1. Uitgangspunten van de toetsconstructie

Bij onderstaande beoordeling van de kwaliteitsaspecten met bijbehorende codes van het voornoemde beoordelingskader worden passages uit de wetenschappelijke verantwoording en de Handleiding veelal letterlijk vermeld. De wetenschappelijke verantwoording heeft betrekking op de uitgangspunten van de toetsconstructie, de normen, de betrouwbaarheid en meetnauwkeurigheid en de validiteit. De Handleiding heeft betrekking op het gebruik van de toets, communicatie over de toetsgegevens en de inhoudsverantwoording.

Algemeen

Het Cito Volgsysteem primair en speciaal onderwijs beoogt de vorderingen van individuele leerlingen, groepen leerlingen en het onderwijs op school van groep 1 tot en met groep 8 te volgen en te evalueren. De toetsen Rekenen-Wiskunde voor groep 4 zijn een onderdeel van het Cito Volgsysteem primair onderwijs en zijn bedoeld voor leerlingen in groep 4 van het primair onderwijs. De toets voor groep 4 is onderdeel van het Cito LVS Rekenen-Wiskunde 3.0 wat fungeert als een systeem om vast te stellen hoe goed leerlingen kunnen rekenen en hoe hun rekenvaardigheid zich in de basisschoolperiode ontwikkelt. Onderstaande beschrijving is gebaseerd op de Handleiding.

Meetpretentie

De toetsen in de toets pakketten Rekenen-Wiskunde 3.0 voor groep 4 van het Cito Volgsysteem primair en speciaal onderwijs zijn bedoeld om vast te stellen hoe goed een leerling kale rekenopgaven en rekenproblemen in contexten kan oplossen en hoe de rekenvaardigheid van de leerling zich in de loop van de jaren ontwikkelt.

Doelgroep

De toetsen Rekenen-Wiskunde 3.0 groep 4 zijn bedoeld voor leerlingen in groep 4 van het primair en speciaal onderwijs, maar kunnen ook gebruikt worden voor leerlingen uit andere jaargroepen die werken op het niveau van groep 4 en voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften. Voor deze groepen speciale leerlingen zijn extra aanwijzingen opgenomen in de handleiding. Er zijn echter geen afzonderlijke normen vastgesteld en de toetsresultaten van deze leerlingen worden geïnterpreteerd met behulp van de gemiddelde vaardigheidsscores voor leerlingen uit het regulier onderwijs.

Gebruiksdoel en functie

Doel van de toetsen Rekenen-Wiskunde 3.0 voor groep 4 is het in kaart brengen van het vaardigheidsniveau en de ontwikkeling van de leerlingen op het gebied van rekenen-wiskunde. Hiervoor wordt de behaalde vaardigheidsscore normgericht geïnterpreteerd op basis van de vaardigheidsverdeling in een adequate, landelijke, referentiegroep. De vaardigheidsscore wordt uitgedrukt in de symmetrische niveau indeling I t/m V en in de asymmetrische niveau indeling A t/m E. De toetsen maken het mogelijk om:

- De rekenvaardigheid van zowel individuele leerlingen als groepen leerlingen (groeps- en schoolniveau) te beoordelen via een vergelijking van de behaalde scores met de scores van een landelijke referentiegroep oftewel niveaubepaling.
- De ontwikkeling van de rekenvaardigheid van zowel individuele leerlingen als groepen leerlingen (groeps- en schoolniveau) door de leerjaren heen te volgen oftewel progressiebepaling.

Inhoudelijke theoretische inkadering:
De inhoud van de toetsen sluit aan bij de kerndoelen Rekenen-Wiskunde van het primair onderwijs zoals die wettelijk zijn vastgesteld. De kerndoelen omvatten de onderwerpen 'wiskundig inzicht en handelen', 'getallen en bewerkingen' en 'meten en meetkunde'. Voor de uitwerking van de kerndoelen tot een domeinbeschrijving is gebruikgemaakt van de inhoud van de referentieniveaus, de tussendoelen van de SLO, de publicaties van het TAL-team en van de leerlijnen zoals die door veelgebruikte methodes zijn uitgewerkt.
De verschillende onderdelen van het domein rekenen-wiskunde vormen een samenhangend geheel van getalbegrip en rekenvaardigheid. Hierin staan inzicht in getallen, maatinzicht, ruimtelijk inzicht en het kunnen uitvoeren van operaties met getallen en het kunnen toepassen van die kennis en inzichten in uiteenlopende situaties centraal. Er worden in de domeinbeschrijving vier domeinen onderscheiden: ' Getallen', 'Verhoudingen', 'Meten en Meetkunde' en 'Verbanden. Deze komen overeen met de referentieniveaus. In de toetsen voor groep 4 zijn alleen opgaven opgenomen voor de domeinen die in die jaargroep op school aan de orde komen. Dit zijn de domeinen 'Getallen' en 'Meten en meetkunde'.
Bij het domein 'Getallen' staan het doorzien van de structuur van de telrij, de structuur van getallen en de relaties tussen getallen centraal.
Bij het domein 'Meten en meetkunde' gaat het om het elementaire begrip van wat meten is, het op beperkte schaal kunnen aflezen en berekenen van tijd en het uitvoeren van berekeningen met geld.

Inhoud van het toetspakket
Het toetspakket Rekenen-Wiskunde 3.0 groep 4 bestaat uit de volgende documenten:

- Handleiding, deze bevat informatie over:
  - de afname van de toets (hfdst. 2),
  - nakijken en verwerken van toetsgegevens (hdfst. 3),
  - interpretatie van de toetsresultaten op leerling- en groepsniveaus (hfdst 4),
  - interpretatie van toetsresultaten op schoolniveau (hfdst 5),
  - theoretisch kader en achtergronden van de toets (hfdst 6),
  - communiceren over toetsresultaten met leerling en ouders (hfdst 7),
  - achtergrondinformatie en veelgestelde vragen (hfdst 8) en
  - enkele bijlagen
- Vier toetsen (van alle toetsen is een papieren en een digitale variant beschikbaar):
  - Toets M4 (Medio groep 4)
  - Toets E4 (Eind groep 4)
  - Toets E3M4 (makkelijke variant van de toets M4)
  - Toets M4E4 (makkelijke variant van de toets E4)
- Afnamekaarten met aanwijzingen voor de papieren of de digitale afname van de toetsen
- Nakijkkaarten
- Antwoordbladen
- Tabellen voor de drie toetsen voor het bepalen van de vaardigheidsscore en –niveau.

2. Beoordeling van de kwaliteitsaspecten

De beoordeling vindt plaats volgens het 'Beoordelingskader voor de psychometrische aspecten van (reeksen van) toetsen uit leerlingvolgsystemen (LOVS)', 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 mevrouw Paulyn K. Berding-Oldersma MSc (secretaris).

De kwaliteit van de dataverzameling

S1.1. Is de steekproef representatief?

Bevindingen:

In januari 2012 zijn in een digitaal en papier-digitaal kalibratieonderzoek M4 en E4 (proefonderzoek) per afnamemoment 150 items voorgelegd aan 125 tot 150 leerlingen van groep 3, verdeeld over taken van 25 opgaven, afgenomen bij 9 tot 12 scholen. Bij leerlingen zijn digitale taken met nieuwe LVS-III opgaven afgenomen in combinatie met de twee papieren taken van LVS-II en in combinatie met de twee digitale taken van LVS-II.

Op grond van het kalibratieonderzoek M4 is voor het normeringsonderzoek M4 een selectie gemaakt van 193 items, verdeeld over 9 boekjes met elk 76, 77 of 78 opgaven verdeeld over 3 taken. Deze zijn opgenomen in een embedded field normeringsonderzoek waarin nieuw ontwikkelde items voor LVS-III meeliepen in de al bestaande en op scholen toegepaste LVS-II toetscyclus. Het embedded field normeringsonderzoek M4 is toegepast op de resultaten van 1969 leerlingen uit groep 4 van 187-103=84 scholen. Voor het bepalen van de normering zijn de gegevens aangevuld met gegevens van 2496 leerlingen uit groep 4 van 103 scholen uit Cito dataretour.

Op grond van het kalibratieonderzoek E4 is voor het normeringsonderzoek E4 een selectie gemaakt van 210 items, verdeeld over 9 boekjes, met elk 81 of 82 opgaven verdeeld over 3 taken. Deze zijn opgenomen in een embedded field normeringsonderzoek waarin nieuw ontwikkelde items voor LVS 3.0 meeliepen in de al bestaande en op scholen toegepaste LVS 2.0 toetscyclus. Het embedded field normeringsonderzoek E4 is toegepast op de resultaten van 1.848 leerlingen uit groep 4 van 177-97=80 scholen. Voor het bepalen van de normering zijn de gegevens aangevuld met gegevens van 2.354 leerlingen uit groep 4 van 97 scholen uit Cito dataretour.

De representativiteit van de steekproeven voor de normeringsonderzoeken M4 en E4 is onderzocht met betrekking tot regio, urbanisatiegraad, schooltype en sekse. Bij regio is uitgegaan van de vier landsdelen / regio's van de CBS-indeling. Bij urbanisatiegraad is uitgegaan van een tweedeling in stad en platteland, afgeleid van de CBS-indeling naar vijf niveaus van verstedelijking. Bij schooltype is uitgegaan van de formatiegewichten volgens OCW. Hierin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders. Bij sekse is een tweedeling gemaakt naar jongens en meisjes. De steekproefverdeling wijkt weinig af van de populatieverdeling. De effectgroottes phi liggen ver onder de 0.10 en zijn daarmee zeer klein. De effectgrootte phi is het grootst voor de variabele schooltype voor het afnamemoment E4 (.042).

Uit de ruwe scores van de individuele leerlingen uit het embedded field normeringsonderzoek en Cito dataretour werden plausible values gegenereerd op de nieuw ontwikkelde vaardigheidsschaal. De normering werd vervolgens gebaseerd op de plausible values van de leerlingen in de normeringssteekproef. De plausible values voor de afnamemomenten M4 en E4 bleken een normale verdeling te vormen. De schoolverdeling werd bepaald met het intercept-only multilevel model. Dit model werd geschat via een

bootstrap procedure. De intraklassecorrelatie (ICC) lag boven de 0.04, wat inhoudt dat een multilevelanalyse zinvol is. Ondanks dat de percentielen van de normgegevens op schoolniveau dichter bij elkaar kwamen te liggen dan in de leerlingverdeling, waren de afstanden groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.

Conclusie:

De steekproeven zijn representatief, zijn adequaat gestratificeerd naar sekse, regio, schooltype en urbanisatiegraad en geven informatie over hoe de steekproeven zich verhouden tot de populatiewaarden. De procedure voor het samenstellen van de steekproeven is onderbouwd en de omstandigheden waaronder data is verzameld, is redelijk vergelijkbaar met de omstandigheden waaronder de toets wordt afgenomen. Daarmee wordt aan aspect S1.1. het oordeel '**voldoende**' toegekend.

S1.2. In geval van een onvolledig dataverzamelingsdesign: is het design adequaat?

Bevindingen:

Om te komen tot een set van psychometrisch en inhoudelijk geschikte items zijn de opgaven uit het proefonderzoek van en de opgaven uit de daaropvolgende normeringsonderzoeken van januari 2012 (M4) en juni 2012 (E4) gekalibreerd. Hiervoor is gebruik gemaakt van het IRT model OPLM. Met dit statistische model zijn de psychometrische kenmerken (moeilijkheidsparameters en discriminatie indices) van de items geschat.

In het kalibratieproces is uitgegaan van een onvolledig maar 'verbonden' design. In het kalibratieproces M4 van januari 2012 zijn 193 items voorgelegd aan 1.969 leerlingen van groep 4. De 193 items waren verdeeld over 9 boekjes (booklets). Elke boekje bestond uit 77, 77 of 78 opgaven verdeeld over 3 taken. Bij leerlingen zijn digitale taken met nieuwe LVS-III opgaven afgenomen in combinatie met de twee papieren taken van LVS-II en in combinatie met de twee digitale taken van LVS-II. De opgaven werden gemiddeld door 125 tot 150 leerlingen gemaakt wat aan het minimum vereiste van 150 voldoet.

In het kalibratieproces E4 van juni 2012 zijn 210 items voorgelegd aan 1.848 leerlingen van groep 4. De 190 items waren verdeeld over 9 boekjes (booklets). Elke boekje bestond uit 81 of 82 opgaven verdeeld over 3 taken. Bij leerlingen zijn digitale taken met nieuwe LVS-III opgaven afgenomen in combinatie met de twee papieren taken van LVS-II en in combinatie met de twee digitale taken van LVS-II. De opgaven werden gemiddeld door 125 tot 150 leerlingen gemaakt wat aan het minimum vereiste van 150 voldoet.

Op basis van inhoudelijke en psychometrische criteria werden 56 items voor elk van de toetsen Rekenen-Wiskunde 3.0 groep 4 geselecteerd (M4, E4, M4E4, E3M4 papier en M4, E4, M4E4, E3M4 digitaal). De 56 items zijn verdeeld naar vaardigheid en inhoudsaspecten van de elementen getallen en getal-relaties, optellen en aftrekken (kaal en context), vermenigvuldigen en delen en naar meten, tijd en geld als de componenten van de latente vaardigheid Rekenen. Van alle opgaven die zijn meegegaan in het normeringsonderzoek zijn de klassieke p-waarde en de r_{it} waarde bepaald en eveneens de IRT-indices. Voor de normeringsonderzoeken M4 en E4 werden na het trekken van een representatieve steekproef, waarbij rekening werd gehouden met verdeling naar regio, urbanisatiegraad, schooltype en sekse, zowel scholen geworven als data gehaald uit Cito dataretour. Voor

het normeringsonderzoek M4 werd gebruik gemaakt van resultaten van 1.969 leerlingen uit groep 4 van 187-103=84 scholen en van dataretour van 2.496 leerlingen uit groep 4 van 103 scholen. Voor het normeringsonderzoek E4 werd gebruik gemaakt van resultaten van 1.848 leerlingen uit groep 4 van 177-97=80 scholen en van dataretour van 2.354 leerlingen uit groep 4 van 97 scholen.

Uit het kalibratieonderzoek (S-toetsing, R1c-waarden en de constante 'c') blijkt dat de items passen bij voornoemd IRT model en dat het model ook past voor de toets als geheel. Dit betekent dat er sprake is van één unidimensionele vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden.

Conclusie:

Het onvolledige maar 'verbonden' design van de proefonderzoeken is adequaat. Het volledige design van de toets M4 en E4 zijn eveneens adequaat. Aan aspect S1.2 wordt het oordeel '**voldoende**' toegekend.

Normering

N1.2.1. Zijn de normgroepen groot genoeg?

Bevindingen:

De toets is genormeerd voor de afnamemomenten M4 en E4.

Op grond van het kalibratieonderzoek M4 (januari 2012) is een selectie gemaakt van 193 items, verdeeld over 9 boekjes met elk 76, 77 of 78 opgaven verdeeld over 3 taken. Deze zijn opgenomen in een embedded field normeringsonderzoek waarin nieuw ontwikkelde items voor LVS 3.0 meeliepen in de al bestaande en op scholen toegepaste LVS 2.0 toetscyclus. Het embedded field normeringsonderzoek M4 is toegepast op de resultaten van 1.969 leerlingen uit groep 4 van 187-103=84 scholen. Voor het bepalen van de normering zijn de gegevens aangevuld met gegevens van 2.496 leerlingen uit groep 4 van 103 scholen uit Cito dataretour.

Op grond van het kalibratieonderzoek E4 (juni 2012) is een selectie gemaakt van 210 items, verdeeld over 9 boekjes, met elk 81 of 82 opgaven verdeeld over 3 taken. Deze zijn opgenomen in een embedded field normeringsonderzoek waarin nieuw ontwikkelde items voor LVS 3.0 meeliepen in de al bestaande en op scholen toegepaste LVS 2.0 toetscyclus. Het embedded field normeringsonderzoek E4 is toegepast op de resultaten van 1.848 leerlingen uit groep 4 van 177-97=80 scholen. Voor het bepalen van de normering zijn de gegevens aangevuld met gegevens van 2.354 leerlingen uit groep 4 van 97 scholen uit Cito dataretour.

Voor de afnamemomenten M4 en E4 werden vaardigheidsverdelingen gepresenteerd op leerlingniveau en op schoolniveau. Dit betreft de gemiddelde score, standaarddeviatie en de percentielen P10, P20, P25, P40, P50, P60, P75 en P80. Van hieruit kunnen de beide niveau indelingen (de symmetrische niveau indeling I t/m V en de asymmetrische niveau indeling A t/m E) worden bepaald.

De normen voor de toetsen Rekenen-Wiskunde 3.0 groep 4 zijn geldig tot en met 2022.

Conclusie:

Er is sprake van relatieve normen, de steekproeven zijn representatief en groot genoeg. Daarmee wordt aan aspect N1.2.1. het oordeel '**voldoende**' toegekend.

N1.2.2. Zijn de normgroepen representatief?

Bevindingen:

De representativiteit van de steekproeven is besproken bij punt S1.1. Hier werd reeds geconstateerd dat deze representatief zijn.

Conclusie:

Aan aspect N1.2.2. wordt het oordeel '**voldoende**' toegekend.

Betrouwbaarheid

B1.1. Zijn of worden de betrouwbaarheidsgegevens correct berekend?

Bevindingen:

Om relevante gegevens bij de toets te genereren, is gebruik gemaakt van het programma OPLAT. Binnen dit programma wordt de coëfficiënt MAcc ('Accuracy of Measurement') berekend. Deze coëfficiënt vertoont qua interpretatie grote overeenkomst met de betrouwbaarheidscoëfficiënt uit de KTT. Deze coëfficiënt wordt in de psychometrische literatuur beschreven en als correct aangemerkt.

Conclusie:

Aan aspect B1.1 wordt het oordeel '**voldoende**' toegekend.

B1.2. Zijn de betrouwbaarheidsgegevens voldoende gezien de beslissingen die met de toets genomen worden?

Bevindingen:

Er wordt verwezen naar de COTAN criteria voor toetsen voor minder belangrijke beslissingen. De interne consistentie betrouwbaarheid is, volgens deze criteria, voldoende bij een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80. Voor de toets Rekenen-Wiskunde 3.0 groep 4 wordt deze coëfficiënt berekend als MAcc (zie B1.1) voor de afnamemomenten M4, E4, M4E4, E3M4 van zowel de papieren als de digitale toetvariant. Aanvullend hierop wordt de standaardmeetfout vermeld. De afnamecontext van de toets leent zich, dankzij een OPLM kalibratie, voor een gesimuleerd test-hertest onderzoek onder ideale condities. De test-hertest coëfficiënt is identiek aan de MAcc voor alle afnamemomenten. De resultaten laten zich lezen als standaardmeetfout (papieren versie: M4 2,948; M4E4 3,086; E4 3,193; E3M4 2,736; digitale versie: M4 2,993; M4E4 3,153; E4 3,230; E3M4 2,791), MAcc (papieren versie: M4 0,93; M4E4 0,94; E4 0,94; E3M4 0,91; digitale versie: M4 0,93; M4E4 0,92; E4 0,93; E3M4 0,92) en test-hertest simulatie (papieren versie: M4

0,93; M4E4 0,94; E4 0,94; E3M4 0,91; digitale versie: M4 0,93; M4E4 0,92; E4 0,93; E3M4 092) en worden aangemerkt als 'voldoende'.

Aanvullend hierop is de lokale meetnauwkeurigheid weergegeven in betrouwbaarheidstabellen (misclassificaties). Uitgaande van de betrouwbaarheids tabellen worden twee indices voor de nauwkeurigheid van de classificaties gerapporteerd: de plus/minus 1 niveau-index en de marginal classification index. Uit de hoogte van de indices blijkt dat de laagst en de hoogst scorende leerlingen accuraat te classificeren zijn, maar dat over het algemeen tussen leerlingen in de niveaugroepen B, C en D, respectievelijk II, III en IV, er een licht minder duidelijk onderscheid te maken is. De marginal classification indices lopen uiteen van 66 tot 75 procent en de resultaten stemmen hiermee eveneens tot tevredenheid.

Conclusie:

De betrouwbaarheid van de toetsen Rekenen-Wiskunde 3.0 groep 4 is 'voldoende' als aangenomen mag worden dat de toets geen zware consequenties voor de leerlingen heeft en ingestemd wordt met de beoordelingscriteria voor de betrouwbaarheid van de COTAN.

Op aspect B.1.2. wordt aan de toets Rekenen-Wiskunde 3.0 groep 4 het oordeel '**voldoende**' toegekend.

Validiteit

V1. 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:

In groep 4 zijn er, voor zowel de papieren als de digitale variant, niet alleen 2 reguliere toetsen M4 en E4 die respectievelijk halverwege en aan het einde van het jaar worden afgenomen, maar ook extra toetsen M4E4 en E3M4 (met kleinere leerstappen dan die in de reguliere toetsen) bedoeld voor leerlingen met een vertraagde ontwikkeling die afgenomen worden op de reguliere afnamemomenten.

De toetsen voor groep 4 bestaan steeds uit twee taken die kunnen worden afgenomen op 2 verschillende dagdelen. Iedere taak bestaat uit 28 opgaven, zijnde een beperkt aantal meerkeuzeopgaven en vooral korte antwoordvragen. De toets kan zowel handmatig als via de computer via het computerprogramma LOVS nagekeken worden.

De toetsontwikkelaars onderscheiden, overeenkomstig de referentieniveaus, vier domeinen:

- Getallen
- Verhoudingen
- Meten en meetkunde
- Verbanden.

In de toetsen voor groep 4 komen alleen de domeinen Getallen en Meten en meetkunde voor. Deze onderwerpen komen aan bod in de meest gebruikte methodes voor rekenen-wiskunde van groep 4. Het is een voor de hand liggende keuze om de domeinen

verhoudingen en verbanden in groep 4 onderbelicht te laten, het is echter niet zo dat inhouden uit deze twee domeinen volledig afwezig zijn.

De toetsontwikkelaars onderscheiden binnen het domein Getallen de onderdelen Getallen en getalrelaties en het onderdeel Bewerkingen. Binnen het domein Meten worden allerlei subonderdelen onderscheiden (het onderdeel meten, waaronder ook meetkunde valt, het onderdeel tijd en het onderdeel geld).

De vaardigheden die getoetst worden zijn gezamenlijk een goede afspiegeling van de rekenvaardigheden waarmee leerlingen in aanraking zijn geweest tijdens het rekenonderwijs dat zij hebben gekregen.

De moeilijkheidsgraad van de toetsen is, gezien het niveau van de doelgroep, aanvaardbaar. De toetsen bevatten open opgaven en meerkeuzeopgaven en korte antwoordopgaven. De opgaven bevatten illustraties waarbij de leerkracht de opgave voorleest. De opgaven zijn duidelijk en ook qua taal voldoende toegankelijk, helder en eenduidig. Het antwoordmodel laat geen ruimte voor interpretatie.

Voor leerlingen met een beperkte aandacht spanne kunnen de toetsen M4 en E4 verdeeld worden in drie delen. Op de afnamekaarten worden de knipmomenten aangegeven. De toetsen M4E3 en M4E4 worden standaard in drie delen afgenomen.

Merk op dat dat de beoordeling van de validiteit zich hieronder beperkt tot het statistisch/psychometrisch onderzoek dat is verricht.

De toets Rekenen-Wiskunde 3.0 groep 4 is niet bedoeld voor voorspellend gebruik. Daarmee is de criteriumvaliditeit niet van toepassing. De (psychometrische) begripsvaliditeit wordt uitgewerkt in unidimensionaliteit, itemkwaliteit, itembias, convergente en divergente validiteit en in verschillen tussen relevante subgroepen.

De resultaten van de uitgevoerde kalibratie maken het aannemelijk dat er sprake is van unidimensionaliteit. Dit betekent dat met elke willekeurige subset van items uit de gekalibreerde itembank dezelfde onderliggende rekenvaardigheid kan worden vastgesteld. Dit wordt tevens bevestigd door de nauwkeurigheid van de itemparameterschattingen (op zes items uit M4 digitaal en vijf items uit E4 digitaal na, is de constante c voor alle items hoger dan 0.20 maar lager dan 0.40).

De gemiddelde moeilijkheidsgraden van de items (als criterium voor de itemkwaliteit) voldoen (papieren versie: p-waarden E3M4 met range van .60 - .92 en een gemiddelde van .79; M4 range .50 - .90 en gemiddelde .70; M4E4 range .57 - .94 en gemiddelde .76; E4 range .45 - .88 en gemiddelde .68; digitale versie: p-waarden E3M4 met range van .55 - .92 en gemiddelde .77; M4 range .40 - .88 en gemiddelde .68; M4E4 range .53 - .93 en gemiddelde .74; E4 range .44 - .90 en gemiddelde .67).

De gemiddelde R_{it} waarde (ook een criterium voor de itemkwaliteit) voldoet eveneens. Voor alle vijf de toetsen is de gemiddelde R_{it} waarde .40 of hoger. Alleen bij de digitale versie van de toets E3M4 is de ondergrens 0.19 (R_{it} moet idealiter boven de 0.20 liggen), maar de gemiddelde R_{it} waarde bij deze toets is met 0.40 uitstekend. De gemiddelden komen uit in de range van .39 tot .43.

In het onderzoek naar itembias is geen sprake van DIF (Differential Item Functioning) naar sekse.

De constructvaliditeit is uitgewerkt in convergente validiteit door de samenhang te onderzoeken met de voorgaande versie Rekenen-Wiskunde 2.0. De correlaties waren voor zowel M4 (r=0,96) en E4 (r=0,97) hoog.

De divergente validiteit is onderzocht door de samenhang met de Schoolvaardigheidstoets rekenen-wiskunde en met de Tempotoets rekenen (beide van Boom testuitgevers) te onderzoeken. Daarnaast is de samenhang met diverse LVS leestoetsen onderzocht. De hoge correlatie met de schoolvaardigheidstoets (M4 r=0,70; E4: r=0,78) en de lagere correlatie met de diverse LVS leestoetsen (r in de range van 0,24 – 0,53) sterken de eerdere COTAN waardering 'goed' of 'voldoende' voor de begripsvaliditeit.

De hoge correlaties tussen Rekenen-Wiskunde 3.0 en vergelijkbare rekentoetsen en de lagere correlatie tussen Rekenen-Wiskunde 3.0 en diverse taaltoetsen vormen een ondersteuning voor de validiteit van de toets.

Wat betreft de verschillen tussen relevante subgroepen scoren jongere leerlingen (de versnelde leerlingen) naar verwachting iets beter dan oudere leerlingen en scoren jongens iets hoger dan meisjes. In termen van effectgrootte is er sprake van een klein effect (M4 0,41 en E4 0,40).

Conclusie:

Op aspect V1.1 wordt aan de toetsen Rekenen-Wiskunde 3.0 groep 4 op dit aspect het oordeel '**voldoende**' toegekend.

Het volg-aspect

VA1.1. Is er een voldoende empirische onderbouwing van de schaal waarop de groei van een leerling wordt uitgedrukt? Wordt groei op een adequate manier gemeten?

Bevindingen:

Het algemene (inhoudelijke) uitgangspunt voor de toets Rekenen-Wiskunde 3.0 groep 4 is dat de (latente) vaardigheid Rekenen kan worden opgevat als een unidimensioneel continuüm en dat elke leerling kan worden voorgesteld als een punt op dit continuüm. Hierbij wordt opgemerkt dat de elementen (1) getallen en (2) meten en meetkunde niet opgevat kunnen worden als te isoleren vaardigheden. Daarmee wordt de rekenvaardigheid, als interactie tussen deze componenten, beschouwd als één unidimensionele vaardigheid.

Uit het kalibratieonderzoek blijkt dat de items passen bij het gehanteerde IRT model en dat het model ook past voor de toetsen M4 en E4 als geheel. Dit betekent dat er sprake is van één unidimensionele vaardigheidsschaal waar items en leerlingen op afgebeeld kunnen worden.

Afhankelijk van het aantal items dat een leerling goed maakt, wordt er een vaardigheidsscore toegekend. Jongere leerlingen scoren iets beter dan oudere leerlingen. Tevens scoren jongens iets hoger dan meisjes.

Voor Rekenen-Wiskunde 3.0 is een nieuwe vaardigheidsschaal ontwikkeld, waarop alle uitgebrachte en nog uit te brengen toetsen uit het Cito Volgsysteem primair en speciaal

basisonderwijs Rekenen-Wiskunde 3.0 worden gekalibreerd. Vanwege het volgmodel en de gevolgde dataverzamelingsstrategie voor de normering, worden de nieuw ontwikkelde toetsen gefaseerd uitgebracht, d.w.z. in elk schooljaar een toetspakket voor een hogere groep.

Conclusie:

Aan aspect VA1.1. wordt het oordeel '**voldoende**' toegekend.

VA1.2. Worden er gegevens verstrekt over hoe groei geïnterpreteerd dient te worden? Wordt de betrouwbaarheid van de groei op die schaal adequaat weergegeven?

Bevindingen:

In hoofdstuk 7 van de handleiding ('Communiceren over toetsresultaten met leerling en ouders') wordt beschreven hoe er met de verschillende gebruikers over de toetsresultaten kan worden gecommuniceerd. Hierin wordt onderscheid gemaakt tussen 'niveau' en 'groei', wat wordt onderbouwd met diverse rapportage mogelijkheden.

In de wetenschappelijke verantwoording wordt toegelicht hoe de toetsen ingezet kunnen worden om de ontwikkeling van leerlingen te volgen in de tijd, namelijk door het toetsresultaat van een leerling te vergelijken met andere leerlingen en door het toetsresultaat van een leerling te vergelijken met diens andere toetsresultaten. Voor alle vergelijkingen geldt dat uitspraken over de voortgang van leerlingen gerelativeerd moeten worden vanwege de (on)betrouwbaarheid van de toetsen. Door betrokkenen bij de toetsen Rekenen-Wiskunde moet beseft worden dat vaardigheidsgroei zich langzaam in de tijd voltrekt.

Conclusie:

Aan aspect VA1.2. wordt 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:

Via de portal van Cito B.V. kan gebruik worden gemaakt van rapportage-/registratieformulieren voor een leerlingrapport, groepsrapport, groepsoverzicht (overzicht van één groep leerlingen tijdens hun schoolperiode) en een alternatief leerlingrapport (voor leerlingen die op een eigen niveau werken). Voor ouders is met name het leerlingrapport of alternatief leerlingrapport informatief omdat deze rapporten van hun kind individueel de vaardigheid en de groei weergeven.

In de Handleiding wordt in hoofdstuk 7 aandacht besteed aan de wijze waarop met ouders over de toetsresultaten gecommuniceerd kan/moet worden. Met name wordt daarbij gewezen op het leerlingrapport, waarin zowel het niveau van de leerling als de progressie van de leerling numeriek en grafisch gepresenteerd worden.

Daarnaast wordt de leraar gewezen op misverstanden die zich bij de interpretatie van de niveau-indelingen bij de ouders kunnen voordoen. Ook moeten zij aan ouders het verschil tussen methode-onafhankelijke en methodegebonden toetsen duidelijk maken en erop wijzen dat deze toetsen leerlingen anders (kunnen) beoordelen. De informatie biedt goede handvatten voor de gesprekken met ouders. In hoofdstuk 8 worden veelgestelde vragen behandeld die weliswaar voor de leraren bestemd zijn maar waar de antwoorden voor een deel ook informatief zijn tijdens bijvoorbeeld de tienminutengesprekken. Over de interpretatie van toetsresultaten is ook een folder ouderinformatie beschikbaar die men via de website van het Cito kan downloaden.

Conclusie:
Op aspect I1.1 wordt aan de toetsen Rekenen-Wiskunde 3.0 groep 4 het oordeel '**voldoende**' toegekend.

3. Verzamelstaat

Kwaliteitsaspect	**Code**	**Oordeel**
De kwaliteit van de steekproef	*S1.1*	**Voldoende**
	S1.2	**Voldoende**
Normering	N1.1	**Voldoende**
	N1.2	**Voldoende**
Betrouwbaarheid	B1.1	**Voldoende**
	B1.2	**Voldoende**
Validiteit	V1.1	**Voldoende**
Volg-aspect	VA1.1	**Voldoende**
	VA1.2	**Voldoende**
Inzicht in leervorderingen	I1.1	**Voldoende**

4. Literatuurlijst

De beoordeling is gebaseerd op het volgende, door Cito B.V. aangeleverde, materiaal:

- Janssen, J., Hop, M., Wouda, J.. (2015). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 4*. Arnhem: Cito B.V.
- Cito B.V. (2014). *Leerkrachtmap Rekenen-Wiskunde 3.0 voor groep 4.* Arnhem: Cito B.V.