Brennan, 2002), waarvan er in de Wetenschappelijke Verantwoording twee worden gerapporteerd: de plus/minus 1 niveau-index en de foutpercentage-index. De eerste index stelt als ambitieniveau dat 95% van de leerlingen dat in een functieniveau valt in werkelijkheid ook in dat functieniveau moet vallen, of één functieniveau daarboven of één functieniveau daaronder (Pillner, 1969; Wheadon en Stockford, 2010). In Tabel 8.9 – 8.12 zijn dit de gearceerde cellen. Dit ambitieniveau is gebaseerd op de veronderstelling dat geen enkel meetinstrument perfect meet en dat er dus altijd sprake is van misclassificaties. In dat licht bezien is de maximale accurateid die op het individuele niveau bereikt kan worden plus of minus één functieniveau. Tabel 8.13 ('Samenvattende indices voor de nauwkeurigheid van de subdomeinen en niveauversies') laat zien dat alle subdomeinen met activiteiten boven dit ambitieniveau liggen. Bij de niveauversie voor groep 1 (*) voor Fonologisch bewustzijn (0,91) en de drie niveauversies voor Luisteren (0,83, 0,88 en 0,90) en Woordenschat (0,91, 0,87 en 0,91) ligt de plus/minus 1 niveau-index iets onder dit ambitieniveau van 0,95. Een mogelijke verklaring hiervoor is dat, onder andere, kortere toetsen en makkelijke toetsen (zoals de kortere en makkelijke niveauversies voor groep 1 (*)) een lagere nauwkeurigheid blijken te hebben dan lange toetsen en toetsen met een grote spreiding in itemmoeilijkheid (zie Wheadon en Stockford, 2010).

Bij de tweede index (foutpercentage-index) wordt de nauwkeurigheid van de classificaties geëvalueerd door te kijken naar de vals-positief en vals-negatief fouten. Het conditionele vals-positief foutpercentage (P+) en het conditionele vals-negatief foutpercentage (P-) is de kans dat een kleuter in respectievelijk een hoger en lager functioneringsniveau valt dan het ware functioneringsniveau van de kleuter. Uit Tabel 8.13 is af te lezen dat misclassificaties vooral in positieve zin zijn: als sprake is van een misclassificatie vallen kleuters vaker in een hoger functieniveau dan hun werkelijke functieniveau en niet in een lager functieniveau. Met andere woorden, kleuters worden vaker niet dan wel benadeeld bij misclassificaties.

Conclusie:

De betrouwbaarheid van het instrument Kleuter in beeld – Taal is 'voldoende' als aangenomen mag worden dat het instrument geen zware consequenties voor de kleuters heeft en er rekening mee gehouden wordt dat er altijd sprake zal zijn van misclassificaties ter grootte van veelal maximaal één functieniveau. Op basis van het voorgaande wordt op aspect B2 aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

Validiteit

V1 Inhoudsvaliditeit: Dragen de items in de toets bij aan de validiteit van de toets (hierbij gaat het om aspecten als relevantie, objectiviteit en efficiëntie van de items)?

Bevindingen:

De Inhoudskaart Nederlandse taal – jonge kind fase 1 (SLO) is met name leidend geweest in de bepaling van de domeinen, subdomeinen en onderdelen. Op de uiteindelijk te meten leerdoelen hebben de inhoudskaart, aanvullende literatuur en feedback van onderwijsprofessionals een rol gespeeld. De gekozen domeinen, uitgewerkt in subdomeinen en onderdelen, zijn belangrijke inhouden en leerdoelen die in de kleuterperiode aan de orde moeten komen. Daar is een goede keuze in gemaakt. De leerkrachtroute en de kindroute gaan uit van dezelfde acht subdomeinen met de daarbij behorende 16 onderdelen.

Per subdomein is ook een toetsmatrijs bij de uitgave van de niveauversies van de opdrachten en van de observatiepunten van de activiteiten weergegeven.

Er is een juiste keuze gemaakt om mondelinge taalvaardigheid en beginnende geletterdheid niet samen te voegen om compensatiegedrag van leerlingen uit te sluiten. Ook i.v.m. het op tijd signaleren van aspecten die vallen onder beginnende geletterdheid en mondelinge taal en preventief te kunnen anticiperen op mogelijke leesproblemen, is het belangrijk dat de 2 betreffende domeinen zijn onderverdeeld in subdomeinen zodat de leerkracht zich een duidelijk beeld kan vormen voor een meer gerichte vervolgaanpak.

Er is een keuze gemaakt in de doelen die aan de orde komen in de groepen 1 en 2 in het observatie-instrument. Enerzijds voor de hanteerbaarheid van het instrument en anderzijds is bijvoorbeeld een bewuste keuze gemaakt om bijvoorbeeld Leesplezier niet op te nemen in de uitwerking met als reden dat de te meten vaardigheid niet objectief is vast te stellen. In de inhoudsverantwoording voor leerkrachten staat in het kort de keuzen die gemaakt zijn. Een leerkracht en schoolteam moeten zich dus wel realiseren dat niet alle doelen die in de groepen 1 en 2 aan de orde komen op deze domeinen gedekt worden. Voor de leerkrachtroute zijn er geen toetsitems uitgewerkt maar zijn doelen op een tweetal niveaus geoperationaliseerd in observatiepunten. In de 2 betreffende niveaus (eind groep 1 en eindgroep 2) is een duidelijke opklimmende moeilijkheidsgraad waar te nemen op de 8 subdomeinen en is ook een herkenbare niveauduiding uitgewerkt van de betreffende doelen (waarbij ook steeds een voorbeeld staat aangegeven).

Bij de kindroute zijn voor het domein Mondelinge taal bij spreken en gesprekken activiteiten uitgewerkt met een observatieformulier waar meer gedetailleerd observatiepunten zijn uitgewerkt dan bij de leerkrachtroute voor het betreffende subdomein. De observatiepunten zijn duidelijk terug te vinden in de leerdoelen van o.a. de inhoudskaart jonge kind van de SLO. Ditzelfde geldt voor het domein Beginnende geletterdheid bij de subdomeinen Oriëntatie op boek en verhaal en Oriëntatie op geschreven taal.

Voor de subdomeinen Woordenschat en Luisteren bij het domein Mondelinge taal en de subdomeinen Fonologische bewustzijn en Receptieve letterkennis bij het domein Beginnende geletterdheid zijn opdrachten geformuleerd via papieren opdrachtboekjes of digitale items. Het zijn in principe toetsitems die via observatie door de leerkracht worden gescoord.

De uitwerking van de items passen in de belevingswereld van de kinderen. Er wordt ook zoveel mogelijk gebruik gemaakt (vooral bij de opdrachten rond woordenschat) van foto's en een aantal items zijn uitgewerkt via hotspots (drie alternatieven in één plaatje). De

uitwerkingen zijn duidelijk uitgewerkt en de afbeeldingen/foto's binnen een item onderscheidend van elkaar.

Volgens de ontwikkelaars zijn de opdrachten voor kinderen door te werken op het platte vlak een voorbereiding op groep 3.

Er zit wel spanning op het principe opdrachten en straks de nieuwe wet PO (met daarin het nieuwe Toetsbesluit), die waarschijnlijk vanaf 1 augustus 2022 van kracht wordt, waar bepaald is dat bij kleuters in het kader van LVS geen schoolse toetsen mogen worden afgenomen. De opdrachten zijn vormgegeven als zijnde een toets. Dit is wel op nadrukkelijk verzoek van onderwijsprofessionals die zowel opdrachten als activiteiten wilden. Het zou goed zijn om ook een alternatief ertegenover te zetten in de vorm van activiteiten-observaties, of observaties vanuit al geplande activiteiten en vanuit betekenisvolle situaties in de groep.

Reactie van Cito op voorgaande alinea: hier zijn we ons van bewust. We hebben bij de keuze voor de opzet van het instrument rekening gehouden met de nieuwe wet én naar het veld geluisterd. Observaties staan centraal. Bij elke leerling en bij elk subdomein kan de leerkracht kiezen voor de leerkrachtroute (volledig gebaseerd op observaties) of de kindroute. De kindroute wordt ingezet als de leerkracht twijfelt over de vaardigheid van een kind op een bepaald subdomein, of als de leerkracht meer informatie wil. In principe wordt de kindroute dus door sommige leerkrachten, op sommige momenten, bij sommige leerlingen ingezet om sommige subdomeinen nader te bekijken. Dit nader bekijken gebeurt aan de hand van activiteiten of opdrachtjes. Het veld gaf aan zowel behoefte te hebben aan opdrachten als aan activiteiten. Zoals beschreven in de wetenschappelijke verantwoording in paragraaf 3.2 hebben opdrachten en activiteiten allebei voordelen. Per subdomein hebben we gekeken wat het meest passend is. Verder kunnen leerkrachten bij de opdrachten kiezen voor papier of digitaal. De regie en flexibiliteit van de leerkracht staan voorop. Bovendien zijn zowel de opdrachten als activiteiten zo speels mogelijk vormgegeven, zodat het aansluit bij de belevingswereld van kleuters en tegelijkertijd een objectief beeld oplevert van het vaardigheidsniveau van kleuters.

De Expertgroep is akkoord met deze reactie van Cito.

Opdrachten woordenschat

Er wordt een terechte keuze gemaakt om meer zelfstandige naamwoorden te toetsen dan werkwoorden en bijvoeglijke naamwoorden afgezet tegen de frequentie ervan in onze gesproken taal.

De selectie van woorden komen uit woordenlijsten die landelijk ook door methode- en toetsontwikkelaars worden gebruikt.

Opdrachten (begrijpend) luisteren

In deze vorm is het een juiste keuze om bij de opdrachten rond verhaalbegrip uit te gaan van korte teksten (verhalend en informatief) van 4 tot 6 zinnen om de concentratie vast te houden. Het is ook een logische keuze gezien de ontwikkelingsfase van een kleuter om alleen items rond figuurlijk taalgebruik aan te bieden op niveau midden groep 2 en eind groep 2.

Opdrachten Fonologisch bewustzijn

Het is belangrijk dat in deze opdrachten ook nadrukkelijk rijm is opgenomen. Omdat dit één van de belangrijke voorwaarden is om te leren lezen. Kinderen met dyslexie hebben moeite met de klankstructuur van een woord. Omdat ze niet goed weten welke klanken

samen een woord vormen, zijn ze veel minder gevoelig voor rijm. Het is voor de leerkracht belangrijk hier al vroeg zicht op te hebben.

Receptieve Letterkennis

Er is gekozen in de opdrachten voor letters die voorkomen in woorden die kleuters beheersen. Dit lijkt een logische keuze als in de opdracht uitgegaan wordt van betekenisvolle woorden. Zo toets je ook wat je wilt toetsen.

Bij het observatieformulier gesprekken en spreken is gebruik gemaakt van de niveaugrids van Praatjes peilen (2014), een aanpak om spelenderwijs mondelinge taalontwikkeling bij jonge kinderen te peilen en te stimuleren ontwikkeld door SLO i.s.m. het Expertisecentrum Nederlands. De peilingsactiviteiten en uitwerkingen van de niveaugrids zijn gebaseerd op de doelen taalontwikkeling jonge kinderen (SLO/UvA 2009). De niveaugrids geven op 5 niveaus (conform de uitwerking van de niveaus in de leerkrachtroute) via welke vragen op een bepaald niveau de peiling van taalontwikkeling (van bijvoorbeeld woordenschat, vloeiend en verstaanbaar vertellen, luisteren etc.) kan plaatsvinden. Dit is een goede inhoudelijke onderlegger geweest voor de uitwerking van de observatiepunten in de kindroute voor de observatiepunten gesprekken en spreken.

Conclusie:

Op aspect V1 wordt aan het instrument Kleuter in beeld – Taal het oordeel '**voldoende**' toegekend.

V2 Constructvaliditeit: Meet de toets in zijn geheel datgene wat hij beoogt te meten?

Bevindingen:

In de WV wordt verslag gedaan van onderzoek naar verschillende aspecten van constructvaliditeit: (1) structuur (unidimensionaliteit), (2) itemkwaliteit (psychometrische kwaliteit van de opdrachten/observatiepunten), (3) itembias, (4) soortgenotenvalliditeit in de vorm van convergente validiteit en (5) verschillen tussen relevante subgroepen.

1. Voor unidimensionaliteit wordt verwezen naar het kalibratieonderzoek van de kindroute in hoofdstuk 5, waarvan de conclusie is dat de modelpassing acceptabel is. Het is dus aannemelijk dat er sprake is van unidimensionaliteit voor elk van de subdomeinen van taalvaardigheid bij kleuters. Dit betekent dat met elke willekeurige subset van opdrachten/observatiepunten uit één subdomein dezelfde onderliggende vaardigheid kan worden vastgesteld. Uit een principale componentenanalyse blijkt dat de leerkrachtroute ook een eendimensionale schaal vormt. Er was één component met een eigenwaarde groter dan 1 en slechts één component boven de knip in een screeplot. Verder laat Tabel 9.2 zien dat alle observatiepunten een hoge componentenlading hadden (0,85 – 0,89). Hiermee is dus voldaan aan de noodzakelijke (doch niet voldoende) voorwaarde voor constructvaliditeit.
2. De beschrijvende gegevens van de opdrachten en observatiepunten (itemkwaliteit) zijn in hoofdstuk 4 gepresenteerd en de resultaten zijn besproken. Het bereik en de gemiddelden voor de p-waarden en de Rit-

waarden van de opdrachten en observatiepunten in de kindroute zijn weergegeven in Tabel 9.3. Alle Rit-waarden liggen boven de 0,23 en zijn daarmee alle voldoende (tussen de 0,20 en 0,29) tot goed (boven de 0,30) volgens het COTAN beoordelingssysteem. De gemiddelde Rit-waarden liggen tussen 0,39 en 0,73. De p-waarden voor de subdomeinen met opdrachten liggen tussen 0,33 en 0,95 en de gemiddelde moeilijkheid past goed bij de groepen (groep 1, midden groep 2 en eind groep 2) met waarden tussen de 0,60 en 0,83. Over de nauwkeurigheid van de itemparameters (constante 'c') is in Hoofdstuk 5 gerapporteerd, waaruit bleek dat de gemiddelde waarde van deze constante goed is in alle niveauversies. Ook de nauwkeurigheid bleek goed te zijn (een $c \leq 0,2$) bij de grote meerderheid van alle itemparameters (86%). Uit de resultaten van de kalibratieanalyses uit Hoofdstuk 4 viel al af te leiden dat de kwaliteit van de items hoog is, hetgeen dus bevestigd wordt door de 'klassieke' itemparameters: zowel de p-waarden als de Rit-waarden zijn goed te noemen volgens de COTAN-criteria.

3. Naar itembias (differentieel functioneren, DIF) is er een onderzoek uitgevoerd met betrekking tot geslacht. Hiertoe is voor alle opdrachten en observatiepunten de itemresponsen van zowel jongens als meisjes berekend en vervolgens is hier een chi-kwadraat toets voor berekend, analoog aan hoe dit is gedaan tijdens de kalibratie kindroute in Hoofdstuk 5 van de WV. Bij de subdomeinen met opdrachten (Fonologisch bewustzijn, Luisteren, Receptieve letterkennis en Woordenschat) bleek er nauwelijks sprake te zijn van itembias. Bij de observaties van het subdomein Gesprekken bleek dat, volgens verwachting, meisjes significant hoger scoren dan jongens op de meeste observatiepunten (bijv. op 'Laat de leerkracht en andere kinderen uitpraten in een groepsgesprek'). Verder werd er ook voor sommige items bias gerapporteerd. Bij de overige activiteiten (Spreken, Oriëntatie op boek en verhaal en Oriëntatie op geschreven taal) is nauwelijks of helemaal geen sprake van itembias. Omdat scholen zelf konden kiezen tussen een digitale afname van de opdrachten of een papieren afname, was een onderzoek naar itembias met betrekking tot afnamemodus niet mogelijk (hoewel er per observatiepunt van de activiteiten duidelijke verschillen waren m.b.t. afnamemodus).
4. Met betrekking tot de convergente validiteit, is de samenhang onderzocht tussen de (indirecte) observaties uit de leerkrachtroute en de vaardigheden zoals gemeten in de kindroute m.b.v. OPLM. Voor de subdomeinen met slechts één observatie (Woordenschat, Luisteren en Letterkennis) is het functioneringsniveau gelijk aan de observatie. In Hoofdstuk 7 van de WV is beschreven hoe het functioneringsniveau is bepaald voor de subdomeinen met meerdere observatiepunten. Uit Tabel 9.4 en Tabel 9.5 blijkt dat de vaardigheid oploopt per functioneringsniveau conform de verwachting. Alleen bij Oriëntatie op boek en verhaal is er tussen functioneringsniveau E2 en boven E2 ($>E2$) geen significante toename in vaardigheid in de kindroute. Ook is de samenhang per subdomein uitgerekend tussen de (indirecte) observaties uit de leerkrachtroute en de vaardigheid in de kindroute door middel van Spearman's rho (de observatie in de leerkrachtroute is van ordinaal meetniveau en de samenhang met de vaardigheid is niet per definitie lineair, maar is wel monotoon stijgend). Tabel 9.6 laat zien dat er een positieve samenhang is

tussen de (indirecte) observaties uit de leerkrachtroute en de vaardigheid zoals gemeten in de kindroute en dat de samenhang bij Beginnende geletterdheid hoger is dan bij Mondelinge taal. Verder blijkt uit Tabel 9.6 dat de samenhang sterker is tussen de indirecte observaties en de vaardigheidsschatting gebaseerd op directe observaties (Oriëntatie op boek en verhaal, Oriëntatie op geschreven taal, Spreken, Gesprekken) dan tussen de indirecte observaties en de vaardigheidsschatting gebaseerd op opdrachten (Fonologisch bewustzijn, Luisteren, Receptieve letterkennis, Woordenschat). Een mogelijke verklaring hiervoor is dat vaardigheden zoals luisteren en woordenschat lastiger in de dagelijkse lespraktijk zijn te observeren en in te schatten dan bijvoorbeeld spreekvaardigheid.

5. Met betrekking tot verschillen tussen relevante subgroepen zijn geboortedatum, geslacht en leerlinggewicht opgevraagd in het kwaliteitsonderzoek. Bij meer dan de helft van de leerlingen is echter geen gewicht opgegeven (1258) en bij de meeste subdomeinen hebben minder dan 15 kleuters met een gewicht 0,3 of 1,2 meegedaan. De verschillen tussen de gewichtscategorieën kunnen om die reden niet getoetst worden en dus zullen alleen verschillen tussen subgroepen voor leeftijd en geslacht worden besproken. De gemiddelde vaardigheid van de subgroepen zal worden vergeleken voor de kindroute en m.b.v. een chi-kwadraat toets zal worden nagegaan of de gerapporteerde functioneringsniveaus in de leerkrachtroute significant verschillen tussen subgroepen. Verder zal worden onderzocht of er sprake is van een samenhang tussen functioneringsniveau enerzijds en geslacht en leeftijdscategorie anderzijds.

Wat betreft verschillen tussen leeftijd is voor alle leerlingen de leeftijd berekend op 1 januari 2019 en zijn vervolgens de kleuters in de volgende vier leeftijdscategorieën ingedeeld: (0) ≤ 5 jaar; (1) 5 – 5,5 jaar; (2) 5,5 – 6 jaar; (3) > 6 jaar. Met uitzondering van het subdomein Spreken laat Tabel 9.7 zien dat de leeftijdscategorie een significant effect heeft op de gemiddelde vaardigheid op de subdomeinen in de kindroute ($\alpha = 0,01$). Tabel 0.1 in Appendix A geeft voor elk subdomein in de kindroute de gemiddelde vaardigheid per leeftijdscategorie weer. Uit deze tabel blijkt dat wanneer er sprake is van een significant effect deze conform verwachting is (i.e., hogere leeftijdscategorieën hebben gemiddeld een hogere vaardigheid). Tabel 9.9 laat zien dat ook in de leerkrachtroute het functioneringsniveau significant verschillend is tussen de leeftijdscategorieën en dat er sprake is van een positieve samenhang tussen functioneringsniveau en leeftijdscategorie. Tabel 0.3 en Tabel 0.4 in Appendix A geeft het percentage kinderen in elke leeftijdscategorie per functioneringsniveau weer in de leerkrachtroute. Uit deze tabellen blijkt dat oudere kinderen gemiddeld een hoger functioneringsniveau hebben conform verwachting.

Wat betreft verschillen tussen jongens en meisjes laat Tabel 9.8 zien dat bij de meeste subdomeinen er geen significante verschillen zijn tussen jongens en meisjes in de kindroute. Alleen bij Luisteren is er sprake van een klein significant effect van geslacht in de kindroute en is de gemiddelde Luistervaardigheid van de meisjes iets hoger dan van de jongens (zie Tabel 0.2

in Appendix A). Tabel 9.10 toont dat er sprake is van een klein significant verschil in functioneringsniveau tussen jongens en meisjes bij twee van de acht subdomeinen (Oriëntatie op geschreven taal en Gesprekken) in de leerkrachtroute. Tabel 0.5 in de Appendix laat zien dat in de leerkrachtroute bij de subdomeinen Oriëntatie op taal en Gesprekken een iets lager percentage meisjes onder functioneringsniveau E1 (<E1) valt en een iets hoger percentage in functioneringsniveau M2 dan bij jongens het geval is.

Conclusie:

De gerapporteerde resultaten in Hoofdstuk 9 (Validiteit) van de WV vormen een psychometrische ondersteuning voor de constructvaliditeit van het instrument Kleuter in beeld – Taal en er wordt dus gemeten wat men beoogt te meten, namelijk taalvaardigheid bij kleuters. Op basis van het voorgaande wordt op aspect V2 aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

Het volg-aspect

Va1 Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een adequate manier gemeten?

Bevindingen:

Uit het kalibratieonderzoek in Hoofdstuk 5 van de WV blijkt dat de items (opdrachten en activiteiten/observatiepunten) van het instrument Kleuter in beeld – Taal op een eendimensionale vaardigheidsschaal afgebeeld kunnen worden. Het instrument Kleuter in beeld – Taal bevat, naast een rapportage van het functioneringsniveau en gemiddelde vaardigheid voor elk van de acht subdomeinen (met afzonderlijke feedback per subdomein), een volgaspect om te kijken of het functioneringsniveau van een individueel kind tijdens de kleuterperiode gelijk blijft (stabiliseert), vooruit of achteruit gaat voor het hoofddomein Taal. Hiertoe kunnen respectievelijk drie signalen gegeven worden: een signaal "ga zo door" (kinderen die het volgens verwachting doen, signaal met een plantje), een "uitstekend" (taalsterke kinderen die het boven verwachting doen en mogelijk extra uitdaging kunnen gebruiken, signaal met een hardloper) of een "wees alert" (taalzwakke kinderen die het onder verwachting doen en mogelijk wat extra hulp kunnen gebruiken i.v.m. risico op latere leesproblemen, signaal met een oogje). Het signaal is afhankelijk van het aantal subdomeinen dat boven of onder het verwachte functioneringsniveau ligt, gegeven de periode dat de kleuter is geobserveerd (drie observatieperiodes: begin, van augustus t/m november; medio, van december t/m maart; eind, van april t/m juli). Hoofdstuk 7 (Beschrijving algoritmes) van de WV beschrijft gedetailleerd de algoritmes waarmee het functioneringsniveau op subdomeinen wordt gerapporteerd en het algoritme waarmee het signaal op het hoofddomein Taal wordt gegeven.

De leerkrachtroute kan eenvoudig via Kleuter in beeld online digitaal ingevuld worden. Op basis van beschrijvingen bij E1 en E2 kiest de leerkracht op een vijfpuntsschaal welk van de vijf niveaus voor elk observatiepunt (de leerkracht vult alle 16 observatiepunten in als er voor alle acht subdomeinen voor de leerkrachtroute gekozen wordt) het beste bij het kind past (zie Figuur 7.1 voor een voorbeeld). Bij de opdrachten uit de kindroute kan de leerkracht na afloop de opdrachten nakijken, aan de hand van de nakijkkaarten uit de

leerkrachtmap en kan daarna het aantal 'goed' invoeren in Kleuter in beeld online. Vervolgens wordt het functioneringsniveau automatisch berekend en wordt dit meteen zichtbaar in de rapportage. Tijdens de activiteiten uit de kindroute vult de leerkracht tijdens de activiteit al zoveel mogelijk het bijbehorend observatieformulier in, bijvoorbeeld direct in Kleuter in beeld online. De leerkracht kan er echter ook voor kiezen eerst het papieren observatieformulier in te vullen en de observaties op een later moment digitaal in te vullen. Bij de kindrapportage wordt per subdomein aangegeven of het kind boven, op, of onder het functioneringsniveau zit en is tevens de groei per subdomein te zien als er meerdere observatieperiodes zijn.

Conclusie:

Op aspect Va1 wordt aan het instrument Kleuter in beeld – Taal het oordeel '**voldoende**' toegekend.

Va2 Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?

Bevindingen:

In een noot geven de auteurs aan dat zij van mening zijn dat dit criterium binnen kleuterobservatie-instrumenten niet passend is en geven hiervoor de volgende beweegredenen:

Bij de kleuters is er geen vast curriculum die voor alle kleuters in Nederland gelijk is en de ontwikkeling van kleuters verloopt vaak grillig. Hierdoor vertonen kleuters vaak nog een heel heterogeen beeld, waar een schaalscore onvoldoende recht aan kan doen. Kleuter in beeld – Taal bevat wel een volgaspect om te kijken of het niveau van een individueel kind tijdens de kleuterperiode gelijk blijft, vooruit of achteruit gaat. In plaats van het verdisconteren van vooruitgang in een vaardigheidsscore voor Taal, geven de auteurs afzonderlijke feedback op acht subdomeinen. Een belangrijk doel van het instrument is om handvaten te geven voor het afstemmen van het onderwijsaanbod. De rapportage geeft een functioneringsniveau voor elk van de subdomeinen. Verfijning middels een vaardigheidsscore per subdomein vinden de auteurs ongepast, omdat een subdomein met te weinig items gemeten wordt om een betrouwbare schaalscore te geven.

Conclusie:

n.v.t.

Va3 Worden er gegevens verstrekt (aan de gebruiker) over hoe groei geïnterpreteerd dient te worden?

Bevindingen:

Wensen uit het veld waren dat het instrument Kleuter in beeld – Taal niet alleen een resultaat op Taal als totaal rapporteert, maar ook informatie over subdomeinen. Dit levert belangrijke informatie voor de leerkracht om het onderwijs op de kinderen af te stemmen, zoals een signaal als het kind het onder, op of boven verwachting doet en inzicht in groei, hoe het kind zich ontwikkelt in de loop van de tijd (progressiebepaling). Op basis van alle

wensen is er gekozen voor drie rapportagevormen, welke alle kunnen worden opgevraagd via Kleuter in beeld online: kindrapportage, groepsrapportage en groepsoverzicht.

Op pagina 102 van de WV wordt een voorbeeld Kindrapportage getoond met hierin aangegeven de observatieperiode(s) en het functioneringsniveau per subdomein. Hierdoor is onmiddellijk te zien op welke subdomeinen het kind boven, op of onder het functieniveau zit en is tevens de groei per subdomein zichtbaar als er meerdere observatieperiodes zijn. Via de online kindrapportage is het mogelijk de gegeven antwoorden van een kind in te zien wanneer de digitale opdrachten gemaakt zijn. Zo is bijvoorbeeld bij het subdomein Fonologisch bewustzijn snel te zien of een kind bijvoorbeeld meer moeite heeft met een bepaald onderdeel (bijvoorbeeld synthese). Op pagina 103 van de WV wordt een voorbeeld Groepsrapportage getoond met hierin de resultaten van de observaties voor de kinderen uit een groep voor elk van de afgeronde subdomeinen. Via verschillende kleuren kan een leerkracht snel zien of een groepje kleuters bijvoorbeeld moeite heeft met een bepaald subdomein en juist op dat gebied wat extra uitleg en oefening nodig heeft. Op pagina 104 van de WV wordt een voorbeeld Groepsoverzicht uit Kleuter in beeld online getoond (niet alleen voor Taal maar ook voor andere domeinen), waarin voor alle kinderen in een groep is te zien welke signalen ze hebben ontvangen voor alle domeinen die zijn ingevuld. De leerkrachtmap Kleuter in beeld – Taal geeft tenslotte tips voor het interpreteren en analyseren van de rapportages, aanvullende signalering bij problemen en hoe de leerkracht de gegevens kan gebruiken voor de onderbouwing van het handelen.

Conclusie:

Op aspect Va3 wordt aan het instrument Kleuter in beeld – Taal het oordeel '**voldoende**' toegekend.

Inzicht in leervorderingen

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

Bevindingen:

Online zijn twee rapportages beschikbaar: een kindrapportage en een groepsrapportage.

In de kindrapportage wordt per subdomein via een balk aangegeven waar het kind zich bevindt qua niveau (begin-midden groep 1, eind groep 1-begin groep 2, midden groep 2 en eind groep 2). Hierdoor zie je de groei per (sub)domein. Een stippellijn geeft de periode aan waarin de observatie heeft plaatsgevonden. De afstand van de balkjes tot de stippellijn geven helder aan of de leerling functioneert op het gewenste niveau, eronder of juist daarboven. Een grafiekje per subdomein geeft aan of er groei heeft plaatsgevonden (dus je kunt bijvoorbeeld onder de stippellijn functioneren maar wel gegroeid zijn). Voor de kindrapportage is een leerkracht- en een ouderversie beschikbaar. Bij de ouderversie zijn het totaalsignaal, de gevolgde route en persoonlijke notities niet zichtbaar gemaakt.

Een ouder/verzorger kan niet makkelijk zelfstandig de rapportage lezen want het vraagt kennis over interpretatie van niveaus, hoe je liggende of opgaande streepjes duidt etc. Het is ook de vraag of dit verstandig is. Het geeft wel een goede onderlegger voor een gesprek tussen leerkracht en ouders/verzorgers, naast nog andere ervaringen, evaluaties

etc. om het over de ontwikkeling van het kind te hebben, de gevolgd en te volgen aanpak en wat ook thuis wenselijk en mogelijk is.

De groepsrapportage is alleen voor de leerkracht toegankelijk. Het is een overzichtelijk schema waar direct duidelijk wordt wat de resultaten van de observaties zijn van de kinderen (met leeftijds aanduiding) op de subdomeinen, indicatie van het niveau op de subdomeinen door verschillend kleurgebruik en aangegeven niveau, resultaten al dan niet gebaseerd op leerkrachtroute of kindroute, gebruik van een pictogrammen om aan te duiden of een kind overall al dan niet volgens verwachting zich ontwikkelt, ruimte via notities (doorklikken op pictogram enveloppe) en observatieperiode. Via het overzichtsschema kun je bij alle leerlingen doorklikken om na te gaan hoe en op welke onderdelen van de subdomeinen het betreffende kind functioneert. Door het schema heb je ook een totaaloverzicht van de hele groep.

Conclusie:

Op aspect I1 wordt aan het instrument Kleuter in beeld – Taal het volgende oordeel toegekend: **'voldoende'**.

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

Bevindingen:

In de kindrapportage is er een evaluatie van de leervorderingen op basis van functioneringsniveaus (boven, op of onder het niveau) van kinderen per subdomein en signalen ("uitstekend", "ga zo door" en "wees alert") voor het totaal van Taal als alle acht subdomeinen zijn gemaakt. Op basis van deze belangrijke informatie uit de evaluatie van de leervorderingen kunnen vervolgstappen worden geformuleerd door het onderwijs beter op het kind af te stemmen. De leerkrachtmap Kleuter in beeld – Taal geeft hiervoor tips in de vorm van het interpreteren en analyseren van de rapportages, aanvullende signalering bij problemen en hoe de leerkracht de gegevens kan gebruiken voor de onderbouwing van het handelen.

In de handleiding is aandacht hoe een leerkracht komt van signaleren naar handelen door globaal een aantal stappen te benoemen. De gegevens die worden verkregen uit Kleuter in beeld worden in een totaalaanpak geplaatst. Hierbij wordt gewezen op de doelen waar de leerkracht mee aan de slag wilt gaan, bevindingen en analyses verkregen via o.a. Kind in beeld te leggen naast de doelen uit bijvoorbeeld de inhoudskaart Taal voor het jong kind (SLO), materialen uit de methoden, eigen observaties etc. Dit is passend bij het uitgangspunt van dit instrument om de regie en ook de te maken keuzen bij de leerkracht neer te leggen. De leerkracht krijgt wel via de uitkomsten van observaties, opdrachten een beeld hoe een leerling zich ontwikkelt, maar moet dit zelf concreet vertalen in een aanpak.

In de hoofdstukken over aanvullende signalering en van signaleren naar handelen wordt het belang van vroegtijdige signalering van een taalontwikkelingsstoornis, problemen met (leren) lezen en/of niet vloeiend spreken aangegeven en wordt o.a. verwezen naar de website van Siméa en de website van het stotterfonds.

Voor kinderen die net in Nederland zijn, wordt wel aangegeven om het eerste halfjaar dat ze in Nederland zijn nog niet het observatie-instrument in te zetten. Het lijkt logisch de uitkomsten van deze kinderen, indien het wel na een half jaar wordt ingezet, anders te

interpreteren en consequenties uit te trekken voor de aanpak dan voor kinderen die de Nederlandse taal al vanaf kleins af aan krijgen aangeleerd. Het zou wenselijk zijn hier iets meer over aan te geven of ook te verwijzen naar sites voor meer inzicht en handvatten.

Conclusie:

Op aspect I2 wordt aan het instrument Kleuter in beeld – Taal het oordeel '**voldoende**' toegekend.

Referentieniveaus

R1 Sluit de inhoud van de toets aan op de kennis en vaardigheden zoals omschreven in de referentieniveaus van het betreffende domein (voor toetsen vanaf groep 6)?

Bevindingen:

Bevindingen:

Het gaat hier om een observatie-instrument voor het jonge kind en daardoor kan het nog niet inhoudelijk rechtstreeks aansluiten op de kennis en vaardigheden zoals omschreven in de referentieniveaus.

De gekozen domeinen en subdomeinen anticiperen hier wel op. Aansluitend op de inhoudskaart jonge kind SLO waar de (sub)domeinen en onderdelen voor een belangrijk deel op gebaseerd zijn is op o.a. op mondelinge taalvaardigheid een inhoudslijn uitgewerkt waarin fase 1 jonge kind, fase 2 en fase 3 (bovenbouw) naast elkaar zijn beschreven. Hierdoor zie je hoe het relateert met de aanbodsdoelen door de leerjaren heen en is ook een duidelijke relatie gemaakt met de referentieniveaus.

Conclusie:

Op aspect R1 wordt aan het instrument Kleuter in beeld – Taal het oordeel '**voldoende**' toegekend.

3. Verzamelstaat

Kwaliteitsaspect	Code	Oordeel
De kwaliteit van de steekproef	S1	Voldoende
	S2	Voldoende
	S3	Voldoende
	S4	Voldoende
Normering	N1.1	Voldoende
	N1.2	Voldoende
	N1.3	Voldoende
	N2.1	n.v.t.
	N2.2	n.v.t.
	N2.3	n.v.t.
Betrouwbaarheid	B1	Voldoende
	B2	Voldoende
Validiteit	V1	Voldoende
	V2	Voldoende
Volg-aspect	Va1	Voldoende
	Va2	n.v.t.
	Va3	Voldoende
Inzicht in leervorderingen	I1	Voldoende
	I2	Voldoende
Referentieniveaus	R1	Voldoende

4. Literatuurlijst

- Schouwstra, S. en Vloedgraven, J. (2020). *Wetenschappelijke verantwoording Kleuter in beeld – Taal*. Arnhem: Cito B.V.
- Fitzpatrick, A.R. (1984, April). *Social influences in standard setting: The effect of group interaction on individuals' judgments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.