1. Algemene informatie

Algemeen en meetpretentie
Kleuter in Beeld – Rekenen 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 rekenen 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 – Rekenen is bedoeld om de rekenvaardigheid 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 vijf subdomeinen binnen Rekenen.
- 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:
De keuze voor de verschillende inhoudsgebieden binnen Kleuter in Beeld – Rekenen is grotendeels gebaseerd op de inhoudskaart Rekenen-Wiskunde kleuters (fase 1) van SLO (2018). Die inhoudskaart hanteert een indeling in zes subdomeinen. Voor Kleuter in Beeld – Rekenen is gekozen om te rapporteren op vijf subdomeinen, waarbij het subdomein Verhoudingen uit de SLO inhoudskaart is ondergebracht bij de subdomeinen Getalbegrip en Meten.
De keuze voor de subdomeinen en de daaronder vallende onderdelen wordt toegelicht in de wetenschappelijke verantwoording.

Inhoud van het toetspakket
Het toetspakket Kleuter in Beeld – Rekenen bestaat uit de volgende documenten:

- Wetenschappelijke verantwoording, deze bevat informatie over
  - Uitgangspunten van de ontwikkeling (hfst 2)
  - Beschrijving meetinstrument (hfst 3)
  - Dataverzameling (hfst 4)
  - Kalibratie (hfst 5)
  - Standaardbepaling voor de kindroute (hfst 6)
  - Beschrijving algoritmes (hfst 7)
  - Betrouwbaarheid en meetnauwkeurigheid (hfst 8)
  - Validiteit (hfst 9)
  - Afname en rapportage (hfst 10)
- Handleiding/Leerkrachtmap Kleuter in Beeld – Rekenen
- Papieren opdrachtenboekjes
- (Handleiding) Kleuter in Beeld online (via Basispoort)
- 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 Liza Kozlowska MA (secretaris).
Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording (WV) en handleiding veelal letterlijk vermeld.

De kwaliteit van de dataverzameling

S1 Is de steekproef representatief?

Bevindingen:
Bij Kleuter in Beeld – Rekenen hebben we te maken met een beoordeling van kleuters via een leerkrachtroute of via een kindroute. De leerkrachtroute is gebaseerd op indirecte observaties van de leerkracht over een langere periode waarop deze het niveau van de kleuter inschat. De leerkracht kiest voor de kindroute wanneer er twijfels bestaan over de beheersing van of meer van de vijf subdomeinen bij rekenen. De kindroute bestaat bij rekenen uit twee opdrachten en drie activiteiten met in totaal 16 onderdelen. Activiteiten betreffen directe observaties waarbij de kleuter en leerkracht een activiteit uitvoeren en de leerkracht de uitvoering waardeert met een observatiepunt op basis van zijn observaties. Er zijn diverse onderzoeken uitgevoerd om de observaties zo gestandaardiseerd en objectief mogelijk te maken. Omdat het niveau waar kleuters mee binnenkomen zeer divers is, past de inhoud van het observatie-instrument 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 2019 en een kwaliteitsonderzoek in februari 2020.

Een belangrijk doel van het proefonderzoek in 2019 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. Al deze scholen hebben per post een uitnodiging ontvangen om te participeren in het onderzoek, waarvoor zich 42 scholen hebben aangemeld. Uiteindelijk hebben 38 scholen en 792 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 2020 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 910 scholen getrokken voor de afname. Uiteindelijk hebben 49 scholen met 1037 leerlingen meegedaan.

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 sekse. 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 sekse is een tweedeling gemaakt naar jongens en meisjes.

Tabel 4.3 en 4.4 laten zien dat het percentage scholen met meer dan 10% achterstandskinderen niet significant afwijkt van de hele populatie scholen. Ook de verdeling over regio's en de urbanisatiegraad in de steekproef wijkt niet af van de verdeling in heel Nederland. De verhouding van jongens en meisjes in de steekproef is ook niet significant verschillend van de verhouding in de populatie.

Conclusie:
voldoende

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

Bevindingen
Voor elk van de twee subdomeinen Getalbegrip en Meten zijn boekjes gemaakt van 15 opdrachten. Voor elk subdomein zijn de zestig opdrachten verdeeld over vier boekjes voor leerjaar 2. Uit elk boekje zijn de drie tot vier makkelijkste opdrachten ook opgenomen in het boekje voor leerjaar 1. De opdrachten die de link zijn tussen de boekjes komen in meer boekjes voor.
Alle kinderen maakten twee subdomeinen uit de kindroute: de twee subdomeinen met opdrachten (Getalbegrip en Meten). Daarnaast deden minstens acht kinderen één activiteit (Bewerkingen, Meetkunde of Verbanden). Het ontwerp voor de combinaties van de verschillende subdomeinen staat in tabel 4.6. In tabel 4.7 staat het aantal observaties

per versie.
De conclusie van het voorgaande is dat de boekjes sterk verankerd zijn en dat we dus te maken hebben met een onvolledig maar verbonden dataverzamelingsdesign.

Uit het kalbratieonderzoek blijkt dat het instrument past bij het itemreponsmodel OPLM. De items passen bij het model, blijkend uit de S-toetsen, en het instrument als geheel past bij het model, blijkend uit de R1c-toetsen. Het voorgaande betekent dat er sprake is van een ééndimensionele vaardigheidsschaal waar items en kleuters op afgebeeld kunnen worden. Zie hoofdstuk 5 van de wetenschappelijke verantwoording voor meer details van het kalbratieonderzoek.

Conclusie:
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:
Bij S1 is opgemerkt dat de 49 scholen uit de steekproef die hebben deelgenomen aan het kwaliteitsonderzoek representatief zijn naar regio, urbanisatiegraad, percentage achterstandsleerlingen en sekse. In de wetenschappelijke verantwoording wordt echter niet ingegaan op de representativiteit van de observatoren/leerkrachten uit de steekproef voor de observatoren die het instrument in de praktijk gaan gebruiken. In het beoordelingskader staat daarover ”……. moeten ook de gebruikte observatoren een goede afspiegeling zijn van de observatoren die het instrument in de praktijk gaan gebruiken, en ook de randvoorwaarden waaronder de observaties worden gemaakt moet een goede afspiegeling zijn van de praktijk waarin het instrument gebruikt gaat worden. Een proefonderzoek met uiterst getrainde onderwijskundigen in een laboratoriumsituatie geldt bijvoorbeeld niet als zodanig.” Het verdient aanbeveling om in het vervolg kort aandacht aan dit criterium te geven.

Conclusie:
voldoende. Daarbij wordt wel aangetekend dat de kwaliteit van de verantwoording ermee gebaat zou zijn als er meer informatie wordt verstrekt over de representativiteit van de observatoren/leerkrachten en de randvoorwaarden waaronder het instrument wordt gebruikt.

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 – Rekenen 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) Veelgestelde 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 zo objectief mogelijk observaties kan uitvoeren en weergeven.

Conclusie:
voldoende

Normering

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

Bevindingen:
Het doel van de standaardbepaling in het onderhavige observatie-instrument is om met leerkrachten en onderwijsprofessionals na te gaan wat met betrekking tot Rekenen 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 domeinen van Rekenen 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.

Voor de standaardbepaling is de 3DC-methode (Data Driven Direct Consensus) gebruikt (Feskens, Keuning, Van Til, & Verheyen, 2014, Keuning, Straat & Feskens, 2017). ondersteund met empirische informatie uit het kwaliteitsonderzoek. Bij de 3DC-standaardbepaling wordt gewerkt met een team van experts op het vakgebied. De experts bekijken 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. Kleuter in Beeld – Rekenen bestaat uit vijf clusters: Getalbegrip, Bewerkingen, Meten, Meetkunde en Verbanden. 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).

Conclusie:
voldoende

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

Bevindingen:
In totaal hebben 30 onderwijsprofessionals meegedaan aan vier standaardbepalingsdagen: 15 kleuterleerkrachten, 7 intern begeleiders, 1 kleuterleerkracht die tevens intern begeleider is en 7 experts op het gebied van rekenen/jonge kind.
Voorafgaand aan elke standaardbepalingsdag kregen de deelnemers informatie over leerlijnen. In verband met de online standaardbepaling kregen de deelnemers ook de opdrachten en observatiepunten toegestuurd en het verzoek een oordeel te vormen hoe een leerling eind groep 1 de activiteiten en opdrachten zou doen. De standaardbepaling verliep via Teams en begon met een uitleg over het ontwikkelingsproces van Kleuter in Beeld – Rekenen en een beschrijving van de inhoud van het instrument. Vervolgens werd het doel van de standaardbepaling en de methode uitgelegd. Daarna gingen de experts in Teams in drie groepjes uiteen om de eerste grens te bepalen. In elk groepje konden de experts onderling overleggen. Een toetsdeskundige van Cito was beschikbaar voor vragen. Tijdens de sessie bekeken de experts voor drie subdomeinen een cluster met opdrachten of observatiepunten en vormden zij zich een oordeel over hoeveel opdrachten/observatiepunten het kind naar verwachting correct zou hebben op dat cluster als zijn/haar vaardigheid zich precies onder E1 van Rekenen bevindt. De experts werd gevraagd de grensleerling voor ogen te houden. De experts gaven hun oordeel aan opeen standaardbepalingsformulier (zie figuur 6.1). Nadat de experts hun oordelen hebben gegeven over het aantal correcte opdrachten/observatiepunten van de grensleerling, werden enkele experts uitgenodigd hun oordeel toe te lichten en volgde een discussie. Nadat de experts op alle clusters een definitief oordeel hebben gegeven over de grens onder E1/E1, gingen ze door met de tweede, derde, vierde en laatste grens.

Conclusie:
voldoende

N1.3 Is er voldoende overeenstemming tussen de beoordelaars?

Bevindingen:
Bij het analyseren van de resultaten van de standaardbepalingssessies 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.
Aan de RSI is te zien dat voor elke expert de oordelen redelijk tot goed samenhangen met de oordelen van de andere experts (RSI = 0.96 – 0,98). De impact van de oordelen van de individuele experts op de domeingrenzen is gering. Als een individuele expert buiten beschouwing wordt gelaten, verandert een domeingrens in alle gevallen met minder dan een punt (op een schaal van 0 tot 90). De gemiddelde interbeoordelaarscorrelatie (*GIR* =

0,88) en de Finn coëfficiënt (Finn = 0,98) laten zien dat de overeenstemming tussen de experts zeer goed is.

Conclusie:
voldoende

N2.1 Zijn de normgroepen groot genoeg?

Bevindingen:
Criterium betreft relatief normeren terwijl hier sprake is van absoluut normeren.

Conclusie:
n.v.t.

N2.2 Zijn de normgroepen representatief?

Bevindingen:
Criterium betreft relatief normeren terwijl hier sprake is van absoluut normeren.

Conclusie:
n.v.t.

N2.3 Zijn de normen correct bepaald?

Bevindingen:
Criterium betreft relatief normeren terwijl hier sprake is van absoluut normeren.

Conclusie:
n.v.t.

Betrouwbaarheid

B1 Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:
In de kindroute is de interbeoordelaarsbetrouwbaarheid tussen het leerkrachtoordeel en het oordeel van een Cito-expert berekend met de correlatie-coëfficiënt en de absolute G-coëfficiënt. Bij alle drie subdomeinen is de interbeoordelaarsbetrouwbaarheid en -overeenstemming goed, bij Meetkunde en Verbanden zelfs uitstekend.
Vanwege het incomplete design kan Cronbach's alpha niet berekend worden maar wel de MAcc (accuracy of measurement). De twaalf coëfficiënten variëren van 0,79 tot 0,92 en voldoen daarmee aan het COTAN criterium voor meetinstrumenten waaraan geen zware consequenties voor leerlingen verbonden zijn.

Conclusie:
voldoende

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

Bevindingen:
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 Meetkunde, 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 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 8.8 en 8.9 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 78% van de kleuters die volgens de resultaten in functieniveau 1 vallen op het subdomein Getalbegrip, ook werkelijk in functieniveau 1 zitten.

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 8.8 en 8.9 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 accuraatheid die op het individuele niveau bereikt kan worden plus of minus één functieniveau. Tabel 8.10 laat zien dat alle subdomeinen en niveauversies boven dit ambitieniveau liggen.

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 8.10 is af te lezen dat de 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:
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:
Bij Kleuter in beeld – Rekenen hebben vier informatiebronnen de ontwikkeling gestuurd: de input van onderwijsprofessionals, wetenschappelijke kennis over rekenen, de expertise van toetsdeskundigen en informatie uit empirisch onderzoek. Vooral de inhoudskaart Rekenen-Wiskunde kleuters (fase 1) van SLO heeft een belangrijke rol gespeeld bij het bepalen van de subdomeinen en onderdelen. Maar ook andere bronnen (zie wetenschappelijke verantwoording p. 10) zijn geraadpleegd.
De items behorende bij een bepaald subdomein vormen gezamenlijk een goede afspiegeling van dat subdomein.
De instructie die aan de leerkrachten gegeven worden bij de items zijn duidelijke zodat het voor de leerkracht helder is en het in principe niet uitmaakt wie de toets afneemt.
Het antwoordmodel laat geen ruimte voor interpretatie.

Conclusie:
De items in de toets dragen **voldoende** bij aan de validiteit van de toets.

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

Bevindingen:
Kleuters in beeld – Rekenen bestaat uit de volgende subdomeinen en onderdelen

Subdomein	**Onderdelen**
Getalbegrip	Telrij Hoeveelheden Getallen Relaties tussen telrij, hoeveelheden en getallen incl Verhoudingen
Bewerkingen	Optellen en aftrekken met hele getallen (tot ten minste 20) (optellen, aftrekken, splitsen) Vermenigvuldigen en delen met hele getallen (tot ten minste 20) ((ver)delen)
Meten	Lengte en omtrek Oppervlakte Inhoud Gewicht Tijd

	Geld Incl Verhoudingen
Meetkunde	Orienteren in de ruimte Construeren Opereren met vormen en figuren
Verbanden	Verbanden

Alle afzonderlijke onderdelen worden bemeten. De afzonderlijke onderdelen samen vormen een goede afspiegeling van het betreffende subdomein.

Vanuit psychometrische analyse wordt opgemerkt dat het bij begripsvaliditeit erom gaat te toetsen of het meetinstrument inderdaad de eigenschap meet die wordt verondersteld. Hiertoe worden gepresenteerd gegevens over de structuur (dimensionaliteit), de psychometrische kwaliteit van de opdrachten en observatiepunten, soortgenoten validiteit en gegevens over verschillen tussen relevante groepen.

Structuur
Uit het kalibratieonderzoek blijkt dat er op enige uitzonderingen na sprake is van niet-significante S-toetsen. Zij vormen een kwantitatieve ondersteuning van de conclusie dat elke set opdrachten/observatiepunten behorend tot één subdomein, een eendimensioneel construct representeren. Ook de R1c waarden laten zien dat de modelpassing acceptabel is en bieden daarmee een ondersteuning van de validiteit. De leerkrachtroute vormt ook een eendimensionale schaal zoals blijkt uit een principale componentanalyse.

Itemkwaliteit
In Tabel 9.3 zijn het bereik en de gemiddelden weergegeven voor de p-waarden en de rit-waarden van de opdrachten en observatiepunten van subdomeinen in de kindroute. Alle rit-waarden liggen boven de 0,23 en zijn daarmee voldoende tot goed volgens het COTAN beoordelingssysteem.

Itembias
Onderzoek naar differentieel functioneren met betrekking tot sekse kon niet uitgevoerd worden vanwege te weinig observaties.

Convergente validiteit
In het kader van de convergente validiteit is de samenhang bekeken tussen de (indirecte) observaties uit de leerkrachtroute en de vaardigheden zoals gemeten in de kindroute. Voor de vijf subdomeinen is Spearman's rho uitgerekend (0,49 – 0,64). Er is een positieve samenhang tussen de observaties uit de leerkrachtroute en de vaardigheid zoals gemeten in de kindroute.

Verschillen tussen relevante subgroepen
In het kwaliteitsonderzoek zijn geboortedatum en geslacht verzameld. Tabel 9.10 laat zien dat hoe ouder het kind, hoe hoger het functioneringsniveau. Er zijn bij alle subdomeinen geen significante verschillen tussen jongens en meisjes.

Conclusie:
De toets in zijn geheel meet wat hij beoogt te meten. Het oordeel is '**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 blijkt dat de items (opdrachten en activiteiten/observatiepunten) van het instrument Kleuter in Beeld – Rekenen op een eendimensionale vaardigheidsschaal afgebeeld kunnen worden. Het instrument Kleuter in Beeld – Rekenenl bevat, naast een rapportage van het functioneringsniveau en gemiddelde vaardigheid voor elk van de vijf 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 Rekenen. Hiertoe kunnen respectievelijk drie signalen gegeven worden: een signaal "ga zo door" (kinderen die het volgens verwachting doen), een "uitstekend" (taalsterke kinderen die het boven verwachting doen en mogelijk extra uitdaging kunnen gebruiken) of een "wees alert" (taalzwakke kinderen die het onder verwachting doen en mogelijk wat extra hulp kunnen gebruiken). 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) beschrijft gedetailleerd de algoritmes waarmee het functioneringsniveau op subdomeinen wordt gerapporteerd en het algoritme waarmee het signaal op het hoofddomein Rekenen 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 vijf subdomeinen voor de leerkrachtroute gekozen wordt) het beste bij het kind past. 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:
voldoende

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

Bevindingen:
In een noot bij Kleuter in Beeld – Taal 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. De Expertgroep ging bij de beoordeling van Kleuter in Beeld – Taal mee in deze argumentatie.

Voorgaande noot ontbreekt in Kleuter in Beeld – Rekenen. Hier is het oordeel van Kleuter in Beeld – Taal overgenomen.

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 – Rekenen niet alleen een resultaat op Rekenen 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. Beide rapportages worden in de online handleiding voldoende uitgelegd.

Op pagina 90 van de wetenschappelijke verantwoording 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. Op pagina 92 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 92 wordt een voorbeeld

Groepsoverzicht uit Kleuter in Beeld online getoond (niet alleen voor Rekenen 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. Ten slotte geeft de leerkrachtmap 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:
Er worden voldoende gegevens verstrekt aan de gebruiker om de groei te kunnen interpreteren. Het oordeel is '**voldoende'.**

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:
De leerkracht kan de kindrapportage opslaan als pdf en printen voor meerdere kinderen. Er is een leerkrachtversie en een ouderversie beschikbaar. Bij de ouderversie zijn het totaalsignaal (uitstekend, ga zo door en wees alert), de gevolgde route en persoonlijke notities niet zichtbaar, bij de leerkrachtversie wel. Voornoemde werkwijze wordt onderbouwd.

Conclusie:
De toetsaanbieder levert een format voor een geschreven toelichting bij de leervorderingen van de leerling die (ook) voor ouders/voogden/verzorgers begrijpelijk is. Het oordeel is '**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 functioneringniveaus (boven, op of onder het niveau) van kinderen per subdomein en signalen (uitstekend, ga zo door en wees alert) voor het totaal van Rekenen als alle vijf 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 – Rekenen 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.

Conclusie:
Er is een evaluatie van de leervorderingen en er worden suggesties gegeven voor de leerkracht om de stap naar het handelen te maken. Het oordeel is '**voldoende'.**

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:
n.v.t.

Conclusie:
n.v.t.

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*	**n.v.t.**

4. Literatuurlijst

- Schouwstra, S., Vloedgraven, J., de Boer, A., Lansink, N. & Nikkels, L. (2020). *Wetenschappelijke verantwoording Kleuter in Beeld – Rekenen.* Arnhem: Cito B.V.