Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 7

Michel Hop, Jan Janssen en Ronald Engelen

Wetenschappelijke verantwoording

Rekenen-Wiskunde 3.0 voor groep 7

Michel Hop
Jan Janssen
Ronald Engelen

Cito
Arnhem, 2017

Inhoud

Bijlagen 71

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de toetsen Rekenen-Wiskunde 3.0 voor groep 7 uit het Cito Volgsysteem primair en speciaal onderwijs. De toetsen van het volgsysteem meten en volgen de algemene rekenvaardigheid van leerlingen in het primair en speciaal onderwijs van groep 3 tot en met 8.
Deze verantwoording doet verslag van de constructie van de toets. Daarbij wordt aandacht besteed aan het theoretisch kader van waaruit de toets is opgezet, het rekendomein, het constructieproces, het afnamedesign en de afname van de proefversie. Ook de wijze waarop de normering van de toets plaatsvond wordt beschreven.
Daarnaast worden alle analyses en resultaten besproken waarmee de gebruiker een uitspraak kan doen over de belangrijkste kenmerken, zoals de normering, de betrouwbaarheid en validiteit van dit instrument.
Ook is er een paragraaf over scoring en interpretatie van de test opgenomen. Meer informatie over scoring en interpretatie staat in de handleiding.
Deze verantwoording is vooral bedoeld voor gebruikers en andere professionals die zich een beeld willen vormen van de kwaliteit van de test. Tezamen met het testmateriaal levert deze wetenschappelijke verantwoording alle informatie die nodig is voor een snelle en efficiënte beoordeling van de kwaliteit op de volgende aspecten:

- Uitgangspunten voor de toetsconstructie
- De kwaliteit van het toetsmateriaal
- De kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Begripsvaliditeit

Criteriumvaliditeit is voor de toetsen Rekenen-Wiskunde 3.0 groep 7 van het Cito Volgsysteem niet van toepassing aangezien de toetsen geen voorspellende meetpretentie hebben.
Daarnaast is de verantwoording bestemd voor docenten die zich nader willen verdiepen in de achtergronden van de toetsen. Ook personen en instanties die in tweede instantie toetsscores van leerlingen in handen krijgen, vinden in deze verantwoording voldoende aanknopingspunten voor de interpretatie van deze scores.
De toetsen Rekenen-Wiskunde 3.0 bestaan – naast deze wetenschappelijke verantwoording – uit de volgende onderdelen:

- Handleiding, met daarin opgenomen een inhoudsverantwoording
- Opgavenboekjes M7 en E7 voor de afnames op papier
- Afnamekaarten en nakijkkaarten voor de afnames op papier
- De digitale toetsen M7en E7 voor afnames achter de computer

De reguliere toetsmomenten zijn medio groep 7 (M7) en eind groep 7 (E7). Voor groep 7 zijn geen tussentoetsen ontwikkeld. Naast een rapportage van het I tot en met V niveau of het A tot en met E niveau, worden ook functioneringsniveaus gerapporteerd die aangeven met welk gemiddeld niveau de vaardigheid van een leerling het best te vergelijken is.
De digitale en papieren varianten zijn op één schaal gebracht en zijn volledig vergelijkbaar. Binnen het Cito Volgsysteem kan het toetsen op maat gerealiseerd worden.

De nieuwe toetsen rekenen-wiskunde voor groep 7 worden de toetsen Rekenen-Wiskunde 3.0 voor groep 7 uit het Cito Volgsysteem primair en speciaal onderwijs genoemd. De toetsen van de vorige uitgave worden de tweede generatie toetsen of LOVS II genoemd. Om de leesbaarheid van dit document te bevorderen worden de toetsen Rekenen-Wiskunde 3.0 uit het Cito Volgsysteem ook met andere benamingen aangeduid. Deze kunnen zijn: derde generatie toetsen, LVS III – toetsen of Rekenen-Wiskunde 3.0.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

De toetsen Rekenen-Wiskunde 3.0 voor groep 7 meten en volgen de algemene rekenvaardigheid. Het is niet de pretentie van de toetsen om schoolsucces te voorspellen of uitspraken te doen die betrekking hebben op de rekenvaardigheid op een toekomstig afnamemoment. Rekenvaardigheid vormt de basis voor de ontwikkeling van schoolse vaardigheden en is onmisbaar in het dagelijkse leven. Bijna elke school in het primair onderwijs hanteert minimaal vanaf groep 3 een methode om rekenvaardigheid te ontwikkelen.
Het onderwijs in rekenen-wiskunde in het basisonderwijs richt zich in de eerste plaats op het verwerven van fundamentele vaardigheden op de terreinen van het rekenen en het meten. Deze fundamentele vaardigheden hebben betrekking op:

– het gebruiken van reken-wiskundetaal;
– het uitvoeren van rekenoperaties;
– het gebruiken van strategieën om rekenproblemen en meetproblemen op te lossen.

De fundamentele vaardigheden moeten door de leerlingen ook gebruikt kunnen worden in praktische toepassingssituaties. Dit betekent dat er verbindingen gelegd worden tussen het onderwijs in rekenen-wiskunde en de alledaagse leefwereld. Verder moeten leerlingen eenvoudige verbanden, regels, patronen en structuren kunnen opsporen. En ten slotte moeten leerlingen redeneerstrategieën en onderzoeksstrategieën kunnen gebruiken.
De toetsen vormen een hulpmiddel om vast te stellen in hoeverre de leerlingen rekenvaardig zijn en hoe deze rekenvaardigheid zich ontwikkelt, door leerlingen te volgen. Alle bovengenoemde aspecten van rekenvaardigheid zijn in de toetsen opgenomen en de toetsen meten in hoeverre leerlingen kale rekenopgaven én rekenproblemen in contexten kunnen oplossen.

2.2 Doelgroep

De toetsen zijn een onderdeel van het Cito Volgsysteem primair en speciaal onderwijs. De toetsen Rekenen-Wiskunde 3.0 groep 7 zijn bedoeld voor leerlingen in groep 7 van het primair onderwijs en leerlingen in het speciaal (basis)onderwijs die functioneren op het niveau van groep 7 in het reguliere basisonderwijs. De toetsen zijn ook te gebruiken voor leerlingen in andere leerjaren die een rekenvaardigheid hebben op het niveau van groep 7. In de handleiding is toegelicht hoe dit toetsen op maat werkt, voor wat betreft het selecteren van toetsen en interpreteren van de resultaten.
Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen.

2.3 Gebruiksdoel en functie

De hoofddoelen van de toetsen Rekenen-Wiskunde 3.0 zijn het in kaart brengen van het rekenvaardigheidsniveau en de ontwikkeling van de rekenvaardigheid van (groepen) leerlingen volgen. Daarnaast brengen de toetsen met behulp van categorieënanalyses in kaart op welke domeinen leerlingen ten opzichte van hun algemene rekenvaardigheid het relatief beter of zwakker doen. De toetsen geven leerkrachten de mogelijkheid om:

- de vaardigheid van individuele leerlingen op het gebied van rekenen-wiskunde te vergelijken met die van andere leerlingen. Dit is mogelijk door vergelijking van behaalde scores met de scores van de landelijke normgroep. Bij de toetsen worden twee systemen gebruikt voor de indeling van de landelijke normgroep, namelijk de indeling in de groepen I tot en met V (met vijf groepen van 20%) en de indeling in de groepen A tot en met E (respectievelijk 25% hoogst scorende, 25%, 25%, 15% en de 10% laagst scorende leerlingen). In paragraaf 3.1 wordt dit verder toegelicht bij ‘verwerking resultaten en interpretatie’;
- de groep als geheel te vergelijken met andere groepen leerlingen. Ook hierbij wordt gebruikgemaakt van een vergelijking met een landelijke normering in I tot en met V en A tot en met E;

- de ontwikkeling in rekenvaardigheid te volgen. Door gebruikmaking van de meettechniek die in paragraaf 2.4.2.2 wordt toegelicht kunnen de scores van leerlingen op verschillende toetsen van het leergebied rekenen-wiskunde onderling met elkaar vergeleken worden. Dit geldt voor de papieren en de digitale toetsen, voor de verschillende afnamemomenten en de verschillende leerjaren;
- resultaten op groeps- en schoolniveau te volgen en te evalueren. Zo kunnen sterke en verbeterpunten vastgesteld worden. Leerkrachten kunnen nagaan of gestelde doelen behaald zijn en waar eventuele hulpprogramma's ingezet moeten worden om verbeteringen tot stand te brengen;
- in kaart te brengen of er rekendomeinen zijn waarbij individuele leerlingen of waarbij de groep als geheel lager of hoger scoort dan verwacht. Hiervoor maakt een leerkracht een categorieënanalyse rekendomeinen. Deze analyse kan alleen uitgevoerd worden met het Computerprogramma LOVS. Bij de analyses krijgen de leerkrachten geen signaal als de leerling volgens verwachting scoort, maar als er duidelijk afwijkend gescoord wordt geeft het programma het signaal 'opvallend' of 'zeer opvallend' aan.

Het volgen van leerlingen in hun groei, ook wel aangeduid als progressiebepaling, is een van de belangrijkste functies van het Cito Volgsysteem primair en speciaal onderwijs (LVS). De toetsen van het LVS geven de leerkracht (en ouders en leerlingen zelf) informatie over de ontwikkeling van de vaardigheden van de leerlingen, individueel en als groep, gedurende (vrijwel) de gehele basisschoolperiode. De toetsen geven antwoord op vragen als: is er sprake van vooruitgang, achteruitgang of van stabilisering? Is de vooruitgang – gelet op de gemiddelde vooruitgang in de populatie – volgens verwachting?
Om leerlingen te kunnen volgen wordt de betreffende vaardigheid, in dit geval rekenvaardigheid, opgevat als een unidimensionale vaardigheid, of 'latente trek'. Het gehanteerde meetmodel (zie paragraaf 2.4.2) maakt het mogelijk om de scores van een leerling op verschillende toetsen, op verschillende momenten afgenomen, onderling te vergelijken. De ruwe scores op de toetsen (de ruwe score is het aantal opgaven goed) zijn daartoe te transformeren in scores op één vaardigheidsschaal. Deze unidimensionele vaardigheidsschaal die aan de toetsen Rekenen-Wiskunde ten grondslag ligt, is ontwikkeld met behulp van het *One Parameter Logistic Model* (Verhelst, 1993; Verhelst & Glas, 1995; Verhelst, Glas & Verstralen, 1995).
Het aantal afnamemomenten per jaar (en het aantal daartoe te construeren verschillende toetsen) wordt bepaald door het tempo waarin een vaardigheid gemiddeld gesproken binnen een leerjaar en over de gehele schoolperiode toeneemt. Meestal is er sprake van twee afnamemomenten per leerjaar ('medio' en 'einde' leerjaar, aangeduid als M en E) en twee – bij het betreffende afnamemoment passende – toetsen. Elke toets wordt geconstrueerd op basis van een gekalibreerde itembank, waarbij een toets zo wordt samengesteld dat deze naar inhoud en moeilijkheidsgraad optimaal past bij het afnamemoment waarvoor deze bedoeld is.

Hoe kunnen we de LVS-toetsen Rekenen-Wiskunde 3.0 inzetten om leerlingen te volgen in de tijd?
Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:
a. We kunnen het toetsresultaat van een leerling vergelijken met die van andere leerlingen op hetzelfde meettijdstip (afnamemoment).
b. We kunnen de toetsresultaten van dezelfde leerling vergelijken met diens eigen toetsresultaten op eerdere of latere meettijdstippen (afnamemomenten).

Bij beide vergelijkingen maken we gebruik van het feit dat de toetsresultaten door toepassing van het IRT-model (OPLM) in de vorm van vaardigheidsscores afgebeeld kunnen worden op dezelfde vaardigheidsschaal. Dat geldt voor zowel individuele leerlingen als voor (gemiddelde) groepsresultaten. Dit alles wordt in meer detail uitgelegd in de leerkrachthandleiding (zie het hoofdstuk 'Interpreteren en analyseren op leerling- en groepsniveau').

Ad a.
Bij de eerstgenoemde vergelijking worden de prestaties van een leerling vergeleken met de prestaties van de hele populatie op een gegeven afnamemoment. Hoe doet een leerling het, bijvoorbeeld, ten opzichte van de gemiddelde leerling? Voor dit doel is de populatie, op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording), ingedeeld in vaardigheidsniveaus (I-V, A-E). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheidsniveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Meriam heeft op afnamemoment medio leerjaar 7 vaardigheidsniveau V behaald". Voor de leerkracht (en voor Meriam en haar ouders) bevat deze uitspraak waardevolle informatie. De leerkracht kan op basis hiervan bijvoorbeeld besluiten om Meriam extra lesstof aan te bieden.

Ad b.
Voor het vergelijken ('volgen') van een leerling op twee verschillende tijdstippen komen twee methodes in aanmerking. Bij de eerste methode worden de **vaardigheidsniveaus** op de twee tijdstippen vergeleken: "op tijdstip M7 had Meriam vaardigheidsniveau V en op tijdstip E7 was het vaardigheidsniveau IV".
Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 214, bijvoorbeeld, op tijdstip M7 en vaardigheidsscore 234 op tijdstip E7. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.
Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Meriam vaardigheids-niveau IV heeft behaald op het eerste tijdstip en niet vaardigheidsniveau V. Bij het vaststellen of het verschil in de prestaties op de twee tijdstippen statistisch significant is (Meriam is vooruitgegaan) of statistisch niet significant is (Meriam is voor- noch achteruitgegaan) speelt het betrouwbaarheidsinterval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5. Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) 'gegroeid' is.
De eenvoudigste manier is om te kijken of de BI's voor de twee tijdstippen overlappen. Als deze twee BI's niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid.
Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven. We geven hier een voorbeeld. Bij de afname Rekenen-Wiskunde M7 behaalde Wout een vaardigheidsscore van 249 met een 67% betrouwbaarheids- interval van 244-254. Bij de afname E7 behaalde Wout een vaardigheidsscore van 254; het bijbehorende betrouwbaarheidsinterval daarbij is 249-258. Aangezien de betrouwbaarheidsintervallen overlappen kunnen we niet zeggen dat Wouts vaardigheid is toegenomen of afgenomen.

Conclusie
De vaardigheidsgroei voor rekenen-wiskunde voltrekt zich geleidelijk in de tijd. De verschillen tussen vaardigheids-scores op achtereenvolgende meettijdstippen zijn betrekkelijk klein, al is de gemiddelde vaardigheidstoename in groep 6 en volgende wat minder groot dan in groep 3 en groep 4. Bovendien is er sprake van meetfouten.
De verschillen in vaardigheidsgroei moeten tegen de achtergrond van die meetfouten worden geïnterpreteerd. Dit betekent dat men weliswaar uitspraken kan doen over de vaardigheidsgroei van een leerling, maar dat deze uitspraken met voorzichtigheid dienen te worden gehanteerd. Dit geldt ook wanneer men de progressie van een leerling volgt in termen van vaardigheidsniveaus of een vergelijking maakt met andere leerlingen in termen van vaardigheidsniveaus. Want ook bij de indeling in vaardigheidsniveaus speelt de nauwkeurigheid van de toets een rol. Hoe de leerkracht hier in de praktijk mee om moet gaan wordt toegelicht in de handleiding voor de leerkracht.

2.4 Theoretische inkadering

2.4.1 Inhoudelijk

De inhoud van de toetsen sluit aan bij de kerndoelen primair onderwijs zoals die wettelijk zijn vastgesteld (Ministerie van Onderwijs, Cultuur en Wetenschap, 2006). De kerndoelen omvatten de onderwerpen 'wiskundig inzicht en handelen', 'getallen en bewerkingen' en 'meten en meetkunde'.

Domeinbeschrijving groep 3 tot en met 8
Voor de uitwerking van de kerndoelen tot een domeinbeschrijving is gebruikgemaakt van de inhoud van het referentiekader taal en rekenen (Expertgroep doorlopende leerlijnen taal en rekenen, 2008; Expertgroep doorlopende leerlijnen taal en rekenen, 2008a; Expertgroep doorlopende leerlijnen taal en rekenen, 2009; Noteboom, 2011), de tussendoelen van de SLO (Buijs, 2008; Buijs, 2008a), de publicaties van het TAL-team (TAL-team 1999; TALteam 2001; TAL-team 2004; TAL-team 2005; TAL-team 2007), de publicaties van de

Periodieke Peilingen van het Onderwijs Niveau (Janssen, 2005; Kraemer, 2005; Kraemer, 2009; Kraemer 2009a; Hop, 2012; Scheltens, 2013) en de Toetswijzers bij de centrale eindtoets PO (College voor Examens, 2012; College voor Toetsen en Examens, 2014) en van de leerlijnen zoals die door veelgebruikte methodes zijn uitgewerkt. Deze informatie is aangevuld met aanwezige expertise op het gebied van rekenen-wiskunde, zowel binnen als buiten Cito. De publicaties en input van vakinhoudelijke deskundigen en vertegenwoordigers vanuit de scholen zijn gebruikt om tot een domeinbeschrijving te komen.
In 2010 is de wet Referentieniveaus Nederlandse taal en rekenen van kracht geworden, met als doel een betere aansluiting te bewerkstelligen tussen het taal- en rekenonderwijs in de verschillende onderwijssectoren en de taal- en rekenvaardigheden van leerlingen te verbeteren. In de wet is vastgelegd wat leerlingen moeten leren als het gaat om Nederlandse taal en rekenen. Dit is geconcretiseerd aan de hand van zogeheten doorlopende leerlijnen en referentieniveaus. De toetsen Rekenen-wiskunde 3.0 groep 7 sluiten aan bij deze nieuwe wetgeving. Voor het rekenonderwijs in het primair en speciaal onderwijs zijn vooral referentieniveaus 1F en 1S van belang. Referentieniveau 1F is een minimum- of drempelniveau dat door ongeveer 87% van de leerlingen aan het eind van het primair en speciaal onderwijs wordt behaald (Hemker, 2017). Het niveau 1S geldt als streefniveau voor het primair en speciaal onderwijs. Dit wordt door ongeveer 44% van de leerlingen behaald. Voor een concretisering van de referentieniveaus 1F/1S verwijzen we naar Noteboom, van Os en Spek (2011).
Het referentiekader rekenen, de genoemde kern- en tussendoelen en leerlijnen voor rekenen en de geformuleerde referentieniveaus zijn als uitgangspunten gebruikt bij de opzet en ontwikkeling van de toetsen Rekenen-Wiskunde voor groep 7 (zie verder paragraaf 3.2).

Kerndoelen Rekenen-Wiskunde (Ministerie van Onderwijs, Cultuur en Wetenschap, 2006)
Wiskundig inzicht en handelen
23. De leerlingen leren wiskundetaal gebruiken.
24. De leerlingen leren praktische en formele rekenwiskundige problemen op te lossen en redeneringen helder weer te geven.
25. De leerlingen leren aanpakken bij het oplossen van rekenwiskundeproblemen te onderbouwen en leren oplossingen te beoordelen.

Getallen en bewerkingen
26. De leerlingen leren structuur en samenhang van aantallen, gehele getallen, kommagetallen, breuken, procenten en verhoudingen op hoofdlijnen te doorzien en er in praktische situaties mee te rekenen.
27. De leerlingen leren de basisbewerkingen met gehele getallen in elk geval tot 100 snel uit het hoofd uitvoeren, waarbij optellen en aftrekken tot 20 en de tafels van buiten gekend zijn.
28. De leerlingen leren schattend tellen en rekenen.
29. De leerlingen leren handig optellen, aftrekken, vermenigvuldigen en delen.
30. De leerlingen leren schriftelijk optellen, aftrekken, vermenigvuldigen en delen volgens meer of minder verkorte standaardprocedures.
31. De leerlingen leren de rekenmachine met inzicht te gebruiken.

Meten en meetkunde
32. De leerlingen leren eenvoudige meetkundige problemen op te lossen.
33. De leerlingen leren meten en leren te rekenen met eenheden en maten zoals bij tijd, geld, lengte, omtrek, oppervlakte, inhoud, gewicht, snelheid en temperatuur.

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. We onderscheiden in de domeinbeschrijving voor het basisonderwijs, overeenkomstig de referentieniveaus, de volgende vier domeinen:
- Getallen;
- Verhoudingen;
- Meten en meetkunde;
- Verbanden.

We bespreken hieronder de onderwerpen die in deze domeinen in groep 3 tot en met 8 aan de orde komen.

Uitwerking domeinbeschrijving

Getallen

1. Getalbegrip

Bij dit onderwerp staat het doorzien van de structuur van de telrij, de structuur van getallen en de relaties tussen getallen centraal. Ook de uitspraak en notatie en het gebruik van wiskundetaal valt onder dit onderwerp.

2. Bewerkingen

Bij bewerkingen wordt er onderscheid gemaakt tussen doelmatig rekenen, rekenen met gebruikmaking van standaardprocedures en globaal/benaderend rekenen. De verschillende bewerkingen optellen, aftrekken, vermenigvuldigen en delen komen aan bod, evenals combinaties hiervan. Bewerkingen met breuken komen in de hogere groepen aan bod en vallen ook onder dit onderwerp.

Verhoudingen

3. Verhoudingen

Bij het onderwerp verhoudingen gaat het om het herkennen en benoemen van verhoudingen en het oplossen van verhoudingsproblemen. Ook breuken, procenten en kommagetallen als manieren om verhoudingen aan te geven komen in de hogere groepen bij dit onderwerp aan bod. In de opgaven wordt leerlingen gevraagd deze getallen in elkaar om te zetten, ze te vergelijken en om ermee te rekenen.

Meten

4.1. Meten: lengte

Bij dit onderwerp gaat het om basiskennis en begrip van lengtematen, aflezen van meetinstrumenten, onderling herleiden van maten, kennis van maten en het toepassen van deze aspecten.

4.2. Meten: oppervlakte

Bij dit onderwerp gaat het om basiskennis en begrip van oppervlaktematen, afpassen met natuurlijke maten, onderling herleiden van enkele veel voorkomende oppervlaktematen, kennis van maten en het toepassen van deze aspecten.

4.3. Meten: inhoud

Bij dit onderwerp gaat het om basiskennis en begrip van inhoudsmaten, afpassen met natuurlijke maten, onderling herleiden van enkele veelvoorkomende inhoudsmaten, kennis van maten en het toepassen van deze aspecten.

4.4. Meten: gewicht

Bij dit onderwerp gaat het om basiskennis en begrip van gewichtsmaten, aflezen van meetinstrumenten, onderling herleiden van maten, kennis van maten en het toepassen van deze aspecten.

4.5. Tijd en snelheid

Bij dit onderwerp gaat het om aflezen van tijden, het rekenen met tijdsaanduidingen, het gebruik van de kalender en datumaanduidingen en om het rekenen met samenstellingen zoals km/uur.

4.6. Geld

Bij dit onderwerp gaat het om rekenen met geld, waarbij specifieke handelingen met munten en bankbiljetten uitgevoerd moeten worden. Denk aan het samenstellen van bedragen, het bepalen van de totale waarde, omwisselen van munten/biljetten, bepalen hoeveel je terugkrijgt en toepassingen zoals prijs per uur bepalen en werken met wisselkoersen.

5. Meetkunde

Hierbij gaat het om eenvoudige kennis en begrippen waarmee de ruimte meetkundig geordend, beschreven en verklaard kan worden. Centraal bij dit onderwerp staat de vaardigheid ‘ruimtelijk redeneren’.

Verbanden

6. Tabellen, diagrammen en grafieken

Onder dit onderwerp vallen opgaven waarbij verschillende soorten tabellen en grafieken aan bod komen, zoals cirkel-, staaf- en beelddiagrammen en lijngrafieken. De leerling moet de gegevens kunnen interpreteren en ermee rekenen.

In tabel 2.1 hieronder is aangegeven welke van de bovenstaande domeinen in de verschillende groepen in de toetsen naar voren komen.

Tabel 2.1 Domeinen die in de toetsen voorkomen in groep 3 tot en met 8

	Groep 3	**Groep 4**	**Groep 5**	**Groep 6**	**Groep 7**	**Groep 8**
Getallen	X	X	X	X	X	X
Verhoudingen				X	X	X
Meten	X	X	X	X	X	X
Verbanden				X	X	X

Voor groep 7 zijn, net als al in groep 6 het geval was, vier domeinen van belang, namelijk het domein Getallen, Verhoudingen, Meten en meetkunde en Verbanden. Nadere informatie over de invulling van deze onderdelen in groep 7 staat in hoofdstuk 3.

2.4.2 Psychometrisch

In deze paragraaf gaan we allereerst in op de procedures die Cito bij de constructie van de LVS-toetsen Rekenen-Wiskunde hanteert; zij komen in paragraaf 2.4.2.1 uitvoerig aan de orde. In deze paragraaf zal ook duidelijk worden dat het gehanteerde IRT-meetmodel in deze procedures een cruciale rol speelt.
In paragraaf 2.4.2.2 wordt uitvoerig op dit meetmodel ingegaan.

2.4.2.1 Opgavenbanken en constructieprocedures

Bij de constructie van opgaven wordt in de regel een veelvoud van het aantal items dat uiteindelijk in de normeringstoets moet worden ingezet afgenomen in een proeftoets. Er moet immers rekening worden gehouden met uitval, bijvoorbeeld wegens meer of minder triviale fouten in de constructie of extreme moeilijkheid of gemakkelijkheid. Ook ontstaat er op deze manier een overschot aan kwalitatief goede opgaven, die aan de opgavenbank worden toegevoegd. Een nieuwe toets wordt samengesteld uit een aantal nieuw geproeftoetste opgaven en uit opgaven die al eerder in de opgavenbank waren opgenomen. Een belangrijk kenmerk van deze opgavenbanken is dat ze gekalibreerd zijn volgens de principes van de IRT: item respons theorie.
Voor Rekenen-Wiskunde 3.0 wordt OPLM gebruikt (Verhelst, Glas en Verstralen, 1995; Verhelst, 1992 en Verhelst en Eggen, 1989; zie verder paragraaf 2.4.2.2), waarbij niet alleen de psychometrische kenmerken (parameters) van de opgaven worden geschat, maar waarbij tevens wordt nagegaan of de opgaven kunnen worden beschreven met een unidimensionele onderliggende vaardigheid.

Opgavenbanken

Bij het samenstellen van toetsen voor het primair onderwijs worden opgaven geselecteerd uit opgavenbanken. Een opgavenbank is nadrukkelijk niet eenvoudigweg een verzameling opgaven of items waaruit een toets-constructeur min of meer naar willekeur een aantal items selecteert om een nieuwe toets te construeren. Hieronder wordt beschreven wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

– *Unidimensioneel continuüm en latente vaardigheid*

Het algemene uitgangspunt is dat de vaardigheid rekenen-wiskunde kan worden opgevat als een unidimensioneel continuüm (de reële lijn), en dat elke leerling voorgesteld kan worden met een getal als een punt op die lijn. Het getal drukt de mate van vaardigheid uit, waarbij een groter getal wijst op een grotere vaardigheid.
De meetprocedure – het afnemen van de toets – heeft tot doel de plaats van de leerling op dit continuüm zo nauwkeurig mogelijk te bepalen. De uitkomst van de meetprocedure bestaat strikt genomen uit twee grootheden: de eerste is de schatting van de plaats van de leerling op het vaardigheidscontinuüm. De tweede grootheid geeft aan hoe nauwkeurig die schatting is, en heeft dus de status van een standaardfout, te vergelijken met de standaardmeetfout uit de klassieke testtheorie. De antwoorden van een leerling op de items worden beschouwd als indicatoren van de vaardigheid, hetgeen ruwweg betekent dat men verwacht dat alle items in de bank deze zelfde vaardigheid meten. De vaardigheid zelf wordt als niet observeerbaar beschouwd, en daarom gewoonlijk omschreven als een latente vaardigheid.

– *'Moeilijkheid' in de Item Respons Theorie*

Hoewel items dezelfde vaardigheid meten, kunnen ze toch systematisch van elkaar verschillen. Het belangrijkste verschil tussen de items is hun moeilijkheidsgraad. In de klassieke testtheorie wordt de moeilijkheidsgraad uitgedrukt met een zogenaamde p-waarde, de proportie correcte antwoorden op het item in een welbepaalde populatie van leerlingen. In de Item Respons Theorie (IRT) die voor het construeren van de opgavenbanken wordt gebruikt, hanteert men echter een andere definitie van moeilijkheid: ruwweg is het de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden. Dit verschil in definitie van de moeilijkheidsgraad tussen klassieke theorie en IRT is uitermate belangrijk: men kan verwachten dat de p-waarde van een item in groep 7 groter zal zijn dan in groep 6, waardoor duidelijk wordt dat de p-waarde een relatief begrip is: ze geeft de moeilijkheid aan van een item in een bepaalde populatie. Binnen de IRT is de moeilijkheid van een item gedefinieerd in termen van de onderliggende vaardigheid, zonder enige verwijzing naar een bepaalde populatie van leerlingen. Zo kan men ook de uitspraak begrijpen dat in de IRT vaardigheid en moeilijkheid op eenzelfde schaal liggen.

– *Kansmodel*

De ruwe omschrijving van de moeilijkheidsgraad die in de vorige alinea werd gehanteerd (de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden) behoeft verdere uitwerking.
Men zou deze omschrijving kunnen opvatten als een drempel: heeft een leerling die mate van vaardigheid niet, dan is hij niet in staat het item juist te beantwoorden; heeft hij die drempel wel gehaald, dan geeft hij (gegarandeerd) het juiste antwoord. Deze interpretatie weerspiegelt een deterministische kijk op het antwoord-gedrag van de leerling, die echter in de praktijk geen stand houdt, omdat eruit volgt dat een leerling die een moeilijk item correct beantwoordt geen fout kan maken op een gemakkelijk(er) item. Daarom wordt in de IRT een kansmodel gebruikt: hoe groter de vaardigheid, des te groter de kans dat een item juist wordt beantwoord.
De moeilijkheidsgraad van een item wordt dan gedefinieerd als de mate van vaardigheid die nodig is om met een kans van precies een half, een juist antwoord te kunnen produceren (zie verder ook de volgende paragraaf over het gehanteerde meetmodel).

– *Kalibratie*

In het voorgaande zijn nogal wat veronderstellingen ingevoerd (unidimensionaliteit; alle items zijn indicatoren voor dezelfde vaardigheid; kansmodel) die niet zonder meer voor waar kunnen worden aangenomen; er moet aangetoond worden dat al die veronderstellingen deugdelijk zijn. Dit 'aantonen' gebeurt met statistische gereedschappen waarop later dieper in wordt gegaan. Maar vóór de items in een toets gebruikt kunnen worden, moet ook geprobeerd worden de waarden van de moeilijkheidsgraden te achterhalen. Dit gebeurt met een statistische schattings-methode die wordt toegepast op de itemantwoorden die bij een steekproef van leerlingen zijn verzameld. Het hele proces van moeilijkheidsgraden schatten en verifiëren of de modelveronderstellingen houdbaar zijn, wordt kalibratie of ijking genoemd. De steekproef van leerlingen (in de boven al aangeduide proeftoets) die hiervoor wordt gebruikt heet kalibratiesteekproef.

– *Afnamedesigns*

Meestal bevat een opgavenbank meer items dan een doorsnee toets, zodat het praktisch niet haalbaar is om alle items aan alle leerlingen voor te leggen. Elke leerling in de kalibratiesteekproef krijgt daarom slechts een gedeelte van de items uit de opgavenbank voorgelegd. Er is dan sprake van een zogenoemd onvolledig design.
Dit gedeeltelijk voorleggen moet met de nodige omzichtigheid gebeuren. Verderop wordt ingegaan op het afnamedesign dat voor de kalibratie is gebruikt. De geïnteresseerde lezer wordt verwezen naar Eggen (1993).

– *Implicaties van gekalibreerde opgavenverzameling*

Als de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In het kalibratieproces worden de items die niet passen bij de verzameling uit de collectie verwijderd. De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken (unidimensionaliteit). Dit houdt onder meer het volgende in:

1 In principe kunnen we met een willekeurige selectie items uit de bank de vaardigheid meten bij een willekeurige leerling. De meting is nauwkeuriger wanneer het niveau van de opgaven beter aansluit bij het niveau van de leerling.

Het voorgaande geldt tevens voor de digitale items. Ook deze items komen uit de itembank Rekenen-Wiskunde. Dus ook met een selectie van digitale items kan de vaardigheid van een leerling bepaald worden. Al hetgeen dat geldt voor de ‘papieren’ items uit de itembank, geldt daarom eveneens voor ‘digitale’ items uit dezelfde itembank.

2 We kunnen een schatting maken van de verdeling van de vaardigheid in een welomschreven populatie, door selecties van items voor te leggen aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van het LVS zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 3 tot medio groep 8. Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van items bij een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde itembanken is immers dat met elke selectie items de vaardigheid van leerlingen kan worden bepaald. In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is.
Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.

3 Aan leerlingen die niet tot de betreffende referentiepopulatie behoren, kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 8 kan een toets maken die normaliter aan groep 6 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 8 ook vergeleken worden met de percentielen in de populatie van groep 6, met bijvoorbeeld de uitspraak: “De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 6.”

4 De vergelijking die bij punt 3 gemaakt is, kan evengoed plaatsvinden als de (achterstands)leerling een andere toets (i.e. een selectie uit de opgavenbank) maakt dan de toets die normaliter aan groep 7 wordt voorgelegd. Immers, het kalibratie-onderzoek heeft ons overtuigd dat alle items dezelfde vaardigheid meten. Met een nieuwe toets meten we dus dezelfde vaardigheid, zodat schattingen die van verschillende toetsen afkomstig zijn zinvol met elkaar kunnen worden vergeleken.
Meer over de kalibratieprocedure en een bespreking van de resultaten daarvan voor toetsen Rekenen-Wiskunde 3.0 is te vinden in hoofdstuk 4 over de normering van de toets.

2.4.2.2 Het gehanteerde meetmodel

In het normeringsonderzoek is gebruikgemaakt van een op de itemresponstheorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst & Kleintjes, 1993a; Verhelst, Glas en Verstralen, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogenaamde ware score, de gemiddelde score die de persoon zou behalen indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Deze klassieke testtheorie zou in dit onderzoek niet gebruikt kunnen worden, aangezien het normeringsonderzoek van de rekenwiskundetoetsen een onvolledig design betrof: niet alle leerlingen hadden alle opgaven gemaakt. Het gebruik van het IRT-model heeft enkele belangrijke voordelen. Op de eerste plaats kunnen de populatie-schattingen onafhankelijk van de schattingen van de itemparameters plaatsvinden. Dat heeft voordelen bij het wegen van de verschillende groepen om te zorgen dat de steekproef geheel overeenkomstig de populatieverdeling is (zie ook paragraaf 4.3). Daarna kan met deze populatieverdeling en kennis over de itemparameters precies bepaald worden welke de item- en toets-karakteristieken zijn voor de populatie. Ook als er ontbrekende waarnemingen zijn aan het einde van een test hebben we bij dergelijke schattingen geen last van de intrinsieke samenhang tussen reeksen van ontbrekende waarnemingen. Voor een overzicht van meer voordelen van IRT boven klassieke testtheorie wordt verwezen naar Hambleton, Swaminathan en Rogers (1991).
In de IRT staat het te meten begrip of de te meten eigenschap centraal. De IRT beschouwt het antwoord op een item als een indicator voor de mate waarin die eigenschap aanwezig is. Het verband tussen eigenschap en itemantwoord is van probabilistische aard en wordt weergegeven in de zogenaamde itemresponsfunctie. Die geeft aan hoe groot de kans is op een correct antwoord als functie van de onderliggende eigenschap of vaardigheid. Formeler: zij X_i de toevalsvariabele die het antwoord op item *i* voorstelt. X_i neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid kiezen we θ (theta). We wijzen erop dat θ niet rechtstreeks observeerbaar is. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd. De itemresponsfunctie *fi*(θ) is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 \mid \theta) \qquad (2.1)$$

Een IRT-model is een speciale toepassing van (2.1) waarbij aan de functie *fi*(θ) een meer of minder specifieke functionele vorm wordt toegekend. Een eenvoudig en zeer populair voorbeeld is het zogenoemde Raschmodel (Rasch, 1960) waarin *fi*(θ) gegeven is door

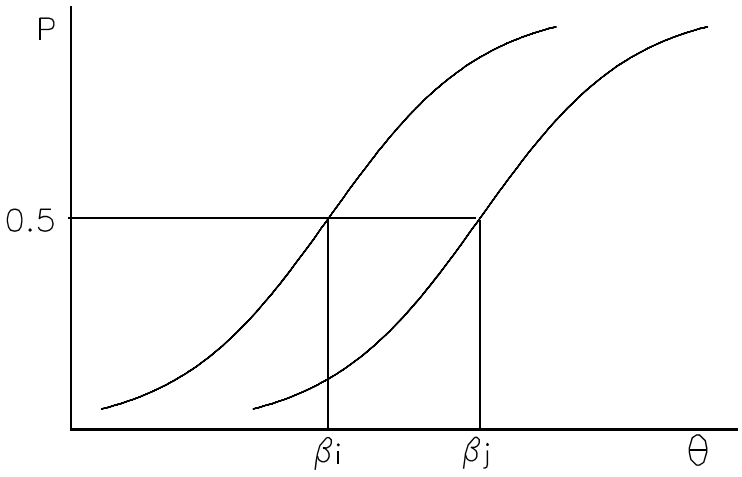
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \qquad (2.2)$$

waarin β*i* de moeilijkheidsparameter van item i is. Dat is een onbekende grootheid die geschat wordt uit de observaties. De grafiek van (2.2) is weergegeven in figuur 2.1 voor twee items, *i* en *j*, die in moeilijkheid verschillen. Deze figuur illustreert dat de itemresponsfunctie een stijgende functie is van θ: hoe groter de vaardigheid, des te groter de kans op een juist antwoord. Indien de latente vaardigheid precies gelijk is aan de moeilijkheidsparameter β*i*, krijgen we

$$f_i(\beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{1}{1 + 1} = \frac{1}{2} \qquad (2.3)$$

Daaruit volgt onmiddellijk een interpretatie voor de parameter β*i*: het is de 'hoeveelheid' vaardigheid die een leerling nodig heeft om een kans van precies een half te hebben om het item *i* juist te beantwoorden. Uit de figuur blijkt duidelijk dat voor item *j* een grotere vaardigheid nodig is om diezelfde kans te bereiken, maar dit is hetzelfde als te zeggen dat item *j* moeilijker is dan item *i*. We kunnen de parameter β*i* dus terecht omschrijven als de moeilijkheids-parameter van item *i*. De implicatie van het bovenstaande is dat 'moeilijkheid' en 'vaardigheid' op dezelfde schaal liggen. Formule (2.2) is geen beschrijving van de werkelijkheid, het is een hypothese over de werkelijkheid die getoetst kan worden op haar houdbaarheid. Hoe zo'n toetsing grofweg verloopt, is te verduidelijken aan de hand van figuur 2.1.

Figuur 2.1 Twee itemresponscurven in het Raschmodel



Daaruit blijkt dat, voor welk vaardigheidsniveau dan ook, de kans om item *j* juist te beantwoorden steeds kleiner is dan de kans op een juist antwoord op item *i*. Daaruit volgt de statistisch te toetsen voorspelling dat de verwachte proportie juiste antwoorden op item *j* kleiner is dan op item *i* in een willekeurige steekproef van personen. Splitst men nu een grote steekproef in twee deelsteekproeven, een 'laaggroep', met de vijftig procent laagste scores, en een 'hooggroep', met de vijftig procent hoogste scores, dan kan men nagaan of de geobserveerde p-waarden van de opgaven in beide deelsteekproeven op dezelfde wijze geordend zijn. Daarvan kan strikt genomen alleen sprake zijn als, in termen van de klassieke testtheorie uitgedrukt, alle opgaven eenzelfde discriminatie-index hebben. Dat echter blijkt lang niet altijd zo te zijn. Ook in het geval van de reken-wiskundetoetsen niet. Veel van de items blijken dan ook niet te kunnen worden beschreven met het Raschmodel. Daarom is bij dit instrument gekozen voor

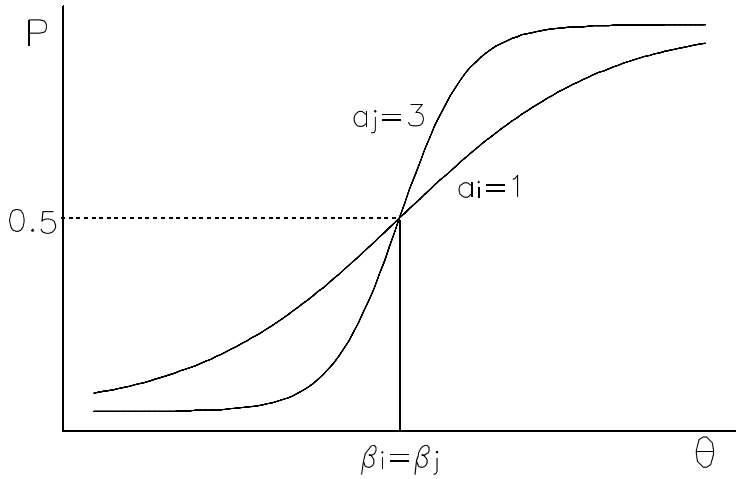
een ander IRT-model, waarbij de discriminatieparameter een rol speelt. De discriminatieparameter geeft de mate aan waarin de itemresponsefunctie verandert in de buurt van de moeilijkheidsparameter.
Alvorens het hier gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML).
Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproefgrootheid (sufficient statistic) bestaat voor de latente variabele θ, namelijk de ruwe score of het aantal correct beantwoorde items. Dat betekent grofweg dat, indien de itemparameters bekend zijn, alle informatie die het antwoord- patroon over de vaardigheid bevat, kan worden samengevat in de ruwe score; het doet er dan verder niet meer toe welke opgaven goed en welke fout zijn gemaakt. Hieruit vloeit voort dat de conditionele kans op een juist antwoord op item i, gegeven de ruwe score, een functie is die alleen afhankelijk is van de itemparameters en onafhankelijk van de waarde van θ. De CML-schattingsmethode maakt van deze functie gebruik. Deze methode maakt geen enkele veronderstelling over de verdeling van de vaardigheid in de populatie, en is ook onafhankelijk van de wijze waarop de steekproef is getrokken.

De CML-schattingsmethode is echter niet bij elk meetmodel toepasbaar. In het zogenaamde één parameter logistisch model (One Parameter Logistic Model, afgekort: OPLM) is CML mogelijk. Dit model is, anders dan het Raschmodel, wel bestand tegen 'omwisseling' van 'proporties juist' in verschillende steekproeven (Glas & Verhelst, 1993; Eggen, 1993; Verhelst & Kleintjes, 1993). De itemresponsfunctie van het OPLM is gegeven door

$$f_i(\theta) = \frac{\exp[\, a_i(\theta - \beta_i)\,]}{1 + \exp[\, a_i(\theta - \beta_i)]} \qquad (2.4)$$

waarin a_i de zogenaamde discriminatie-index van het item is. Door deze indices te beperken tot (positieve) gehele getallen, en door ze a-priori als constanten in te voeren, is het mogelijk CML-schattingen van de itemparameters β_i te maken. In figuur 2.2 is de itemresponscurve weergegeven van twee items *i* en *j*, die even moeilijk zijn maar verschillend discrimineren.

Figuur 2.2 Twee itemresponscurven in het OPLM: zelfde moeilijkheid, verschillende discriminatie



De schattingen worden berekend met het computerprogramma OPLM (Verhelst, Glas en Verstralen, 1995).
Dit programma voert ook statistische toetsen uit op grond waarvan kan worden bepaald of het model de gegevens adequaat beschrijft. Omdat een aantal van deze toetsen bijzonder gevoelig is voor een verkeerde specificatie van de discriminatie-indices, zijn de uitkomsten van deze toetsen bruikbaar als modificatie-indices: ze geven een aanwijzing in welke richting deze discriminatie-indices moeten worden aangepast om een betere overeenkomst tussen model en gegevens te verkrijgen. Kalibratie van items volgens het OPLM is dan ook een iteratief proces

waarin alternerend de modelfit van items wordt onderzocht door middel van statistische toetsing en de waarden van de discriminatie-indices worden aangepast op grond van de resultaten van deze toetsen. Deze aanpassingen geschieden in de praktijk op basis van een en hetzelfde gegevensbestand. Er kan dus kanskapitalisatie optreden.

De kans dat er een significant verband tussen variabelen wordt gevonden stijgt dan, terwijl het verband eigenlijk op toeval berust. Indien een steekproef een voldoende grootte heeft, is het effect van deze kanskapitalisatie echter gering (Verhelst, Verstralen en Eggen, 1991).
Hoewel het OPLM aanzienlijk flexibeler is dan het Raschmodel, heeft het met dit model toch een nadeel gemeen, waardoor het bij het kalibreren van meerkeuze-opgaven niet zonder meer bruikbaar is. Uit de formules (2.2) en (2.4) volgt dat, indien θ zeer klein is, de kans op een juist antwoord zeer dicht in de buurt van nul komt. Maar een aantal items in het normeringsonderzoek zijn meerkeuze-items, zodat blind gokken een zekere kans op een juist antwoord impliceert. Er bestaan modellen die rekening houden met de raadkans (Lord & Novick, 1968), maar die laten geen CML-schattingsmethode toe. De ongeschiktheid van het Raschmodel of OPLM voor meerkeuzevragen is echter relatief: indien de items in vergelijking met de vaardigheid van de leerling niet al te moeilijk zijn, blijkt dat het effect van het raden op de overeenkomst tussen model en gegevens klein is. Slechts een zeer beperkt aantal opgaven in de Reken-Wiskundetoetsen zijn meerkeuze-opgaven. Alleen bij opgaven die anders scorings-problemen geven en bij doelen die op andere wijze moeilijk te toetsen zijn is gebruikgemaakt van de meerkeuzevorm. Daarnaast zijn de pure gokkansen bij de meerkeuze-opgaven in de Reken-Wiskundetoetsen niet zeer groot: bij het willekeurig invullen meestal .25. Hierdoor en door een verstandige dataverzamelingsprocedure toe te passen en met name niet te moeilijke opgaven te selecteren in de test, kan het OPLM toch toegepast worden op meerkeuzevragen, waarbij de overeenkomst tussen model en data de uiteindelijke doorslag over die geschiktheid moet geven. Indien het meetmodel op grond van de kalibratieresultaten aanvaard kan worden, dat wil zeggen dat er na onderzoek geen praktische reden meer is om aan het meetmodel te twijfelen, dan kan men het meetmodel gebruiken om te gaan meten. Bij deze meetprocedure worden de itemparameters vastgezet op hun geschatte waarde uit de kalibratie.

Voor de schatting van de populatieverdeling wordt gebruikgemaakt van de ‘marginale grootste aannemelijkheids-methode’ (in het Engels: Marginal Maximum Likelihood, verder afgekort als MML).
Deze schattingsmethode veronderstelt naast (2.2) ook nog dat de vaardigheid θ in de populatie een bepaalde verdeling heeft. De meeste computerprogramma’s die IRT-analyses kunnen uitvoeren, veronderstellen een normale verdeling. Bovendien stelt deze methode de voorwaarde dat de steekproef die voor de schatting gebruikt wordt uit die verdeling een aselecte steekproef is. Omdat leerlingen bovendien gevolgd worden is het mogelijk gelijktijdig de verdelingen op de verschillende normeringsmomenten te schatten.

Geldigheid van de normen
De toetsen van het Cito Volgsysteem primair en speciaal onderwijs worden elke acht tot tien jaar vernieuwd. Niet alleen de inhoud wordt volledig vernieuwd en aangepast aan de ontwikkelingen in het onderwijs, ook worden de normen opnieuw vastgesteld. Omdat er enige tijd verloopt tussen de dataverzameling in het normeringsonderzoek en het moment waarop een vernieuwde toets wordt uitgebracht, kan men voor de toetsen Rekenen-Wiskunde 3.0 groep 7 een geldigheid aanhouden tot en met 2025.
Daarnaast monitort Cito periodiek de normering: jaarlijks wordt aan de hand van representatieve afnamedata nagegaan of er systematisch verschuivingen in het prestatieniveau plaatsvinden. Indien nodig wordt de normering aangepast.

3 Beschrijving van de toetsen

3.1 Opbouw en structuur van de toetsen

Het toetspakket Rekenen-Wiskunde 3.0 voor groep 7 uit het Cito Volgsysteem primair en speciaal onderwijs bevat in totaal twee papieren toetsen: M7 en E7 en twee digitale toetsen: M7 en E7. De toetsen M7 en E7 zijn bedoeld voor afname op de reguliere afnamemomenten medio (M) en einde (E) schooljaar. Voor groep 7 zijn geen tussentoetsen ontwikkeld. Aan een leerling waarbij de rekenvaardigheid zich minder snel ontwikkelt kan medio groep 7 of eind groep 7 één van de toetsen van groep 6 voorgelegd worden.

Opbouw
In totaal bevat de M7 toets 96 opgaven en ook de E7 toets bevat 96 opgaven. De reguliere toets M7 bestaat uit 3 taken van elk 32 opgaven. De reguliere taak E7 bestaat ook uit 3 taken van elk 32 opgaven. De taken dienen bij voorkeur te worden afgenomen op verschillende dagdelen zodat de leerlingen geconcentreerd aan alle taken kunnen werken.

Vorm
De toetsen van groep 7 bevatten naast meerkeuzeopgaven ook open opgaven. De opgaven van de toetsen M7, en E7 bestaan voor een deel uit opgaven met een context en voor een deel uit opgaven zonder context.
De leerlingen lezen zelf de opgaven en maken de toets zelfstandig. Bij de digitale taken kan de leerling ervoor kiezen de opgave te laten voorlezen door de computer. Bij de papieren afname beantwoorden de leerlingen de vragen door het antwoord op het antwoordblad te noteren. Hierdoor zijn de boekjes meerdere jaren te gebruiken. Bij een digitale afname toetsen de leerlingen hun antwoord op een bedieningspaneel in dat op het computer-scherm staat afgebeeld. Bij meerkeuzeopgaven klikken ze het antwoord van hun keuze aan met de muis.

Keuze van een passende toets: Toetsen op maat
De rekenvaardigheid van leerlingen in een groep loopt vaak sterk uiteen. Als gevolg daarvan zal eenzelfde rekentoets voor een deel van de leerlingen goed op niveau zijn, maar voor andere leerlingen erg moeilijk of erg gemakkelijk. Met name voor een aantal leerlingen van niveau IV en voor de leerlingen van niveau V (of de leerlingen van niveau D en E) zijn de toetsen van het eigenlijke afnamemoment (bijvoorbeeld de M7-toets voor leerlingen medio groep 7) aan de moeilijke kant. Voor een aantal leerlingen van niveau I (of niveau A) zijn de toetsen echter aan de gemakkelijke kant. De gehanteerde meettechniek maakt het mogelijk de toetsen op het niveau van de leerlingen af te stemmen. Omdat de toetsscores op verschillende rekentoetsen telkens naar eenzelfde schaal worden omgezet is het mogelijk leerlingen die verschillende toetsen maken toch met elkaar te vergelijken. Leerlingen kunnen daardoor bijvoorbeeld een toets maken die hoort bij een vorig afnamemoment (een M7-leerling maakt een toets E6) of een volgend afnamemoment (een M7-leerling maakt de toets E7). Voor zeer zwakke leerlingen of extreem vaardige leerlingen maakt de leerkracht, op basis van eigen observaties, resultaten op methodetoetsen en indien aanwezig eerdere resultaten op de LVS-toetsen een inschatting van de best passende toets. Hierbij vormt het onderwijsaanbod een belangrijk uitgangspunt: de toets dient zoveel mogelijk aan te sluiten bij de lesstof waar de leerling op dat moment aan werkt.
Een leerling die eind groep 7 nog met de lesstof van medio groep 7 bezig is, kan aan het einde van groep 7 bijvoorbeeld de toets M7 maken.

Het afnemen van de toetsen
De papieren toetsen kunnen zowel klassikaal als individueel afgenomen worden. De digitale versie maakt de leerling zelfstandig. De leerling kan ervoor kiezen de opgaven te laten voorlezen door de computer door te klikken op het icoontje van een oortje in het beeldscherm.

Van alle toetsen is zowel een papieren als een digitale variant beschikbaar. In de praktijk is gebleken dat de leerlingen voor het maken van de digitale versies minder tijd nodig hebben dan voor het maken van de papieren versies. In de toetsmappen is één handleiding opgenomen behorend bij zowel de papieren als de digitale toetsen. Deze handleiding richt zich op de organisatorische kant van de afname en op de verwerking en interpretatie van

de toetsresultaten. Meer technische aspecten van de digitale afname komen aan bod in een aparte handleiding voor de digitale toetsen.

Correctie van de toetsen
De papieren toetsen Rekenen-Wiskunde 3.0 zijn zowel handmatig na te kijken en te verwerken als via de computer, met behulp van het Computerprogramma LOVS of een leerlingadministratiepakket van een andere partij dan Cito. Voor het handmatig nakijken van de toets kan gebruikgemaakt worden van een lijst met goede antwoorden die in de bijlage van de handleiding is opgenomen. Indien gewenst kan de leerkracht in het Computerprogramma LOVS de goede antwoorden aanklikken.
Op basis van het aantal goede antwoorden (de toetsscore), wordt een inschatting gemaakt van de algemene rekenvaardigheid van de leerlingen. De leerkracht kan het aantal goede antwoorden invoeren in het administratiepakket dat de school gebruikt. De toetsscore wordt dan automatisch omgezet naar de bijbehorende vaardigheidsscore met een score-interval. Een andere optie is om met behulp van de omzettingstabellen op Cito Portal de vaardigheidsscore bij de behaalde toetsscore op te zoeken.
Bij de digitale versies van de toetsen worden de antwoorden van de leerlingen door de computer gescoord en hoeft de leerkracht de toetsen dus niet zelf na te kijken. Via het Computerprogramma LOVS worden de toetsscores omgezet naar de bijbehorende vaardigheidsscores.

Verwerking resultaten en interpretatie
De resultaten van de digitale afnames worden standaard met de computer verwerkt. De resultaten van de papieren afnames kan de leerkracht zowel met de computer als handmatig verwerken. Voor de handmatige verwerking zijn rapportageformulieren ontwikkeld die beschikbaar zijn via Cito portal.

Er zijn zowel rapportages op leerling-, groeps- als schoolniveau. Op leerlingniveau kan gekozen worden tussen het leerlingrapport en het alternatief leerlingrapport om te kunnen signaleren en om leerlingen in de tijd te volgen. Voor het signaleren op groepsniveau kan handmatig een groepsoverzicht gemaakt worden. Bij digitale verwerking zijn meerdere soorten (grafische) weergaven mogelijk van de resultaten, zoals het groepsrapport en het alternatief groepsrapport. Op schoolniveau kunnen resultaten nader bestudeerd worden door middel van een dwarsdoorsnede, een trendanalyse leerlingen, een trendanalyse jaargroepen en de rapportage vaardigheidsgroei.
In de handleiding voor de leerkrachten worden in hoofdstuk 4 de interpretatie- en analysemogelijkheden op leerling- en groepsniveau behandeld. In hoofdstuk 5 van de handleiding komt de interpretatie op schoolniveau aan bod. De handleiding gaat in op de inhoudelijke interpretatie van de (papieren en digitale) rapportages. In de handleiding bij het Computerprogramma LOVS staan de aanwijzingen over de wijze waarop de rapportages op te vragen zijn en welke keuzemogelijkheden de school hierbij heeft.
In de rapportagematerialen zijn twee niveau-indelingen opgenomen, waarmee de leerkracht de scores van een leerling kan vergelijken met die van een grote groep leerlingen. De leerkracht kan een keuze maken uit een indeling in de niveaus:
– I tot en met V;
– A tot en met E.
Daarnaast heeft de leerkracht de mogelijkheid om functioneringsniveaus op te vragen.

In het overzicht op pagina 21 wordt de betekenis van de vaardigheidsniveaus I tot en met V en A tot en met E toegelicht.

I – V		A – E	
20% hoogst scorende leerlingen	I 20%	A 25%	25% hoogst scorende leerlingen
20% boven het landelijk gemiddelde	II 20%	B 25%	25% ruim boven tot net boven het landelijk gemiddelde
20% landelijk gemiddelde	III 20%	C 25%	25% net tot ruim onder het landelijk gemiddelde
20% onder het landelijk gemiddelde	IV 20%	D 15%	15% ruim onder het landelijk gemiddelde
20% laagst scorende leerlingen	V 20%	E 10%	10% laagst scorende leerlingen

Bij de indeling in I tot en met V worden op de overzichten de laagste groep en de hoogste groep nog onderverdeeld in twee groepen die ieder 10% van de leerlingen bevatten. Deze groepen worden van elkaar gescheiden door een stippellijn. In de eerste versie van de LVS-toetsen werd alleen de indeling A tot en met E gebruikt. In de praktijk bleek deze enkele nadelen te hebben. Ten eerste is deze indeling niet symmetrisch. Bovendien zien sommige leerkrachten C als de gemiddelde groep. Het belangrijkste nadeel is echter dat er geen gemiddelde groep is. Daarom is bij de tweede versie van de toetsen een indeling toegevoegd met de niveaus I tot en met V. De indeling in de niveaus I tot en met V is symmetrisch opgebouwd en heeft als voordeel dat er een gemiddelde groep is. Deze indeling sluit aan bij de niveau-indeling van andere Cito-toetsinstrumenten zoals de Entreetoets. Vanaf afnamemoment E6 worden op het leerlingrapport niet alleen bovengenoemde niveau-indelingen – als vormen van relatieve normering – vermeld, maar ook – bij wijze van absolute, domeingerichte normering – de onderscheiden referentieniveaus. Er is zichtbaar vanaf welke vaardigheidsscore het 1F-niveau en het 1S-niveau is behaald (zie over referentieniveaus eerder p. 10). De bijbehorende cesuren in de vaardigheidsverdeling zijn in het normeringsonderzoek bepaald door een koppeling (ankering) aan te brengen met de zogenoemde referentiesets. Deze zijn ontwikkeld door het College voor Toetsen en Examens (CvTE). Door middel van openbare Ankersets Rekenen 1F en 1S zijn toetsontwikkelaars, uitgeverijen en onderzoekers in de gelegenheid gesteld de landelijke prestatiestandaard van de referentieniveaus over te brengen op hun eigen producten. Cito heeft hiervoor een apart onderzoek verricht in groep 8 om zo over een eigen referentieset te kunnen beschikken. In dit anker-onderzoek van april 2014 zijn 30 opgaven 1F, 20 opgaven 1S en 30 opgaven 2F van de Openbare Set opgenomen. Daarnaast zijn 295 LOVS-opgaven opgenomen. Deze items zijn in een 12 boekjes middels een onvolledig maar gelinkt design afgenomen bij 2390 leerlingen. Bijna elk item kwam in twee boekjes voor; het minimum aantal waarnemingen voor een item bedroeg 180, gemiddeld waren er 383 waarnemingen en het maximale aantal waarnemingen bedroeg 435. De range van p-waarden was .001-.955 met als gemiddelde .63. De fit van de items was met een R1c van 2910 met 2438 vrijheidsgraden goed te noemen. Doordat de OS-items en de LOVS-items op een schaal gebracht zijn kunnen we de referentie-lijntjes 1F en 1S nu overbrengen naar de LOVS3-schaal. Voor de uitgave van groep 6, 7 en 8 zijn ankeritems vanuit dat onderzoek meegenomen in het design van de normeringsonderzoeken. Daardoor kunnen vanaf groep 6 de lijnen van de referentieniveaus al weergegeven worden.

Naast de niveaus I tot en met V en A tot en met E kan de leerkracht functioneringsniveaus opvragen.
De functioneringsniveaus geven aan met welke gemiddelde leerling de vaardigheidsscore van de getoetste leerling vergelijkbaar is. Een functioneringsniveau M7 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling heel dicht ligt bij de score van de gemiddelde leerling medio groep 7. De indeling in functioneringsniveaus is oorspronkelijk ontwikkeld voor het speciaal (basis)onderwijs zodat de school op deze manier voor leerlingen met

forse leerachterstanden meer inzicht kregen in hun niveau. Mede dankzij de komst van het ‘passend onderwijs’ ontstond ook bij regulier onderwijs de wens om functioneringsniveaus te gebruiken en zijn de functioneringsniveaus opgenomen in de rapportages.

Analyse van resultaten: Categorieënanalyse
Voor analyses op leerlingniveau (van zowel de toetsresultaten van de papieren versies als de digitale versies) is een speciale rapportage ontwikkeld: de categorieënanalyse.
Bij elke toets kunnen de opgaven onderverdeeld worden in een relatief klein aantal didactisch zinvolle categorieën. Uit de vaardigheidsscore die de leerling behaalt en het toegekende niveau (I t/m V, A t/m E of functioneringsniveau) weten we hoe de score van de leerling zich verhoudt tot die van andere leerlingen. De categorieënanalyse is bedoeld als hulpmiddel voor de leerkracht om na te gaan of de leerling, gegeven zijn vaardigheidsniveau, evenwichtig presteert op de verschillende categorieën van de toets.
Met een categorieënanalyse kan nagegaan worden of leerlingen op een bepaald onderdeel meer of minder fouten maken dan op grond van hun vaardigheidsniveau verwacht mag worden. Wanneer een categorieënanalyse aanwijzingen geeft dat een leerling bij één of meerdere categorieën zwakker scoort dan verwacht, dan is voor de leerkracht het bijbehorende advies om aan de hand van bijvoorbeeld eigen observaties en de resultaten op methodetoetsen dit beeld te verifiëren en om de antwoorden bij de opgaven uit de betreffende categorie nader te bekijken. Indien nodig kan de leerkracht een diagnostisch gesprek voeren om meer informatie te verkrijgen over welke fouten de leerling bij deze categorie opgaven maakt.

Niet in alle leerjaren zijn dezelfde categorieën van toepassing. In de hoogste leerjaren worden de categorieën onderscheiden die ook bij de Centrale Eindtoets gebruikt worden. In de lagere leerjaren zijn andere categorieën van toepassing. In groep 7 worden de categorieën getallen, verhoudingen, meten en verbanden gebruikt. Niet elke categorie is met evenveel items vertegenwoordigd, dat zou immers geen recht doen aan de relatieve belangrijkheid van de categorieën in het onderwijs.
De verwachte aantallen worden vergeleken met het daadwerkelijk aantal goede antwoorden. Met behulp van een statistische toets (Chi-kwadraat) wordt aangegeven of het verschil significant is op het 10% niveau of het 5% niveau. Binnen het computerprogramma worden die verschillen aangeduid als ‘opvallend’ of ‘zeer opvallend’. In de grafische presentatie bij de categorieënanalyse kan een leerkracht zien of een leerling evenwichtig scoort op de domeinen en of het verschil als niet opvallend, opvallend of zeer opvallend moet worden geïnterpreteerd.

Naast een ‘categorieënanalyse leerling’ is er ook een zogenoemde ‘categorieënanalyse groep’. In deze laatste rapportage staan eerst per leerling alle resultaten uit de ‘categorieënanalyse leerling’ overzichtelijk onder elkaar voor de hele groep. Vervolgens staat in een tabel aangegeven hoeveel leerlingen uit die groep per categorie beneden en hoeveel leerlingen boven verwachting scoren, inclusief de gemiddelde afwijking naar beneden/boven. Op basis van de gemiddelde verschillen wordt met een t-toets nagegaan of er sprake is van significantie (tweezijdig). Significantie op het 10% niveau maar niet op het 5% niveau levert de kwalificatie opvallend op. Significantie op het 5% niveau levert de kwalificatie zeer opvallend op. Wanneer een groot deel van de leerlingen uit de groep beneden verwachting scoort en/of wanneer de gemiddelde afwijking naar beneden groot is, wordt een signaal (zeer) opvallend gegeven. Dit geeft aan dat deze categorie op groepsniveau om extra aandacht vraagt. Het signaal (zeer) opvallend kan echter ook betekenen dat in de groep juist opvallend veel leerlingen zijn die boven verwachting scoren en dat het dus een categorie betreft die de groep makkelijk af gaat.
In de handleiding bij het Computerprogramma LOVS is voor de leerkrachten een uitvoerige beschrijving opgenomen van de categorieënanalyse en de interpretatie van de uitkomsten. Ook in hoofdstuk 4 van de handleiding in de toetsmap staan aanwijzingen over de interpretatie en het gebruik van de ‘categorieënanalyse leerling’ en de ‘categorieënanalyse groep’.
Naar de categorieënanalyse is geen empirisch onderzoek verricht en deze moet dan ook puur gezien worden als een handreiking naar de leerkracht. De statistiek geeft aan hoe groot de verschillen zijn tussen verwacht en geobserveerd en of op basis van kansrekening aan de verschillen belang kan worden gehecht. Of er daadwerkelijk conclusies voor het onderwijs uit afgeleid kunnen worden hangt af van nadere analyse en interpretatie van de antwoorden van de leerling.
De toetsen en rapportagemogelijkheden maken deel uit van een systeem van leerlingenzorg waarbij een school werkt volgens de cyclus signaleren, analyseren, plannen en handelen. De rapportages richten zich hierbij op de eerste twee fases van deze cyclus.

3.2 Inhoudsverantwoording

In het ontwikkelproces van de toetsen is een aantal fasen te onderscheiden:
- uitwerking domeinbeschrijving;
- itemconstructie;
- proeftoetsing, normeringsonderzoek en kalibratie-analyses;
- samenstelling toetsen.

Deze fases worden hieronder nader toegelicht.
Deze informatie vormt een aanvulling op de inhoudsverantwoording die opgenomen is in de handleiding van het toetspakket Rekenen-Wiskunde 3.0 voor groep 7. In hoofdstuk 6 daarvan staat een uitgebreide inhoudsbeschrijving per afnamemoment en een serie overzichten die leerkrachten zicht geven op de doorgaande lijn bij de verschillende te onderscheiden onderwerpen. Met behulp van die overzichten kunnen de leerkrachten de scores van leerlingen inhoudelijk interpreteren. De paragrafen bestaan uit grafieken waarop de p50- en p80-kanspunten van de items in de toetsen, geordend op basis van p-waarde, zijn afgebeeld. Achter de portal staan de doorgaande lijnen met de balkjes behorende bij de vaardigheids-niveaus. Deze zijn gebaseerd op de meest recente normeringsgegevens. Bij de grafieken horen overzichten waarbij de opgaven eveneens zijn geordend op basis van p-waarden. Met een willekeurige vaardigheidsscore als uitgangspunt kan de leerkracht uit de overzichten afleiden welke opgaven van dat onderdeel bij die vaardigheidsscore goed beheerst worden, welke matig en welke onvoldoende.

Uitwerking domeinbeschrijving
Op basis van de domeinbeschrijving (zie paragraaf 2.4.1) zijn de domeinen en onderwerpen geselecteerd die relevant zijn voor groep 7. Deze onderwerpen komen aan bod in de meest gebruikte methodes voor rekenen-wiskunde in groep 7. Bij het construeren van de opgaven en het samenstellen van de toets is gekeken naar de wijze waarop en de mate waarin deze onderwerpen in de methodes naar voren komen. Doordat de leerlijnen van de methodes op hoofdlijnen aan elkaar gelijk zijn, en bij het construeren van de opgaven is nagegaan of groepen leerlingen die met een andere reken-wiskundemethode werken de betreffende stof aangeboden hebben gekregen, kunnen de toetsen bij elke rekenen-wiskunde methode van groep 7 gebruikt worden.
In hoofdstuk 2 zijn de globale doelen aangegeven voor het reken-wiskundeonderwijs voor de groepen 3 tot en met 8. Hier zullen we nu verder per onderwerp uitwerken welke leerstof in de toetsen voor groep 7 bij die onderwerpen globaal aan bod komt. De toetsen voor groep 7 bevatten opgaven uit vier domeinen: 1. Getallen, 2. Verhoudingen, 3. Meten en 4. Verbanden. Bij de verschillende onderwerpen binnen die domeinen zijn vervolgens doelen opgesteld die aansluiten bij de leerlijnen en leerstof van leerlingen in groep 7.
In het volgende overzicht staat per onderwerp aangegeven welke leerstof in de toetsen voor groep 7 aan bod komt.

Uitwerking doelen voor toetsen groep 7

Getallen
Bij dit onderdeel staan centraal het doorzien van de structuur van de telrij, de structuur van getallen en de relaties tussen getallen. Naast het getalbegrip vallen ook de bewerkingen binnen het domein Getallen.

1. Getalbegrip

1.1 Positiewaarde en positioneren:
- Het omzetten van aanduidingen in spreektaal naar getallen met cijfers, bijvoorbeeld 0,4 miljoen is 400 000.
- Bepalen van de waarde van cijfers in gehele getallen en kommagetallen, bijvoorbeeld weten en begrijpen dat in het getal 246,8 de 8 niet 8 maar 0,8 voorstelt.
- Inzicht in de plaats van gehele getallen, kommagetallen en breuken in de telrij onder andere door het plaatsen van getallen op de getallenlijn (zowel precies als globaal).
- De plaats van hele getallen, kommagetallen en breuken op de getallenlijn herkennen.
- Getallen plaatsen tussen andere getallen in de telrij, bijvoorbeeld door aan te geven waar 60 006 ligt op de getallenlijn tussen 60 000 en 60 010.

1.2 Tellen:
- Verder tellen en terugtellen met eenheden en bijvoorbeeld sprongen van 2, 10, 100, 1000 (vanaf een willekeurig getal), 5 (vanaf een veelvoud van 5), 25 (vanaf een veelvoud van 25), 50 (vanaf een veelvoud van 25) en van 0,1, 0,01 en 0,001.
- Samenstellen, gebruikmakend van de structuur van het positiesysteem, bijvoorbeeld 0,9 + 200 + 60 =.

1.3 Splitsen
- Splitsen op basis van positiewaarde.

1.4 Vergelijken
- Vergelijken en ordenen van gehele getallen, kommagetallen en breuken.
- Omzetten van kommagetallen in breuken.

1.5 Wiskundetaal gebruiken
- Vanuit een context naar een kale opgave.

1.6 Afronden van gehele getallen, kommagetallen en breuken
- Bepalen of naar boven of naar beneden afgerond moet worden.
- Aangeven wat het dichtst bij een bepaald getal ligt.

2. Bewerkingen

We onderscheiden binnen het onderdeel Bewerkingen vier deelgebieden namelijk optellen en aftrekken, vermenigvuldigen en delen, combinaties van berekeningen en bewerkingen met breuken. Er komen zowel contextopgaven als kale opgaven voor. De leerling moet bij het uitvoeren van bewerkingen gebruikmaken van basiskennis van getallen, inzicht in relaties tussen getallen en eigenschappen van bewerkingen. Ze mogen hierbij notities maken en tussenuitkomsten opschrijven.

Optellen en aftrekken met hele getallen en kommagetallen in contextopgaven en kale opgaven

Bij dit deelgebied gaat het om de vaardigheid van leerlingen om opgaven doelmatig of met behulp van een standaardprocedure of globaal schattend op te lossen.
- Doelmatig optellen en aftrekken gebruikmakend van eigenschappen van getallen en bewerkingen, waaronder ook het rekenen met nullen.
- Standaardprocedures gebruiken bij optellen en aftrekken van grotere getallen met meer cijfers.
- Globaal (benaderend) optellen en aftrekken met grotere hele getallen.

Vermenigvuldigen en delen met hele getallen en kommagetallen in contextopgaven en kale opgaven

Ook bij dit deelgebied gaat het om de vaardigheid van leerlingen om opgaven doelmatig of met behulp van een standaardprocedure of globaal schattend op te lossen.
- Doelmatig vermenigvuldigen en delen gebruikmakend van eigenschappen van getallen en van bewerkingen, waaronder ook rekenen met nullen.
- Standaardprocedures gebruiken bij vermenigvuldigen en delen van grotere gehele getallen en kommagetallen. Bij delen met rest in contexten de rest juist interpreteren of verwerken.
- Globaal (benaderend) vermenigvuldigen en delen met grotere hele getallen en kommagetallen zoals 49 x 198,95 is ongeveer 50 x 200.

Combinaties van berekeningen

Bij dit deelgebied gaat het om het uitvoeren van meerdere bewerkingen in één opgave.
- Berekeningen uitvoeren met combinaties van bewerkingen.
- Berekenen van gemiddelde.
- De geldende regels kennen voor de volgorde waarin rekenbewerkingen moeten worden uitgevoerd.

Bewerkingen met breuken

Bij bewerkingen met breuken gaat het onder andere om het vereenvoudigen en gelijknamig maken, optellen en aftrekken en een deel van een geheel getal nemen.
- Vereenvoudigen en gelijknamig maken van breuken.

- Breuken als gemengd getal schrijven.
- Optellen en aftrekken van gelijknamige en ongelijknamige breuken en gemengde getallen in opgaven met en zonder context.
- Deel nemen van een geheel getal, heel getal vermenigvuldigen met een breuk en omgekeerd, zowel in contextopgaven als in opgaven zonder context.
- Breuk vermenigvuldigen met een breuk of een deel nemen van een breuk.
- Een heel getal delen door een breuk/gemengd getal en een breuk/gemengd getal delen door een breuk.

3. Verhoudingen

Bij dit onderdeel gaat het erom dat leerlingen structuur en samenhang van verhoudingen op hoofdlijnen leren te doorzien en dat zij er in praktische situaties mee rekenen. Verhoudingen kunnen beschreven worden in verhoudingentaal (zoals bij één op de tien Nederlanders) maar soms ook in breukentaal (bijvoorbeeld driekwart van de inwoners is ouder dan 25 jaar) of met procenten (80 procent van de mensen stemde voor). Binnen verhoudingen gaat het dus soms ook om het rekenen met breuken en procenten. Daar waar het om breuken als getallen gaat of om bewerkingen met breuken hebben we de doelen bij het domein getallen beschreven.

Verhoudingen herkennen, benoemen en gebruiken

- Verhoudingen herkennen, benoemen, beschrijven en gebruiken als zoveel op de zoveel, als deel van een geheel.
- Oplossen van eenvoudige verhoudingsproblemen met hele getallen, kommagetallen en breuken, bijvoorbeeld bij gebruik van recepten, snelheid, prijs per stuk/liter/kg, mengen, afstanden vergelijken, vergroten en verkleinen.

Werken met schaallijn en schaal

- Het begrip schaal kennen en rekenen met schaallijnen en schaalnotaties.
- Lengte of afstand bepalen op basis van gegeven schaal.

Procenten herkennen en gebruiken

- Bij verdelingen de ontbrekende percentages berekenen op basis van de kennis dat het geheel 100% is.
- Rekenen met percentages via het deel nemen van, via rekenen met breuken, via verhoudingen, via vermenigvuldigen met een bijbehorend kommagetal of via 1% in opgaven met en zonder context.
- Berekenen hoeveel procent het deel is, of hoeveel de toename of afname is.
- Betekenis geven aan percentages boven 100% en hiermee rekenen in opgaven met en zonder context.
- Verhoudingen en procenten in elkaar omzetten. Procenten en kommagetallen in elkaar omzetten. Breuken en procenten in elkaar omzetten. Verhoudingen en breuken in elkaar omzetten.

4. Meten

Bij dit onderdeel gaat het erom greep te krijgen op de werkelijkheid om ons heen. Het gaat om:

- basiskennis en begrip van verschillende grootheden (lengte, omtrek, oppervlakte, inhoud, gewicht, tijd, snelheid en geld);
- leren meten met verschillende instrumenten;
- leren meetresultaten af te lezen en te interpreteren;
- het omzetten van maateenheden en de opbouw en decimale structuur van het metrieke stelsel.

Bij meetkunde ligt de nadruk op het beschrijven van en het greep krijgen op ruimtelijke aspecten van de werkelijkheid.

Lengte en omtrek

- Meetinstrumenten hanteren en aflezen: liniaal, meetlint, rolmaat, kilometerteller.
- Orde van grootte van lengte en omtrek in een situatie inschatten op basis van referentiematen.
- Kiezen van de juiste lengtemaat bij een situatie.
- Omtrek berekenen van rechthoekige figuren.
- Herleidingen uitvoeren met de lengtematen kilometer, hectometer, decameter, meter, decimeter, centimeter, millimeter.
- Oplossen van toepassingsproblemen waarbij herleidingen en berekeningen met lengtematen uitgevoerd moeten worden.

Oppervlakte

- Orde van grootte van oppervlakte in een situatie inschatten op basis van referentiematen (de oppervlakte van een deur is ongeveer 2 m^2).
- Kiezen van de juiste oppervlaktemaat bij een situatie.
- Oppervlakte bepalen met natuurlijke maten (bijvoorbeeld het aantal tegels dat op een vloer komt te liggen).
- Oppervlakte berekenen van rechthoekige figuren.
- De samenhang tussen standaardmaten kennen en herleidingen uitvoeren (1 m^2 = 10 000 cm en 1 hectare is 100 are). Het betreft km^2, hm^2, dam^2, m^2 dm^2 cm^2 mm^2 hectare en are.
- Oplossen van toepassingsproblemen waarbij berekeningen met oppervlaktematen moeten worden uitgevoerd.

Inhoud

- Meetinstrumenten hanteren en aflezen.
- Kiezen van de juiste inhoudsmaat bij een situatie.
- Notie van inhoudsmaten en het gebruik hiervan.
- Bepalen inhoud met een natuurlijke maat, bijvoorbeeld het bepalen van het aantal blikken in een stapel.
- Samenhang tussen standaardmaten kennen en herleidingen uitvoeren. Het betreft liter, deciliter, centiliter en milliliter, m^3, dm^3, cm^3 en mm^3
- Oplossen van toepassingsproblemen waarbij herleidingen en berekeningen met inhoudsmaten uitgevoerd moeten worden.

Gewicht

- Aflezen van het weegresultaat op een weegschaal.
- Notie van gram en kilogram en het gebruik daarvan.
- Kiezen van de juiste maat bij een situatie.
- Samenhang tussen de standaardmaten gram en kilogram kennen en herleidingen uitvoeren.
- Oplossen van toepassingsproblemen waarbij herleidingen en berekeningen met gewichtsmaten uitgevoerd moeten worden.

Meetkunde

- Aangeven hoe een vooraanzicht er van bovenaf of vanaf de zijkant uitziet.
- Aangeven welk figuur van gegeven stukjes kan worden gemaakt.
- Omvormen van figuren en het in gedachten reconstrueren van bouwplaten.
- Symmetrie herkennen en het spiegelbeeld van een figuur herkennen.
- Interpreteren van plattegronden en bouwtekeningen.
- Mentaal innemen van standpunten (meetkundig).

Tijd en snelheid

- Aflezen van digitale en analoge tijden.
- Omzetten van digitale tijdsaanduidingen in analoge tijden en omgekeerd.
- Herleidingen uitvoeren met de tijdmaten uur, kwartier, minuut en seconde.
- Het gebruiken van gegevens van een kalender.
- Rekenen met tijdsaanduidingen (tijdsduur en tijdsverschillen).

Geld

- Gepast samenstellen van bedragen met munten/biljetten.
- Totale waarde van munten/biljetten bepalen.
- Omwisselen van munten/biljetten in andere munten/biljetten.
- Rekenen met geldbedragen in euro’s (betalen/wisselgeld/bijbetalen).
- Oplossen van toepassingsproblemen waarin met euro’s en andere valuta gerekend wordt.
- Samenstelling gebruiken als prijs/uur, prijs/meter en prijs/liter.

5. Verbanden

Binnen dit onderdeel gaat het om het omgaan met tabellen, diagrammen en grafieken.
Tabellen, diagrammen en grafieken worden tegenwoordig in veel toepassingen frequent gebruikt om getalsmatige (kwantitatieve) gegevens op een compacte en overzichtelijke manier weer te geven. Op de televisie, op sites, in dagbladen en tijdschriften en schoolboeken worden allerlei soorten tabellen, diagrammen en grafieken gebruikt om met name kwantitatieve informatie over te brengen. Aflezen, interpreteren en combineren van de verschillende representatievormen, grafisch weergeven van informatie en het herkennen en beschrijven van een onderliggend verband of voorspellingen doen zijn belangrijke vaardigheden voor kinderen om te leren.

- Lezen en interpreteren van gegevens uit tabellen.
- Lezen en interpreteren van gegevens uit diagrammen zoals cirkeldiagram, staafdiagram (horizontaal, verticaal) en beelddiagram.
- Lezen en interpreteren van gegevens uit lijngrafieken.
- Verschillende informatiebronnen met elkaar in verband brengen (tabellen, diagrammen en grafieken) en hieruit gegevens lezen, vergelijken en interpreteren.

Bij de verschillende doelen voor groep 7 op een afnamemoment zijn opgaven geconstrueerd die een operationalisering vormen van die doelen.
Voor een beschrijving van de inhoud van de toetsen M7 en E7 verwijzen we ook naar de Inhoudsverantwoording in het toetspakket (Cito, 2017). Daarin is een uitgebreide inhoudsbeschrijving opgenomen die geïllustreerd wordt met voorbeeldopgaven uit de toetsen, alsmede een aanduiding van de moeilijkheidsgraad van die opgaven op de zogenaamde doorgaande lijnen.

In tabel 3.1 staan de aantallen opgaven per domein, per toets weergegeven. De verdeling is in grote lijnen hetzelfde voor de papieren toetsen en de digitale toetsen. Deze opgaven vormen de basis voor de categorieënanalyse.

Tabel 3.1a Verdeling opgaven over domeinen van de papieren toetsen van de uitgave Rekenen-Wiskunde groep 7

	M7	**E7**
Getallen	42	40
Verhoudingen	12	16
Meten	30	28
Verbanden	12	12

Tabel 3.1b Verdeling opgaven over domeinen van de digitale toetsen van de uitgave Rekenen-Wiskunde groep 7

	M7	**E7**
Getallen	42	40
Verhoudingen	14	16
Meten	28	28
Verbanden	12	12

Itemconstructie

De toetsen bestaan uit meerkeuzeopgaven en open opgaven waarbij de leerling een kort antwoord in de vorm van een getal geeft. Bij de meerkeuzeopgaven geeft de leerling zijn antwoord door een kruisje bij één van de alternatieven te zetten (papier) of door op het betreffende alternatief te klikken (digitaal).
De opgaven, aansluitend bij de domeinbeschrijving, zijn geconstrueerd door itemschrijfcommissies die bestonden uit leerkrachten basisonderwijs, schoolbegeleiders en pabodocenten. De constructeurs kregen een opdracht, opgesteld door toetsdeskundigen van Cito. In deze opdracht stond omschreven voor welke categorieën opgaven geconstrueerd moesten worden. In een meegeleverde toetswijzer kregen de constructeurs voorbeeldopgaven ter

illustratie van de gebruikte categorieën. Ook kregen de constructeurs de belangrijkste richtlijnen waar de opgaven aan moesten voldoen, zoals bijvoorbeeld aanwijzingen over taal en gebruik van afbeeldingen. Geconstrueerde items zijn in commissievergaderingen onder leiding van een toetsdeskundige besproken en zo nodig bijgesteld. Na de uitwerking van de opgaven door toetsdeskundigen van Cito en door tekenaars zijn de opgaven nogmaals gescreend. Bij die screening zijn naast leerkrachten basisonderwijs ook leerkrachten uit het Speciaal (Basis) Onderwijs, die veel werken met leerlingen met extra onderwijsbehoeften, betrokken geweest. Door al deze activiteiten wordt voorkomen dat dubbelzinnigheden of onvolkomenheden in de opgaven zitten. Dit zorgt ervoor dat leerlingen de inhoud van de items juist interpreteren.

Proeftoetsing, normeringsonderzoek en kalibratie-analyses
Bij een proeftoetsing zijn in 2011 halverwege en aan het einde van leerjaar 7 nieuw ontwikkelde opgaven voorgelegd aan leerlingen. Daarbij zijn voor het afnamemoment medio groep 7 ongeveer 240 nieuwe opgaven en voor het afnamemoment eind groep 7 ongeveer 200 nieuwe opgaven geproeftoetst. Elke deelnemende school maakte 1 taak met ongeveer 32 opgaven. Voor elke opgave zijn ongeveer 100 à 150 responsen verzameld.
In 2013 is opnieuw een proeftoetsing voor leerjaar 7 uitgevoerd. Voor M7 zijn toen ongeveer 370 opgaven geproeftoetst en voor E7 380 opgaven. In totaal waren er na de proeftoetsingen voor zowel M7 als voor E7 ruim 500 unieke opgaven beschikbaar. Na de proeftoetsingen zijn opgaven geselecteerd op basis van moeilijkheidsgraad en discriminerend vermogen. Deze items zijn opgenomen in het normeringsonderzoek. De opgaven zijn bij het normeringsonderzoek op basis van het afnamedesign voorgelegd aan een steekproef van leerlingen en scholen in 2015. Het afnamedesign werd zo ingericht dat a) de nieuw geconstrueerde opgaven bij de kalibratie konden worden gekoppeld aan de opgaven van de bestaande toetsen van het leerlingvolgsysteem (LVS-II) en b) de opgaven van het betreffende afnamemoment (per categorie) konden worden gekoppeld aan de opgaven van zowel een eerder als een later afnamemoment en c) alle nieuwe opgaven onderling konden worden gekoppeld.

Het afnamedesign voor het normeringsonderzoek voor afnamemoment medio groep 7 (M7) bestond uit in totaal twaalf opgavenboekjes met elk 32 of 33 opgaven. Alle leerlingen maakten de bestaande M7-toets van LVS II (3 taken) en 1 taak met nieuw geconstrueerde opgaven voor M7 en nieuwe ankeropgaven. Daarnaast werden in een aantal nieuw geconstrueerde taken ook referentieset opgaven opgenomen. De nieuw geconstrueerde ankeropgaven hadden deels betrekking op het latere afnamemoment E7 en deels betrekking op het eerdere afnamemoment E6.
Het normeringsonderzoek voor afnamemoment einde groep 7 (E7) kende een soortgelijk afnamedesign, bestaande uit twaalf opgavenboekjes met elk 34 opgaven, met dien verstande dat de toegevoegde ankeropgaven hier betrekking hadden op het eerdere afnamemoment M7, het latere afnamemoment E7, en de items van de Referentieset.
Voor beide afnamedesigns geldt dat alle nieuwe opgaven in 2 taken zijn opgenomen om deze onderling te kunnen verbinden. Deze taken werden zo over de opgavenboekjes verdeeld dat deze koppeling mogelijk was.
De (nieuwe) ankeropgaven zijn slechts in 1 taak opgenomen.

In het normeringsonderzoek M7 zijn op deze wijze in totaal 243 verschillende nieuwe items voorgelegd aan 2279 leerlingen van groep 7 verdeeld over 12 boekjes. Elk boekje bestond uit 128 of 129 opgaven verdeeld over 4 taken. De 96 opgaven (verdeeld over 3 taken) uit de bestaande LVS-II taken werden door alle leerlingen gemaakt. In de nieuwe taken kwamen de E6-ankeropgaven en de E7-ankeropgaven 1 keer voor. Elke nieuwe opgave voor medio groep 7 kwam in 2 boekjes voor. De nieuwe opgaven werden minimaal door 155 leerlingen en maximaal door 410 leerlingen en gemiddeld door 301 leerlingen gemaakt. Van de 2279 hebben 67 leerlingen geen volledig boekje gemaakt. De positie van de nieuwe taak in de boekjes varieerde.

Tabel 3.2 Design medio groep 7 met aantallen opgaven en aantallen leerlingen per boekje (B1-B12)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
E6 anker	0	3	3	3	3	0	3	3	4	4	4	4
M7 nieuwe opgaven	30	27	27	27	30	27	27	27	21	18	18	21
E7 anker	0	0	0	3	0	3	0	0	4	7	10	7
Refset	3	3	3	0	0	3	3	3	3	3	0	0
Nieuwe taak totaal	33	33	33	33	33	33	33	33	32	32	32	32
LVSM7 gen II taak 1,2 3	96	96	96	96	96	96	96	96	96	96	96	96
Totaal aantal opg	129	129	129	129	129	129	129	129	128	128	128	128
Aantal leerlingen	163	224	186	166	155	201	193	201	193	172	193	163
Aantal scholen	5	7	6	8	6	7	8	7	8	8	7	7

De taak met nieuw geconstrueerde opgaven voor M7 bevatte in elk van de boekjes opgaven uit alle vier onderscheiden opgavencategorieën. Uit het hierna volgende overzicht is af te lezen hoe de verdeling van de opgaven over de categorieën is gerealiseerd.

Tabel 3.3 Aantallen opgaven per domein in de nieuwe taken van medio groep 7

	B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**	**B11**	**B12**
getallen	15	18	15	15	15	15	15	14	14	14	14	17
verhoudingen	6	3	5	6	3	6	6	4	3	3	6	3
meten	9	9	7	9	9	7	9	9	9	9	9	9
verbanden	3	3	6	3	6	5	3	6	6	6	3	3
Totaal	33	33	33	33	33	33	33	33	32	32	32	32

In het normeringsonderzoek E7 zijn in totaal 346 verschillende items voorgelegd aan 1187 leerlingen van eind groep 7 verdeeld over 12 boekjes. In totaal daarvan zijn 250 items nieuw vanuit LVS generatie III. 11 leerlingen hebben geen volledig boekje gemaakt. Elk boekje bestond uit 130 opgaven verdeeld over 4 taken.
De 96 opgaven (verdeeld over 3 taken) in de LVS-II taken werden aan alle 1187 leerlingen voorgelegd. In de nieuwe taken kwamen de M7-ankeropgaven, M8- ankeropgaven, en de referentiesetopgaven 1 keer voor. Elke nieuwe opgave voor eind groep 7 kwam in 2 boekjes voor. De nieuwe opgaven werden minimaal door 53 leerlingen en maximaal door 219 leerlingen en gemiddeld door 160 leerlingen gemaakt. De positie van de nieuwe taak in de boekjes varieerde.

Tabel 3.4 Design eind groep 7 met aantallen opgaven en aantallen leerlingen per boekje (B1-B12)

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
M7 anker	3	3	3	0	0	3	4	4	4	4	3	3
E7 nieuwe opgaven	22	25	21	27	27	24	27	27	27	27	31	31
M8 anker	3	3	4	4	4	4	3	3	3	3	0	0
Refset anker	6	3	6	3	3	3	0	0	0	0	0	0
Nieuwe taak totaal	34	34	34	34	34	34	34	34	34	34	34	34
LVS E7 gen II taak 1	32	32	32	32	32	32	32	32	32	32	32	32
LVS E7 gen II taak 2	32	32	32	32	32	32	32	32	32	32	32	32
LVS E7 gen II taak 3	32	32	32	32	32	32	32	32	32	32	32	32
Totaal aantal opgaven	130	130	130	130	130	130	129	130	130	130	130	130
Aantal leerlingen	106	53	108	110	108	109	130	103	106	108	68	92
Aantal scholen	2	2	4	4	4	4	4	4	4	4	3	4

De taak met nieuw geconstrueerde opgaven voor E7 bevatte in elk van de 12 boekjes opgaven uit alle 4 onderscheiden opgavencategorieën. Uit het hierna volgende overzicht is af te lezen hoe de verdeling van de opgaven over de categorieën is gerealiseerd.

Tabel 3.5 Aantallen opgaven per domein in de nieuwe taken van eind groep 7

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
Getallen	14	12	16	15	16	15	16	16	14	18	14	15
Verhoudingen	6	5	4	6	3	7	4	4	4	4	6	6
Meten	10	10	8	6	11	8	10	10	10	8	8	10
Verbanden	4	7	6	7	4	4	4	4	6	4	6	3
Totaal	34	34	34	34	34	34	34	34	34	34	34	34

Na de normeringsafnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket OPLM (Verhelst, 1993; Verhelst en Glas, 1995). Voor een algemene technische beschrijving van dit model zie paragraaf 2.4.2.
Bij de analyses is de kwaliteit van de afzonderlijke items en de totale verzameling voor een afnamemoment in kaart gebracht. Moeilijkheidsparameters en discriminatieparameters zijn geschat. Bij de analyses van de antwoorden van de leerlingen op de opgaven is nagegaan of de verschillende onderdelen een beroep doen op hetzelfde complex aan vaardigheden. Dat bleek het geval te zijn. Voor uitgebreidere informatie over de kalibratie verwijzen wij naar hoofdstuk 4.
Op basis van de informatie uit de analyses is voor eind groep 6 tot en met medio groep 8 een schaal geconstrueerd en is een selectie van opgaven gemaakt voor de papieren uitgave.

Het kalibratie-onderzoek digitaal en papier-digitaal
De geselecteerde opgaven voor de papieren uitgave van medio groep 7 en eind groep 7 (van LVS generatie III) zijn, met voor elke categorie enkele extra opgaven, gedigitaliseerd. De digitale opgaven zijn in een digitaal onderzoek en een (vergelijkend) papier-digitaal onderzoek, afgenomen in januari 2016 bij leerlingen van medio groep 7 en in juni 2016 bij leerlingen van eind groep 7. Het doel van deze afnames was om gegevens te verzamelen voor het schatten van itemparameters van de digitale itemvarianten en om de digitale items op dezelfde schaal onder te brengen als de papieren items.

Bij zowel leerlingen van medio groep 7 als eind groep 7 zijn digitale taken met nieuwe opgaven afgenomen in combinatie met de papieren taken van LVS generatie II en in combinatie met de digitale taken van LVS generatie II.

Zowel bij de afnames voor medio groep 7 als bij de afnames voor eind groep 7 waren 6 groepen leerlingen betrokken. In de tabellen hieronder is te zien welke taken die groepen leerlingen gemaakt hebben.

Tabel 3.6 Design medio groep 7 digitaal onderzoek (B2, B4 en B6) en papier-digitaal onderzoek (B1, B3 en B5)

	B1	**B2**	**B3**	**B4**	**B5**	**B6**
LVS M7 gen III digitaal taak 1 (nieuw) (36 items)						
LVS M7 gen III digitaal taak 2 (nieuw) (35 items)						
LVS M7 gen III digitaal taak 3 (nieuw) (35 items)						
LVS M7 gen II papier (96 items)						
LVS M7 gen II digitaal (96 items)						
Aantal leerlingen	176	73	1514	105	130	102
Aantal scholen	9	4	7	6	6	6
Aantal items	132	132	131	131	131	131

Gemiddeld zijn er 256 waarnemingen voor elk nieuw digitaal item. Het minimaal aantal waarnemingen per item is 232.

Tabel 3.7 Design eind groep 7 digitaal onderzoek (B2, B4 en B6) en papier-digitaal onderzoek (B1, B3 en B5)

	B1	**B2**	**B3**	**B4**	**B5**	**B6**
LVS E7 gen III digitaal taak 1 (nieuw) (36 items)						
LVS E7 gen III digitaal taak 2 (nieuw) (35 items)						
LVS E7 gen III digitaal taak 3 (nieuw) (35 items)						
LVS E7 gen II papier (96 items)						
LVS E7 gen II digitaal (96 items)						
Aantal leerlingen	107	87	93	75	93	101
Aantal scholen	6	4	5	4	4	5
Aantal items	132	132	131	131	1312	131

Gemiddeld zijn er 185 waarnemingen per item voor de nieuwe items E7 generatie III.

Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programmapakket OPLM (Verhelst, 1992; Verhelst, Glas en Verstralen, 1995) en is één itembank met items voor papieren en digitale afnames samengesteld. Omdat de papieren en digitale opgaven op de vaardigheidsschaal rekenen-wiskunde passen kunnen we zeggen dat de papieren en digitale opgaven dezelfde vaardigheid meten. Zowel de papieren toetsen als de digitale toetsen bevatten opgaven met psychometrisch goede eigenschappen. Afnames van papieren en digitale toetsversies leveren vergelijkbare resultaten op bij de vaststelling van het vaardigheidsniveau van leerlingen.

Samenstelling toetsen

De opgaven voor de toetsen zijn geselecteerd uit de verzameling opgaven van de gekalibreerde itembank.
Die bank bevat opgaven van LVS generatie II en LVS generatie III voor papieren en digitale toetsen.
In totaal zijn er vier toetsen samengesteld. Voor de papieren variant twee toetsen: M7 en E7 en voor de digitale variant twee toetsen: M7 en E7.
De toetsen voor M7 en E7 van de papieren variant zijn samengesteld uit opgaven die in de nieuwe taken bij het normeringsonderzoek zaten. De toetsen voor M7 en E7 van de digitale variant zijn samengesteld uit opgaven die in de nieuwe taken zaten van het digitale en papier-digitale onderzoek.

Alle toetsen zijn samengesteld op basis van inhoudelijke en psychometrische criteria. Voor de samenstelling van de toetsen is gekeken naar de verdeling van de opgaven over de verschillende categorieën en het belang van de betreffende onderdelen voor het onderwijs. De aantallen zoals vermeld in de toetsmatrijs van tabel 3.1 in paragraaf 3.2 geven de verdeling over de categorieën aan die in de uitgegeven toetsen is toegepast. Bij het selecteren van opgaven is gekeken naar de gewenste verdeling over de domeinen en naar een adequate verdeling van de moeilijkheidsgraad van de opgaven en het discriminerend vermogen van de opgaven.

De toetsen zijn geschikt om verschillen in rekenvaardigheid tussen leerlingen in beeld te brengen. Dit komt doordat opgaven van verschillende moeilijkheidsgraad zijn opgenomen. In de toetsen worden makkelijke en moeilijke opgaven afwisselend aangeboden. Een goede illustratie hiervan en van de samenstelling van de toetsen zijn de figuren in bijlage 1: p50 en p80-kanspunten van de opgaven in de papieren toetsen en digitale toetsen M7, en E7 in relatie tot de vaardigheidsverdelingen van E6, M7, E7 en M8.
In deze figuren is zichtbaar dat de toetsen opgaven bevatten van uiteenlopende moeilijkheidsgraad. In de figuren is de verdeling van de opgaven over de toetsen M7 en E7 visueel weergegeven. De balkjes in de figuren geven het p50- (onderkant van het balkje) en p80-kanspunt (bovenkant van het balkje) van elke opgave aan. Het p50-punt geeft de vaardigheidsscore aan waarbij er sprake is van een kans van 50% om een opgave goed te beantwoorden.
Bij de toets M7 ligt het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 7. Bij de toets E7 liggen de opgaven veelal hoger op de vaardigheidsschaal, rond de gemiddelde vaardigheidsscore eind groep 7. In de figuren is zichtbaar dat bij alle toetsen geldt dat naast relatief ‘gemakkelijke’ opgaven ook opgaven van een gemiddeld moeilijkheidsniveau en moeilijkere opgaven voorkomen. Deze keuze is gemaakt om te zorgen dat van zowel goede, gemiddelde als minder vaardige leerlingen het vaardigheidsniveau goed kan worden ingeschat. Er is gezocht naar een optimale balans tussen nauwkeurig uiteenlopende vaardigheidsniveaus in beeld brengen en zorgen voor een prettige toetservaring voor leerling en leerkrachten.

3.3 Statistische beschrijving

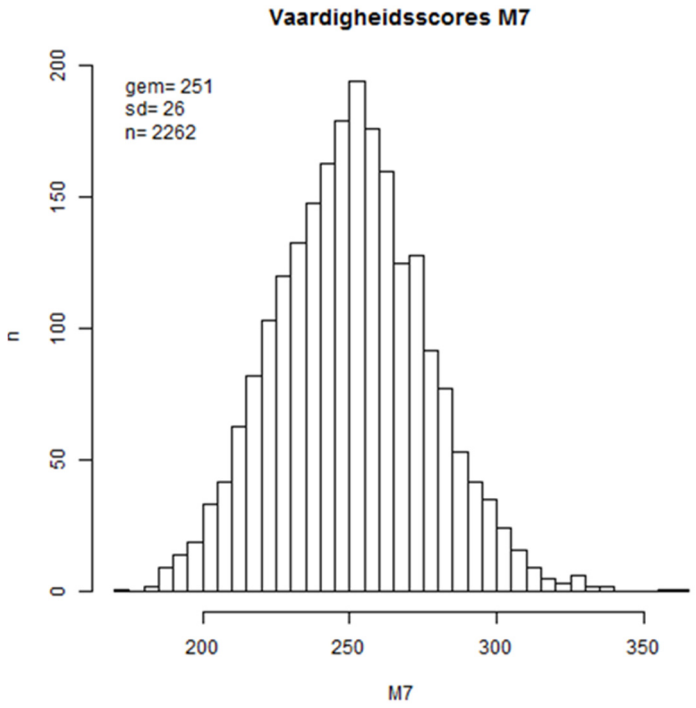
In hoofdstuk 4 zullen de kalibratie en normering uitgebreid worden beschreven. Voorafgaand aan deze uitgebreide beschrijving geven we hier een samenvattend overzicht van de beschrijvende gegevens van de M7 en E7 toetsen, zowel op de ruwe scoreschaal als op de vaardigheidsschaal.
De gegevens zijn gebaseerd op de prestaties van de leerlingen in de normeringssteekproeven: 2262 leerlingen medio groep 7 en 1175 leerlingen eind groep 7.

De waarden in tabel 3.8 en de figuren 3.1 en 3.2 laten zien dat de vaardigheidsverdelingen medio groep 7 en eind groep 7 bij benadering normaal verdeeld zijn.

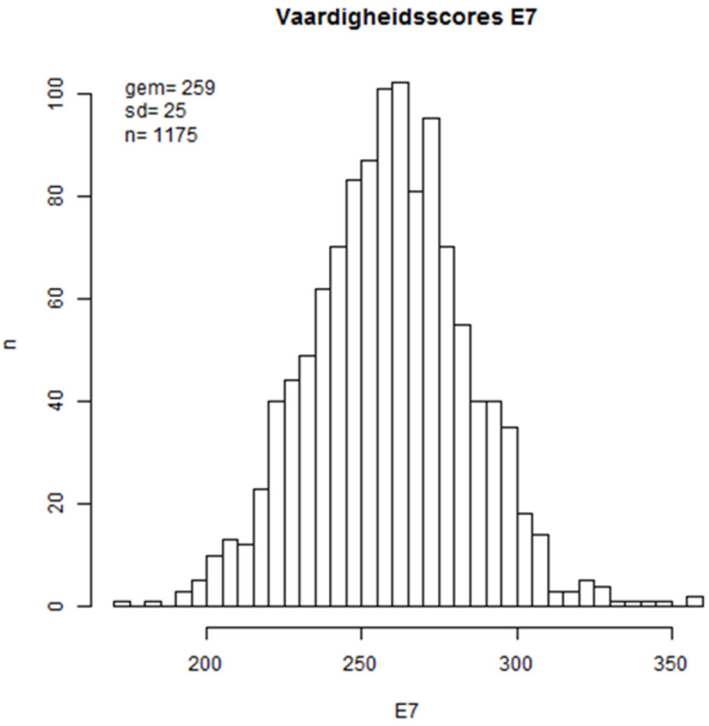
Tabel 3.8 Beschrijvende gegevens toetsen M7 en E7 op de ruwe scoreschaal

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
M7 papier	61,3	19,16	-0,46	-0,57
E7 papier	61,5	19,35	-0,48	-0,55
M7 digitaal	60,4	18,53	-0,41	-0,61
E7 digitaal	58,1	17,88	-0,29	-0,64

Figuur 3.1 Weergave van behaalde vaardigheidsscores M7



Figuur 3.2 Weergave van behaalde vaardigheidsscores E7



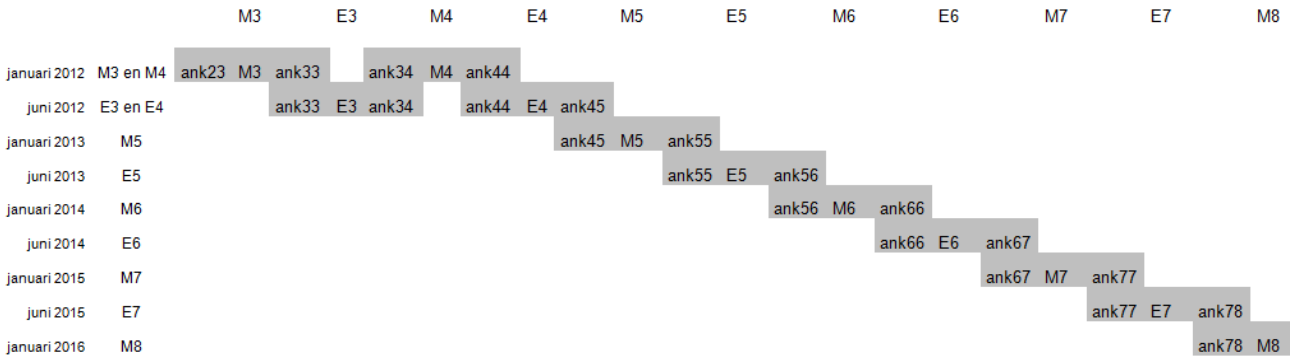
4 Kalibratie en normering

4.1 Opzet normeringsonderzoeken LVS: het macro design

Het opzetten van een leerlingvolgsysteem in het basisonderwijs is een complexe onderneming, en het verzamelen van de gegevens om het systeem te ijken en normeren moet met de nodige zorg gebeuren. Immers, het is niet voldoende om voor elke halfjaargroep (M3, E3, M4, E4, M5, E5, M6, E6, M7, E7, M8) over normen te beschikken, er moet ook voor gezorgd worden dat de prestaties over de jaren heen met elkaar vergelijkbaar zijn. Hiertoe dienen de prestaties van leerlingen over alle leerjaren heen te worden afgebeeld op een gemeenschappelijke vaardigheidsschaal.
Om zo'n gemeenschappelijke schaal te realiseren kunnen we niet volstaan met het ontwikkelen van afzonderlijke toetsen voor de meetmomenten en elke toets afzonderlijk ijken en normeren. Prestaties van bijvoorbeeld de populatie M7 moeten vergelijkbaar zijn met die van andere populaties, bijvoorbeeld E6 en E7, oftewel het dataverzamelingsdesign dient verbonden te zijn. Hiertoe dient een longitudinale opzet gebruikt te worden.

De verbondenheid van het design
Het idee van een gemeenschappelijke schaal impliceert strikt genomen dat men iemands vaardigheid zou kunnen schatten aan de hand van een willekeurig samengestelde toets. Het spreekt echter vanzelf dat het een zinloze onderneming is een toets die geconstrueerd is voor groep 7 voor te leggen aan leerlingen van groep 3, omdat zo'n toets ongetwijfeld opgaven zal bevatten die een beroep doen op kennis van leerstof die in groep 3 niet is onderwezen. Dit betekent dat we door de algemene kenmerken van het curriculum in het rekenonderwijs tamelijk beperkt zijn in het voorleggen van itemmateriaal aan leerlingen voor wie het niet specifiek is geconstrueerd. Daarom is er besloten dat het overlapmateriaal dat aan een bepaalde (half-)jaargroep kan worden voorgelegd alleen itemmateriaal mag bevatten dat specifiek voor die halfjaargroep is geconstrueerd en voor de twee belendende halfjaargroepen. Voor M7 betekent dit dat de leerlingen in het normeringsonderzoek items voorgelegd krijgen die specifiek voor M7 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor E6 en E7. Het macro-design is weergegeven in figuur 4.1.

Figuur 4.1 Het macro-design voor de normeringsafnames



De items die voor de overlap of verankering zorgen, duiden we in het macro-design aan met ank, gevolgd door 2 cijfers. Zo duidt ank34 de groep items aan die enerzijds bestaat uit items geconstrueerd voor E3 en anderzijds uit items geconstrueerd voor M4. Die items zijn dus zowel eind groep 3 als medio groep 4 afgenomen. De groep items ank33 bevat items voor M3 en E3, die dus zowel M3 als E3 zijn afgenomen. Een item kan hoogstens in één overlapgroep voorkomen, dat wil zeggen: de ank-blokjes hebben geen gemeenschappelijke items.

Longitudinale opzet
Een volledig longitudinaal design impliceert dat een cohort leerlingen gevolgd wordt van M3 tot en met M8.
Een dergelijk design heeft een aantal zwaarwegende nadelen. Het is onvermijdelijk dat er uitval plaats zal vinden. Bij een hoog percentage uitval wordt het steeds ingewikkelder, zo niet onmogelijk, om betrouwbare normen op te

stellen. Bovendien is een longitudinale studie belastend voor de deelnemende scholen en leerlingen. Dit brengt het risico mee van ongewenste en moeilijk controleerbare neveneffecten.
Daarom is ervoor gekozen het longitudinale karakter van het onderzoek in te perken, en aan de deelnemende scholen te vragen deel te nemen op maximaal drie opeenvolgende meetmomenten, waarbij het startmoment verspreid is voor verschillende scholen. Bijvoorbeeld: school A start met groep 3 op het mediomoment van schooljaar x en zal eveneens deelnemen aan de opvolgende momenten E3 (schooljaar x) en M4 (schooljaar x+1). School B zal starten op moment E3 (schooljaar x) en zal eveneens deelnemen aan de opvolgende momenten M4 (schooljaar x+1) en E4 (schooljaar x+1). Op deze manier wordt rekening gehouden met de belasting voor scholen en worden toch de benodigde longitudinale data verkregen. Aansluitend bij de verbondenheid van het design via opeenvolgende toetsmomenten en de longitudinale opzet zal de kalibratie per leerjaar worden uitgevoerd op een beperkt deel van de gemeenschappelijke schaal. De kalibratie zal plaatsvinden op basis van de verzamelde data voor dat leerjaar op de afnamemomenten, aangevuld met de gegevens van het voorgaande en het opvolgende afnamemoment. Voor groep 3 vindt de kalibratie plaats op basis van de gegevens die op de afnamemomenten M3, E3 en M4 verzameld zijn. In het geval van leerjaar 7 vindt de kalibratie plaats op basis van de verzamelde gegevens op de afnamemomenten E6, M7, E7 en M8. Dit sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven, een sterke leerling in groep 7 zal wel opgaven uit groep 8, maar een zwakke leerling zal die opgaven nog niet kunnen maken. Op deze manier kan dus beter rekening gehouden worden met de uitbreidingen in het onderwijsaanbod.
Voor kalibratie en normering van de toetsen van elke jaargroep zal op een gedeelte van het eerder vermelde design (zie figuur 4.1) worden gefocust. In het geval van groep 7 betreft het dus het gedeelte uit figuur 4.1, dat in figuur 4.2 is weergegeven.

Figuur 4.2 Design groep 7

januari 2015	M7	Ank67	M7	Ank77		
juni 2015	E7			Ank77	E7	Ank78

Opgemerkt dient te worden dat de normering onafhankelijk is van de aangeboden items, mits deze qua inhoud passen bij de jaargroep. De opzet van de kalibratie en de normering zullen in de volgende paragrafen verder worden beschreven.

Om de prestaties van leerlingen en groepen te kunnen blijven volgen, zullen deze op een overkoepelende schaal worden geplaatst door gebruik te maken van een transformatie. Deze transformatie wordt afgeleid uit de overlappende populaties op de kalibratie en normering per jaargroep. De overlappende jaargroepen op opvolgende schalen bestaan uit dezelfde leerlingen in beide kalibraties en hebben per definitie dezelfde vaardigheidsverdeling. Om deze reden kan uit de vaardigheidsverdelingen van die jaargroepen de transformatie berekend worden.

4.2 De kalibratie

4.2.1 De opzet van de kalibratie

Normeringssteekproef
Prestaties van leerlingen blijken al snel na publicatie van een toets te verschuiven, omdat bij het onderzoek dat ten grondslag ligt aan de normering sprake is van low stakes afnamesituaties (Keuning et al. 2014).
Bij de ontwikkeling van LVS-III is geprobeerd om bias in de normen te vermijden door de afnamesituatie waarin de toets wordt afgenomen zoveel mogelijk te laten lijken op de situatie na uitgave. Bij deze vorm van *embedded field* onderzoek maken leerlingen de gehele LVS-II toets en een vierde taak met LVS-III items als onderdeel van de reguliere afname. Hierdoor zijn ze gedurende de gehele toets waarschijnlijk even gemotiveerd als wanneer ze

alleen de reguliere LVS-II-toets hadden gemaakt. Een belangrijk tweede voordeel van deze aanpak is dat (zie Keuning et al., 2014) de normeringssteekproef aangevuld kan worden met resultaten uit dataretour van LVS-II. Bij de normering van Rekenen-Wiskunde 3.0 groep 7 bestond de uiteindelijke normeringssteekproef voor de helft uit resultaten van leerlingen uit het *embedded field* normeringsonderzoek en voor de andere helft uit resultaten uit dataretour. Doordat er tijdens de selectie van dataretourdata rekening gehouden is met relevante achtergrond-variabelen, werd het mogelijk om de totale normeringssteekproef representatief te maken voor deze variabelen. Bij de normering van LVS-III wordt rekening gehouden met de variabelen regio, urbanisatiegraad, schooltype, en sekse. In paragraaf 4.3.1 wordt de selectieprocedure uitgebreid toegelicht.

LVS-schaling
De LVS-schaling is tot stand gekomen op basis van de data die over de opgaven verzameld zijn bij de kalibratie- en normeringsonderzoeken en de digitale en papieren-digitale kalibratieonderzoeken.
In hoofdstuk 3 staan de tabellen met de designs voor het kalibratie- en normeringsonderzoek van papieren afnames voor medio groep 7 en eind groep 7 (tabel 3.2 en tabel 3.4). In het *embedded field* onderzoek bij de papieren afnames zijn in totaal twaalf boekjes voor medio groep 7 en twaalf boekjes voor eind groep 7 afgenomen. Naast de drie taken van de reguliere uitgave van LVS II maakte elke leerling één taak met nieuwe opgaven. Elke nieuwe taak bevatte naast nieuwe opgaven ook ankeropgaven uit het voorafgaande en het volgende afnamemoment. In bijlage 2: Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M7 en Bijlage 3: Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek E7 zijn de uitgewerkte microdesigns met de aantallen opgaven per categorie opgenomen.
Op basis van de data van deze afnames is een kalibratie uitgevoerd en een schaal geconstrueerd. Op die schaal zijn ook digitale items ondergebracht. De data voor het uitvoeren van de kalibratie-analyses zijn verzameld bij de digitale en papier-digitale onderzoeken. Zie hiervoor in hoofdstuk 3 de paragraaf over het kalibratie-onderzoek digitaal en papier-digitaal. In de volgende paragraaf wordt het kalibratieproces beschreven.

4.2.2 De stappen in de kalibratie

Met kalibratie wordt bedoeld dat we kengetallen zoeken bij de items die de antwoorden van de leerlingen goed representeren. Hoe de kengetallen gezocht worden ligt deels vast door het gekozen model (zie paragraaf 2.4.2.2). Hoe succesvol deze operatie is, kan statistisch getoetst worden. Eenvoudig gezegd, schatten we in OPLM met de CML-methode de itemparameters en controleren we of deze de data goed voorspellen. Voor een exacte beschrijving van de statistische toetsen die in OPLM gebruikt worden, hun eigenschappen en feitelijke implementatie in OPLM verwijzen we naar Verhelst (1993). Hier beperken we ons tot een korte beschrijving van de principes van de statistische toetsen die gebruikt zijn in de kalibratie-procedure.
De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen aangezien het OPLM behoort tot de exponentiële familie, met de gewogen somscore:

$$s = \sum_{i=1}^{k} a_i x_i \,, \tag{4.1}$$

als een ‘afdoende statistiek’ (*sufficient statistic*) voor de vaardigheid θ. Dit betekent dat alle informatie in de data met betrekking tot de vaardigheid in deze statistiek aanwezig is. Hiervan wordt gebruikgemaakt bij de statistische toetsen in OPLM. Het basisprincipe van de statistische toetsen in OPLM is dat op grond van de afdoende statistiek s de personen in de data kunnen worden gegroepeerd. En binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+\mid s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+\mid s)$. Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+\mid s)$ evalueren, $prop(+\mid s)$ volgt uit de data. Discrepanties tussen $p(+\mid s)$ en $prop(+\mid s)$ duiden op schendingen van het model. Deze discrepanties vormen de basis voor de diverse statistische toetsen in OPLM. De toetsingsgrootheid voor de veronderstelde discriminatie-indices is gegeven door

$$M = f_{s\in H}(p(+\mid s) - prop(+\mid s)) + f_{s\in L}(prop(+\mid s) - p(+\mid s)). \tag{4.2}$$

Deze zogeheten M-toetsen verdelen de scoregroepen in een laag deel (L) en een hoog deel (H) en f is een monotone functie. M-toetsen hebben een duidelijke interpretatie: is M significant positief dan is de veronderstelde steilheid van de ICC (item karakteristieke curve) overschat in het model, is M daarentegen erg laag dan is de index te klein. Verhelst laat zien voor welke functie, f , $M \approx N(0,1)$. In OPLM zijn drie verschillende M-toetsen geïmplementeerd die verschillen in de definitie van de hoge en lage scoregroepen. Naast deze M-toetsen is er een algemene itemtoets die de volgende vorm heeft

$$S = f(p(+ \mid s) - prop(+ \mid s)) .$$

Deze zogeheten S-toets heeft een χ^2 verdeling onder het model. Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie van de rechteroverschrijdingskansen van alle afzonderlijke S-toetsen komt hiervoor in aanmerking. Als we deze S-toetsen opvatten als onafhankelijk, wat ze strikt genomen niet zijn, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn op het (0,1) interval.

Kortom, als we afzien van de formeel-statistische achtergrond van de gehanteerde toetsen, kan de kalibratieprocedure als volgt worden samengevat:

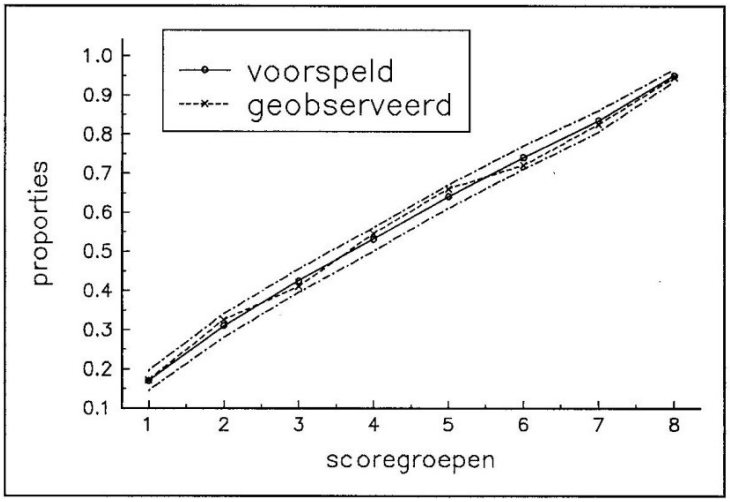
1. Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2. Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3. Met behulp van de M-toetsen controleren we of de discriminatie-indices goed zijn ingesteld.
4. Een volgende controle betreft de overschrijdingskansen van de S-toetsen en een grafische modelcontrole door middel van het programma WOPPLOT (grafische inspectie van de ICC's).
5. Vervolgens vindt een globale modelcontrole plaats in de vorm van een R1c-toets en de verdeling van de overschrijdingskansen van de S-toetsen.

De stappen 1 tot en met 5 worden een aantal malen doorlopen tot het resultaat bevredigend is. Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen spelen een rol in dit beslissingsproces (zie hiervoor hoofdstuk 2 over de achtergronden van de toetsinhoud).

4.2.3 Toetsing van het IRT-model

Het is niet eenvoudig om de kwaliteit van de kalibratie aan te tonen. De belangrijkste statistische instrumenten om de passing van een opgave in het IRT-model te bewerkstellingen en uiteindelijk te documenteren betreffen de hierboven al besproken S-toetsen. Het lastige daarvan is, dat de toetsing voor een groot deel visueel gebeurt. Dit kunnen we illustreren aan de hand van figuur 4.3 (zie Staphorsius, 1994, blz. 239). Figuur 4.3 beeldt voor een opgave de gegevens af waarop de betreffende S-toetsen gebaseerd zijn (zie handleiding OPLM: Verhelst; 1992). Ten behoeve van deze toetsing wordt de totale groep van leerlingen die een verzameling opgaven gemaakt heeft, ingedeeld in een aantal scoregroepen (meestal acht, maar minder als de variatie in scores kleiner is). Elke groep bestaat uit leerlingen met een ongeveer even hoge score. De geobserveerde proporties juiste antwoorden van deze groepen (telkens gesymboliseerd door een x) zijn door de middelste stippellijn verbonden. De volle lijn daarentegen verbindt de proporties die op grond van de parameterschattingen voorspeld kunnen worden. De twee buitenste lijnen geven het 95%-betrouwbaarheidsinterval aan. De breedte van dit interval is in belangrijke mate afhankelijk van het aantal leerlingen dat de opgave heeft beantwoord. Uit de figuur blijkt heel duidelijk dat de geobserveerde proporties, zoals bedoeld, binnen het 95%-betrouwbaarheidsinterval van de (geschatte) voorspelde proporties liggen, en dit komt in grote lijnen overeen met een niet-significante S-toetsings- grootheid (Verhelst et al., 1994).

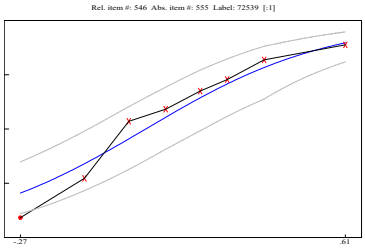
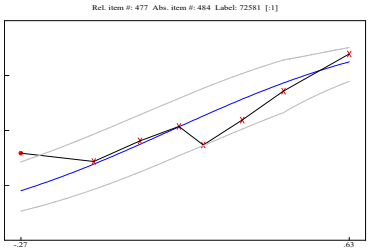
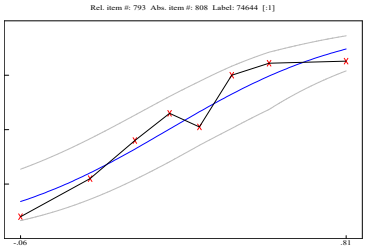
Figuur 4.3 Grafische voorstelling van een S_r-toets



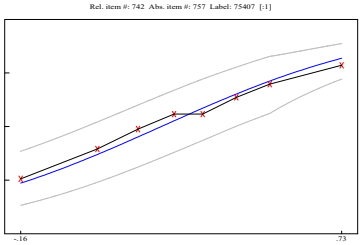
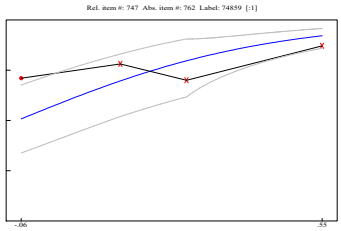
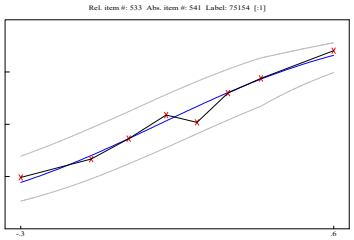
Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.4 illustreren dat voor de toetsen M7 en E7 zelfs bij de minst passende opgave sprake is van een zeer aanvaardbaar beeld. Er wordt in dit geval voor een deel (van de onderscheiden scoregroepen) niet beantwoord aan de eis dat de geobserveerde proportie binnen het 95%-betrouwbaarheidsinterval van de geschatte proporties ligt. Dit beeld doet zich slechts bij enkele opgaven voor die dan ook een uitzondering vormen. De overige opgaven voldoen voor alle scoregroepen wel aan die eis. De afbeeldingen voor de representatieve en best passende opgaven illustreren dit. Dit leidt tot de conclusie dat bij vrijwel alle opgaven in de toetsen Rekenen-Wiskunde 3.0 een grafische voorstelling van de S-toetsing hoort die in grote lijnen met figuur 4.3 overeenkomt. Dit is, zeker gezien de relatief grote aantallen observaties die in dit onderzoek gedaan zijn, een zeer sterke aanwijzing dat het meetinstrument en het meetmodel dat ontwikkeld is, respectievelijk gebruikt is, adequaat zijn om het gedrag van de leerlingen te verklaren. Bovendien blijkt, en dat is vanuit theoretisch oogpunt nog belangrijker, dat gemeten verschillen in gedrag tussen de leerlingen te verklaren zijn door één unidimensionaal concept.

Figuur 4.4 Voorbeelden van S-toetsen voor de toetsen Rekenen-Wiskunde M7 en E7 met de best passende, de slechtst passende en een qua passing representatieve opgave

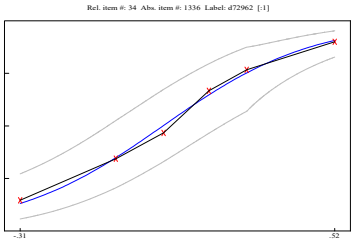
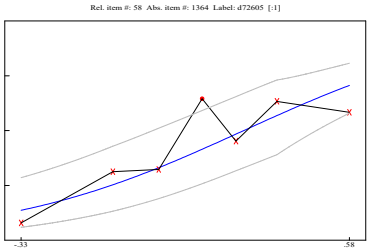
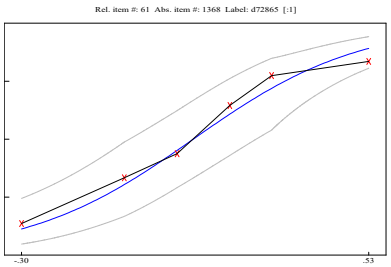
M7

		
Best passend	Slechtst passend	Representatieve passing

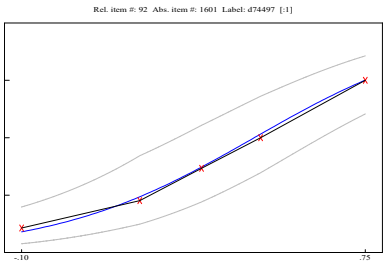
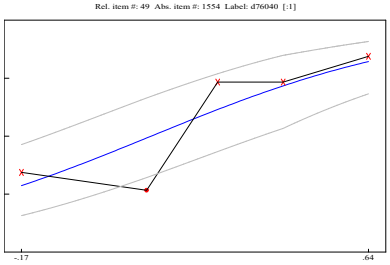
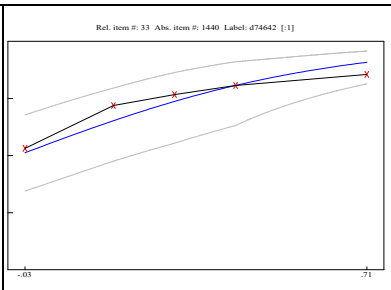
E7

		
Best passend	Slechtst passend	Representatieve passing

M7 digitaal

		
Best passend	Slechtst passend	Representatieve passing

E7 digitaal

		
Best passend	Slechtst passend	Representatieve passing

In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Niettemin zijn er bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij zijn we vooral geïnteresseerd in de distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Als we de S-toetsen opvatten als onafhankelijk, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.1 waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten bij een kalibratie van alle opgaven van de toets LVS-III Rekenen-Wiskunde voor groep 7. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor de toets de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Hierbij valt op te merken dat het aantal opgaven waarbij de S-toets significant was op het niveau lag dat te verwachten valt onder een nul-model, zoals bij een significantie niveau van 5% te verwachten valt dat 5% van de resultaten significant is, zonder dat dit betekenis heeft. De resultaten zoals die hier gevonden worden passen in dat beeld. Al met al vormen zij een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren.

Tabel 4.1 Verdeling van overschrijdingskansen bij S-toetsen voor de toetsen M7 en E7

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M7	1	4	5	10	11	14	7	8	6	9	11	10

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
E7	1	2	4	11	10	9	12	7	14	6	7	13

M7	**0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.**											
digitaal	0	0	1	3	8	3	10	9	11	12	13	26

E7	**0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.**											
digitaal	0	0	0	4	2	9	5	13	8	14	11	30

In tabel 4.2 zijn de R1c-waarden weergegeven voor dezelfde afnames waarvoor in tabel 4.1 de resultaten van de S-toetsen zijn weergegeven. R1c is een statistiek die zicht geeft op de modelpassing van de toets als geheel. Voor een acceptabele modelfit geldt als vuistregel dat R1c bij voorkeur niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df) is.
De modelpassing van de toetsen voldoen aan deze voorwaarden. Voor alle toetsen M7 en E7 geldt zowel voor de papieren versie als de digitale versie dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant bij een aantal toetsen. Aan dit laatste moet bij steekproeven met een dergelijke omvang niet te veel waarde worden gehecht.

Tabel 4.2 R1c-waarden voor M7–E7

Toetsversie	**R1c**	**df**	**p**
M7-papier	911,8	724	<0,005
E7-papier	675,3	530	<0,005
M7-digi	547,0	605	0,955
E7-digi	443,9	498	0,9597

Ten slotte bespreken we nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer & Sijtsma, 2010, p. 40). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN-Beoordelingssysteem met 'c' aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratie-populatie. Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze 'c'.
Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd.
In tabel 4.3 zijn gemiddelde en range van deze waarden voor alle toetsitems weergegeven. De gemiddelde waarde van de constante is uitstekend te noemen. De gemiddelde c-waarde ligt voor de papieren toetsen ruim onder de 0,20. Slechts voor 12 opgaven van de 96 gekalibreerde opgaven die in de uitgave M7 zitten en voor 10 opgaven die in de E7 uitgave zitten is de waarde hoger dan 0,20. Voor E7 geldt daarbij dat één waarde groter is dan 0,30. Voor de digitale toetsen geldt ook dat deze c-waarden gemiddeld onder de 0,20 liggen. Voor 14 opgaven bij M7 digitaal van de 96 opgaven van de uitgave is de waarde hoger dan 0,20. Voor E7 digitaal geldt dat 26 opgaven van de 96 opgaven in de uitgave een c- waarde hebben hoger dan 0,20, waarvan er twee groter dan 0,30 zijn.
Over het algemeen kan de nauwkeurigheid van de schattingen als goed beoordeeld worden. Voor een beperkt aantal opgaven is de c-waarde hoger dan 0,20, maar ook deze waarden kunnen nog als voldoende beoordeeld worden.

Tabel 4.3 Nauwkeurigheid van de itemparameterschattingen (constante 'c')

Toetsversie	Constante 'c'	
	Range	Gemiddelde
M7-papier	0,062-0,218	0,113
E7-papier	0,076-0,364	0,146
M7-digi	0,087-0,218	0,155
E7-digi	0,102-0,318	0,191

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen LVS-III Rekenen-Wiskunde groep 7 de kalibratie geslaagd is. Hiermee is het laatste woord nog niet gezegd over de validiteit, maar het kalibratieonderzoek brengt in ieder geval een essentieel aspect van het validiteitsvraagstuk naar voren: de rechtvaardiging van wat in de meeste toetstoepassingen gebruikelijk is, namelijk het reduceren van alles wat de leerling heeft geantwoord tot een enkele toetsscore (of afgeleid daarvan, een enkele schatting van zijn onderliggende vaardigheid). De kalibratie-analyse, als puur formeel proces, kan geen uitspraken doen over de inhoudsvaliditeit of over de constructvaliditeit als antwoord op de vraag: hoe kan worden aangetoond dat het concept dat de items in de bank meten, dekkend is voor en samenvalt met het construct dat we in de toetsen LVS Rekenen-Wiskunde 3.0 proberen te meten (zoals dat in het didactisch en het wetenschappelijk forum wordt bedoeld)? In hoofdstuk 6 over validiteit zal worden nagegaan of de gemeten concepten inderdaad overeenkomen met het begrip zoals bedoeld. De vraag is dan in het geval van het onderdeel Rekenen-Wiskunde: kan het unidimensionale concept onder de opgaven in de opgavenbank Rekenen-Wiskunde inderdaad worden opgevat als de vaardigheid 'Rekenen- Wiskunde'? Een geslaagde kalibratie op een unidimensionaal construct beschouwen we als een noodzakelijke voorwaarde voor deze begripsvaliditeit.

4.3 De normering

Sinds schooljaar 2013/2014 wordt door Cito een nieuwe werkwijze voor het normeren van leerlingvolgsysteem-toetsen gevolgd. Deze werkwijze wordt gebruikt bij het monitoren van de normering van eerder uitgegeven toetsen, maar wordt ook gebruikt bij de normering van de nieuw uit te geven toetsen, zo ook LVS-III Rekenen-Wiskunde. De werkwijze die we hieronder beschrijven, komt uit Keuning et al. (2014). Allereerst besteden we aandacht aan de opzet van het normeringsonderzoek, de gehanteerde procedures en de aantallen leerlingen per afnamemoment (paragraaf 4.3.1). Vervolgens komt in paragraaf 4.3.2 de representativiteit van de normsteek-proeven aan de orde. De paragraaf wordt afgerond met een presentatie van de resultaten van de normering (i.e. de kenmerken van de vaardigheidsverdelingen op de onderscheiden afnamemomenten; paragraaf 4.3.3).

4.3.1 Opzet

Tijdens het *embedded field* normeringsonderzoek zoals omschreven in paragraaf 4.2.1 worden data verzameld. Voor het embedded field normeringsonderzoek is een representatieve steekproef getrokken uit de verzameling van alle basisscholen in Nederland. Dit is gedaan vanuit het bij Cito gebruikelijke steekproefkader dat bepaald wordt door regio, urbanisatiegraad en schooltype (zie verderop voor een omschrijving van deze achtergrond-variabelen). De dekkingsgraad van de LVS toetsen Rekenen-Wiskunde is bijzonder hoog: de toetsen worden door 90% tot 95% van de scholen toegepast. Voor de deelnemende scholen aan het normeringsonderzoek is nagegaan of zij als groep afweken van wat men voor de totale populatie van scholen zou mogen verwachten.

Vanzelfsprekend worden de data die via Cito dataretour binnenkomen opgeschoond voordat ze gebruikt worden. Uit de bestanden worden de volgende categorieën leerlingen verwijderd:
- Leerlingen uit het speciaal onderwijs en leerlingen voor wie het onderwijstype onbekend is.
- Leerlingen van scholen die het LVS selectief inzetten. In de hogere leerjaren blijken sommige scholen het LVS namelijk alleen in te zetten bij zwakkere leerlingen (zie Keuning, 2011).
- Leerlingen die op hetzelfde afnamemoment meerdere toetsen van dezelfde vaardigheid maken. Alleen de gegevens van de toets die bij het afnamemoment hoort, worden behouden. Daarnaast worden de scholen verwijderd die ook aan de *embedded field* normeringsonderzoeken deelnemen.

Er is voor gekozen om alleen data te selecteren van het schooljaar waarin ook het normeringsonderzoek heeft plaatsgevonden. Er wordt naar gestreefd om de uiteindelijke normeringssteekproef voor ongeveer 50 procent te baseren op gegevens uit het *embedded field* normeringsonderzoek en voor 50 procent op gegevens uit Cito dataretour. De streefverhouding kan desgewenst ook anders gekozen worden, maar het ligt niet voor de hand om het aandeel van het ene gegevensbestand veel groter te maken dan het aandeel van het andere gegevensbestand. Door Cito dataretour een groter gewicht te geven, neemt het percentage leerlingen dat de nieuwe LVS-III toetsen maakt namelijk verhoudingsgewijs af. Met het oog op de constructie en validering van LVS-III is dit onwenselijk. Door het *embedded field* normeringsonderzoek een groter gewicht te geven, neemt de hoeveelheid data af die volledig in de feitelijke toetssituatie verzameld zijn. Dit is een gemiste kans. Juist het combineren van het *embedded field* normeringsonderzoek met Cito dataretour biedt grote voordelen ten opzichte van alternatieve onderzoeksdesigns. Enerzijds wordt er op deze manier voor gezorgd dat de toetsresultaten die gebruikt worden bij het bepalen van de normen zoveel mogelijk in de feitelijke toetssituatie verzameld zijn. Anderzijds is het mogelijk om via Cito dataretour de “kwaliteit” van het *embedded field* normeringsonderzoek te checken. Een belangrijke randvoorwaarde is wel dat de uiteindelijke normeringsteekproef representatief is voor de landelijke populatie van scholen en leerlingen. Representativiteit van de normeringssteekproef zoals die samengesteld wordt op basis van het *embedded field* normeringsonderzoek (± 50 procent) en Cito dataretour (± 50 procent) is te realiseren door bij de selectie van data uit Cito dataretour rekening te houden met relevante achtergrondvariabelen. Bij de normering van LVS-III wordt rekening gehouden met de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse*. De verschillende variabelen zijn als volgt gedefinieerd:

- ***Regio***. Bij de definitie van de variabele *regio* is uitgegaan van de CBS-indeling naar landsdeel. Dit betekent dat er vier regio’s onderscheiden zijn. Regio *noord* omvat de provincies Groningen, Friesland en Drenthe; regio *oost* de provincies Overijssel, Gelderland en Flevoland; regio *west* de provincies Utrecht, Noord-Holland, Zuid-Holland en Zeeland en regio *zuid* de provincies Noord-Brabant en Limburg.

- ***Urbanisatiegraad***. Bij de definitie van de variabele *urbanisatiegraad* is ervoor 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 & Hemker, 2009).

- ***Schooltype***. Bij de definitie van de variabele *schooltype* is gebruikgemaakt van de formatiegewichten van de leerlingen binnen een school volgens de meest recente regeling van OCW. Daarin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders:
  - 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 heeft of hebben 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: 1 = maximaal basisonderwijs of (V)SO-ZMLK, 2 = maximaal LBO/VBO, praktijkonderwijs of VMBO basis- of kaderberoepsgerichte leerweg, en 3 = overig VO en hoger. Leerlingen met een formatiegewicht van 0.3 of 1.2 zijn te definiëren als achterstandsleerlingen. Scholen zijn ingedeeld naar het percentage achterstands-leerlingen volgens een indeling in vier typen: (1) percentage achterstandsleerlingen [0, .10] (2) percentage achterstandsleerlingen [.10, .25] (3) percentage achterstandsleerlingen [.25, .40] (4) en percentage achterstandsleerlingen [.40, 1].

– ***Sekse***. Bij de variabele *sekse* is een tweedeling naar jongens en meisjes gehanteerd.

Het is niet mogelijk om expliciet rekening te houden met de variabele *etniciteit*, omdat (a) er geen eenduidige referentiegegevens voor de populatie bekend zijn, en (b) Cito dataretour weinig tot geen informatie bevat over de etnische herkomst van leerlingen. Onderzoek heeft echter laten zien dat de verdeling naar etnische herkomst sterk samenhangt met de verdeling naar urbanisatiegraad en schooltype (Hemker, Kordes en Van Weerden, 2011). Om deze reden is aangenomen dat de uiteindelijke normeringsteekproef voldoende representatief is naar etnische herkomst als de verdeling naar urbanisatiegraad en schooltype overeenkomt met de verdeling in de landelijke populatie.

Bij het selecteren van data uit Cito dataretour wordt rekening gehouden met vier achtergrondvariabelen die samen 4 × 2 × 4 × 2 = 64 verschillende categorieën representeren. De variabelen *regio*, *urbanisatiegraad* en *schooltype* zijn op het niveau van de school gedefinieerd. De variabele *sekse* is op het niveau van de leerling gedefinieerd. Het is niet goed mogelijk om bij het selecteren van data tegelijkertijd rekening te houden met school- én leerlingvariabelen. Daarom vindt de dataselectie in twee stappen plaats. In de eerste stap worden iteratief scholen uit Cito dataretour toegevoegd aan de dataset met normeringsgegevens. Niet elke school heeft daarbij evenveel kans om geselecteerd te worden. Bij de selectie wordt namelijk rekening gehouden met de regio en de urbanisatiegraad van de school en het aantal achterstandsleerlingen. De kans w_{ijk} dat een school met regio i, urbanisatiegraad j en schooltype k geselecteerd wordt, hangt af van het reeds geselecteerde aantal leerlingen N_{S}, het gewenste aantal leerlingen N_{T}, en het beschikbare aantal leerlingen in Cito dataretour N_{D}:

$$w_{ijk} = \frac{(n_{\mathrm{T},\,ijk} - n_{\mathrm{S},\,ijk}) \div (N_{\mathrm{T}} - N_{\mathrm{S}})}{n_{\mathrm{D},\,ijk} \div N_{\mathrm{D}}} = \frac{N_{\mathrm{D}}(n_{\mathrm{T},\,ijk} - n_{\mathrm{S},\,ijk})}{n_{\mathrm{D},\,ijk}(N_{\mathrm{T}} - N_{\mathrm{S}})}, \tag{4.3}$$

waarbij vereist is dat $n_{\mathrm{S},\,ijk}$ $n_{\mathrm{T},\,ijk}$. Zoals we kunnen zien, wordt het percentage leerlingen dat (nog) gewenst is voor een bepaalde categorie (in dit geval de populatie) gedeeld door het percentage leerlingen dat via Cito dataretour beschikbaar is voor opname in die categorie (in dit geval de steekproef).

In geval $n_{\mathrm{S},\,ijk} > n_{\mathrm{T},\,ijk}$ is de kans w_{ijk} die uit de formule volgt negatief en niet toe te passen. Dat kan in twee situaties gebeuren. Ten eerste kan een bepaalde categorie in het licht van de gekozen N_{T} en de via de landelijke gegevens van DUO en/of CBS te bepalen $n_{\mathrm{T},\,ijk}$ oververtegenwoordigd zijn in de dataset met normeringsgegevens. In dat geval kan het selectiealgoritme niet gestart worden. De oplossing is om enkele scholen te verwijderen totdat voor alle categorieën geldt dat $n_{\mathrm{S},\,ijk} \leq n_{\mathrm{T},\,ijk}$. Ten tweede kan tijdens de selectie blijken dat een categorie oververtegenwoordigd raakt als we een bepaalde school vanuit Cito dataretour toevoegen aan de dataset met normeringsgegevens. Dit risico wordt groter naarmate het reeds geselecteerde aantal leerlingen N_{S} dichterbij het gewenste aantal leerlingen N_{T} komt te liggen.
De oplossing is om N_{T} bij de berekening van de gewichten te vermenigvuldigen met een vrij te kiezen constante C en het algoritme te beëindigen in de eerste iteratie waarbij geldt dat $N_{\mathrm{S}} \geq N_{\mathrm{T}}$. Als constante C groot gekozen wordt, heeft het selectiealgoritme veel ruimte om scholen te kiezen. Het voordeel is dat het selectiealgoritme snel voorziet in een oplossing. Het nadeel is dat de verdeling naar *regio*, *urbanisatiegraad* en *schooltype* zoals we die na toepassing van het selectiealgoritme observeren in de normeringssteekproef nogal kan afwijken van de verdeling zoals we die wensen op basis van de landelijke gegevens van DUO en/of CBS. Als constante C klein gekozen wordt, zal het selectiealgoritme minder snel een oplossing vinden. Het eindresultaat zal doorgaans wel een grotere gelijkenis vertonen met de landelijke gegevens van DUO en/of CBS.

Tot nu toe is bij de selectie van data uitsluitend rekening gehouden met de schoolvariabelen *regio*, *urbanisatiegraad* en *schooltype.* De leerlingvariabele *sekse* is nog niet in beschouwing genomen.
Dat gebeurt in de tweede stap. Als blijkt dat de normeringssteekproef die is samengesteld in de eerste stap niet representatief is met betrekking tot de variabele *sekse*, dan wordt een tweede steekproeftrekking uitgevoerd. Eerst wordt op basis van de landelijke gegevens van CBS en de geobserveerde aantallen in de normerings-steekproef de kans w_q bepaald dat een leerling met sekse *q* in een representatieve normeringssteekproef zit:

$$w_q = \frac{n_{\mathrm{T},q} \div N_{\mathrm{T}}}{n_{\mathrm{S},q} \div N_{\mathrm{S}}} = \frac{n_{\mathrm{T},q} N_{\mathrm{S}}}{N_{\mathrm{T}} n_{\mathrm{S},q}}. \tag{4.4}$$

Zoals we kunnen zien, wordt het gewenste percentage leerlingen in categorie q gedeeld door het geobserveerde percentage leerlingen in categorie q. Als w_q voor alle leerlingen in de normeringssteekproef bepaald is, wordt binnen elke school een steekproef met teruglegging getrokken.
Bij het trekken van de steekproef wordt rekening gehouden met w_q. De trekking wordt beëindigd op het moment dat het geselecteerde leerlingaantal gelijk is aan het oorspronkelijke leerlingaantal. De steekproeftrekking wordt per school uitgevoerd, omdat het met het oog op de schoolnormering noodzakelijk is dat de scholen qua omvang en samenstelling zoveel mogelijk intact blijven (zie paragraaf 3.6). Dit is ook de reden dat in de eerste stap uitsluitend gehele scholen geselecteerd worden en geen individuele leerlingen.
Het algoritme is toegepast bij de ontwikkeling van Rekenen-Wiskunde 3.0. Samenvattend gaat het bij het algoritme om het genereren van een representatieve normeringssteekproef op basis van het normerings-onderzoek en Cito dataretour. Het uitgangspunt was om de data die tijdens het *embedded field* normerings-onderzoek verzameld zijn te verdubbelen met behulp van data uit Cito dataretour. Het gewenste aantal leerlingen werd dus voor afnamemoment medio groep 7 ingesteld op $N_{\mathrm{T}} = 2 \times 2262 = 4524$.
In tabel 4.4 is te zien in welke aantallen scholen en leerlingen het selectiealgoritme heeft geresulteerd.
De conclusie is dat het voor afnamemoment medio groep 7 tot de gewenste oplossing heeft geleid. De aantallen leerlingen die via het *embedded field* normeringsonderzoek en uit dataretour bij de normering zijn betrokken wijken weliswaar enigszins af van de nagestreefde 50:50 verhouding, maar dit is een gevolg van het exacte verloop van het algoritme gegeven de verdeling van scholen over de categorieën in de achtergrondvariabelen. Lichte afwijkingen zijn daarbij te verwachten. Dat geldt ook voor de eventuele afwijkingen in de steekproef van de populatieverdelingen voor de variabelen *regio*, *urbanisatiegraad*, *schooltype* en *geslacht*. Ook in een volledig aselecte steekproef zijn dit soort afwijkingen immers per definitie toe te schrijven aan toeval. Niettemin is in een vervolgstap de landelijke representativiteit van de normerings-steekproef ter controle onderzocht. Deze controleanalyses worden gerapporteerd in paragraaf 4.3.2.
In tabel 4.4 zijn ook de aantallen scholen en leerlingen vermeld voor het normerings-onderzoek voor afname-moment E7. De aantallen voor het normeringsonderzoek betreffen uitsluitend nieuw verzamelde data en geen gegevens uit dataretour.

Zowel voor afnamemoment medio groep 7 als eind groep 7 heeft het selectiealgoritme tot de gewenste oplossing geleid. Wel valt op dat het selectiealgoritme relatief veel scholen ongeschikt verklaart. Dat komt doordat een erg kleine waarde voor constante *C* is gekozen. Het gevolg is dat het selectiealgoritme weinig ruimte heeft gekregen om af te wijken van de gewenste aantallen in elke categorie. Vooral in de laatste iteraties waarin het geobserveerde leerlingaantal al dicht bij het gewenste leerlingaantal ligt, kan toevoeging van een school leiden tot een oververtegenwoordiging van bepaalde categorieën. In beginsel is het geen probleem dat de selectie van relatief veel scholen na de berekening van w_{ijk} ongedaan gemaakt wordt. Wel is het de vraag in hoeverre het zinvol is om te streven naar steekproeven die volledig representatief zijn voor de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse*. Ook in aselecte steekproeven kan de verdeling van leerlingen over de verschillende categorieën immers afwijken van de verdeling in de populatie. In een aselecte steekproef is deze afwijking per definitie het gevolg van toeval. Statistische weging is in een aselecte steekproef dan ook niet op zijn plaats. Door bij de normering van LVS-III de benodigde data representatief te trekken, zijn de afwijkingen die we vinden in relatie tot de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse* in zekere zin ook toe te schrijven aan toeval. Afwijkingen tussen de steekproef en de populatie kunnen in dat geval verdedigbaar zijn. Niettemin wordt in een vervolgstap de landelijke representativiteit van de normeringssteekproef ter controle onderzocht.
In tabel 4.4 wordt weergegeven welke aantallen van de steekproef en van dataretour uiteindelijk zijn meegenomen in de normering.

Tabel 4.4 Aantal leerlingen en scholen per afnamemoment die meegenomen zijn in de normering

Afnamemoment	Aantal leerlingen			Aantal scholen
	Steekproef	Dataretour	Normering	Normering
M7	2262*	2248	4510	181
E7	1175*	1175	2350	123

4.3.2 Representativiteit

Door de werkwijze die wordt gevolgd tijdens de normering is representativiteit van de normeringssteekproeven in principe gegarandeerd. Niettemin wordt er een controle uitgevoerd op de representativiteit door de populatieverdelingen te vergelijken met de steekproefverdelingen. In tabel 4.5 tot en met 4.8 worden de resultaten van de representativiteitsanalyses getoond. De steekproef is geanalyseerd in relatie tot de variabelen regio, urbanisatiegraad, schooltype en sekse.

$$phi = \sqrt{\frac{X^2}{N}} \tag{4.5}$$

Tabel 4.5 Aantal en percentage leerlingen in de populatie en de steekproef naar schooltype

	Populatie	Steekproef			
Stratum	**%**	**M7**	**%**	**E7**	**%**
0-10%	67,3	3095	68,6	1613	68,6
10-25%	22,2	1001	22,2	517	22,0
25-40%	6,6	264	5,9	144	6,1
>40%	3,8	150	3,3	76	3,2

M7 χ^2 (3, N = 4510) = 8,260; p = 0,041; phi = 0,043

E7 χ^2 (3, N = 2350) = 3,743; p = 0,291; phi = 0,040

Tabel 4.6 Aantal en percentage leerlingen in de populatie en de steekproef naar regio

	Populatie	Steekproef			
Regio	**%**	**M7**	**%**	**E7**	**%**
Noord	10,0	445	9,9	225	9,6
Oost	22,4	999	22,2	533	22,7
West	47,9	2207	48,9	1154	49,1
Zuid	19,7	859	19,1	428	18,6

M7 χ^2 (3, N = 4510) = 2,094; p = 0,553; phi = 0,022

E7 χ^2 (3, N = 2350) = 2,478; p = 0,479; phi = 0,032

Tabel 4.7 Aantal en percentage leerlingen in de populatie en de steekproef naar urbanisatiegraad

	Populatie	Steekproef			
Urbanisatie	**%**	**M7**	**%**	**E7**	**%**
Platteland	54,8	2487	55,1	1308	55,7
Stad	45,2	2023	44,9	1042	44,3

M7 χ^2 (1, N = 4510) = 0,272; p = 0,602; phi = 0,008

E7 χ^2 (1, N = 2350) = 0,771; p = 0,380; phi = 0,018

Tabel 4.8 Aantal en percentage leerlingen in de populatie en de steekproef naar geslacht

	Populatie	Steekproef			
Geslacht	**%**	**M7**	**%**	**E7**	**%**
jongen	50,5	2180	49,5	1201	51,5
meisje	49,5	2224	50,5	1130	48,5

M7 χ^2 (1, N = 4510) = 1,702; p = 0,192; phi = 0,020
E7 χ^2 (1, N = 2331) = 1,008; p = 0,315; phi = 0,021

De χ^2 -waarden zijn laag en in een enkel geval significant. Bij grotere steekproeven zegt significantie echter niet zoveel. Het is beter om de effectgrootte phi als uitgangspunt te nemen. We zien dat de effectgroottes ver onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). De conclusie is daarom dat de normeringssteekproeven een zeer goede afspiegeling vormen van de populatie.

4.3.3 Normeringsresultaten

Na de normeringssteekproef te hebben samengesteld, konden de normen worden bepaald. Naast het gemiddelde werden de percentielen berekend. Dat gebeurde op basis van de verdeling van scores in de normeringssteekproef zoals die is samengesteld op basis van het *embedded field* normeringsonderzoek en Cito dataretour. Om de scores die leerlingen behalen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het *embedded field* normeringsonderzoek en Cito dataretour worden zogeheten *plausible values* gegenereerd op de nieuw ontwikkelde vaardigheidsschaal. Deze *plausible values* representeren de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven de scores. De *plausible values* geven niet alleen informatie over de geschatte vaardigheid maar ook over de onzekerheid die bij die schatting hoort (Keuning et al.,2014). De normering wordt vervolgens gebaseerd op de *plausible values* van de leerlingen in de normeringssteekproef.
De *plausible values* voor dit afnamemoment vormen bij benadering een normale verdeling. Op basis van deze scoreverdeling worden de percentielen berekend die horen bij de vaardigheidsindelingen A tot en met E en I tot en met V zoals beschreven in paragraaf 3.1.
Tabel 4.9 geeft de normgegevens voor LVS-III Rekenen-Wiskunde M7 en E7.

Tabel 4.9 Normtabel op leerlingniveau voor LVS-III Rekenen-Wiskunde M7-E7

Tijdstip	***M***	***SD***	***Kurt.***	***Skew.***	***P*10**	***P*20**	***P*25**	***P*40**	***P*50**	***P*60**	***P*75**	***P*80**	***P*90**
M7	251,9	24,6	-0,015	0,081	220,3	231,3	235,6	245,82	251,8	257,7	267,9	272,2	284,1
E7	260,3	23,4	0,127	042	230,5	240,7	245,0	254,3	260,3	266,2	275,6	279,0	290,1

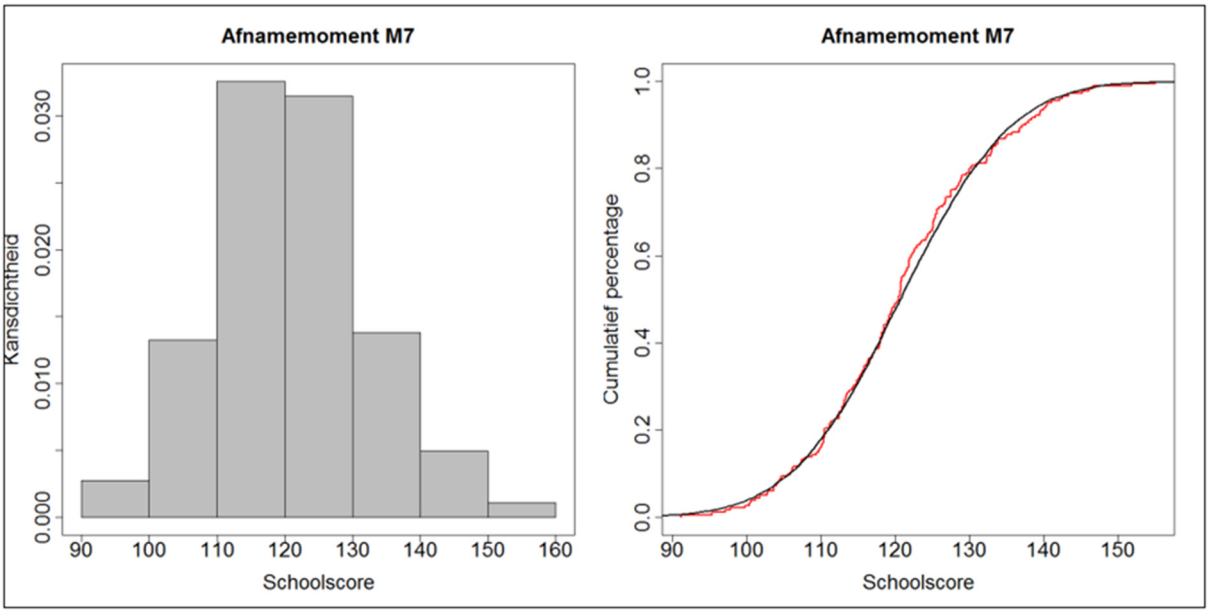
Naast een normering op leerlingniveau kent Cito ook een normering op schoolniveau. Om de schoolverdeling te bepalen wordt het intercept-only multilevel model gebruikt met een gemiddelde per school en een variantie op school- en leerlingniveau. De schatting van het model verloopt via een bootstrapprocedure. Dit betekent dat het multilevel model meerdere keren wordt geschat, steeds op basis van een andere selectie van scholen en leerlingen uit de normeringssteekproef. Bij elke replicatie wordt het aantal te selecteren scholen gelijkgesteld aan het aantal scholen dat in de normeringssteekproef zit. Vervolgens worden binnen een school leerlingen geselecteerd. Ook dit aantal wordt gelijkgesteld aan het aantal leerlingen dat feitelijk op de betreffende school zit. De scholen en leerlingen worden geselecteerd met teruglegging. Als de selectie is afgerond, wordt het multilevel model geschat en de intraklassecorrelatie en het design effect uitgerekend. Tabel 4.10 laat de samenvatting van de resultaten van de bootstapprocedure zien. De uitkomsten zijn behoorlijk stabiel. De intraklassecorrelatie (ICC) ligt boven de .04, wat inhoudt dat een multilevelanalyse zinvol is (Snijders & Bosker, 1999).

Tabel 4.10 Samenvatting uitkomsten multilevel analyse LVS-III Rekenen-Wiskunde M7-E7

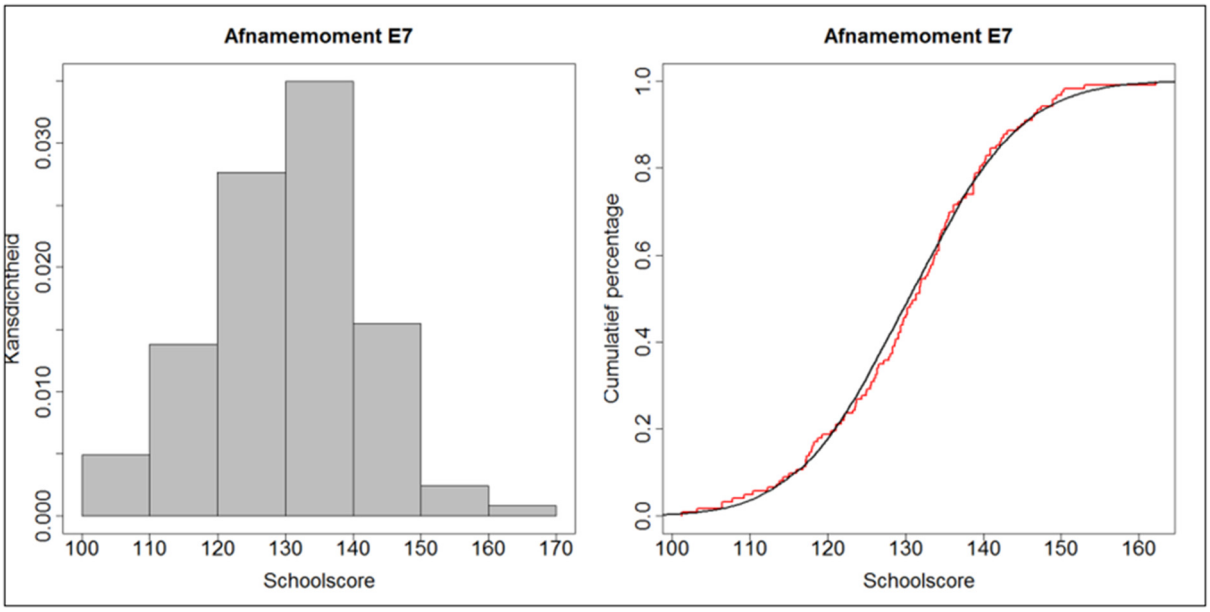
Afname-moment	Aantal replicaties	Aantal scholen	Gemiddelde	SD School	SD Leerling	ICC
M7	20	181	251,4	9,0	24,0	0,090
E7	20	123	260,2	8,2	22,8	0,114

Figuur 4.4 en 4.5 laat de verdeling van schoolgemiddelden zien op de oorspronkelijke schaal. Deze is nog niet getransformeerd naar LVS schaal. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 181 en 123. Op het eerste gezicht lijkt de verdeling redelijk normaal verdeeld. Op basis van het eindresultaat uit de bootstrapprocedure zijn de percentielen voor de vaardigheids-verdeling A tot en met E en I tot en met V berekend. Tabel 4.11 geeft de normgegevens op schoolniveau. De percentielen komen dichter bij elkaar te liggen dan in de leerlingverdeling. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.

Figuur 4.4 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde M7



Figuur 4.5 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde E7



Tabel 4.11 Normtabel op schoolniveau voor LVS-III Rekenen-Wiskunde M7-E7

Afname-moment	***M***	***SD***	***P*10**	***P*20**	***P*25**	***P*40**	***P*50**	***P*60**	***P*75**	***P*80**	***P*90**
M7	251,4	9,0	239,8	243,8	245,3	249,1	251,4	253,7	257,5	259,0	263,0
E7	260,2	8,2	249,8	253,4	254,7	258,2	260,2	262,3	265,7	267,1	270,7

5 Betrouwbaarheid en meetnauwkeurigheid

5.1 Methoden om de betrouwbaarheid te bepalen

In hoofdstuk 4 is aangegeven dat elke leerling die deelgenomen heeft aan het normeringsonderzoek slechts een deel van de items gemaakt heeft die uiteindelijk in de toetsen Rekenen-Wiskunde 3.0 opgenomen zijn.
De betrouwbaarheid van de uitgegeven toetsen in klassieke zin is dan ook niet rechtstreeks te bepalen.
Het is echter mogelijk om de betrouwbaarheid van elke uiteindelijke toets te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toetsen OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toetsen volledig bestaan uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de verschillende toetsen te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).
In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool Var(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt).
De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden.
De variantie van de geobserveerde toetsscore wordt aangegeven met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie, kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend gaan worden.

$$Var(t) = E[Var(t \mid \tau(\theta))] \tag{5.1}$$

Deze variantie kan opgevat worden als de (gemiddelde) meetfoutvariantie in de metriek van de geobserveerde scores (t). In analogie met de theorie over de betrouwbaarheid volgt dan

$$MAcc = \frac{Var(\tau)}{Var(\tau) + Var(t)} \tag{5.2}$$

waarin MAcc staat voor 'Accuracy of Measurement'.

5.2 Betrouwbaarheid: resultaten

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Rekenen-Wiskunde.
In de tweede kolom staat de maximumscore, voor iedere toets is deze gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde scores van de leerlingen op de verschillende toetsen. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe score van iedere toets. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de verschillende toetsen (of toetsonderdelen) is. De schattingen van de gemiddeldes, de standaardmeetfouten en de betrouwbaarheden zijn gebaseerd op de data van het normeringsonderzoek.
De betrouwbaarheidscoëfficiënten zijn goed te noemen. Voor toetsen van het type waar geen zware consequenties voor leerlingen aan verbonden zijn (zoals de toetsen LVS Rekenen-Wiskunde) geeft de COTAN (Commissie TestAangelegenheden Nederland van het Nederlands Instituut van Psychologen) aan dat een betrouwbaarheids-coëfficiënt lager dan 0,70 onvoldoende is, een betrouwbaarheidscoëfficiënt tussen 0,70 en 0,80 voldoende, en een

betrouwbaarheidscoëfficiënt hoger dan 0,80 goed (Evers, Lucassen, Meijer & Sijtsma, 2010, p. 33). Op grond van dit criterium is de meetnauwkeurigheid van alle toetsen goed te noemen.

Tabel 5.1 Betrouwbaarheden, gemiddelden en standaardmeetfouten bij de toetsen Rekenen-Wiskunde

Toets	Maximum score	Gemiddelde	Standaard-meetfout	Macc	Test-hertest (simulaite
M7	96	61,3	4,1	0,95	0,95
E7	96	61,5	4,1	0,96	0,96
M7 digitaal	345	215,6	15,4	0,95	0,95
E7 digitaal	349	209,4	15,6	0,95	0,95

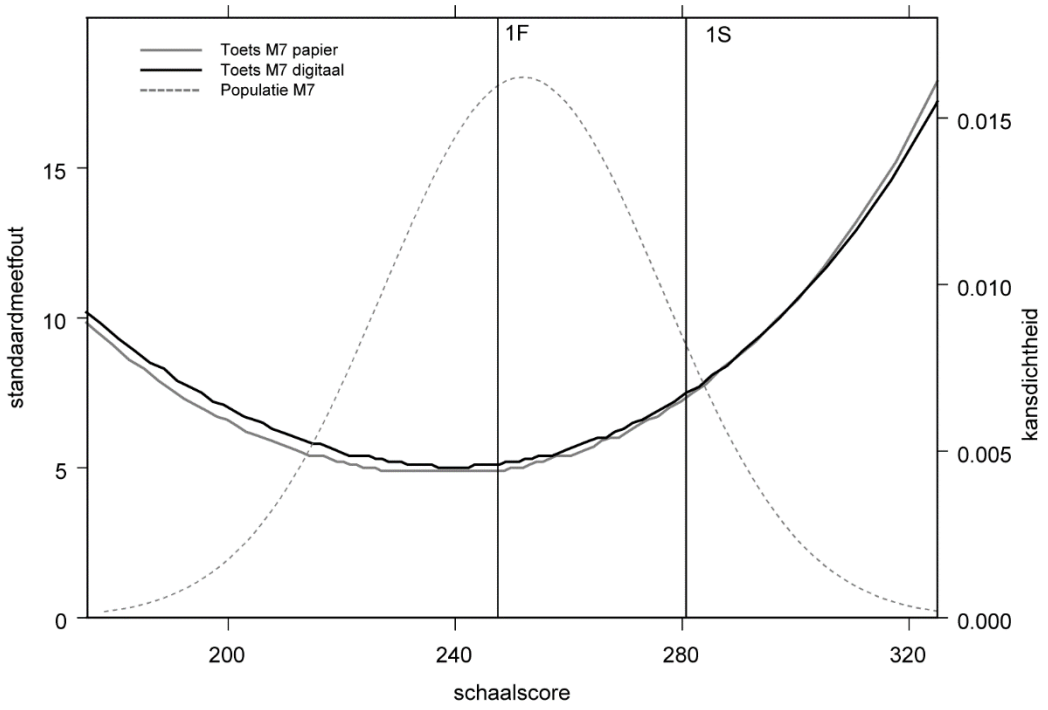
digitale toetsen werken met gewogen score in het Computerprogramma LOVS

Het feit dat alle items OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1 000 000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1 000 000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toetsafname is immers volledig onafhankelijk van de eerste toetsafname en wordt niet beïnvloed door de kennis die de leerling mogelijk verworven heeft via de eerste toetsafname. Daarnaast is er sprake van invloed van een test-hertest-interval: beide afnames worden gesimuleerd alsof zij op hetzelfde moment plaats zouden vinden. De resultaten zijn weergegeven in tabel 5.1 (zie kolom 6). De uitkomsten komen overeen met eerder berekende coëfficiënten en leiden dan ook tot dezelfde conclusies met betrekking tot de betrouwbaarheid van de toetsen Rekenen-Wiskunde 3.0.

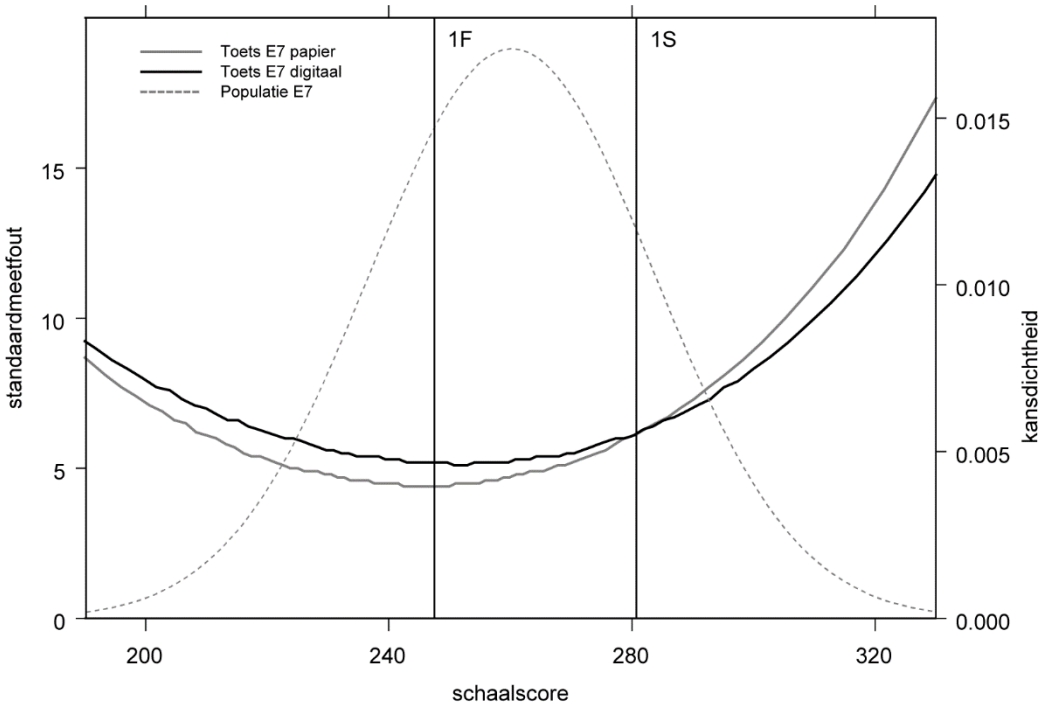
5.3 Lokale betrouwbaarheid en meetnauwkeurigheid

De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid van de verschillende toetsen Rekenen-Wiskunde. De figuren 5.1 en 5.2 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de papieren en digitale toetsen M7 en E7. In deze figuren staat voor iedere toets de grootte van de meetfout op de vaardigheidsschaal afgebeeld.
Ook zijn de kansdichtheidfuncties voor de normgroepen op de verschillende afnamemomenten opgenomen. Deze laten zien hoe de vaardigheid van de leerlingen verdeeld is over de vaardigheidsschaal in de populaties die de toets gemaakt hebben. De figuren maken duidelijk dat de meetfout kleiner is in de lagere en gemiddelde vaardigheidsregionen dan in de hogere vaardigheidsregionen.

Figuur 5.1 Grootte van de meetfouten voor de papieren en digitale toetsen M7 en de kansdichtheidsfuncties voor de M7-populatie



Figuur 5.2 Grootte van de meetfouten voor de papieren en digitale toetsen E7 en de kansdichtheidsfuncties voor de E7-populatie



Betrouwbaarheidstabellen
De betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden is af te leiden uit de betrouwbaarheidstabellen 5.2 tot en met 5.5. De betrouwbaarheidstabellen laten het effect van de lokale meetnauwkeurigheid zien.
Zo laat tabel 5.2 bijvoorbeeld zien dat 87,1 procent van de leerlingen die bij de M7-toets (papier) in scoregroep I vallen met hun geschatte vaardigheidsscore ook met hun werkelijke vaardigheidsscore in deze scoregroep vallen. Anders gezegd: de kans dat een leerling van niveau I terecht als een leerling van niveau I wordt bestempeld is ongeveer 87,1 procent. Verder laat de tabel zien dat 12,8 procent van de leerlingen in niveaugroep I een vaardigheidsscore heeft die in werkelijkheid in scoregroep II valt.
Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toetsen is te vinden in de handleiding van het toetspakket (Cito, 2014). In de tabel van toetsscore naar vaardigheidsscore en niveau staat het score-interval vermeld. In deze kolom staat voor iedere ruwe score op elke toets het 67-procents- betrouwbaarheidsinterval voor de bijbehorende vaardigheidsschatting.

In de onderzoeksliteratuur is weinig geschreven over de beoordeling van betrouwbaarheidstabellen. Wanneer een betrouwbaarheidstabel als goed of voldoende kan worden beschouwd is onduidelijk en wat verwacht mag worden onder ideale omstandigheden is, voor zover ons bekend, niet onderzocht. Daarom worden betrouwbaarheids-tabellen vaak samengevat in één of meerdere indices. Wij gebruiken de *plus/minus 1 niveau-index* en de *Marginal Classification Accuracy*. De eerste maat is bedacht door Pilliner (1969). Hij stelt als ambitieniveau dat 95 procent van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, ***of 1*** scoregroep daarboven ***of 1*** scoregroep daaronder. In de tabellen zijn dit de gearceerde cellen. Dit ambitieniveau is gebaseerd op de veronderstelling dat geen enkele toets perfect meet en dat er dus altijd sprake is van foutieve classificaties. In dat licht is de maximale accuraatheid die op het individuele niveau bereikt kan worden plus of minus één scoregroep. De tweede maat wordt op verschillende plekken in de literatuur beschreven. De maat laat zien hoe vaak de classificatie op basis van de geschatte vaardigheidsscore gemiddeld gezien overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. Bij een ideale toetsafname lijkt de *Marginal Classification Accuracy* rond 0,75 - 0,80 uit te komen. In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

De samenvattende indices voor de afnamemomenten medio groep 7 en einde groep 7 zijn te vinden in de tabellen 5.2a tot en met 5.5a. In deze tabellen laten de Marginal Classification Accuracy waarden zien dat de meeste leerlingen op basis van hun geschatte vaardigheidsscore geplaatst worden in de niveaugroep waar ze werkelijk thuishoren en de Accuracy plus/minus 1 niveau waarden maken aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969).

Tabel 5.2 Betrouwbaarheidstabel toets M7 papier voor afnamemoment medio 7

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-Groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	90,2	9,8	0,0	0,0	0,0	E	86,9	13,1	0,0	0,0	0,0
IV	11,8	73,3	14,8	0,1	0,0	D	10,9	74,8	14,2	0,0	0,0
III	0,0	16,1	68,0	15,9	0,0	C	0,0	10,5	76,6	12,8	0,0
II	0,0	0,1	15,7	70,7	13,5	B	0,0	0,0	13,6	74,1	13,2
I	0,0	0,0	0,1	12,8	87,1	A	0,0	0,0	0,0	11,3,	88,7

Tabel 5.2a Samenvattende indices toets M7 papier voor afnamemoment medio 7

	Toets M7 papier, afnamemoment M7	
	scoregroep I t/m V	**scoregroep A t/m E**
Marginal classification accuracy	77,9	79,9
Accuracy plus/minus 1 niveau	99,9	100

Tabel 5.3 Betrouwbaarheidstabel toets M7 digitaal voor afnamemoment medio 7

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-Groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	91,3	8,7	0,0	0,0	0,0	E	91,1	8,9	0,0	0,0	0,0
IV	12,9	72,2	14,8	0,1	0,0	D	11,4	73,2	15,5	0,0	0,0
III	0,0	16,9	66,7	16,4	0,0	C	0,0	10,6	75,5	13,9	0,0
II	0,0	0,2	17,2	70,6	12,0	B	0,0	0,0	12,7	74,8	12,5
I	0,0	0,0	0,1	15,6	84,3	A	0,0	0,0	0,0	13,2	86,8

Tabel 5.3a Samenvattende indices toets M7 digitaal voor afnamemoment medio 7

	Toets M7 digitaal, afnamemoment M7	
	scoregroep I t/m V	**scoregroep A t/m E**
Marginal classification accuracy	77,7	80,1
Accuracy plus/minus 1 niveau	99,9	100

Tabel 5.4 Betrouwbaarheidstabel toets E7 papier voor afnamemoment E7

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-Groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	90,2	9,8	0,0	0,0	0,0	E	88,1	11,9	0,0	0,0	0,0
IV	10,6	75,3	14,0	0,0	0,0	D	10,3	75,2	14,5	0,0	0,0
III	0,0	15,3	69,8	14,8	0,0	C	0,0	9,5	78,4	12,1	0,0
II	0,0	0,1	17,1	68,3	14,4	B	0,0	0,0	13,6	74,3	12,1
I	0,0	0,0	0,2	13,2	86,5	A	0,0	0,0	0,0	11,7	88,2

Tabel 5.4a Samenvattende indices toets E7 papier voor afnamemoment eind 7

	Toets E7 papier, afnamemoment E7	
	scoregroep I t/m V	**scoregroep A t/m E**
Marginal classification accuracy	78,0	80,4
Accuracy plus/minus 1 niveau	99,9	100

Tabel 5.5 Betrouwbaarheidstabel toets E7 digitaal voor afnamemoment eind 7

	Scoregroep waarin de ware score valt						**Scoregroep waarin de ware score valt**				
Score-groepen V t/m I	**V**	**IV**	**III**	**II**	**I**	**Score-Groepen E t/m A**	**E**	**D**	**C**	**B**	**A**
V	91,7	8,3	0,0	0,0	0,0	E	88,3	11,6	0,0	0,0	0,0
IV	12,4	72,8	14,8	0,1	0,0	D	15,9	70,8	13,2	0,0	0,0
III	0,0	16,7	67,0	16,3	0,1	C	0,0	13,1	74,4	12,5	0,0
II	0,0	0,2	15,9	66,2	17,7	B	0,0	0,0	11,1	72,8	16,1
I	0,0	0,0	0,1	9,1	90,8	A	0,0	0,0	0,0	10,9	89,1

Tabel 5.5a Samenvattende indices toets E7 digitaal voor afnamemoment eind 7

	Toets E7 digitaal, afnamemoment E7	
	scoregroep I t/m V	**scoregroep A t/m E**
Marginal classification accuracy	78,7	79,7
Accuracy plus/minus 1 niveau	99,9	100

Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 99,9 tot 100 procent van de leerlingen in een scoregroep ook in werkelijkheid in die scoregroep, ***of 1*** scoregroep daarboven ***of 1*** scoregroep daaronder. De *Marginal Classification Accuracy* loopt uiteen van 77,7 tot 80,4 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in ruim twee derde van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten stemmen hiermee tot tevredenheid: het percentage misclassificaties is beperkt.
Op basis van bovenstaande gegevens concluderen we dat op basis van de toetsen Rekenen-Wiskunde 3.0 groep 7 de leerlingen op een betrouwbare manier ingedeeld kunnen worden in normgroepen. Deze indeling voldoet uitstekend gegeven het doel van de toets. Uiteraard dienen de gebruikers rekening te houden met het gegeven dat er altijd sprake kan zijn van misclassificatie; veelal van maximaal 1 niveau verschil.

6 Validiteit

De begripsvaliditeit van een toets heeft betrekking op de vraag in hoeverre de toetsscores toe te schrijven zijn aan verklarende concepten en constructen die deel uitmaken van het theoretische kader dat aan de ontwikkeling van de toets ten grondslag ligt. Met inhoudsvaliditeit wordt bedoeld de representativiteit van de opgaven qua leerstofgebied. Bij leervorderingentoetsen, zoals deze toets Rekenen-Wiskunde 3.0 voor groep 7, speelt de inhoudsvaliditeit een relatief belangrijke rol. Dat houdt echter niet in dat begripsvaliditeit onbelangrijk is. Vandaar dat daar ook grondig onderzoek naar is verricht.
In paragraaf 6.1 wordt beschreven waarop de inhoudsvaliditeit van de toets gebaseerd is. De paragrafen 6.2 tot en met 6.6 zijn gewijd aan een aantal aspecten van begripsvaliditeit. In paragraaf 6.2 wordt het unidimensionele karakter van de toets aangegeven en worden gegevens over de structuur van de toets gepresenteerd. In paragraaf 6.3 wordt de kwaliteit van het itemmateriaal behandeld. Paragraaf 6.4 gaat over onderzoek naar vraagpartijdigheid (itembias). Paragraaf 6.5 behandelt het soortgenootonderzoek dat in het kader van de ontwikkeling van deze toets is uitgevoerd. Dit onderzoek levert data op voor de convergente en divergente validiteit. Als laatste komen in paragraaf 6.6 verschillen tussen relevante groepen aan bod.

6.1 Inhoudsvaliditeit

De samenstelling van de toets is bepaald door inhoudelijke criteria en psychometrische criteria. Voor de inhoudsvaliditeit zijn de inhoudelijke criteria relevant. Inhoudelijk zijn richtinggevend geweest de beschrijving van de kerndoelen van de SLO (Ministerie van Onderwijs, Cultuur en Wetenschappen, 2006), en de uitwerking daarvan zoals deze zijn terug te vinden in de inhoud van de referentieniveaus en de tussendoelen van de SLO (Buijs, 2008). De concrete domeinbeschrijving is per referentiedomein in hoofdstuk 3 weergegeven net als de verdere inhoudelijke verantwoording van de toets en de verdeling van de opgaven over de verschillende domeinen. De constructie van de opgaven is eveneens afgeleid van deze domeinindeling en ook de definitieve selectie van opgaven in de toets is gebaseerd op een gewenste verdeling van verschillende typen opgaven binnen en over de verschillende domeinen. Beoogd is de toetsen onafhankelijk samen te stellen van de verschillende onderwijsmethoden in die zin dat de getoetste stof in alle methodes aan bod is gekomen en de inhoud van de toets niet in meerdere mate tot uitdrukking komt in één van de methoden. Bij de constructie van de opgaven zijn leerkrachten uit het onderwijs betrokken zodat de opgaven voor wat betreft rekeninhoud/getallen en voor wat betreft context aansluiten bij leerlingen van groep 7.

6.2 Unidimensionaliteit, respectievelijk structuur

Zoals in hoofdstuk 4 al aangegeven, zijn bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij is duidelijk geworden dat voor de toets de verdeling gelijkmatig is over het gehele interval van overschrijdingskansen. Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat sprake is van niet-significante S-toetsen. Het aantal significante S-toetsen was te verwaarlozen onder het nulmodel. Zij vormen een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren (zie tabel 4.1).
Ook in hoofdstuk 4 zijn als maat voor de modelpassing de R1c-waarden gepresenteerd. Omdat deze eveneens ondersteuning bieden voor de validiteit refereren we daar nogmaals aan. R1c is een statistiek die zicht geeft op de modelpassing van de toets als geheel. Voor een acceptabele modelpassing geldt als vuistregel dat R1c bij voorkeur niet significant zou moeten zijn en niet groter dan ongeveer anderhalf maal het aantal vrijheidsgraden (df). In tabel 4.2 zijn deze waarden te vinden.
De modelpassing van de toetsen voldoet aan de voorwaarde dat voor de momenten M7 en E7 de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt.
Voor wat betreft de kwaliteit van de kalibratie verwijzen we naar hoofdstuk 4. Daar wordt onder andere in tabel 4.2 aangetoond dat de kalibratie geslaagd genoemd mag worden.
Het bovenstaande geeft aan dat de rekenvaardigheid in voldoende mate unidimensioneel gemeten kan worden op een rekenschaal en dat het dus mogelijk is om op basis van de verkregen scores de vaardigheid van de leerlingen

op een enkele schaal weer te geven. De toetsopgaven van de verschillende domeinen zorgen voor een goede dekking van het begrip rekenvaardigheid in groep 7. De toetsopgaven richten zich op inhoudelijk relevante domeinen als Getallen, Verhoudingen, Meten en meetkunde en Verbanden. De schaal zou wellicht een nog fraaiere passing vertonen als slechts één van deze onderdelen gekozen zou zijn – dit ligt dan dichter tegen het theoretische ideaal van unidimensionaliteit aan –, maar zou inhoudelijk te eenzijdig zijn om praktisch zinvolle uitspraken over rekenvaardigheid te doen.
Ook de grote onderlinge samenhang tussen de vier verschillende inhoudelijke domeinen geeft aan dat in de praktijk deze vier domeinen samengenomen kunnen worden om een unidimensionele uitspraak te doen. De correlaties tussen de subschalen, zoals gegeven in tabel 6.1 variërend van 0,94 tot 0,99 geven aan dat voor praktisch gebruik de schaal zeer goed gebruikt kan worden om leerlingen op een enkele rekenschaal te plaatsen, met een goede relevante inhoudelijke dekking van de verschillende domeinen.

Correlatie scores deeltaken met elkaar en totaal

Tabel 6.1 Latente correlaties en aantallen leerlingen tussen score op deeltaken en totaalscore van de toetsen M7 en E7

M7

	Getallen	**Verhoudingen**	**Meten**	**Verbanden**
Getallen		2212	2212	2212
Verhoudingen	0,96		2212	2212
Meten	0,96	0,95		2212
Verbanden	0,96	0,99	0,94	

E7

	Getallen	**Verhoudingen**	**Meten**	**Verbanden**
Getallen		1176	1176	1176
Verhoudingen	0,98		1176	1176
Meten	0,99	0,99		1176
Verbanden	0,97	0,96		

Ten slotte bespreken we, in het kader van de structuur van de toetsen, nog een methode om de modelpassing te verantwoorden die wordt besproken in het COTAN Beoordelingssysteem (Evers, Lucassen, Meijer & Sijtsma, 2010, p. 40). Het betreft hier een poging om de nauwkeurigheid van de itemparameterschattingen te beoordelen op basis van een constante (in het COTAN-Beoordelingssysteem met ‘c’ aangeduid) die weergeeft hoe de relatie is tussen de standaardfout van de moeilijkheidsparameter van een item en de standaarddeviatie van de vaardigheidsverdeling van de kalibratiepopulatie.
Het beoordelingssysteem geeft ook richtlijnen voor het beoordelen van de grootte van deze ‘c’. Deze dient te worden beoordeeld als goed als de waarde lager is dan of gelijk aan 0,20. Waarden tussen 0,30 en 0,40 kunnen nog als voldoende worden beschouwd. In hoofdstuk 4 is deze informatie al weergegeven maar omdat deze ook relevant is voor de validiteit wordt deze informatie hier nog aangehaald. In tabel 4.3 zijn gemiddelde en range van deze waarden voor alle opgaven per toets weergegeven. De gemiddelde waarde van de constante, is met een waarde rond 0,13 voor de papieren toetsen uitstekend te noemen. Voor de digitale toetsen liggen de waarden een fractie hoger maar ook deze is met een gemiddelde van rond de 0,17 prima. Zowel bij de papieren toetsen als de digitale toetsen zijn er slechts enkele opgaven met een c-waarde boven de 0,30. Bij de papieren toetsen slechts 1 van de 192 en bij de digitale toetsen slechts 2 van de 192. Bij geen enkele toets zijn er items met een waarde boven de 0,40.
Over het algemeen kan de nauwkeurigheid van de schattingen als goed beoordeeld worden. De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen.

6.3 Itemkwaliteit

Tabel 6.2 Range en gemiddelde van p- en R_{ir}-waarden naar toetsmoment

	P-waarden		Rit-waarden		N-tems
	Range	**Gemiddelde**	**Range**	**Gemiddeld**	
M7	0,42 – 0,86	0,64	0,22 – 0,61	0,41	96
E7	0,41 – 0,90	0,64	0,21 – 0,55	0,41	96
M7 dig	0,40 – 0,87	0,63	0,11 – 0,56	0,39	96
E7 dig	0,37 – 0,91	0,60	0,21 – 0,61	0,37	96

In tabel 6.2 zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de R_{ir} waarden van de items van de papieren en digitale toetsen, M7 en E7. Bij de toetsen is te zien dat de p-waarden liggen tussen de 0,37 en 0,91. Er is gezorgd voor een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de papieren en digitale toetsen voor M7 en E7 ligt tussen de 0,60 en 0,64. Er wordt in het algemeen voor groep 7 gestreefd naar een gemiddelde p-waarde tussen de 0,60 en 0,75. Daarmee zijn de toetsen niet te moeilijk en wordt voorkomen dat de leerling gefrustreerd raakt tijdens de toetsafname. De gemiddelde R_{ir} is bij alle toetsen uitstekend. Door de COTAN wordt een R_{it}-waarde van boven de 0,30 gekwalificeerd als goed. Met een gemiddelde R_{it} van 0,37 of hoger is de itemkwaliteit van de toetsen uitstekend te noemen. Er wordt gestreefd alleen items op te nemen met een R_{it} van 0,20 of hoger. De R_{ir} , die hier gegeven is, ligt altijd iets lager dan de $R_{it.}$ Er is in de toetsen M7 en E7 papier en M7 en E7 digitaal slechts 1 item met een R_{it} van lager dan 0,20. In totaal dus 1 van de 384 items. Bijlage 4 bevat een volledig overzicht van de p-waarden en de R_{it}-waarden van de items van de toetsen.
In tabel 6.3 zijn de verdelingskarakteristieken gegeven van de ruwe scores op de verschillende toetsmomenten. De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. Omdat deze gemiddelde moeilijkheidsgraad voor alle onderdelen rond de 0,65 ligt, zijn de verdelingen linksscheef (vergelijk de negatieve waarden in de kolom ‘skewness’), de ene wat meer dan de andere. De verdelingen zijn ééntoppig.

Tabel 6.3 Verdelingskenmerken van de toetsen Rekenen-Wiskunde groep 7

	Gemiddelde	**Standaarddeviatie**	**Skewness**	**Kurtosis**
M7 Ruwe score	61,3	19,2	-0,46	-0,57
E7 Ruwe score	61,5	19,4	-0,48	-0,55
M7 digitaal Ruwe score	60,4	18,5	-0,41	-0,61
E7 digitaal Ruwe score	58,1	17,9	-0,29	-0,64

6.4 Itembias

Er is onderzoek uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning,* DIF) met betrekking tot sekse. Hiervoor is de papieren kalibratie als uitgangspunt genomen voor zowel M7 als E7. Voor alle toetsopgaven zijn geobserveerde en verwachte scores voor zowel jongens als meisjes in verschillende scoregroepen berekend. Vervolgens is hier een S-statistiek voor berekend, analoog aan hoe dit gebeurt tijdens de kalibratie (zie hoofdstuk 4). Het onderzoek naar DIF over sekse per item liet bij 14 van de 339 items van de totale schaal van E6-E7 lichte vorm van DIF zien. 8 daarvan waren in het voordeel van jongens, de andere 6 in het voordeel van meisjes. Wanneer we kijken naar de items van de kalibratie van de schaal M7-M8 dan zien we daar 4 items van de 346 items met een lichte vorm van DIF. 3 daarvan zijn in het voordeel voor meisjes, 1 in het voordeel van jongens. Al met al kunnen we daarom concluderen dat er nauwelijks sprake is van DIF met betrekking tot sekse.

6.5 Soortgenootonderzoek

Soortgenootonderzoek wordt uitgevoerd vanuit de gedachte dat de correlatie van de ontwikkelde rekentoets met andere rekentoetsen zoals de toets LVS II Rekenen-Wiskunde en de SVT Rekenen Boom hoog is, aangezien hetzelfde onderliggende construct wordt gemeten, en hoger ligt dan de correlatie met andere leervorderingstoetsen zoals Spelling, Begrijpend lezen en Woordenschat. Zowel de rekentoetsen onderling als de rekentoetsen met andere leervorderingen worden verondersteld te correleren aangezien in beide gevallen een beroep wordt gedaan op een zelfde onderliggende vaardigheid die omschreven kan worden als vermogen tot leren/ intelligentie en waarbij ook motivatie en inzet een rol spelen. Echter zal de te vinden correlatie van rekentoetsen onderling hoger zijn dan van rekentoetsen met toetsen Spelling, Woordenschat en Begrijpend lezen omdat deze toetsen gelijksoortig zijn en inhoudelijk overeenkomen. Hierna worden de resultaten van het soortgenootonderzoek gepresenteerd. Samenvattend kan worden aangegeven dat de gevonden resultaten in lijn zijn met de verwachtingen en daarmee een onderdeel zijn van de bewijsvoering van de validiteit van de ontwikkelde rekentoets LVS Rekenen-Wiskunde 3.0 groep 7.

Correlatie scores LVS II met LVS III

De leerlingen van de normeringssteekproef hebben zowel opgaven van de LVS generatie II als opgaven van de LVS generatie III gemaakt en daardoor kan een correlatie worden bepaald tussen de score op de LVS generatie II toetsen en de LVS generatie III toetsen. De latente correlaties gebaseerd op items uit LVS II en die van LVS III waren hoog voor M7 (r=0,96; N=2210) en E7 (r=0,98; N=1183). Aangezien de toetsen Rekenen-Wiskunde LVS generatie II door de COTAN op alle onderdelen (met criteriumvaliditeit niet van toepassing) met een goed is beoordeeld is, vormt deze hoge correlatie een belangrijke bewijsvoering voor de validiteit van de LVS generatie III toetsen.

In het kader van het soortgenootonderzoek is bij 2 scholen de Schoolvaardigheidstoets Rekenen-Wiskunde van Boom testuitgevers afgenomen. De scholen werden benaderd op basis van het feit dat zij hadden deelgenomen aan het normeringsonderzoek voor de 3de generatietoetsen Rekenen-Wiskunde LVS van groep 7. De scholen waren gelegen in Den Haag en Prinsenbeek.

Tabel 6.4 Correlatie van Cito toetsen Rekenen-Wiskunde met de Schoolvaardigheidstoets (Boom)

	Cito Rekenen-WiskundeM7 generatie 3	
	correlatie	**N**
Schoolvaardigheidstoets Rekenen-Wiskunde (Boom)	0,91	153

In tabel 6.4 staat de correlatie van de nieuwe toets Rekenen-Wiskunde M7 weergeven met de Schoolvaardigheidstoets Rekenen-Wiskunde. De Schoolvaardigheidstoets Rekenen-Wiskunde is positief beoordeeld door de COTAN. De correlatie met de Schoolvaardigheidstoets Rekenen-Wiskunde van Boom testuitgevers is 0,91 voor M7. Deze waarde is gecorrigeerd voor attenuatie. De gevonden correlatie ligt wat hoger dan de correlatie die eerder gevonden is tussen de 3de generatietoets LVS Rekenen-Wiskunde M3, E3, M4, M5 en de Schoolvaardigheidstoets Rekenen-Wiskunde van Boom. Deze waren 0,70 voor M3 en 0,78 voor E3 en 0,71 voor M4. Voor groep M5 was de correlatie 0,77. En de correlatie bij groep 6 was 0,79 M6 en 0,78 E6. Echter waren deze niet gecorrigeerd voor attenuatie.

Tabel 6.5 Correlatie van de derde generatie Cito LVS toetsen Rekenen-Wiskunde M7 met diverse toetsen op het gebied van leervorderingen

	Rekenen-Wiskunde*	Aantal leerlingen
Cito Spelling M7	0,30	336
Cito Begrijpend Lezen M7	0,81	375
Cito Woordenschat M7	0,71	305
Cito DMT	0,36	371

*Deze correlaties zijn gecorrigeerd voor attenuatie

Aan de scholen die hebben deelgenomen aan het normeringsonderzoek is gevraagd of dataretour van de betreffende leerlingen van andere LVS-onderdelen gebruikt mocht worden. Met de dataretour functie zijn van de leerlingen van het normeringsonderzoek ook de scores op andere LVS-toetsen beschikbaar. In de tabel hierboven is de correlatie tussen de toets Rekenen-Wiskunde M7 3.0 en de toetsen Spelling, Begrijpend lezen, Woordenschat en DMT weergegeven. In de tabel hieronder zijn de correlaties voor E7 weergegeven.

Tabel 6.6 Correlatie van de derde generatie Cito LVS toetsen Rekenen-Wiskunde E7met diverse toetsen op het gebied van leervorderingen

	Rekenen-Wiskunde*	Aantal leerlingen
Cito Woordenschat E7	0,68	325
Cito Spelling werkwoorden	0,65	325

*Deze correlaties zijn gecorrigeerd voor attenuatie

De hoogte van de correlaties bij zowel M7 als E7 van de 3de generatie LVS toetsen Rekenen-Wiskunde met de leervorderingstoetsen van Taal zijn allen in de lijn der verwachting en komen overeen met wat is gerapporteerd in de wetenschappelijke verantwoordingen van groep 3 tot en met 6 LVS Rekenen-Wiskunde generatie III.

6.6 Verschillen tussen relevante subgroepen

In de tabellen 6.7 en 6.8 worden voor afnamemoment M7 en E7 de gemiddelde scores van de leerlingen per lesmethode weergegeven. Er is sprake van aanzienlijke verschillen tussen de gemiddelden van de methoden. Het aantal leerlingen is echter bij 3 van de 5 methoden te laag om daar verantwoord uitspraken over te kunnen doen. Tabel 6.9 en tabel 6.10 geven de effectgrootte tussen de verschillende methoden. Daarbij zijn alleen de methoden opgenomen met meer dan 150 leerlingen. Dit omdat anders individuele scholen een te grote invloed hebben op de berekening van het effect.

Tabel 6.7 Gemiddelde score per lesmethode

Moment	Lesmethode	M	SD	Aantal
	Wizwijs	245,7	22,3	66
M7	Pluspunt	247,7	24,4	168
	Alles telt	252,5	19,2	81
	Rekenrijk	264,7	24,0	68
	WIG	252,0	24,8	638

Tabel 6.8 Gemiddelde score per lesmethode

Moment	Lesmethode	M	SD	Aantal
	Wizwijs	248,8	22,9	64
E7	Pluspunt	257,2	24,3	156
	Alles telt	262,8	18,2	76
	Rekenrijk	271,2	25,7	68
	WIG	261,6	24,3	623

Tabel 6.9 Effectgrootte tussen methoden voor M7

	Pluspunt
WIG	0.17

Tabel 6.10 Effectgrootte tussen methoden voor E7

	Pluspunt
WIG	0.18

Er is sprake van geen effect bij de methoden Pluspunt en WIG zowel bij M7 als bij E7.

In tabel 6.11 wordt de score per halfjaar groep gepresenteerd voor zowel M7 als E7. Gebruikelijk is een patroon te vinden waarbij de jongste leerlingen het hoogst scoren en de oudste leerlingen het laagst. (Janssen, 2015; Hop, 2016)

Tabel 6.11 Gemiddelde score per halfjaargroep

M7

Halfjaargroep	M	SD	Aantal
10	263,9	23,4	115
10,5	252,1	24,0	762
11	251,8	25,6	780
11,5	242,5	25,3	281
12	235,0	22,9	102

E7

Halfjaargroep	M	SD	Aantal
10,5	275,5	19,4	54
11	261,7	24,7	440
11,5	259,6	24,1	408
12	247,6	25,5	147
12,5	241,2	23,6	54

Het patroon in elk van de tabellen is naar verwachting. De jongste groep leerlingen in de groep scoort hoog in vergelijking met de andere halfjaargroepen. Deze leerlingen zijn de versnelde leerlingen die op grond van hun cognitieve capaciteiten en/of leerprestaties een groep hebben overgeslagen. Aan de andere kant scoren de oudste twee groepen leerlingen in elk van de afnamemomenten het laagst. Ook hier is dat naar verwachting aangezien deze leerlingen in veel gevallen op grond van hun leerprestaties een jaar gedoubleerd hebben.

Tabel 6.12 Gemiddelde score jongen-meisje

Moment	Geslacht	Aantal	M	SD
M7	jongen	1058	255,6	24,6
	meisje	1098	246,7	25,6

Moment	Geslacht	Aantal	M	SD
E7	jongen	542	263,0	24,7
	meisje	576	256,6	25,1

Per afnamemoment is in de bovenstaande tabellen de gemiddelde score van jongens en meisjes weergegeven. Uit onderzoek is bekend dat de effectgrootte rond de 0,3 is halverwege de basisschool (Hop, 2012). Op alle afnamemomenten scoren jongens hoger dan meisjes. In termen van effectgrootte is er sprake van een klein effect (M7 0,35 en E7 0,26).

Tabel 6.13 Gemiddelde score leerlinggewicht

Moment	Leerlinggewicht	Aantal	M	SD
M7	0	1441	251,8	25,2
	0.30	79	239,0	25,5
	1.20	66	240,0	24,1
E7	0	783	259,6	24,8
	0.30	52	246,6	26,4
	1.20	36	249,9	25,3

Zowel bij M7 als bij E7 is een matig effect te zien tussen leerlingen met een gewicht van 0 en leerlingen met een gewicht van 0.30 of 1.20. Dit effect varieert van 0,51 tot 0,48 bij M7 en 0,51 tot 0,39 bij E7.
Zowel bij M7 als bij E7 is er geen effect waar te nemen bij leerlingen met een gewicht van 0.30 of 1.20 (zie hoofdstuk 4 voor een beschrijving van de leerlinggewichten).

7 Samenvatting

In dit hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken.
De toetsen Rekenen-Wiskunde 3.0 voor groep 7 vormen een hulpmiddel om vast te stellen in hoeverre leerlingen rekenvaardig zijn en hoe deze rekenvaardigheid zich ontwikkelt door leerlingen te volgen. Met behulp van categorieënanalyses kan in kaart worden gebracht op welk domein leerlingen ten opzichte van hun algemene rekenvaardigheid het relatief beter of zwakker doen. We beschrijven in hoofdstuk 2 dat de inhoud van de toetsen aansluit bij de kerndoelen primair onderwijs en bij de referentieniveaus. In de domeinbeschrijving onderscheiden we de domeinen Getallen, Verhoudingen, Meten en meetkunde en Verbanden. Voor groep 7 zijn alle domeinen van belang. Dat zijn ook de onderwerpen waarop bij de categorieënanalyses bij M7 en E7 op gerapporteerd wordt. We geven in het tweede hoofdstuk ook aan dat we met opgavenbanken werken en dat het algemene uitgangspunt is dat de vaardigheid rekenen-wiskunde kan worden opgevat als een unidimensioneel continuüm. Verder wordt in hoofdstuk 2 het gehanteerde meetmodel beschreven dat op de itemresponstheorie is gebaseerd.
Nadat we in hoofdstuk 2 de uitgangspunten bij de toetsconstructie beschreven hebben, hebben we in hoofdstuk 3 de inhoud van de toetsen uitgewerkt. Daarbij zijn de doelen voor de toetsen van groep 7 uitvoerig beschreven. Ook is in dit hoofdstuk verslag gedaan van de itemconstructie, de opzet van de normeringsonderzoeken en de kalibratieonderzoeken digitaal en papier – digitaal. Omdat de papieren en digitale opgaven op de vaardigheids-schaal rekenen-wiskunde passen, kunnen we zeggen dat de papieren en digitale opgaven dezelfde vaardigheid meten.

In hoofdstuk 4 rapporteerden we over de kalibratie en normering. We beschreven de opzet, de gevolgde stappen bij de kalibratie en de toetsing van het IRT-model dat gebruikt is bij de analyses. Uit de S-toetsing kan geconcludeerd worden dat het meetinstrument en het meetmodel adequaat is om het gedrag van leerlingen te verklaren. Bovendien blijkt dat verschillen in gedrag tussen de leerlingen zijn te verklaren door een unidimensioneel concept. Uit de resultaten van de analyses met betrekking tot R1c- waarden en de constante ‘c’ trekken we de conclusie dat de kalibratie geslaagd is.
In paragraaf 4.3.2 wordt aangetoond dat de normeringssteekproef op basis van de variabelen regio, urbanisatie-graad, schooltype en sekse een zeer goede afspiegeling vormt van de populatie. In de laatste paragraaf van hoofdstuk 4 presenteren we de normeringsresultaten.

In hoofdstuk 5 staan de betrouwbaarheden van de toetsen. De betrouwbaarheidscoëfficiënten van de toetsen zijn met 0,95 en hoger, goed te noemen. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien.
In het laatste hoofdstuk, hoofdstuk 6, wordt de validiteit van de toetsen behandeld. Zowel de begripsvaliditeit als de inhoudsvaliditeit komt aan bod. De inhoudsvaliditeit wordt aangetoond door te verwijzen naar de verschillende bronnen die in het Nederlands onderwijs richtinggevend zijn voor de inhoud van het rekenen-wiskunde domein. Vervolgens komt in hoofdstuk 6 de begripsvaliditeit aan bod. In eerste instantie door te verwijzen naar het unidimensionele karakter van de toets zoals dat in hoofdstuk 4 is aangetoond met de s-toetsen. Een goede modelfit wordt vervolgens aangetoond door te verwijzen naar de verhouding van de R1c ten opzichte van de vrijheidsgraden (df). Ook de gepresenteerde correlaties van de verschillende domeinen met elkaar en de verschillende domeinen met de totaalscore bieden ondersteuning voor het unidimensionale karakter van de toetsen. Als laatste in het kader van bewijsvoering van de structuur van de toets wordt in hoofdstuk 6 verwezen naar de constante c, die op enkele uitzonderingen na, er voor de verschillende items zeer goed uitzien. In hoofdstuk 6 wordt vervolgd met de gegevens over de kwaliteit van de items. De kwaliteit van de items is zeer goed te noemen. Zowel in termen van p-waarden als R_{it}-waarden. In hoofdstuk 6 worden gegevens gepresenteerd over DIF-onderzoek. Voor wat betreft sekse is vastgesteld dat er nauwelijks sprake is van DIF.

In hoofdstuk 6 wordt uitgebreid aandacht besteed aan de soortgenootvaliditeit. Als eerste bewijsvoering voor de validiteit van de toetsen wordt de hoge correlatie van de tweede generatie toetsen (LVS toetsen) met de toetsen Rekenen-Wiskunde 3.0 uit het Cito Volgsysteem opgevoerd. Daarmee is een belangrijk bewijsstuk geleverd voor de validiteit van de toetsen Rekenen-Wiskunde 3.0. Er wordt vervolgd met de presentatie van de correlatie-gegevens van een onderzoek waarbij een groep leerlingen naast de toets Rekenen-Wiskunde M7 een (Cotan positief beoordeelde) toets van een andere uitgever heeft gemaakt. Ook uit dat onderzoek blijkt dat de correlatie

tussen deze toets met de toetsen Rekenen-Wiskunde 3.0 groep 7 hoog is. De correlatie met andere toetsen op het gebied van leervorderingen blijkt lager te zijn dan de correlatie van de rekentoetsen onderling. Ook dat kan als bewijs van (divergente) validiteit worden opgevoerd. Als laatste worden verschillen tussen relevante subgroepen gepresenteerd. De scores van de verschillende halfjaargroepen zijn volgens verwachting en laten een bekend en verklaarbaar patroon zien. De verschillen in scores tussen jongens en meisjes zijn ook onderzocht. Jongens scoren hoger dan meisjes. In termen van effectgrootte is er sprake van een klein effect (M7 0,35 en E7 0,26).
Al met al kunnen we concluderen dat de validiteit van de toetsen Rekenen-Wiskunde 3.0 goed te noemen is.

8 Literatuur

Albert, J.H. (1992). *Bayesian estimation of normal ogive item response curves using Gibbs sampling.* Journal of Educational Statistics, 17, 251-269.

Béguin, A. A., & Glas, C. A. W. (2001). *MCMC estimation and some fit analysis of multidimensional IRT models.* Psychometrika, 66, 471-488.

*Besluit Kerndoelen basisonderwijs (*1993*).* 's Gravenhage, Sdu.

Boxtel, H. van, & B.T. Hemker (2009). *Wetenschappelijke verantwoording van de Intelligentietest Eindtoets Basisonderwijs*. Arnhem: Cito.

Buijs, K., Klep, J., & Noteboom, A. (2008). *Tule – Rekenen/Wiskunde*. SLO, Enschede.

Buijs, K., Scherpenzeel, P. van, Voorde, M. ten, & Zwaart, P. van der (2008a). *Werken aan de doorlopende leerlijn rekenen wiskunde van po naar vo. Enschede*. SLO, Enschede.

Cito (2014) Primair en speciaal onderwijs. *Cito Volgsysteem. Rekenen-Wiskunde 3.0. Groep 4*. Arnhem: Cito.

Cito (2013) Primair en speciaal onderwijs. *Cito Volgsysteem. Rekenen-Wiskunde 3.0. Groep 3*. Arnhem: Cito.

Cito (2006). *Leerling- en onderwijsvolgsysteem, Rekenen-Wiskunde, groep 5.* Arnhem: Cito.

Cito (2002). *Rekenen hulpboek groep 5 medio. Arnhem*: Cito.

Cito (2002). *Rekenen hulpboek groep 5 eind.* Arnhem: Cito.

Cito (2002). Entreetoets Groep 5. Arnhem: Cito.

Cito (z.j.). *Computerprogramma LOVS.* Arnhem: Cito.

Cito (z.j.). *Handleiding Computerprogramma LOVS*. Arnhem: Cito.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Erlbaum.

College voor Examens. (2012). *Toetswijzer bij de centrale eindtoets PO taal en rekenen.* Utrecht: College voor Examens.

College voor Toetsen en Examens. (2014). *Toetswijzer bij de centrale eindtoets PO taal en rekenen.* Utrecht: College voor Toetsen en Examens.

Eggen, T.J.H.M., (1993). *Itemresponstheorie en onvolledige gegevens.* In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk*. (pp. 239-284). Arnhem: Cito.

Embretson, S.E. (1983). *Construct validity: Construct representation versus nomothetic span.* Psychological Bulletin 93, *179-197.*

Evers, A., Lucassen, W., Meijer, R. & Sijstma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests.* Amsterdam, NIP/COTAN.

Expertgroep doorlopende leerlijnen taal en rekenen (2008). *Over de drempels met rekenen en taal. Hoofdrapport.* Enschede: Expertgroep doorlopende leerlijnen taal en rekenen.

Expertgroep doorlopende leerlijnen taal en rekenen (2008a). *Over de drempels met rekenen.* Enschede: Expertgroep doorlopende leerlijnen taal en rekenen.

Expertgroep doorlopende leerlijnen taal en rekenen (2009). Referentiekader taal en rekenen. De referentieniveaus. Enschede: SLO.

Glas, C.A.W. & Verhelst, N.D. (1993*). Een overzicht van itemresponsmodellen.* In: T.J.H.M. Eggen & P.F. Sanders (red.)*. Psychometrie in de praktijk. (*pp. 179-238). Arnhem: Cito*.*

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Heuvel-Panhuizen, M. van den, K. Buys & A. Treffers (red.) (2000). *Jonge kinderen leren rekenen.* Tussendoelen Annex Leerlijnen. Hele getallen. Bovenbouw basisschool. Groningen: Wolters-Noordhoff.

Hemker, B. (2017). *Jaarlijkse meting Taal en rekenen.* Arnhem, Cito.

Hemker, B.T., J. Kordes & J.J. van Weerden (2011): *Peiling van de rekenvaardigheid en de taalvaardigheid in jaargroep 8 en jaargroep 4 in 2010 - Jaarlijks Peilingsonderzoek van het Onderwijsniveau*. Arnhem: Cito.

Hop, M. (Eindredactie) (2012) *Balans van het reken-wiskundeonderwijshalverwege de basisschool 5. Uitkomsten van de vijfde peiling in 2010*. PPON-reeks nummer 47. Arnhem: Cito.

Hop, M., Janssen, J. & Engelen, R. (2016). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 5*. Arnhem: Cito.

Janssen, J., Hop, M., Wouda, J. & Hollenberg, J. (2015). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 3*. Arnhem: Cito.

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

Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8*. Arnhem: Cito.

Janssen, J., F. van der Schoot en B. Hemker (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4*. PPON-reeks nummer 32. Arnhem: Cito.

Janssen, J. en Engelen, R. (2001). *Verantwoording van de toetsen Rekenen-Wiskunde 1, 2 en 3*. Arnhem: Citogroep.

Keuning, J. (2011). *Normeren op schoolniveau met Cito dataretour*. Arnhem: Cito.

Keuning, J. (2014). *Actualiteit en kwaliteit van normen. Een werkwijze voor het normeren van een leerlingvolgsysteem.* Arnhem: Cito.

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

Kraemer, J-M. (2011). *Oplossingsmethoden voor aftrekken tot 100.* Proefschrift. Arnhem, Cito B.V.

Kraemer, J-M., F. van der Schoot & P. van Rijn (2009). *Balans van het reken-wiskundeonderwijs in het speciaal basisonderwijs. Uitkomsten van de derde peiling in 2006.* PPON-reeks nummer 39. Arnhem: Cito.

Kraemer, J-M. (2009a). *Balans over de strategieën en procedures bij het hoofdrekenen halverwege de basisschool. Uitkomsten van de peiling in 2005.* PPON-reeks nummer 40. Arnhem: Cito.

Kraemer, J-M (2008). *Diagnosticeren en plannen in de onderbouw*. Arnhem: Cito.

Kraemer, J-M., J. Janssen, F. van der Schoot & B. Hemker (2005). *Balans van het reken-wiskundeonderwijs halverwege de basisschool 4. Uitkomsten van de vierde peiling in 2003*. PPON-reeks nr. 31. Arnhem: Cito.

Koninklijke Nederlandse Akademie van Wetenschappen (2009). *Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering.* KNAW, Amsterdam.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McIntosh, A., Reys, B., & Reys, A. (1992). *A proposed framework for examining basic number sense.* In: For the Learning of Mathematics (1992, 12, 3, pag. 2-9).

Ministerie van Onderwijs, Cultuur en Wetenschap*. (2006). Kerndoelen primair onderwijs.* Op 4 januari 2009 ontleend aan http:// www.slo.nl/primair/kerndoelen/Kerndoelenboekje.pdf.

Ministerie van Onderwijs, Cultuur en Wetenschappen (2004). *Voorstel herziene kerndoelen basisonderwijs.*

Ministerie van Onderwijs, Cultuur en Wetenschappen (2004a).*Voorstel herziene kerndoelen basisonderwijs.* SLO (z.j.). Tule inhouden & activiteiten Rekenen-Wiskunde. Op 11 januari 2009 ontleend aan http://tule.slo.nl/RekenenWiskunde/F-KDRekenenWiskunde.html.

Ministerie van Onderwijs, Cultuur en Wetenschappen (1998*). Kerndoelen basisonderwijs* (1998). *Over de relatie tussen de algemene doelen en kerndoelen per vak*. Den Haag, Sdu.

Noteboom, A., Os, S. van & Spek, W. (2011). *Concretisering referentieniveaus 1F/1S*. Enschede: SLO.

Pilliner, A. (1969). *Estimation of number of grades to be awarded in an exanination by consideration of its reliability coefficient.* Edinburgh: The Godfrey Thomson Unit for Educational Research.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhage: Danish Institute for Educational Research.

Sanders, P. & Verstralen, H. (2010). *Het beoordelen van toetsscores.* In P. Sanders (ed.), *Toetsen op school* (pp. 143-155). Arnhem: Cito.

Scheltens, F., B. Hemker, J. Vermeulen (2013). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 5.* PPON-reeks nummer 51. Arnhem: Cito.

Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Newbury Park/London/New Delhi: Sage Publications.

Snijders, T.A.B. & Bosker, R.J. (1993). *Standard errors and sample sizes for two-level research*. Journal of Educational Statistics, 18, 237-260.

Staphorsius, G. (1994). Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument. Universiteit Twente, 1994.

TAL-team (2007). *Meten en meetkunde in de bovenbouw*. Groningen: Wolters Noordhoff.

TAL-team (2005). *Breuken, procenten, kommagetallen en verhoudingen. Tussendoelen Annex Leerlijnen.* Groningen: Wolters Noordhoff.

TAL-team. (2004). *Jonge kinderen leren meten en meetkunde.* Groningen: Wolters Noordhoff.

TAL-team. (2001). *Kinderen leren rekenen.* Groningen: Wolters Noordhoff.

TAL-team. (1999). *Jonge kinderen leren rekenen: hele getallen*. Groningen: Wolters Noordhoff.

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Verhelst, N. (2007). *Profielanalyse met Item Respons Theorie*. Arnhem: Cito.

Verhelst, N. D. & Verstralen, H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL): Theory and a computer program.* Arnhem: Cito.

Verhelst, N.D., Glas C.A.W. & Verstralen H.H.F.M. (1995). *OPLM: One Parameter Logistic Model. Computerprogram and manual*. Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (1995a). *The one parameter logistic model*. In: G.H. Fischer & I.W. Molenaar (Eds.). *Rasch models: Foundations, recent developments and applications* (pp. 215-239). New York: Springer.

Verhelst, N.D., (1993). *Itemresponstheorie.* In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk.* (pp. 83-178). Arnhem: Cito.

Verhelst, N.D. & Kleintjes, F.G.M. (1993a). Toepassingen van itemresponsetheorie. In: T.J.H.M. Eggen en P.F.Sanders (red.). *Psychometrie in de praktijk.* Arnhem: Cito.

Verhelst, N.D. (1992). *Het één parameter model (OPLM). Een theoretische inleiding en een handleiding bij het computerprogramma.* Arnhem: Cito.

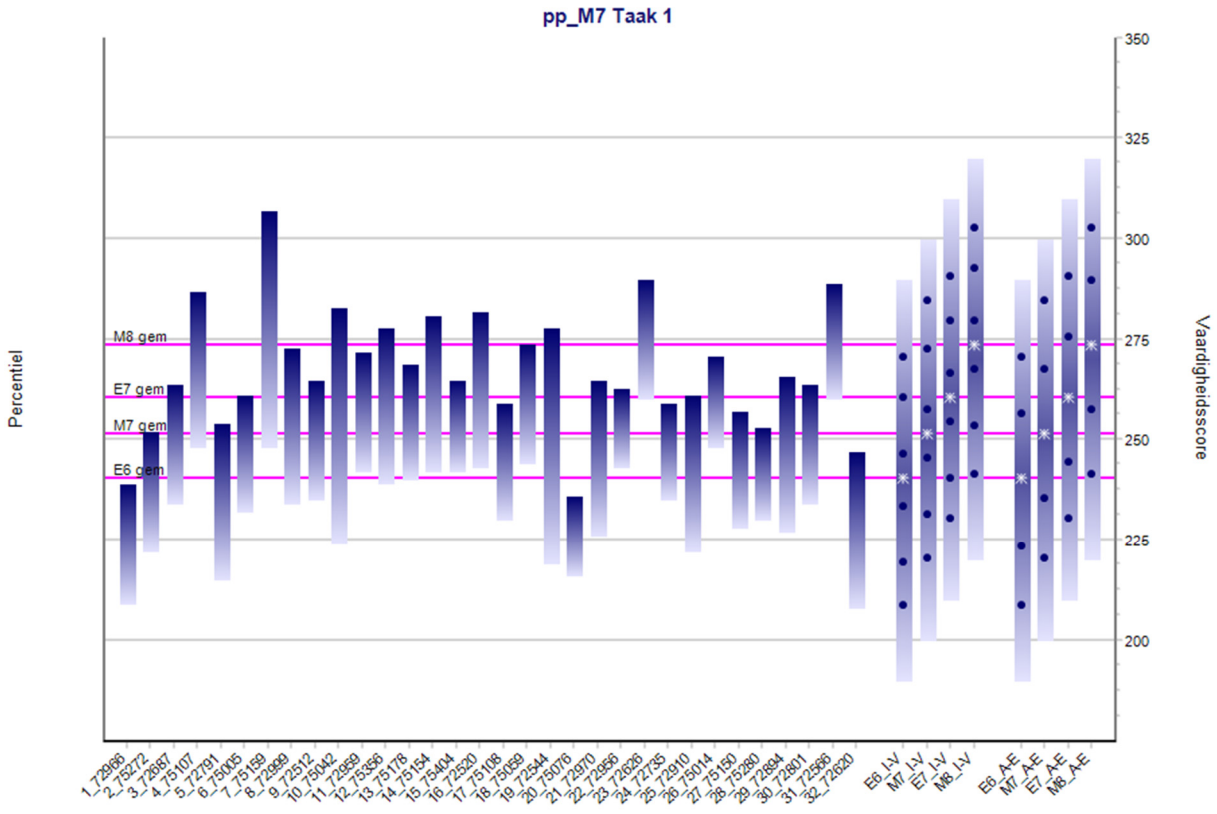
Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.H.J.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* Measurement and Research Department Reports 91-10. Arnhem: Cito.

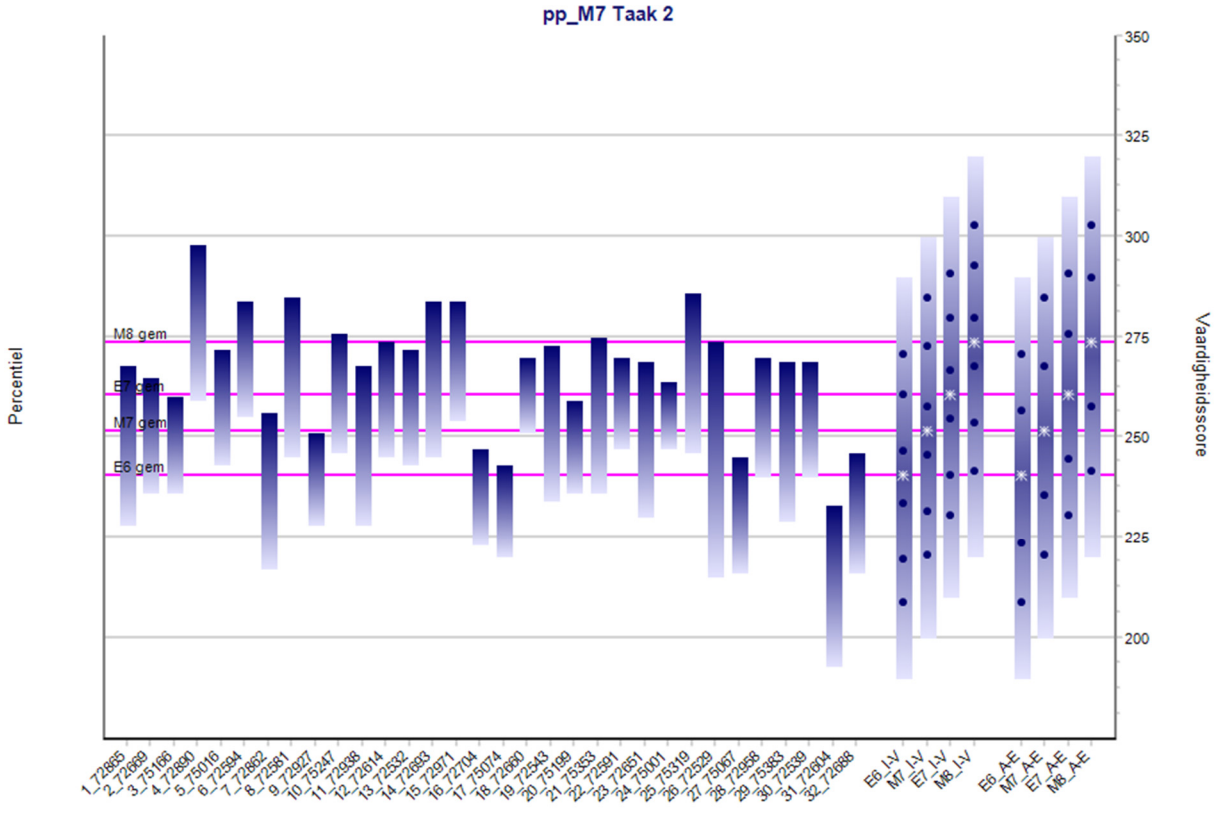
Verhelst, N., Eggen, T. (1989). *Psychometrische aspecten van peilingsonderzoek* (PPON-rapport, nr 4). Arnhem: Cito.

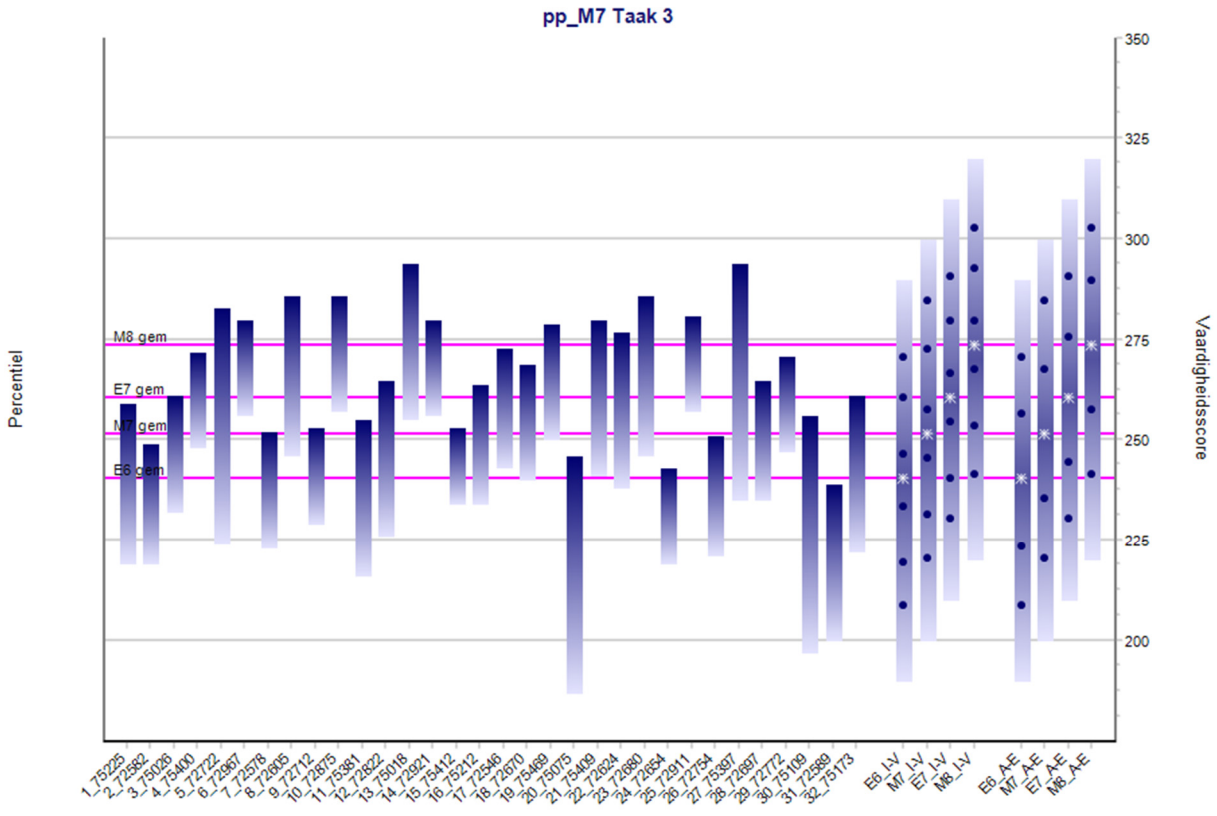
Verstralen, H.H.F.M. (1997*). OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem: Cito.

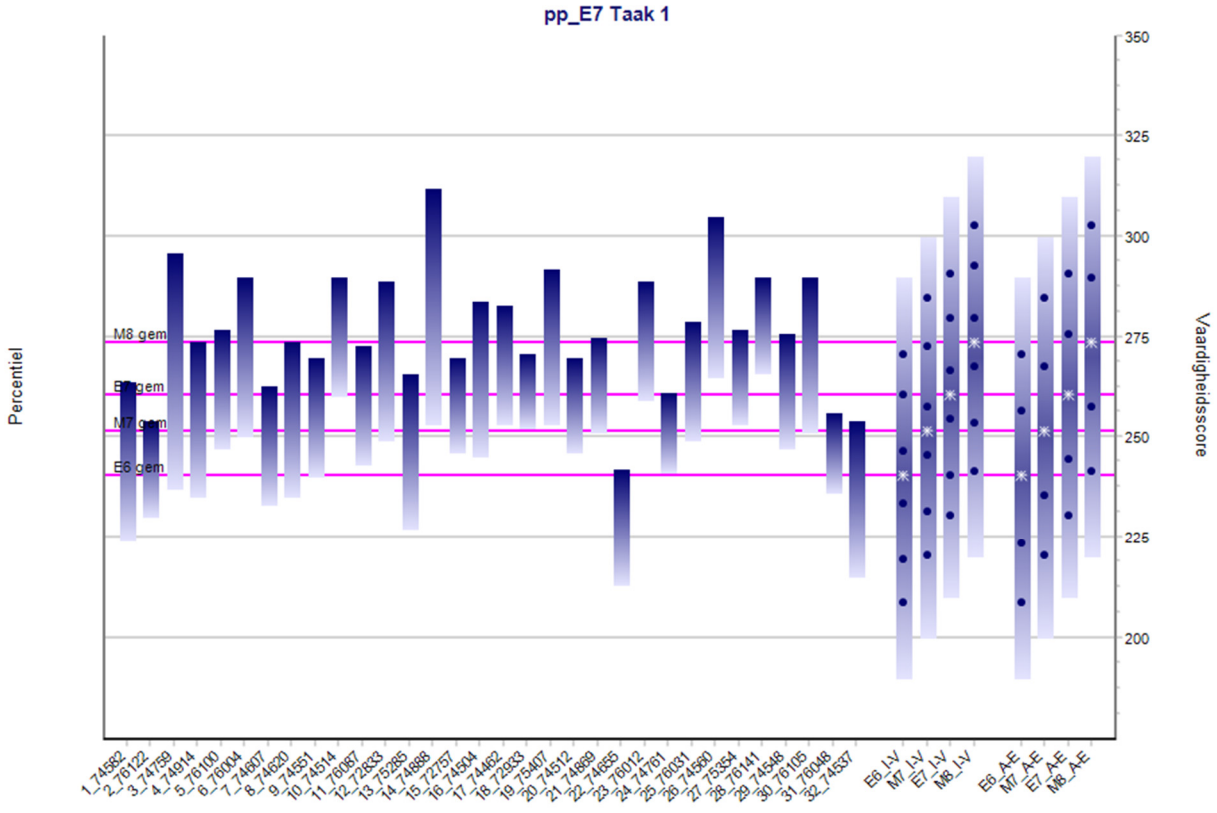
Bijlagen

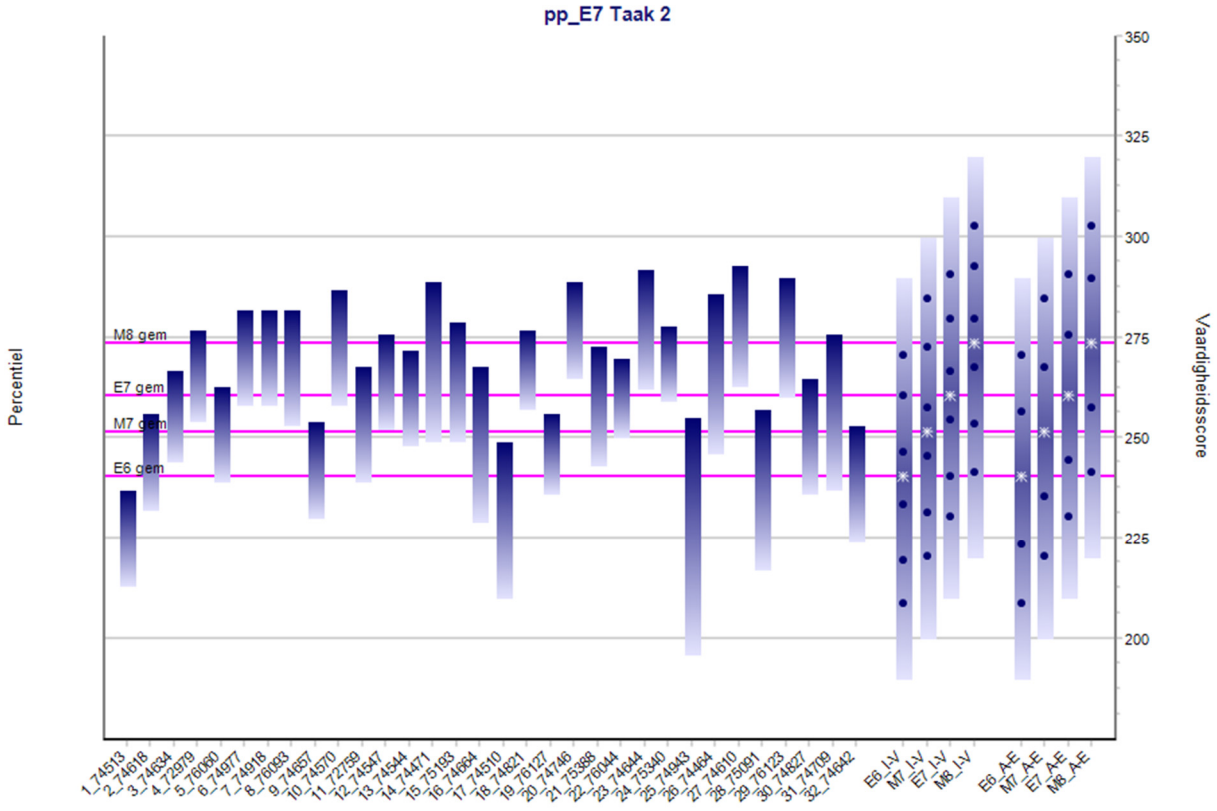
Bijlage 1 p50 en p80-kanspunten van de opgaven in de papieren toetsen en digitale toetsen M7 en E7 in relatie tot de vaardigheidsverdelingen van E6, M7, E7 en M8

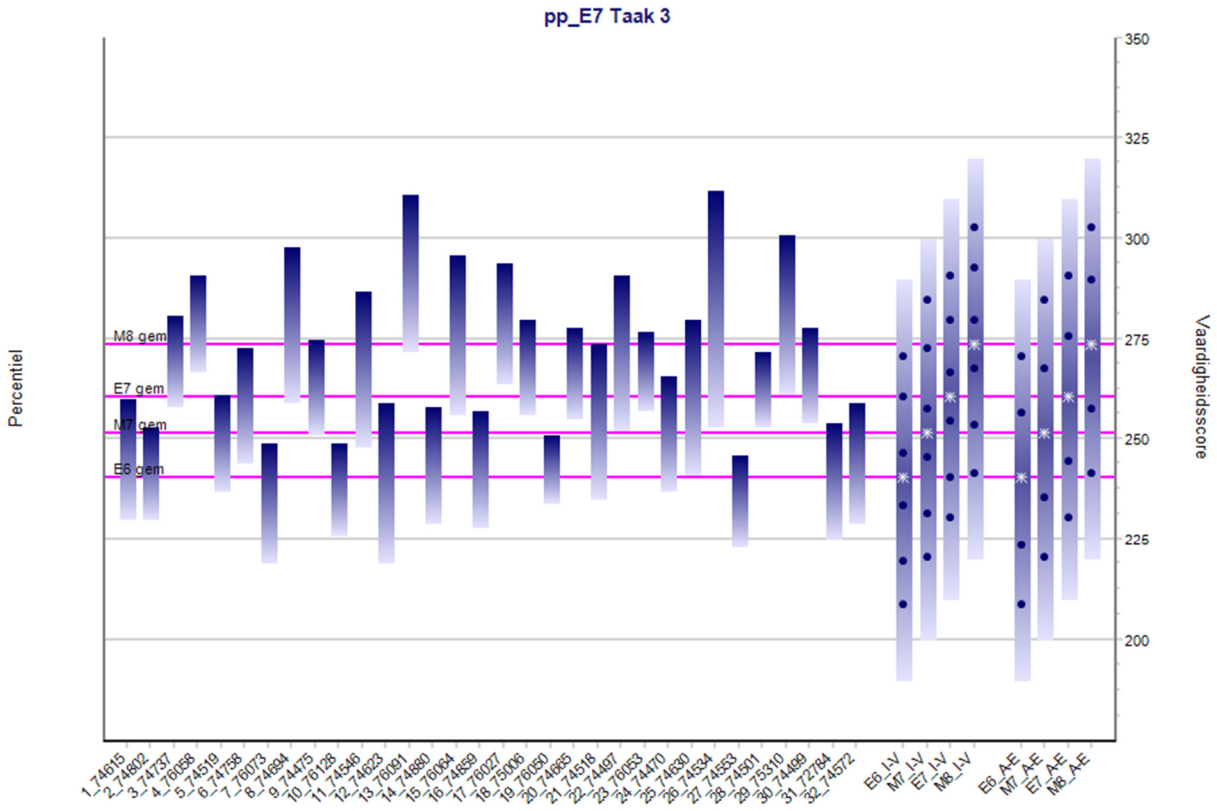


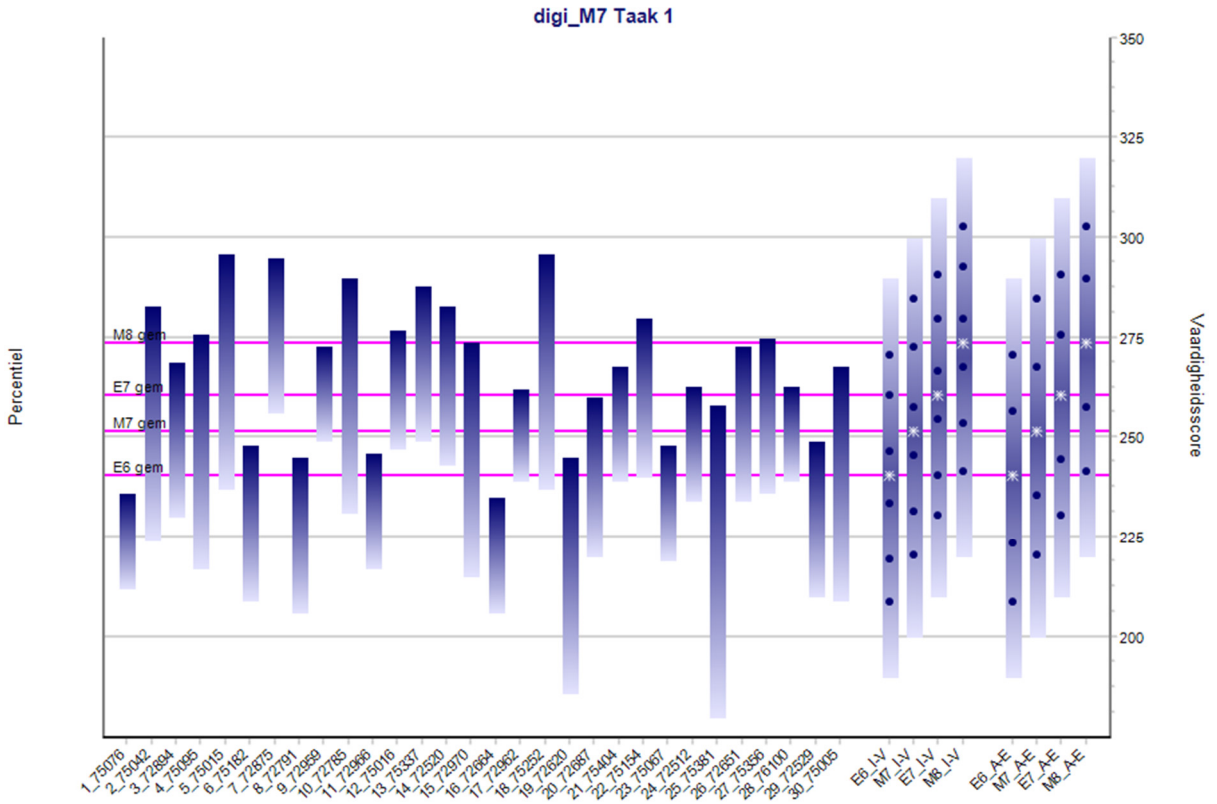


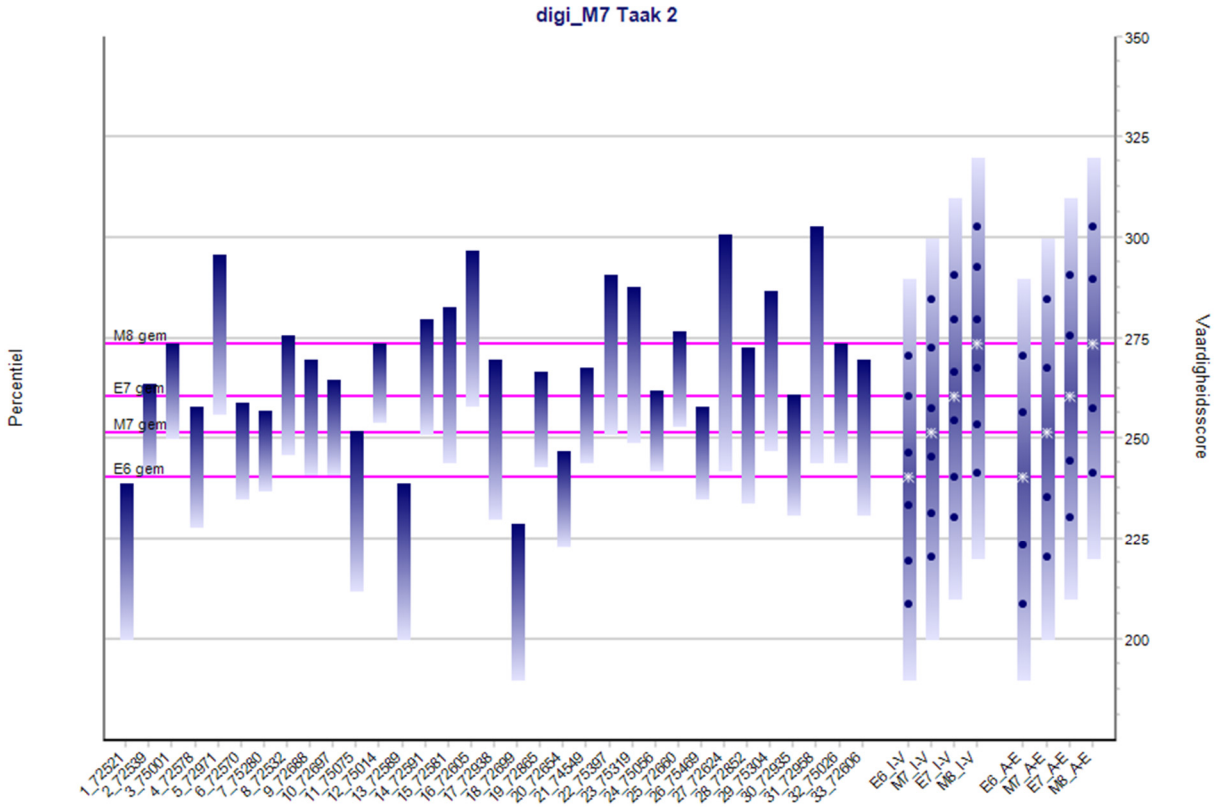

pp_M7 Taak 3
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_75225
2_72582
3_75026
4_75400
5_72722
6_72967
7_72578
8_72605
9_72712
10_74875
11_75381
12_72822
13_75018
14_72921
15_75412
16_75212
17_72546
18_72670
19_75469
20_75075
21_75409
22_72624
23_72680
24_72654
25_72911
26_72754
27_75397
28_72697
29_72772
30_75109
31_72589
32_75173
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E

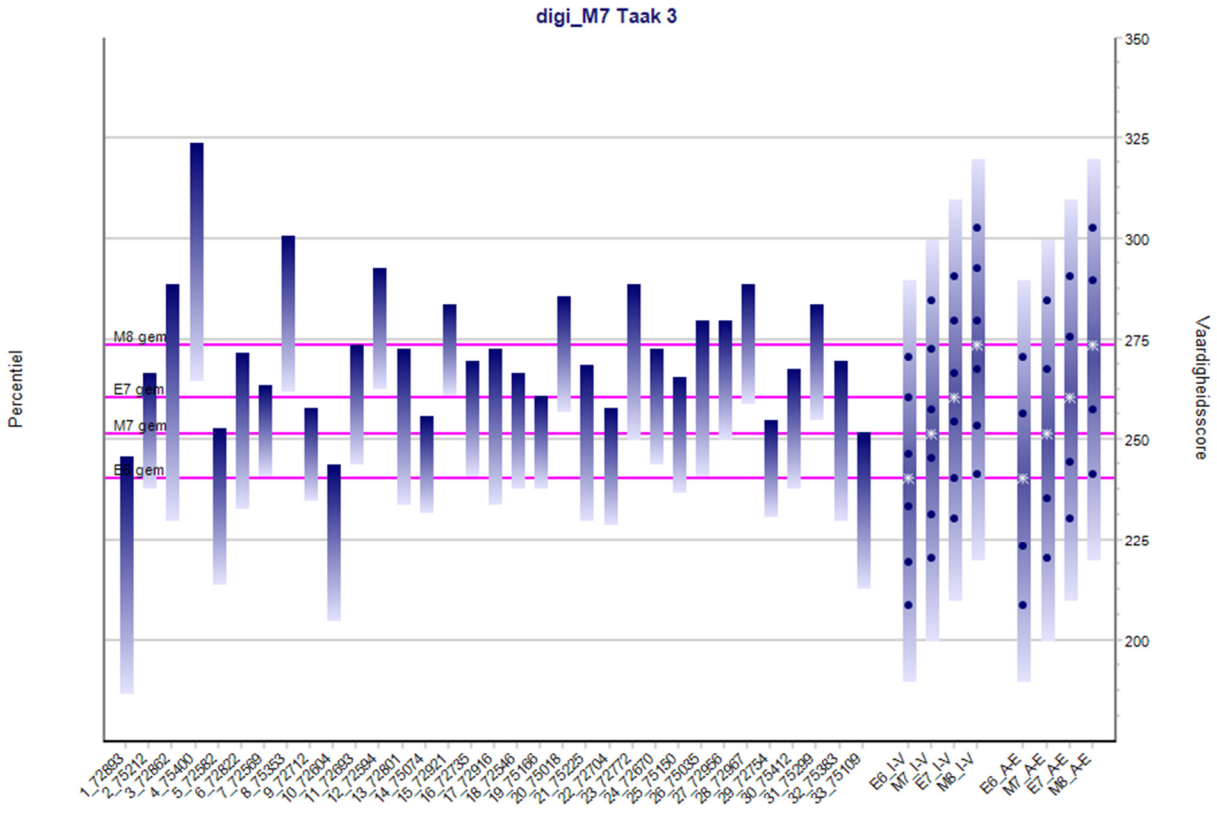

pp_E7 Taak 1
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74582
2_76122
3_74759
4_74914
5_76100
6_76004
7_74607
8_74620
9_74551
10_74514
11_76087
12_72833
13_75285
14_74888
15_72757
16_74504
17_74462
18_72933
19_75407
20_74512
21_74869
22_74655
23_76012
24_74761
25_76031
26_74560
27_75354
28_76141
29_74548
30_76105
31_76048
32_74537
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E

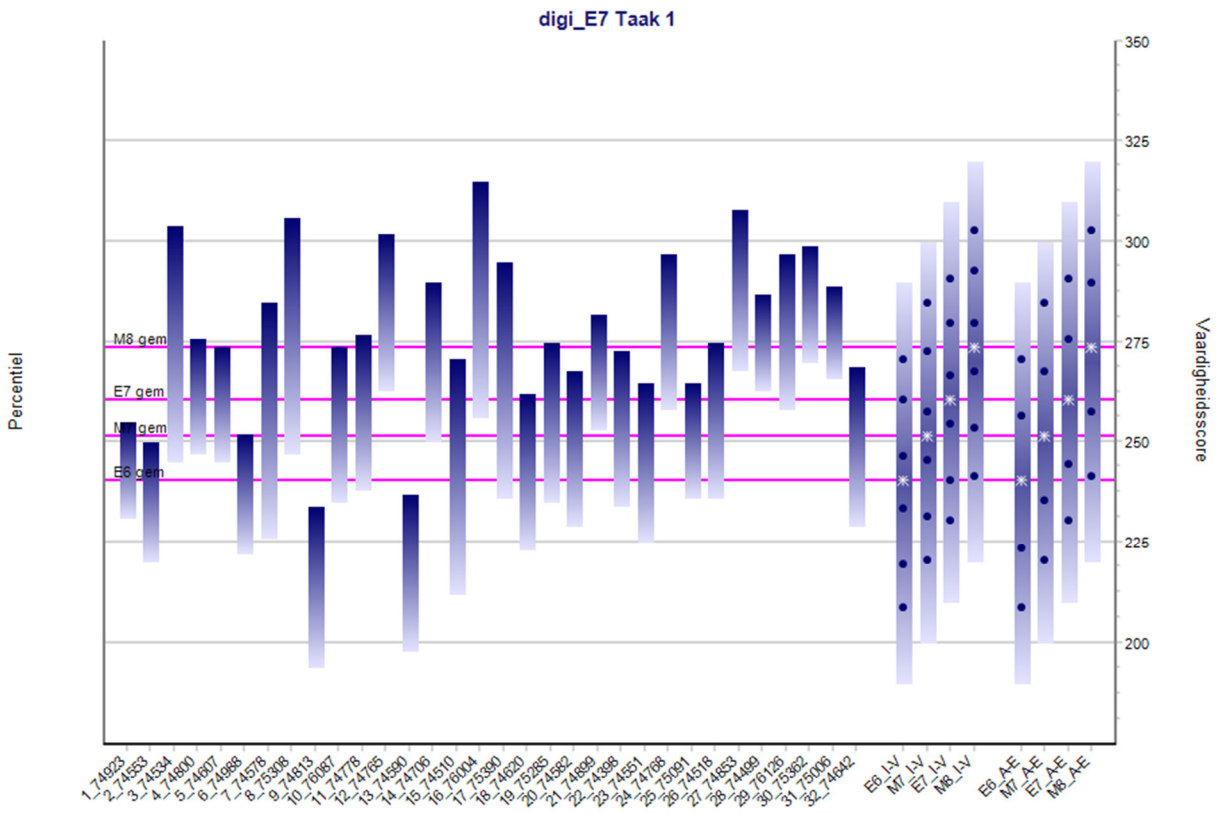

pp_E7 Taak 2
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74513
2_74618
3_74634
4_72979
5_76060
6_74977
7_74918
8_76083
9_74657
10_74570
11_72759
12_74547
13_74544
14_74471
15_75193
16_74664
17_74510
18_74821
19_76127
20_74746
21_75388
22_76044
23_74644
24_75340
25_74943
26_74464
27_74610
28_75091
29_76123
30_74827
31_74709
32_74642
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E

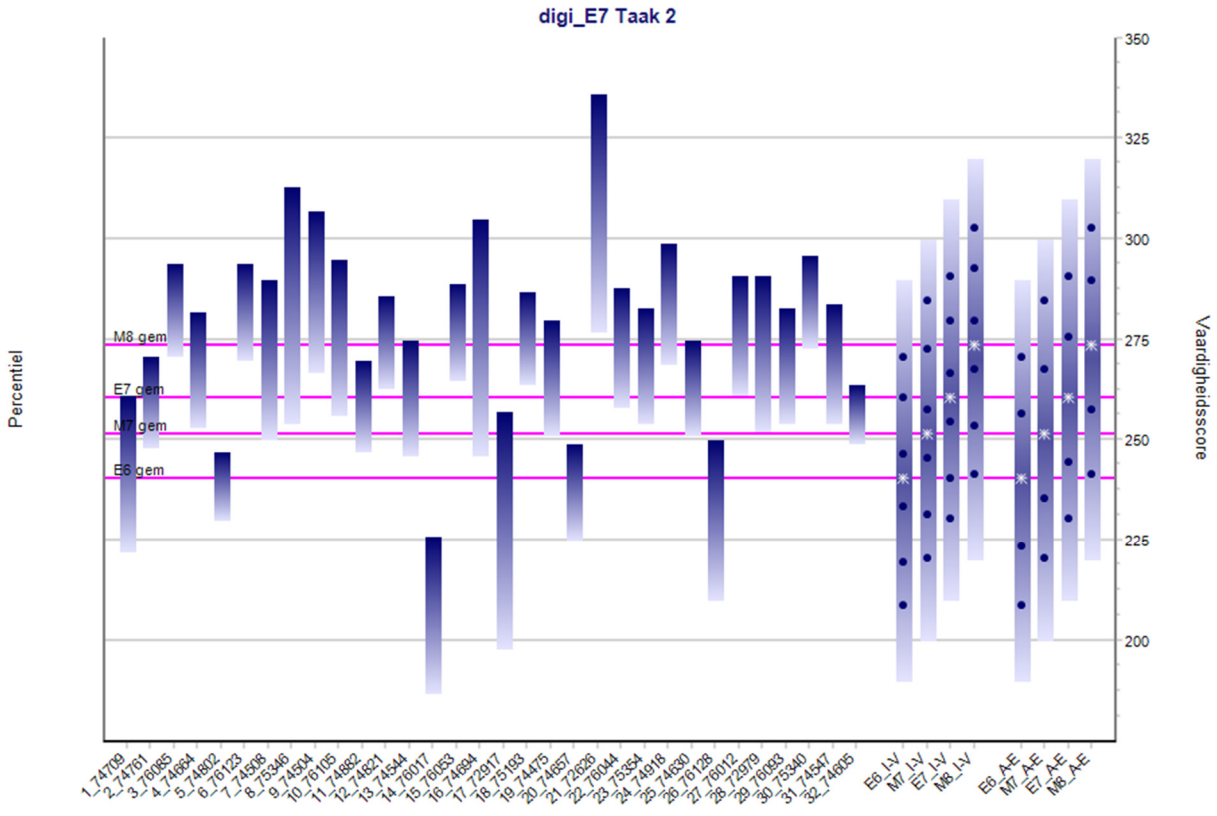

pp_E7 Taak 3
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74615
2_74802
3_74737
4_76058
5_74519
6_74758
7_76073
8_74694
9_74475
10_76128
11_74546
12_74623
13_76091
14_74680
15_76064
16_74859
17_76027
18_75006
19_76050
20_74665
21_74518
22_74497
23_76053
24_74470
25_74630
26_74534
27_74553
28_74591
29_75310
30_74499
31_72784
32_74572
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E

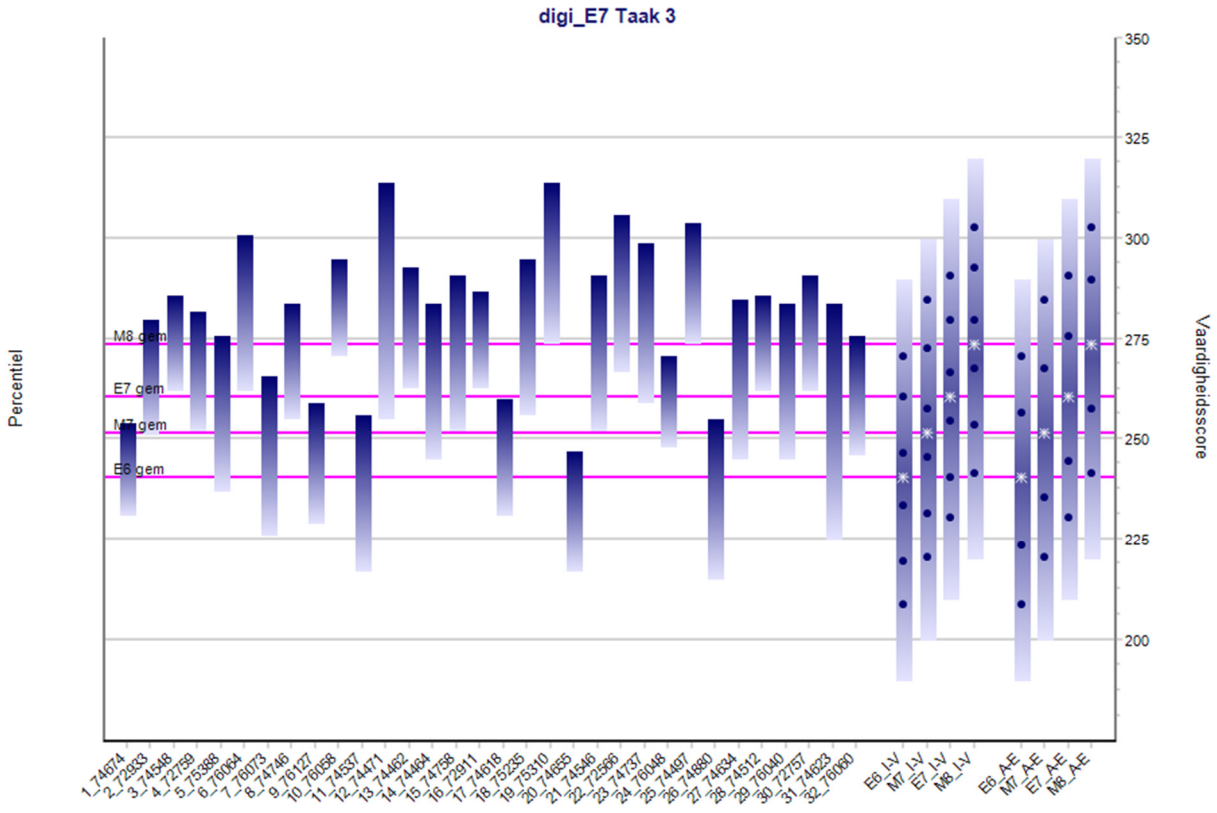

digi_M7 Taak 1
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_75076
2_75042
3_72684
4_75095
5_75015
6_75182
7_72675
8_72791
9_72959
10_72785
11_72966
12_75016
13_75337
14_72520
15_72970
16_72664
17_72962
18_75252
19_72620
20_72687
21_75404
22_75154
23_75067
24_72512
25_75381
26_72651
27_75356
28_76100
29_72529
30_75005
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E


digi_M7 Taak 2
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_72521
2_72539
3_75001
4_72576
5_72971
6_72570
7_75280
8_72532
9_72686
10_72697
11_75075
12_75014
13_72569
14_72591
15_72581
16_72695
17_72698
18_72699
19_72865
20_72654
21_74549
22_75397
23_75319
24_75456
25_72860
26_75469
27_72624
28_72852
29_75304
30_72935
31_72968
32_75026
33_72806
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E


digi_M7 Taak 3
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_72893
2_75212
3_72862
4_75400
5_72582
6_72822
7_72569
8_75353
9_72712
10_72604
11_72683
12_72594
13_72801
14_75074
15_72921
16_72735
17_72916
18_72546
19_75166
20_75018
21_75225
22_72704
23_72772
24_72570
25_75150
26_75035
27_72956
28_72967
29_72754
30_75412
31_75299
32_75383
33_75109
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E


digi_E7 Taak 1
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74923
2_74553
3_74534
4_74800
5_74607
6_74988
7_74578
8_75308
9_74813
10_76087
11_74778
12_74765
13_74590
14_74706
15_74510
16_76004
17_75390
18_74620
19_75285
20_74582
21_74699
22_74396
23_74551
24_74768
25_75091
26_74518
27_74653
28_74499
29_76126
30_75362
31_75006
32_74642
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E


digi_E7 Taak 2
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74709
2_74761
3_76085
4_74664
5_74802
6_76123
7_74508
8_75346
9_74594
10_76105
11_74882
12_74821
13_74544
14_76017
15_76053
16_74694
17_72917
18_75193
19_74475
20_74657
21_72626
22_76044
23_75354
24_74918
25_74630
26_76128
27_76012
28_72979
29_76093
30_75340
31_74547
32_74605
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E


digi_E7 Taak 3
Percentiel
Vaardigheidsscore
M8 gem
E7 gem
M7 gem
E6 gem
350
325
300
275
250
225
200
1_74674
2_72933
3_74548
4_72759
5_75388
6_76064
7_76073
8_74746
9_76127
10_76058
11_74537
12_74471
13_74462
14_74464
15_74758
16_72911
17_74616
18_75235
19_75310
20_74655
21_74546
22_72586
23_74737
24_76048
25_74497
26_74680
27_74634
28_74512
29_76040
30_72757
31_74623
32_76060
E6_I-V
M7_I-V
E7_I-V
M8_I-V
E6_A-E
M7_A-E
E7_A-E
M8_A-E

Bijlage 2 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M7

Item	Blok	Domein	taak 1	taak 2	taak 3	taak 4	taak 5	taak 6	taak 7	taak 8	taak 9	taak 10	taak 11	taak 12
M7	1	GB	3	3										
M7	2	GB		3	3									
M7	3	GB			3	3								
M7	4	GB				3	3							
M7	5	GB					3	3						
M7	6	GB						3	3					
M7	7	GB							3	3				
M7	8	GB								3	3			
M7	9	GB										3	3	
M7	10	GB	3											3
M7	11	GB	3		3									
M7	12	GB		3		3								
M7	13	GB					3		3					
M7	14	GB						3		3				
M7	15	GB	3			3								
M7	16	GB		3			3							
M7	17	GB			2			2						
M7	17	ME			1			1						
M7	18	ME			3		3							
M7	19	ME				3		3						
M7	20	ME							3		3			
M7	21	ME								3		3		
M7	22	ME									3		3	
M7	23	ME										3		3
M7	24	ME	3											
M7	25	ME		3									3	
M7	26	ME				3								3
M7	27	ME					3							3
M7	28	ME	3					3						
M7	29	ME		3					3					
M7	30	ME			3						3			
M7	31	VH	3				3							
M7	32	VH		3				3						
M7	33	VH			3				3					
M7	34	VH				3						3		
M7	35	VH							3				3	
M7	36	VH								3				3
M7	37	VH	3								3			
M7	38	VB		3					3					
M7	39	VB			3					3				
M7	40	VB				3					3			
M7	41	VB					3					3		
M7	42	VB						3					3	
M7	43	VB					3							3
M7	44	VB	3							3				
M7	45	KL	3									3		
M7	46	KL		3									3	
M7	47	KL			3									3
M7	48	KL				3			3					

M7	49	KL					3			3				
M7	50	KL						3			3			
E6	51	GB												4
E6	52	GB											4	
E6	53	GB										4		
E6	54	GB									4			
E6	55	ME								3				
E6	56	ME							3					
E6	57	ME					3							
E6	58	VH				3								
E6	59	VB			3									
E6	60	KL		3										
E7	61	GB												4
E7	62	GB											4	
E7	63	GB										4		
E7	64	GB									4			
E7	65	ME											3	
E7	66	ME										3		
E7	67	ME				3								
E7	68	VH											3	
E7	69	VB						3						
E7	70	KL												3
RS	71	GB							3					
RS	72	GB								3				
RS	73	ME	3											
RS	74	ME		3										
RS	75	VH			3									
RS	76	VH						3						
RS	77	VB									3			
RS	78	VB										3		
totaal			33	33	33	33	33	33	33	33	32	32	32	32

RS = Referentieset

GB = Getallen

ME = Meten en meetkunde

VH = Verhoudingen

VB = Verbanden

KL = Kaal

Bijlage 3 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek E7

item	blok	domein	taak 1	taak 2	taak 3	taak 4	taak 5	taak 6	taak 7	taak 8	taak 9	taak 10	Taak 11	Taak 12
E7	1	GB										3		3
E7	2	GB									3			3
E7	3	GB								1			1	
E7	4	GB							1				1	
E7	5	GB											1	1
E7	6	GB							3				3	
E7	7	GB						3			3			
E7	8	GB					3			3				
E7(M7)	9	GB				4				4				
E7(M7)	10	GB			4				4					
E7(M7)	11	GB		4				4						
E7(M7)	12	GB		4			4							
E7(M8)	13	GB	4											4
E7(M8)	14	GB	4										4	
E7(M8)	15	GB				4						4		
E7(M8)	16	GB			4							4		
M7	17	GB							4					
M7	18	GB								4				
M7	19	GB									4			
M7	20	GB										4		
M8	21	GB			4									
M8	22	GB				4								
M8	23	GB					4							
M8	24	GB						4						
RSA	25	GB	3											
RSA	26	GB		3										
E7	27	VH		4						4				
E7	28	VH			4						4			
E7	29	VH	3										3	
E7	30	VH				3								3
E7	31	VH					3						3	
E7(M7)	32	VH						3						3
E7(M8)	33	VH							4			4		
M7	34	VH						3						
M8	35	VH	3											
RSA	36	VH				3								
RSA	37	VH					3							
E7	38	ME	3	3										
E7	39	ME	2								2			
E7	40	ME									1		1	
E7	41	ME					1				1			
E7	42	ME					4	4						
E7	43	ME							4	4				
E7	44	ME			3	3								
E7	45	ME		3	3									
E7	46	ME				3						3		
E7	47	ME		1								1		
E7(M7)	48	ME								3	3			
E7(M7)	49	ME						3			3			

E7(M7)	50	ME							3					3
E7(M8)	51	ME										4		4
E7(M8)	52	ME					4						4	
M7	53	ME											3	
M7	54	ME												3
M7	55	ME	3											
M8	56	ME		3										
M8	57	ME							3					
M8	58	ME								3				
RSA	59	ME			3									
RSA	60	ME						3						
E7	61	VB				4	4							
E7	62	VB						4	4					
E7	63	VB											3	3
E7	64	VB	3	3										
E7	65	VB			3								3	
E7(M7)	66	VB				3					3			
E7(M8)	67	VB								4		4		
M7	68	VB		3										
M8	69	VB									3			
RSA	70	VB	3											
RSA	71	VB			3									
E7	72	KL							4				4	
E7	73	KL					4			4				
E7	74	KL	3	3										
E7(M7)	75	KL				3		3						
E7(M8)	76	KL									4			4
M7	77	KL												
M8	78	KL			3							3		
Totaal			34	34	34	34	34	34	34	34	34	34	34	34

E7(M7) = E7 items ook in normeringsonderzoek M7

M7 = M7 items ook in normeringsonderzoek E7

E7(M8) = E7 items ook in normeringsonderzoek M8

M8 = M8 items ook in normeringsonderzoek E7

GB = getallen

ME = Meten en meetkunde

VB = Verbanden

VH = Verhoudingen

KL = Kaal

Bijlage 4 Klassieke en IRT-indices van de opgaven in de M7 en E7 papieren en digitale toetsen

M7 papier

nr	P-Val	RIT	RIR	disc	diff	se
1	0,835	0,384	0,368	3	-0,443	0,038
2	0,752	0,430	0,412	2	-0,641	0,034
3	0,658	0,457	0,438	4	-0,097	0,024
4	0,532	0,385	0,363	4	-0,280	0,037
5	0,758	0,348	0,329	4	-0,279	0,033
6	0,679	0,453	0,434	4	-0,131	0,030
7	0,519	0,284	0,259	3	-0,837	0,055
8	0,632	0,379	0,357	3	-0,310	0,039
9	0,650	0,459	0,439	2	-0,693	0,028
10	0,648	0,275	0,252	3	-0,406	0,042
11	0,587	0,466	0,446	2	-0,223	0,030
12	0,596	0,383	0,361	2	-0,116	0,039
13	0,612	0,464	0,444	4	-0,125	0,029
14	0,576	0,384	0,362	2	-0,617	0,033
15	0,606	0,530	0,511	3	-0,320	0,025
16	0,567	0,385	0,362	3	-0,838	0,039
17	0,696	0,449	0,430	3	-0,236	0,028
18	0,570	0,467	0,446	3	-0,448	0,031
19	0,671	0,272	0,249	4	-0,092	0,058
20	0,850	0,480	0,466	2	-0,387	0,028
21	0,687	0,370	0,348	3	-0,312	0,041
22	0,601	0,581	0,564	3	-0,050	0,022
23	0,423	0,455	0,434	3	-0,325	0,028
24	0,668	0,521	0,503	3	-0,308	0,026
25	0,713	0,363	0,342	4	-0,137	0,038
26	0,546	0,532	0,513	3	-0,120	0,021
27	0,710	0,445	0,426	3	-0,454	0,023
28	0,719	0,507	0,489	3	-0,051	0,028
29	0,682	0,371	0,349	4	0,012	0,041
30	0,657	0,458	0,438	3	-0,106	0,031
31	0,429	0,456	0,436	3	-0,637	0,032
32	0,795	0,333	0,314	2	-0,602	0,035
33	0,672	0,373	0,351	3	-0,152	0,039
34	0,648	0,459	0,439	3	-0,586	0,026
35	0,659	0,523	0,504	3	-0,591	0,022
36	0,448	0,380	0,358	4	-0,234	0,036
37	0,584	0,466	0,446	2	-0,476	0,030
38	0,472	0,463	0,442	4	-0,438	0,028
39	0,744	0,353	0,333	4	-0,713	0,041
40	0,551	0,385	0,363	4	-0,269	0,034
41	0,739	0,500	0,483	3	-0,194	0,028
42	0,551	0,467	0,447	4	-0,075	0,028
43	0,673	0,372	0,351	4	-0,422	0,030
44	0,565	0,467	0,446	4	-0,442	0,028
45	0,586	0,466	0,446	3	-0,344	0,030
46	0,552	0,385	0,363	3	-0,405	0,033
47	0,479	0,463	0,443	4	-0,116	0,027
48	0,775	0,483	0,467	3	-0,260	0,029
49	0,801	0,469	0,453	3	-0,125	0,024
50	0,512	0,580	0,562	3	-0,423	0,022
51	0,636	0,379	0,357	3	-0,408	0,034

52	0,665	0,521	0,503	4	-0,346	0,026
53	0,618	0,381	0,359	3	-0,125	0,034
54	0,556	0,532	0,513	3	-0,312	0,025
55	0,664	0,374	0,353	3	0,005	0,039
56	0,563	0,622	0,607	3	-0,095	0,020
57	0,543	0,385	0,363	2	-0,287	0,037
58	0,690	0,268	0,246	4	-0,307	0,061
59	0,798	0,408	0,391	5	-0,222	0,027
60	0,609	0,464	0,444	3	0,058	0,026
61	0,665	0,374	0,353	2	-0,189	0,030
62	0,612	0,464	0,444	4	-0,240	0,029
63	0,860	0,294	0,277	3	0,101	0,039
64	0,794	0,410	0,393	3	-0,118	0,027
65	0,730	0,358	0,337	4	-0,165	0,043
66	0,775	0,420	0,402	4	-0,935	0,025
67	0,681	0,453	0,433	3	-0,169	0,024
68	0,539	0,531	0,513	4	-0,163	0,021
69	0,646	0,276	0,252	3	-0,472	0,059
70	0,455	0,523	0,504	4	-0,197	0,026
71	0,750	0,431	0,413	3	-0,132	0,030
72	0,542	0,385	0,363	3	-0,004	0,038
73	0,726	0,505	0,487	4	-0,318	0,023
74	0,456	0,460	0,440	4	-0,264	0,029
75	0,751	0,351	0,331	4	-0,340	0,045
76	0,691	0,369	0,348	4	-0,457	0,042
77	0,478	0,383	0,360	6	-0,158	0,037
78	0,454	0,522	0,504	2	-0,929	0,022
79	0,701	0,564	0,547	3	0,038	0,018
80	0,661	0,457	0,437	3	-0,523	0,029
81	0,578	0,467	0,446	3	-0,450	0,027
82	0,611	0,464	0,444	3	-0,302	0,028
83	0,522	0,467	0,446	3	-0,483	0,030
84	0,807	0,235	0,215	3	-0,423	0,052
85	0,583	0,384	0,362	3	-0,355	0,036
86	0,606	0,382	0,360	2	-0,350	0,033
87	0,542	0,385	0,363	3	-0,554	0,038
88	0,803	0,468	0,451	5	-0,344	0,023
89	0,448	0,521	0,502	4	-0,285	0,026
90	0,760	0,427	0,409	6	-0,174	0,033
91	0,592	0,281	0,257	3	-0,400	0,055
92	0,652	0,459	0,439	4	-0,092	0,029
93	0,552	0,532	0,513	5	-0,422	0,021
94	0,771	0,248	0,227	7	-0,071	0,063
95	0,832	0,312	0,294	5	-0,181	0,036
96	0,714	0,363	0,342	4	-0,702	0,032

E7 papier

nr	P-Val	RIT	RIR	disc	diff	se
1	0,752	0,337	0,317	3	-0,106	0,044
2	0,790	0,461	0,445	2	-0,038	0,045
3	0,625	0,267	0,243	4	0,042	0,056
4	0,684	0,356	0,335	4	0,020	0,051
5	0,620	0,448	0,427	4	0,162	0,042
6	0,576	0,370	0,348	4	0,198	0,051
7	0,737	0,421	0,403	3	-0,002	0,031
8	0,687	0,355	0,334	3	0,015	0,040
9	0,682	0,437	0,418	2	0,078	0,045
10	0,502	0,450	0,429	3	0,313	0,031
11	0,657	0,443	0,422	2	0,114	0,041
12	0,584	0,369	0,347	2	0,185	0,051
13	0,739	0,341	0,321	4	-0,081	0,055
14	0,540	0,272	0,247	2	0,229	0,057
15	0,649	0,509	0,491	3	0,149	0,027
16	0,618	0,367	0,345	3	0,132	0,039
17	0,565	0,452	0,431	3	0,233	0,040
18	0,603	0,567	0,550	3	0,212	0,022
19	0,555	0,371	0,348	4	0,231	0,037
20	0,649	0,509	0,491	2	0,149	0,039
21	0,595	0,515	0,497	3	0,210	0,033
22	0,864	0,349	0,333	3	-0,243	0,039
23	0,508	0,450	0,430	3	0,305	0,039
24	0,718	0,545	0,528	3	0,087	0,022
25	0,602	0,450	0,429	4	0,186	0,031
26	0,461	0,367	0,344	3	0,375	0,057
27	0,578	0,516	0,497	3	0,229	0,033
28	0,436	0,504	0,485	3	0,385	0,036
29	0,626	0,447	0,427	4	0,155	0,031
30	0,570	0,370	0,348	3	0,208	0,055
31	0,765	0,526	0,510	3	0,029	0,024
32	0,806	0,314	0,295	2	-0,222	0,043
33	0,895	0,373	0,359	3	-0,242	0,044
34	0,776	0,469	0,452	3	-0,016	0,037
35	0,675	0,504	0,486	3	0,118	0,035
36	0,569	0,516	0,497	4	0,240	0,034
37	0,718	0,493	0,475	2	0,064	0,039
38	0,524	0,515	0,496	4	0,289	0,025
39	0,521	0,515	0,496	4	0,292	0,032
40	0,568	0,451	0,431	4	0,230	0,039
41	0,790	0,462	0,445	3	-0,037	0,039
42	0,523	0,451	0,430	4	0,286	0,042
43	0,694	0,434	0,415	4	0,062	0,041
44	0,585	0,516	0,497	4	0,222	0,024
45	0,627	0,513	0,494	3	0,175	0,027
46	0,582	0,370	0,347	3	0,188	0,054
47	0,604	0,449	0,429	4	0,184	0,032
48	0,726	0,345	0,325	3	-0,055	0,064
49	0,830	0,301	0,283	3	-0,281	0,075
50	0,535	0,567	0,550	3	0,280	0,021
51	0,761	0,527	0,512	3	0,034	0,034
52	0,444	0,506	0,486	4	0,376	0,024
53	0,655	0,443	0,423	3	0,116	0,046

54	0,616	0,566	0,549	3	0,199	0,031
55	0,482	0,448	0,427	3	0,337	0,030
56	0,517	0,566	0,548	3	0,298	0,032
57	0,806	0,225	0,205	2	-0,441	0,100
58	0,605	0,368	0,346	4	0,152	0,040
59	0,472	0,447	0,426	5	0,350	0,039
60	0,792	0,321	0,302	3	-0,189	0,042
61	0,498	0,450	0,429	2	0,317	0,042
62	0,717	0,428	0,409	4	0,028	0,049
63	0,672	0,358	0,337	3	0,041	0,039
64	0,803	0,391	0,374	3	-0,114	0,033
65	0,757	0,414	0,395	4	-0,035	0,047
66	0,793	0,460	0,443	4	-0,043	0,042
67	0,528	0,516	0,497	3	0,285	0,033
68	0,426	0,502	0,483	4	0,396	0,024
69	0,734	0,487	0,469	3	0,042	0,025
70	0,653	0,443	0,423	4	0,119	0,041
71	0,830	0,375	0,358	3	-0,167	0,047
72	0,509	0,370	0,348	3	0,302	0,037
73	0,597	0,515	0,496	4	0,208	0,027
74	0,823	0,440	0,425	4	-0,092	0,048
75	0,597	0,369	0,346	4	0,165	0,057
76	0,781	0,325	0,306	4	-0,167	0,045
77	0,411	0,360	0,338	6	0,453	0,057
78	0,771	0,408	0,389	2	-0,058	0,031
79	0,531	0,371	0,348	3	0,268	0,050
80	0,777	0,405	0,387	3	-0,068	0,047
81	0,460	0,446	0,425	3	0,364	0,029
82	0,543	0,516	0,497	3	0,268	0,026
83	0,794	0,550	0,536	3	0,012	0,021
84	0,561	0,517	0,497	3	0,249	0,033
85	0,685	0,356	0,334	3	0,019	0,052
86	0,565	0,370	0,348	2	0,216	0,050
87	0,535	0,567	0,550	3	0,281	0,033
88	0,711	0,430	0,410	5	0,036	0,033
89	0,647	0,363	0,341	4	0,083	0,041
90	0,541	0,272	0,247	6	0,228	0,071
91	0,843	0,426	0,411	3	-0,127	0,047
92	0,592	0,568	0,550	4	0,224	0,029
93	0,492	0,369	0,347	5	0,327	0,049
94	0,565	0,517	0,497	7	0,245	0,024
95	0,796	0,396	0,378	5	-0,101	0,033
96	0,767	0,410	0,391	4	-0,050	0,033

M7 digitaal

nr	P-Val	RIT	RIR	disc	diff	se
1	0,853	0,427	0,412	5	-0,255	0,038
2	0,652	0,275	0,251	2	-0,121	0,066
3	0,660	0,375	0,353	3	-0,039	0,046
4	0,685	0,270	0,246	2	-0,200	0,067
5	0,579	0,283	0,258	2	0,046	0,069
6	0,790	0,334	0,314	3	-0,290	0,052
7	0,462	0,383	0,359	3	0,276	0,046
8	0,808	0,325	0,306	3	-0,332	0,055
9	0,533	0,531	0,512	5	0,181	0,030
10	0,612	0,280	0,256	2	-0,028	0,064
11	0,783	0,414	0,396	4	-0,184	0,043
12	0,544	0,468	0,446	4	0,161	0,037
13	0,524	0,386	0,363	3	0,180	0,044
14	0,567	0,385	0,362	3	0,112	0,048
15	0,696	0,268	0,244	2	-0,227	0,068
16	0,857	0,365	0,349	4	-0,336	0,047
17	0,634	0,526	0,507	5	0,065	0,031
18	0,583	0,282	0,258	2	0,036	0,069
19	0,811	0,233	0,213	2	-0,560	0,078
20	0,723	0,359	0,338	3	-0,153	0,050
21	0,622	0,463	0,442	4	0,060	0,038
22	0,589	0,384	0,361	3	0,078	0,048
23	0,770	0,421	0,402	4	-0,161	0,041
24	0,660	0,456	0,436	4	0,008	0,037
25	0,785	0,136	0,114	1	-1,103	0,160
26	0,629	0,380	0,357	3	0,013	0,047
27	0,625	0,380	0,358	3	0,019	0,047
28	0,630	0,526	0,507	5	0,070	0,040
29	0,783	0,337	0,317	3	-0,274	0,051
30	0,719	0,263	0,240	2	-0,286	0,072
31	0,833	0,311	0,292	3	-0,394	0,056
32	0,616	0,528	0,509	5	0,086	0,030
33	0,516	0,531	0,511	5	0,199	0,029
34	0,706	0,445	0,425	4	-0,059	0,037
35	0,465	0,383	0,360	3	0,272	0,044
36	0,674	0,518	0,499	5	0,016	0,032
37	0,669	0,570	0,553	6	0,040	0,027
38	0,544	0,468	0,446	4	0,161	0,036
39	0,600	0,465	0,444	4	0,088	0,036
40	0,606	0,529	0,510	5	0,097	0,032
41	0,771	0,342	0,322	3	-0,249	0,055
42	0,473	0,576	0,558	6	0,244	0,025
43	0,831	0,312	0,293	3	-0,389	0,058
44	0,510	0,467	0,445	4	0,205	0,037
45	0,563	0,385	0,362	3	0,119	0,048
46	0,454	0,382	0,359	3	0,288	0,046
47	0,660	0,375	0,353	3	-0,039	0,046
48	0,870	0,284	0,268	3	-0,503	0,066
49	0,593	0,530	0,511	5	0,112	0,031
50	0,773	0,482	0,465	5	-0,118	0,036
51	0,584	0,531	0,512	5	0,123	0,039
52	0,505	0,385	0,362	3	0,208	0,044
53	0,520	0,386	0,363	3	0,186	0,044

54	0,610	0,580	0,562	6	0,104	0,026
55	0,487	0,528	0,509	5	0,232	0,029
56	0,671	0,519	0,500	5	0,020	0,033
57	0,557	0,284	0,259	2	0,093	0,069
58	0,637	0,379	0,356	3	-0,001	0,047
59	0,535	0,386	0,363	3	0,163	0,044
60	0,686	0,451	0,430	4	-0,029	0,039
61	0,546	0,284	0,259	2	0,117	0,066
62	0,568	0,467	0,446	4	0,130	0,036
63	0,653	0,376	0,354	3	-0,027	0,047
64	0,804	0,236	0,215	2	-0,536	0,084
65	0,630	0,462	0,441	4	0,048	0,038
66	0,618	0,280	0,255	2	-0,041	0,070
67	0,431	0,280	0,255	2	0,366	0,069
68	0,767	0,344	0,324	3	-0,240	0,054
69	0,640	0,378	0,356	3	-0,006	0,049
70	0,620	0,528	0,509	5	0,081	0,030
71	0,426	0,378	0,355	3	0,332	0,046
72	0,675	0,518	0,499	5	0,015	0,033
73	0,811	0,323	0,304	3	-0,340	0,054
74	0,568	0,467	0,446	4	0,130	0,036
75	0,396	0,451	0,430	4	0,351	0,035
76	0,631	0,380	0,357	3	0,010	0,045
77	0,701	0,511	0,493	5	-0,018	0,034
78	0,408	0,514	0,495	5	0,321	0,031
79	0,600	0,465	0,444	4	0,088	0,036
80	0,637	0,379	0,356	3	-0,001	0,047
81	0,626	0,462	0,441	4	0,054	0,038
82	0,642	0,525	0,506	5	0,055	0,031
83	0,452	0,461	0,440	4	0,278	0,035
84	0,660	0,375	0,353	3	-0,039	0,046
85	0,703	0,446	0,426	4	-0,055	0,038
86	0,520	0,386	0,363	3	0,187	0,047
87	0,579	0,467	0,445	4	0,116	0,038
88	0,637	0,461	0,440	4	0,040	0,036
89	0,586	0,384	0,361	3	0,083	0,046
90	0,518	0,467	0,446	4	0,194	0,037
91	0,432	0,458	0,437	4	0,304	0,038
92	0,712	0,508	0,489	5	-0,032	0,031
93	0,626	0,462	0,441	4	0,054	0,038
94	0,471	0,463	0,442	4	0,254	0,037
95	0,665	0,374	0,352	3	-0,047	0,048
96	0,768	0,343	0,323	3	-0,243	0,050

E7 digitaal

nr	P-Val	RIT	RIR	disc	diff	se
1	0,784	0,457	0,439	5	-0,029	0,043
2	0,823	0,373	0,355	4	-0,154	0,054
3	0,582	0,272	0,246	2	0,138	0,076
4	0,626	0,446	0,424	4	0,155	0,043
5	0,642	0,443	0,421	4	0,133	0,043
6	0,813	0,380	0,361	4	-0,134	0,054
7	0,680	0,260	0,235	2	-0,088	0,080
8	0,571	0,272	0,246	2	0,161	0,076
9	0,889	0,255	0,238	3	-0,461	0,083
10	0,684	0,355	0,332	3	0,020	0,059
11	0,667	0,358	0,335	3	0,049	0,056
12	0,480	0,372	0,347	3	0,346	0,056
13	0,877	0,265	0,248	3	-0,418	0,075
14	0,576	0,371	0,347	3	0,197	0,056
15	0,743	0,245	0,222	2	-0,250	0,085
16	0,522	0,274	0,248	2	0,267	0,075
17	0,633	0,267	0,241	2	0,024	0,078
18	0,762	0,331	0,309	3	-0,127	0,062
19	0,684	0,355	0,332	3	0,021	0,057
20	0,726	0,343	0,321	3	-0,056	0,058
21	0,572	0,452	0,429	4	0,225	0,044
22	0,691	0,353	0,330	3	0,008	0,049
23	0,748	0,336	0,314	3	-0,097	0,059
24	0,516	0,373	0,348	3	0,291	0,056
25	0,717	0,424	0,403	4	0,028	0,048
26	0,679	0,356	0,333	3	0,030	0,057
27	0,439	0,368	0,344	3	0,409	0,052
28	0,468	0,513	0,492	5	0,350	0,037
29	0,516	0,373	0,348	3	0,291	0,056
30	0,414	0,443	0,421	4	0,424	0,044
31	0,439	0,509	0,489	5	0,381	0,035
32	0,723	0,344	0,322	3	-0,050	0,059
33	0,769	0,328	0,307	3	-0,140	0,061
34	0,632	0,509	0,489	5	0,168	0,037
35	0,390	0,500	0,479	5	0,436	0,034
36	0,570	0,452	0,429	4	0,227	0,042
37	0,832	0,508	0,494	7	-0,041	0,036
38	0,395	0,501	0,481	5	0,430	0,034
39	0,575	0,371	0,347	3	0,199	0,053
40	0,533	0,274	0,248	2	0,245	0,075
41	0,446	0,369	0,345	3	0,398	0,053
42	0,532	0,373	0,348	3	0,265	0,053
43	0,644	0,507	0,487	5	0,155	0,037
44	0,473	0,514	0,493	5	0,344	0,035
45	0,633	0,445	0,423	4	0,146	0,045
46	0,912	0,232	0,217	3	-0,551	0,091
47	0,449	0,511	0,490	5	0,370	0,037
48	0,580	0,272	0,246	2	0,142	0,076
49	0,800	0,227	0,205	2	-0,421	0,092
50	0,462	0,512	0,492	5	0,357	0,035
51	0,590	0,450	0,428	4	0,202	0,044
52	0,826	0,430	0,413	5	-0,096	0,046
53	0,414	0,269	0,243	2	0,501	0,068

54	0,520	0,453	0,430	4	0,290	0,042
55	0,560	0,452	0,430	4	0,240	0,044
56	0,415	0,443	0,421	4	0,421	0,042
57	0,598	0,514	0,493	5	0,208	0,035
58	0,826	0,300	0,281	3	-0,271	0,066
59	0,491	0,452	0,429	4	0,326	0,041
60	0,564	0,372	0,347	3	0,216	0,053
61	0,561	0,452	0,430	4	0,239	0,041
62	0,367	0,494	0,474	5	0,462	0,036
63	0,555	0,452	0,430	4	0,246	0,041
64	0,649	0,629	0,613	8	0,182	0,028
65	0,785	0,456	0,438	5	-0,031	0,043
66	0,588	0,450	0,428	4	0,205	0,042
67	0,480	0,514	0,494	5	0,337	0,037
68	0,576	0,451	0,429	4	0,220	0,042
69	0,672	0,358	0,334	3	0,041	0,059
70	0,491	0,372	0,348	3	0,330	0,052
71	0,740	0,339	0,317	3	-0,082	0,061
72	0,553	0,452	0,430	4	0,248	0,044
73	0,767	0,405	0,385	4	-0,050	0,049
74	0,384	0,498	0,478	5	0,443	0,036
75	0,795	0,317	0,296	3	-0,197	0,063
76	0,529	0,274	0,248	2	0,252	0,075
77	0,470	0,450	0,428	4	0,352	0,042
78	0,618	0,367	0,343	3	0,131	0,057
79	0,564	0,372	0,347	3	0,216	0,053
80	0,468	0,513	0,492	5	0,350	0,030
81	0,755	0,410	0,390	4	-0,031	0,049
82	0,534	0,373	0,348	3	0,263	0,056
83	0,393	0,362	0,338	3	0,482	0,052
84	0,841	0,361	0,343	4	-0,190	0,056
85	0,564	0,372	0,347	3	0,216	0,056
86	0,448	0,369	0,345	3	0,395	0,045
87	0,506	0,373	0,348	3	0,305	0,052
88	0,632	0,509	0,489	5	0,168	0,037
89	0,372	0,435	0,413	4	0,478	0,043
90	0,802	0,313	0,293	3	-0,214	0,067
91	0,613	0,367	0,343	3	0,139	0,054
92	0,480	0,514	0,494	5	0,337	0,037
93	0,618	0,367	0,343	3	0,131	0,057
94	0,487	0,452	0,429	4	0,332	0,044
95	0,687	0,258	0,234	2	-0,104	0,086
96	0,631	0,445	0,423	4	0,148	0,043

Cito helpt je inzicht te krijgen in je ontwikkeling en mogelijkheden. Door kennis, vaardigheden en competenties objectief meetbaar te maken en de ontwikkeling er van te volgen, kun je het beste uit jezelf halen, verantwoorde keuzes maken en beter richting geven aan je toekomst. Cito draagt daaraan bij door wereldwijd werk te maken van goed en eerlijk toetsen, vanuit de kernwaarden kundig, toonaangevend, integer, innovatief en betrokken.

Cito
Amsterdamseweg 13
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11
www.cito.nl

Fotografie: Ron Steemers