

Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 8

Michel Hop, Floor Scheltens en Ronald Engelen

cito.nl



Wetenschappelijke verantwoording

Rekenen-Wiskunde 3.0 voor groep 8

Michel Hop
Floor Scheltens
Ronald Engelen

© Cito B.V. Arnhem (2019)

Niets uit dit werk mag zonder voorafgaande schriftelijke toestemming van Cito worden openbaar gemaakt en/of verveelvoudigd door middel van druk, fotokopie, scanning, computersoftware of andere elektronische verveelvoudiging of openbaarmaking, microfilm, geluidskopie, film- of videokopie of op welke wijze dan ook.

Inhoud

1	Inleiding	5
2	Uitgangspunten van de toetsconstructie	7
2.1	Meetpretentie	7
2.2	Doelgroep	7
2.3	Gebruiksdoel en functie	7
2.4	Theoretische inkadering	9
2.4.1	Inhoudelijk	9
2.4.2	Psychometrisch	12
2.4.2.1	Opgavenbanken en constructieprocedures	12
2.4.2.2	Het gehanteerde meetmodel	14
3	Beschrijving van de toetsen	19
3.1	Opbouw en structuur van de toetsen	19
3.2	Inhoudsverantwoording	23
3.3	Statistische beschrijving	31
4	Kalibratie en normering	33
4.1	Opzet normeringsonderzoeken LVS: het macro design	33
4.1.1	Opzet normeringsonderzoek B8	34
4.2	De kalibratie	35
4.2.1	De opzet van de kalibratie	35
4.2.2	De stappen in de kalibratie	35
4.2.3	Toetsing van het IRT-model	36
4.3	De normering	40
4.3.1	Opzet	40
4.3.2	Representativiteit	43
4.3.3	Normeringsresultaten	45
5	Betrouwbaarheid en meetnauwkeurigheid	47
5.1	Methoden om de betrouwbaarheid te bepalen	47
5.2	Betrouwbaarheid: resultaten	47
5.3	Lokale betrouwbaarheid en meetnauwkeurigheid	48
6	Validiteit	53
6.1	Inhoudsvaliditeit	53
6.2	Unidimensionaliteit, respectievelijk structuur	53
6.3	Itemkwaliteit	55
6.4	Itembias	55
6.5	Soortgenootonderzoek	56
6.6	Verschillen tussen relevante subgroepen	57
7	Samenvatting	61
8	Literatuur	63

Bijlagen 67

- 1 p50- en p80-kanspunten van de opgaven in de papieren toetsen en digitale toetsen B8/M8 in relatie tot de vaardigheidsverdelingen 68
- 2 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M8 72
- 3 Klassieke en IRT-indices van de opgaven in de papieren en digitale toetsen B8/M8 74

1 Inleiding

Deze wetenschappelijke verantwoording heeft betrekking op de toetsen Rekenen-Wiskunde 3.0 voor groep 8 uit het Cito Volgstelsysteem primair en speciaal onderwijs. De toetsen van het volgstelsysteem 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
- Kwaliteit van het toetsmateriaal
- Kwaliteit van de handleiding
- Normen
- Betrouwbaarheid
- Begripsvaliditeit

Criteriumvaliditeit is voor de toetsen Rekenen-Wiskunde 3.0 groep 8 van het Cito Volgstelsysteem 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 B8/M8 voor de afnames op papier
- Afnamekaarten en nakijkkaarten voor de afnames op papier
- Digitale toetsen B8/M8 voor afnames achter de computer

De toets heeft een normering voor twee afnamemomenten, aan het begin van leerjaar 8 en halverwege leerjaar 8. De leerkracht of school bepaalt zelf op welk van de momenten de toets wordt afgenomen. 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 Volgstelsysteem kan het toetsen op maat gerealiseerd worden.

De nieuwe toetsen rekenen-wiskunde voor groep 8 worden de toetsen Rekenen-Wiskunde 3.0 voor groep 8 uit het Cito Volgstelsysteem 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 Volgstelsysteem 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 8 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 verbanden gelegd worden tussen het onderwijs in rekenen-wiskunde en de alledaagse leefwereld. Verder moeten leerlingen eenvoudige verbanden, regels, patronen en structuren kunnen opsporen. 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 Volgstelsel primair en speciaal onderwijs. De toetsen Rekenen-Wiskunde 3.0 groep 8 zijn bedoeld voor leerlingen in groep 8 van het primair onderwijs en leerlingen in het speciaal (basis)onderwijs die functioneren op het niveau van groep 8 in het reguliere basisonderwijs. De toetsen zijn ook te gebruiken voor leerlingen in andere leerjaren die een rekenvaardigheid hebben op het niveau van groep 8. 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. Daartoe bevat de toetsmap extra toetsboekjes en zijn alternatieve rapportages ontwikkeld.

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 Volgstelsel 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 unidimensionale 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 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 E7 had Meriam vaardigheidsniveau V en op tijdstip M8 was het vaardigheidsniveau IV".

Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 238, bijvoorbeeld, op tijdstip E7 en vaardigheidsscore 256 op tijdstip M8. 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 bij ad a. als bij 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 vaardigheidsniveau 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 elkaar overlappen. Als deze twee BI's elkaar niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze elkaar 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 E7 behaalde Wout een vaardigheidsscore van 277 met een 67% betrouwbaarheidsinterval van 271-283. Bij de afname M8 behaalde Wout een vaardigheidsscore van 283; het bijbehorende betrouwbaarheidsinterval daarbij is 277-288. Aangezien de betrouwbaarheidsintervallen elkaar 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 vaardigheidsscores op achtereenvolgende meettijdstippen zijn betrekkelijk klein. 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; TAL-team 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 8 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 8 (zie verder paragraaf 3.2).

Kerdoelen 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, maatzicht, 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. In groep 8 is eveneens het werken met de rekenmachine relevant.

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 voorkomen.

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
Rekenmachine*						X

*facultatief

Voor groep 8 zijn, net als al in groep 6 en 7 het geval was, vier domeinen van belang, namelijk de domeinen Getallen, Verhoudingen, Meten en meetkunde en Verbanden. Optioneel is in de toets B8/M8 een taak Rekenmachine opgenomen. Scholen kunnen ervoor kiezen deze taak wel of niet te maken. Er zijn dan ook 2 normeringen beschikbaar, een voor de afname zonder rekenmachine, en een voor de afname met rekenmachine. Nadere informatie over de invulling van deze onderdelen in groep 8 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 gemakkelijkerheid. 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 geproefde 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 Item Respons Theorie (IRT). 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 8 groter zal zijn dan in groep 7, 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 'papierene' 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 aselechte 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 7 wordt voorgelegd, en zijn vaardigheidsschatting kan behalve met de populatie van groep 8 ook vergeleken worden met de percentielen in de populatie van groep 7, met bijvoorbeeld de uitspraak: "De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 7.";
- 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 8 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 Item Respons Theorie (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 populatieschattingen 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 toetskarakteristieken 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 $f_i(\theta)$ is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 | \theta) \quad (2.1)$$

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

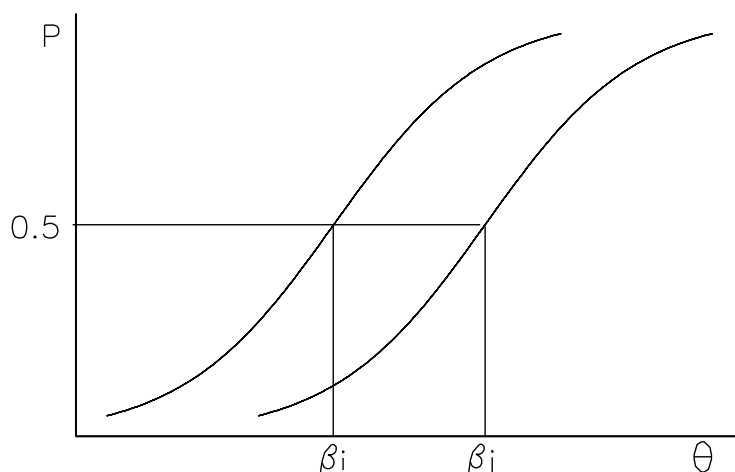
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad (2.2)$$

waarin β_i de moeilijkheidsparameter van item i is. Dat is een onbekende grootte 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

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

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 moeilijkheidsparameter 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.

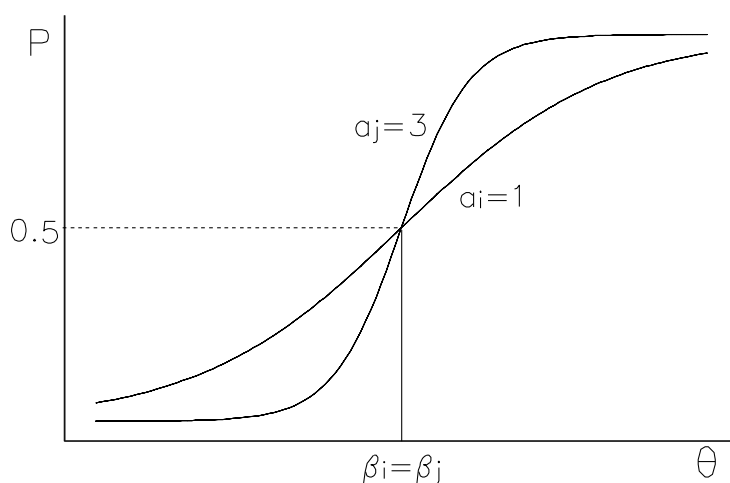
Alvorens het hier gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid als CML). Die maakt gebruik van het feit dat in het Raschmodel een afdoende steekproefgrootte (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

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

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 dataverzamingsprocedure 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 aselechte 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 8 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 8 uit het Cito Volgsysteem primair en speciaal onderwijs bevat in totaal een papieren toets B8/M8 en een digitale toets B8/M8. De toets B8/M8 is genormeerd op de afnamemomenten B8 en M8. Beide toetsen hebben een variant met en zonder rekenmachine. Voor groep 8 zijn geen tussentoetsen ontwikkeld. Aan een leerling waarbij de rekenvaardigheid zich minder snel ontwikkelt kan begin groep 8 of medio groep 8 één van de toetsen van groep 6 of 7 voorgelegd worden.

Opbouw

In totaal bevat de B8/M8 toets 96 opgaven. Daarnaast is er een optionele taak Rekenmachine beschikbaar. De Rekenmachinetaak bevat 20 opgaven. De reguliere toets B8/M8 bestaat uit drie 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 8 bevatten naast meerkeuzeopgaven ook open opgaven. De opgaven van de toets B8/M8 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 computerscherm 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 M8-toets voor leerlingen medio groep 8) 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 E7) 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.

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 een 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 dan ook 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%		
20% laagst scorende leerlingen	V 20%	D 15%	15% ruim onder het landelijk gemiddelde
		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. 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 M8 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling heel dicht ligt bij de score van de gemiddelde leerling medio groep 8. 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 functionerings-niveaus 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 8 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. Wanneer de rekenmachinetaak in de afname is meegenomen wordt ook deze in de categorieënanalyse meegenomen. De categorie Rekenmachine is opgenomen als extra categorie naast Getallen, Verhoudingen, Meten en Verbanden.

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 scores, 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 scores 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 die 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 fasen 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 8. 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-waarden, zijn afgebeeld. Achter de portal staan de doorgaande lijnen met de balkjes behorende bij de vaardigheidsniveaus. 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 8. Deze onderwerpen komen aan bod in de meest gebruikte methodes voor rekenen-wiskunde in groep 8. 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-wiskundemethode van groep 8 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 8 bij die onderwerpen globaal aan bod komt. De toetsen voor groep 8 bevatten opgaven uit vier domeinen: 1. Getallen, 2. Verhoudingen, 3. Meten en 4. Verbanden. Daarnaast is er een optionele taak Rekenmachine met opgaven waarbij de rekenmachine gebruikt wordt. In deze taak komen de opgaven ook uit de vier genoemde domeinen. Bij de verschillende onderwerpen binnen die domeinen zijn vervolgens doelen opgesteld die aansluiten bij de leerlijnen en leerstof van leerlingen in groep 8.

In het volgende overzicht staat per onderwerp aangegeven welke leerstof in de toetsen voor groep 8 aan bod komt.

Uitwerking doelen voor toetsen groep 8

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) en 50 (vanaf een veelvoud van 25), 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 $49 \times 198,95$ is ongeveer 50×200 .

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.
- 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.
- 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 ($1 \text{ m}^2 = 10\,000 \text{ cm}^2$ en $1 \text{ hectare} = 100 \text{ are}$). 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 8 op een afnamemoment zijn opgaven geconstrueerd die een operationalisering vormen van die doelen.

Voor een beschrijving van de inhoud van de toetsen B8/M8 verwijzen we ook naar de Inhoudsverantwoording in het toetspakket (Cito, 2018). 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 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 B8/M8. Rekenmachine-opgaven zijn optioneel.

	B8/M8
Getallen	40
Verhoudingen	16
Meten	28
Verbanden	12
Rekenmachine	20

Tabel 3.1b Verdeling opgaven over domeinen van de digitale toetsen van de uitgave Rekenen-Wiskunde groep B8/M8. Rekenmachine-opgaven zijn optioneel.

	B8/M8
Getallen	40
Verhoudingen	16
Meten	28
Verbanden	12
Rekenmachine	20

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 een 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 moeten 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 2012 halverwege van leerjaar 8 nieuw ontwikkelde opgaven voorgelegd aan leerlingen. Daarbij zijn voor het afnamemoment medio groep 8 ongeveer 210 nieuwe opgaven geproeftoetst.

Elke deelnemende school maakte een taak met ongeveer 32 opgaven. Voor elke opgave zijn ongeveer 100 à 150 responsen verzameld.

In 2014 is opnieuw een proeftoetsing voor leerjaar 8 uitgevoerd. Voor M8 zijn toen ongeveer 370 opgaven geproeftoetst. In totaal waren er na de proeftoetsingen voor M8 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 afname-moment (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.

Normeringsonderzoek B8

Op basis van signalen uit het veld is in latere instantie besloten om ook een B8-normering vast te stellen. Scholen bleken er behoefte aan te hebben om al in oktober/november in groep 8 te toetsen. We hebben daarom besloten om alsnog (in oktober 2018) een B8-normeringsonderzoek uit te voeren. Op deze manier hebben we gezorgd voor één toets Rekenen-Wiskunde voor groep 8 die naar wens in oktober/november of in januari/februari af te nemen is. Hoe het normeringsonderzoek B8 is uitgevoerd, staat beschreven in hoofdstuk 4.

Het afnamedesign voor het normeringsonderzoek voor afnamemoment medio groep 8 (M8) bestond uit in totaal 14 opgavenboekjes met 30 tot 33 opgaven. Alle leerlingen maakten de bestaande M8-toets van LVS II (3 taken en rekenmachinetaak) en een taak met nieuw geconstrueerde opgaven voor M8 inclusief ankeropgaven en 12 opgaven rekenmachine. De ankeropgaven waren opgaven van het eerdere afnamemoment E7 en opgaven uit de eigen referentieset. Voor het afnamedesign geldt dat alle nieuwe opgaven in twee taken zijn opgenomen om deze onderling te kunnen verbinden. Deze taken werden zo over de opgavenboekjes verdeeld dat deze koppeling mogelijk was. De ankeropgaven zijn slechts in één taak opgenomen.

In het normeringsonderzoek M8 zijn op deze wijze in totaal 198 verschillende nieuwe items voorgelegd aan 2206 leerlingen van groep 8 verdeeld over 14 boekjes. Elk boekje bestond uit 157 tot en met 160 opgaven verdeeld over 5 taken. De 96 opgaven (verdeeld over 3 taken) uit de bestaande LVS II-taken werden door alle leerlingen gemaakt. De rekenmachinetaak uit LVS II werd ook door alle leerlingen gemaakt (20 opgaven). In de nieuwe taken kwamen de E7-ankeropgaven één keer voor. Vrijwel elke nieuwe opgave voor medio groep 8 kwam in twee boekjes voor. De nieuwe opgaven werden minimaal door 151 leerlingen leerling gemaakt. De positie van de nieuwe taak in de boekjes varieerde.

Tabel 3.2 Design medio groep 8 met aantallen opgaven en aantallen leerlingen per boekje (B1-B14)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
M8 nieuwe opgaven(incl. rekenmachine)	32	25	33	31	27	29	27	24	25	20	22	15	36	36
E7 anker	4	8	0	3	7	7	8	11	10	15	14	18	0	0
Refset	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Nieuwe taak totaal	32	31	30	31	31	32	31	31	32	32	32	31	32	32
LVS M8 gen II taak 1,2 3	96	96	96	96	96	96	96	96	96	96	96	96	96	96
Rekenmachinetaak gen II	20	20	20	20	20	20	20	20	20	20	20	20	20	20
Nieuwe opgaven rekenmachine	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Totaal aantal opgaven	160	157	157	158	158	160	159	159	159	159	160	157	160	160
Aantal leerlingen	172	151	150	115	171	176	116	141	162	175	174	168	160	175
Aantal scholen	6	6	8	7	8	7	6	6	6	7	8	7	6	7

De taak met nieuw geconstrueerde opgaven voor M8 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 8

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
getallen	12	11	15	13	13	13	14	14	12	16	15	14	23	23
verhoudingen	6	6	4	4	4	4	4	4	6	6	7	3	7	4
meten	7	7	7	10	10	11	7	7	11	7	7	7	2	5
verbanden	7	7	4	4	4	4	6	6	3	3	3	7	0	0
	32	31	30	31	31	32	31	31	32	32	32	31	32	32

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 B8/M8 (van generatie LVS 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 8. 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 de leerlingen van medio groep 8 zijn digitale taken met nieuwe opgaven afgenomen in combinatie met de papieren taken van generatie LVS II en in combinatie met de digitale taken van generatie LVS II. Bij de afnames voor medio groep 8 waren 6 groepen leerlingen betrokken. In de tabellen hieronder is te zien welke taken die groepen leerlingen gemaakt hebben.

Daarnaast is voor de link tussen de papieren en digitale toetsen gebruik gemaakt van dataretourgegevens van de afnamen digitaal M8 2019.

Tabel 3.6 Design medio groep 8 digitaal onderzoek (B2, B4 en B6) en papier-digitaal onderzoek (B1, B3 en B5)

	B1	B2	B3	B4	B5	B6
LVS M8 papier						
LVS M8 generatie III digitaal taak 1 (nieuw, 46 items)						
LVS M8 generatie III digitaal taak 2 (nieuw, 45 items)						
LVS M8 generatie III digitaal taak 3 (nieuw, 45 items)						
LVS M8 digitaal						
aantal leerlingen	69	52	19	48	170	19
aantal items	162	162	161	161	161	161
gemiddeld 63 waarnemingen voor elk nieuw digitaal item						
minimaal aantal waarnemingen 19						

Tabel 3.6a Design medio groep 8 digitaal onderzoek (B2, B4 en B6) en papier-digitaal onderzoek (B1, B3 en B5) aangevuld met dataretour

	B1	B2	B3	B4	B5	B6	Dataretour M8 1-3	Dataretour M8 1-4
LVS M8 papier								
LVS M8 generatie III digitaal taak 1 (nieuw, 46 items)								
LVS M8 generatie III digitaal taak 2 (nieuw, 45 items)								
LVS M8 generatie III digitaal taak 3 (nieuw, 45 items)								
LVS M8 digitaal								
aantal leerlingen	69	52	19	48	170	19	1829	546
aantal items	162	162	161	161	161	161	96	116
gemiddeld 1076 waarnemingen voor digitaal item in uitgave								
minimaal waarnemingen 306								

Na de afnames zijn de antwoorden van de leerlingen op de toetsen geanalyseerd met behulp van het programma-pakket OPLM (Verhelst, 1992; Verhelst, Glas en Verstralen, 1995) en is een 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 generatie LVS II en generatie LVS III voor papieren en digitale toetsen. In totaal zijn er twee toetsen samengesteld. Voor de papieren variant een toets B8/M8 en voor de digitale variant een toets B8/M8.

De toets B8/M8 van de papieren variant is samengesteld uit opgaven die in de nieuwe taken bij het normeringsonderzoek zaten. De toets B8/M8 van de digitale variant is 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 B8/M8, 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 B8/M8 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 B8/M8 ligt het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 8. 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 toetsen B8/M8, zowel op de ruwe scoreschaal als op de vaardigheidsschaal.

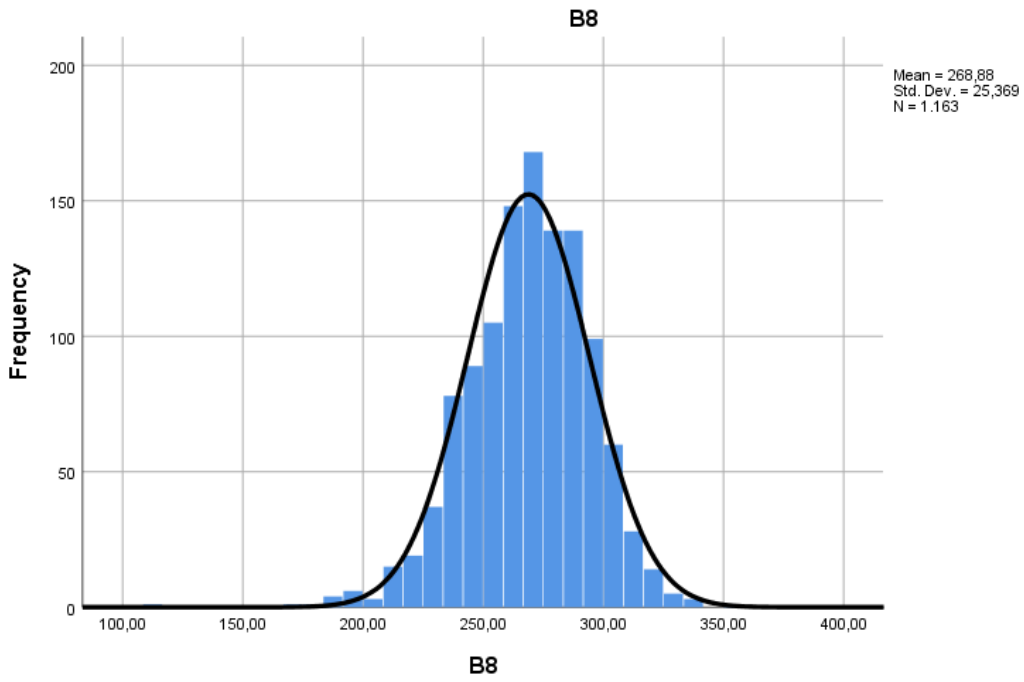
De gegevens zijn gebaseerd op de prestaties van de leerlingen in de normeringssteekproeven.

De waarden in tabel 3.8 en de figuren 3.1 en 3.2 laten zien dat de vaardigheidsverdelingen medio groep 8 bij benadering normaal verdeeld is.

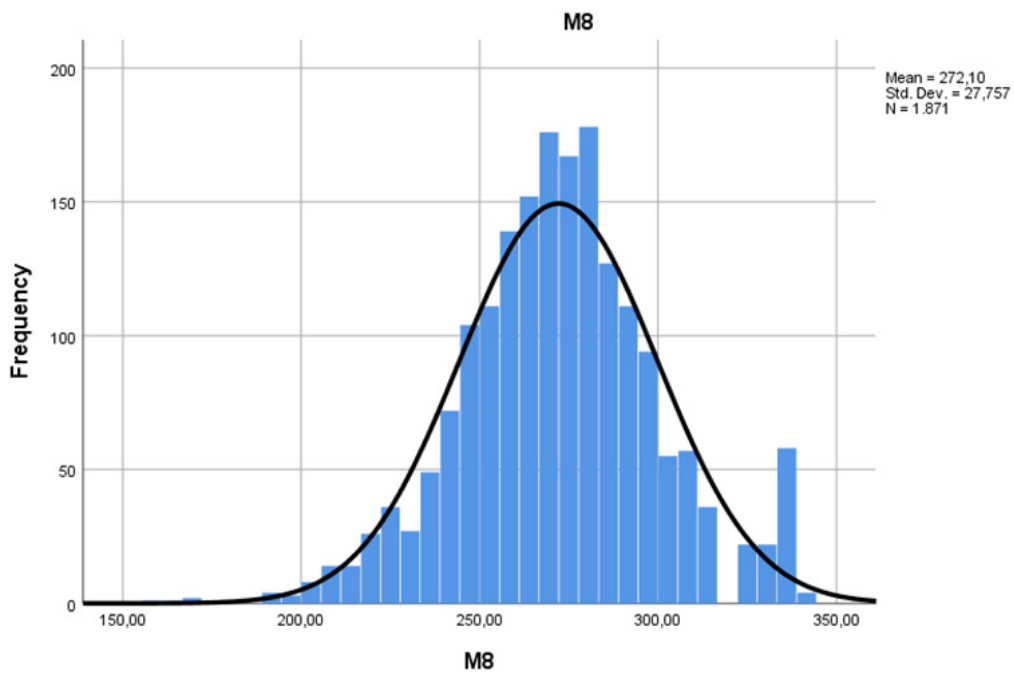
Tabel 3.8 Beschrijvende gegevens toets B8/M8 op de ruwe scoreschaal op afnamemoment B8 en M8

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
B8 papier	60,6	20,39	-0,64	-0,45
B8 papier met rekenmachinetaak	73,8	24,25	-0,61	-0,46
B8 digitaal	59,2	19,83	-0,69	-0,38
B8 digitaal met rekenmachinetaak	71,1	23,91	-0,70	-0,37
M8 papier	63,7	19,04	-0,56	-0,42
M8 papier met rekenmachinetaak	77,7	22,61	-0,58	-0,38
M8 digitaal	62,2	18,59	-0,50	-0,50
M8 digitaal met rekenmachinetaak	74,8	22,43	-0,49	-0,51

Figuur 3.1 Weergave van behaalde vaardigheidsscores B8



Figuur 3.2 Weergave van behaalde vaardigheidsscores M8



4 Kalibratie en normering

4.1 Opzet normeringsonderzoeken LVS: het macro-design

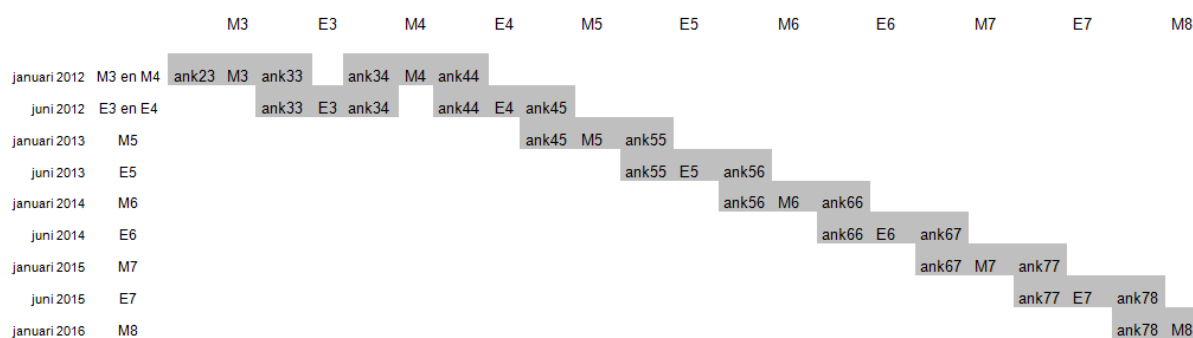
Het opzetten van een leerlingvolgsysteem in het basisonderwijs is een complexe onderneming, en het verzamelen van de gegevens om het systeem te ijken en normeren moet met de nodige zorg gebeuren. Immers, het is niet voldoende om voor elke halfjaargroep (M3, E3, M4, E4, M5, E5, M6, E6, M7, E7, M8) over 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 8 voor te leggen aan leerlingen van groep 3, omdat een dergelijke 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 dat 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 M8 betekent dit dat de leerlingen in het normeringsonderzoek items voorgelegd krijgen die specifiek voor M8 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor 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 twee 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 8 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 8 betreft het dus het gedeelte uit figuur 4.1, dat in figuur 4.2 is weergegeven.

Figuur 4.2 Design groep 8

januari 2015	Ank76	M7	Ank77			
juni 2015			Ank77	E7	Ank78	
januari 2016					Ank78	M8

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.

4.1.1 Opzet normeringsonderzoek B8

Zoals eerder beschreven, is pas in latere instantie besloten om ook een B8-normering vast te stellen. Scholen bleken er behoefte aan te hebben om al in oktober/november in groep 8 te toetsen. We hebben daarom besloten om alsnog (in oktober 2018) een B8-normeringsonderzoek uit te voeren.

Bij dit normeringsonderzoek maakten leerlingen van groep 8 de inmiddels al uitgegeven toets voor groep 8. Het design van het onderzoek voor B8 bestond dus uit de drie of vier taken (in totaal 96 of 116 items) van de definitieve toets voor groep 8. Deze toets werd in oktober 2018 afgenomen door scholen en de data werden aan Cito geleverd in de vorm van dataretour.

Om deelnemers te werven voor het normeringsonderzoek B8 zijn scholen aangeschreven die het Computerprogramma LOVS gebruiken. Er hadden zich 15 scholen opgegeven, met 515 leerlingen. Met behulp van dataretour hebben we de data kunnen aanvullen tot 934 leerlingen, omdat al meer scholen de uitgebrachte toets voor groep 8 hadden afgenomen in oktober. Voor de normering B8 worden ook de leerlingen meegenomen die een makkelijker toets (bijvoorbeeld E7) gemaakt hebben; dit zijn veelal de iets minder vaardige leerlingen.

In totaal bestaat de normeringssteekproef voor B8 dan uit 1163 leerlingen verdeeld over 41 scholen.

De opgaven in het normeringsonderzoek B8 waren al gekalibreerd. De leerlingen maakten immers de al uitgegeven toets voor groep 8. Alles wat in de volgende paragraaf staat over de kalibratie is van toepassing op de toetsen B8/M8, onafhankelijk van het afnamemoment.

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 8 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 achtergrondvariabelen, 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 8 (tabel 3.2). In het *embedded field* onderzoek bij de papieren afnames zijn in totaal 14 boekjes voor medio groep 8 afgenomen. Naast de vier 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 afnamemoment.

In bijlage 2: Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek B8/M8 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 kalibratieprocedure.

De statistische toetsen in OPLM hebben goede statistische en asymptotische eigenschappen aangezien het OPLM behoort tot de exponentiële familie, met de gewogen somscore:

$$s = \sum_{i=1}^k a_i x_i, \quad (4.1)$$

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.

Binnen deze groepen kan de verwachte proportie goede antwoorden op een item onder het model, $p(+|s)$, vergeleken worden met de feitelijk geobserveerde proportie goede antwoorden, $prop(+|s)$. Via de basisvergelijking van OPLM kunnen we eenvoudig de conditionele kans op het goed beantwoorden van de items afleiden en daarmee kunnen we $p(+|s)$ evalueren, $prop(+|s)$ volgt uit de data. Discrepancies tussen $p(+|s)$ en $prop(+|s)$ 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

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

Deze zogenaamde 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

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

Deze zogenaamde S-toets heeft een χ^2 verdeling onder het model.

Als globale modeltoets is de R1c-toets (Glas, 1988) geschikt. Ook de distributie 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 WOPLOT (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.

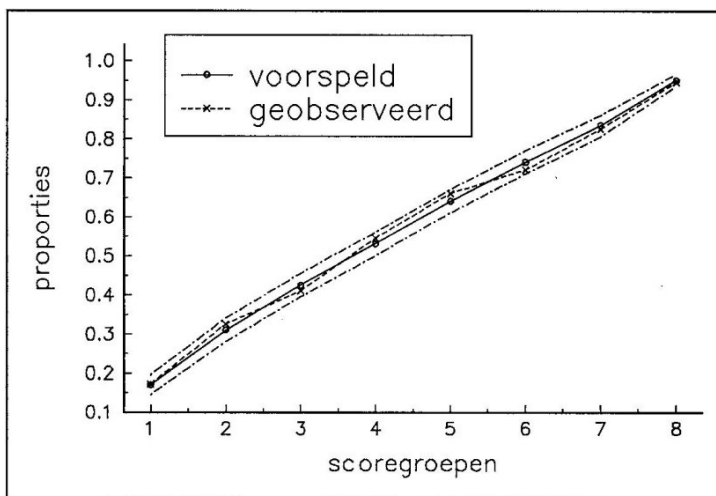
Afhankelijk van de uitkomsten kunnen items worden verwijderd. Ook inhoudelijke overwegingen spelen een rol in dit beslissingsproces (zie hiervoor hoofdstuk 2 over de achtergronden van de toetsinhoud).

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 bewerkstelligen 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).

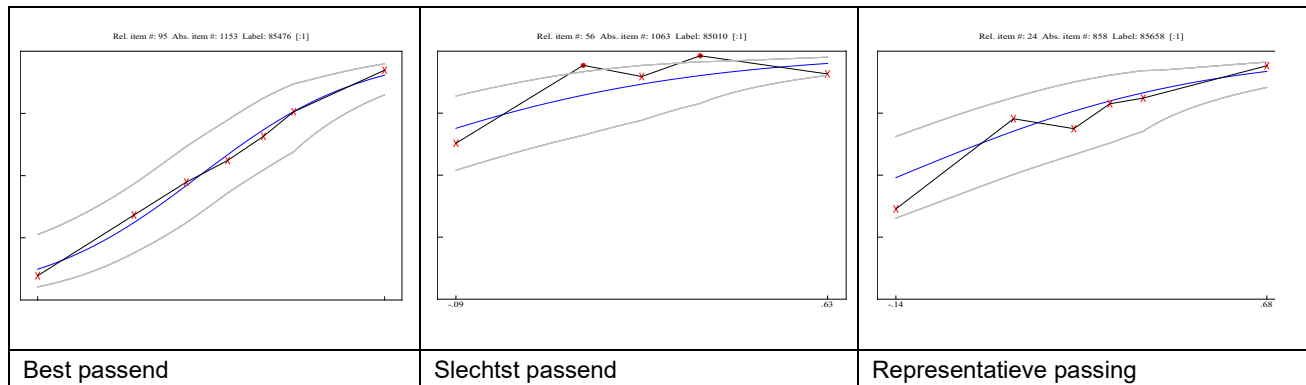
Figuur 4.3 Grafische voorstelling van een S-toets



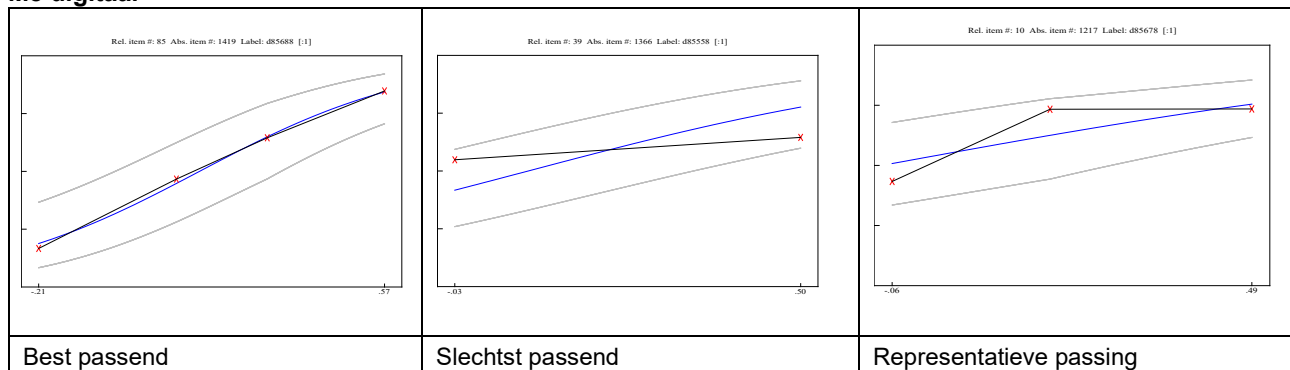
Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.4 illustreren dat voor de toets B8/M8 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 M8 met de best passende, de slechtst passende en een qua passing representatieve opgave

M8



M8 digitaal



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 8. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor de papieren toetsen de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Bij de digitale toetsen is de verdeling minder gelijkmatig. Echter zien we daar nagenoeg geen significante resultaten. Dit als geheel 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 significantieniveau 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 B8/M8

B8/M8	0	5	4	7	8	6	11	10	9	13	10	13
B8/M8 met RM	0	6	5	7	11	7	12	11	12	14	15	16
B8/M8 digitaal	0	0	1	2	3	6	2	2	8	8	12	52
B8/M8 Digitaal RM	0	1	4	5	0	2	3	7	13	10	12	59

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 B8/M8 geldt zowel voor de papieren versie als de digitale versie dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootte 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 B8/M8

Toetsversie	R1c	df	p
B8/M8	755,9	694	0,05
B8/M8RM	873,9	806	<0,005
B8/M8 digitaal	114,1	124	0,73
B8/M8 digitaal RM	130,8	150	0,88

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 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 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 1 opgave van de 96 gekalibreerde opgaven die in de uitgave B8/M8 zitten en voor 6 opgaven die in de B8/M8 uitgave rekenmachine (116 opgaven) zitten is de waarde hoger dan 0,20. Voor de digitale toetsen geldt ook dat de c-waarden gemiddeld onder de 0,20 liggen, het gemiddelde komt uit op 0,05 voor de digitale toets met 96 opgaven en de digitale toets met rekenmachinetaak. Over het algemeen kan de nauwkeurigheid van de schattingen voor de papieren toetsen 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. Voor de digitale toetsen is het beeld eveneens positief met een gemiddelde van 0,05.

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

Toets	Constance 'c'	
	Range	Gemiddelde
B8/M8	0,084 – 0,213	0,129
B8/M8 rm	0,073 – 0,254	0,130
B8/M8 digitaal	0,03 – 0,09	0,05
B8/M8 digitaal rm	0,03 – 0,013	0,05

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen LVS-III Rekenen-Wiskunde groep 8 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 leerlingvolgysteemtoetsen 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 normsteekproeven 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 achtergrondvariabelen). De dekkinggraad 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 afwijken 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 toetsituatie 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 toetsituatie 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 Eén 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 Eén 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 achterstandsleerlingen volgens een indeling in vier typen: (1) percentage achterstandsleerlingen [0, .10), (2) percentage achterstandsleerlingen [.10, .25), (3) percentage achterstandsleerlingen [.25, .40) en (4) 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 $4 \times 2 \times 4 \times 2 = 64$ verschillende categorieën representeren. De variabelen *regio*, *urbanisatiegraad* en *schooltype* zijn op het niveau van de school gedefinieerd. De variabele *seks*e 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 :

$$w_{ijk} = \frac{(n_{T,ijk} - n_{S,ijk}) \div (N_T - N_S)}{n_{D,ijk} \div N_D} = \frac{N_D(n_{T,ijk} - n_{S,ijk})}{n_{D,ijk}(N_T - N_S)}, \quad (4.3)$$

waarbij vereist is dat $n_{S,ijk} \leq n_{T,ijk}$. 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 $n_{S,ijk} \leq n_{T,ijk}$ 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 $n_{T,ijk}$ 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 $n_{S,ijk} \leq n_{T,ijk}$. 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 $N_S \geq N_T$.

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 gelijkheid 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 *seks*e 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 *seks*e, dan wordt een tweede steekproeftrekking uitgevoerd.

Eerst wordt op basis van de landelijke gegevens van CBS en de geobserveerde aantallen in de normeringssteekproef de kans w_q bepaald dat een leerling met seks q in een representatieve normeringssteekproef zit:

$$w_q = \frac{n_{T,q} \div N_T}{n_{S,q} \div N_S} = \frac{n_{T,q} N_S}{N_T n_{S,q}}. \quad (4.4)$$

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 8 ingesteld op $N_T = 2 \times 2200 = 4400$.

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 8 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 *seks*. Ook in een volledig aselechte 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 normeringsonderzoek voor afnamemoment M8. De aantallen voor het normeringsonderzoek betreffen uitsluitend nieuw verzamelde data en geen gegevens uit dataretour.

Zowel voor afnamemoment medio groep 8 als begin groep 8 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 *seks*. Ook in aselechte steekproeven kan de verdeling van leerlingen over de verschillende categorieën immers afwijken van de verdeling in de populatie. In een aselechte steekproef is deze afwijking per definitie het gevolg van toeval. Statistische weging is in een aselechte 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 *seks* 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
M8	2200	2413	4613	195
B8		1163	1163	40

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.

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (4.5)$$

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

Stratum	Populatie	Steekproef			
	%	M8	%	B8	%
0-10%	67,3	3222	69,8	818	70,3
>10%	32,7	1391	30,2	345	29,7

M8 $\chi^2(1, N = 4613) = 13,588; p = 0,000; \phi = 0,108$

B8 $\chi^2(1, N = 1163) = 4,869; p = 0,027; \phi = 0,065$

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

Regio	Populatie	Steekproef			
	%	M8	%	B8	%
Noord	10,0	449	9,7	42	3,6
Oost	22,4	1010	21,9	454	39,0
West	47,9	2254	48,9	347	29,8
Zuid	19,7	900	19,5	320	27,5

M8 $\chi^2(3, N = 4613) = 1,764; p = 0,623; \phi = 0,020$

B8 $\chi^2(3, N = 1163) = 306,543; p = 0,00; \phi = 0,510$

De regioverdeling voor B8 is niet representatief is ($\phi=0.510$). Maar aangezien de gemiddelden per regio niet veel verschillen is een weging niet noodzakelijk. In onderstaande tabel staan de basisgegevens die noodzakelijk zijn voor de berekeningen.

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

Regio	Populatie	Steekproef		
	%	B8	%	gem.
Noord	10	42	3.6	259.6
Oost	22.4	454	39.0	266.3
West	47.9	347	29.8	272.3
Zuid	19.7	320	27.5	270.1

In de eerste vijf kolommen staan dezelfde waarden zoals ze te vinden zijn in Tabel 4.6. In de laatste kolom staan de **geobserveerde** gemiddelden per regio. We kunnen nu het populatiegemiddelde op twee manieren schatten. Bij de eerste manier negeren we het feit dat de steekproef voor regio niet representatief is. We vinden dan $(3.6*259.6+39.0*266.3+29.8*272.3+27.5*270.1)/100=268.9$. Bij de tweede manier wegen we naar de populatie; de steekproef-gewichten in bovenstaande formule worden dan vervangen door de populatie-gewichten. Zo vinden we 269.2. Een verschil in gemiddelde schaalscores van 0.3 hetgeen (gezien de spreiding van de schaalscores van 23.9) een te verwaarlozen verschil is. De conclusie kan als volgt geformuleerd worden: 'Alhoewel de steekproef voor regio voor B8 niet representatief is ($\phi=.0510$) is een weging naar de populatie niet noodzakelijk' .

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

Populatie		Steekproef			
Urbanisatie- graad	%	M8	%	B8	%
Platteland	54,8	2542	55,1	756	65,0
Stad	45,2	2071	44,9	407	35,0

M8 $\chi^2(1, N = 4613) = 0,225$; $p = 0,635$; $\phi = 0,007$

B8 $\chi^2(1, N = 1163) = 48,891$; $p = 0,000$; $\phi = 0,205$

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

Populatie		Steekproef			
Sekse	%	M8	%	B8	%
jongen	50,5	2237	49,4	579	49,8
meisje	49,5	2296	50,7	584	50,2

M8 $\chi^2(1, N = 4533) = 2,332$; $p = 0,127$; $\phi = 0,023$

B8 $\chi^2(1, N = 1163) = ,238$; $p = 0,626$; $\phi = 0,014$

Van 80 leerlingen was de sekse niet bekend.

De χ^2 -waarden zijn in een aantal gevallen 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 zeer acceptabel zijn behalve voor de B8 regioverdeling (cf. Cohen, 1988). Daar is sprake van een groot effect.

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 B8 en M8.

Tabel 4.9 Normtabel op leerlingniveau voor LVS-III Rekenen-Wiskunde B8-M8

Tijdstip	M	SD	Kurt.	Skew.	P10	P20	P25	P40	P50	P60	P75	P80	P90
B8	268,9	25,4	-0,39	-0,48	236,5	248,5	252,5	264,52	269,52	276,5	286,52	289,5	298,5
M8	272,7	24,0	-0,40	-0,49	241,5	253,5	257,5	267,5	273,5	279,5	289,5	292,5	302,5

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 B8-M8

Afname-moment	Aantal replicaties	Aantal scholen	Gemiddelde	SD School	SD Leerling	ICC
B8	20	40	267,4	8,4	23,9	0,111
M8	20	195	273,5	8,1	26,7	0,113

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.

Tabel 4.11 Normtabel op schoolniveau voor LVS-III Rekenen-Wiskunde B8-M8

Afname-moment	M	SD	P10	P20	P25	P40	P50	P60	P75	P80	P90
B8	267,4	8,4	256,6	260,3	261,7	265,3	267,4	269,6	273,1	274,5	278,2
M8	273,5	8,1	263,1	266,7	268,0	271,5	273,5	275,6	279,0	280,3	283,9

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 $\tau(\theta)$. 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 $Var(\tau)$. Tussen θ en $\tau(\theta)$ 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 $\tau(\theta)$ 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 $Var(t|\tau(\theta))$, 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.

$$Var(t) = E[Var(t | \tau(\theta))] \quad (5.1)$$

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

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

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 betrouwbaarheidscoëfficiënt 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 (simulatie)
B8_3 papier	96	60,6	4,046	0,96	0,96
B8_4 papier (rm ^{**})	116	73,8	4,443	0,97	0,97
M8_3 papier	96	63,7	4,007	0,96	0,96
M8_4 papier (rm)	116	77,7	4,398	0,96	0,96
B8_3 digitaal	363	222,4	15,977	0,96	0,96
B8_4 digitaal (rm)	439	268,0	17,559	0,97	0,97
M8_3 digitaal	363	234,6	15,867	0,95	0,95
M8_4 digitaal (rm)	439	282,7	17,443	0,96	0,96

* digitale toetsen werken met gewogen score in het Computerprogramma LOVS

**rm is met rekenmachinetaak

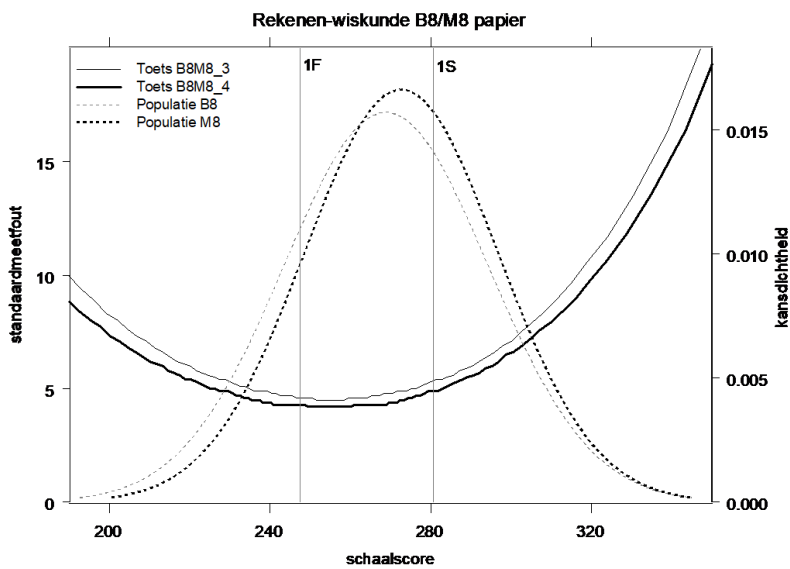
Het feit dat alle items OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1000 000 leerlingen. Daarbij hebben we enerzijds de vaardigheidsverdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselekt uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1000 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 nagenoeg 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

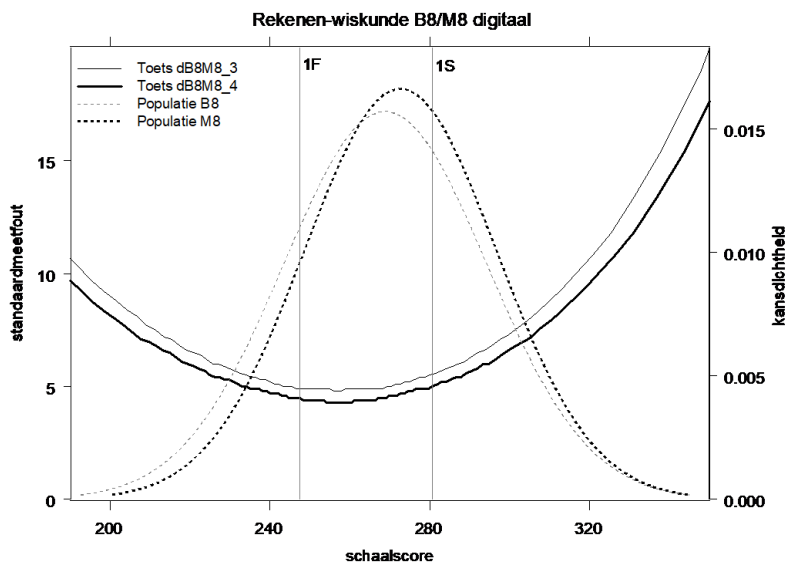
De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid van de verschillende toetsen Rekenen-Wiskunde. De figuren 5.1 en 5.2 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de papieren en digitale toetsen B8/M8. 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 B8/M8 en de kansdichtheidsfuncties voor de B8-populatie



Figuur 5.2 Grootte van de meetfouten voor de papieren en digitale toetsen B8/M8 en de kansdichtheidsfuncties voor de M8-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 5.9. De betrouwbaarheidstabellen laten het effect van de lokale meetnauwkeurigheid zien.

Zo laat tabel 5.3 bijvoorbeeld zien dat 85,4 procent van de leerlingen die bij de M8-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 85,4 procent. Verder laat de tabel zien dat 14,4 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 beschreven door Wheadon & Stockford (2011). 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 8 en einde groep 8 zijn te vinden in de tabellen 5.2 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 is beschreven door Wheadon & Stockford (2011).

Tabel 5.2 Betrouwbaarheidstabel toets B8 papier voor afnamemoment begin 8

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	90,7	9,3	0,0	0,0	0,0	E	88,3	11,7	0,0	0,0	0,0
IV	10,4	74,8	14,8	0,0	0,0	D	10,9	76,1	13,1	0,0	0,0
III	0,0	13,5	72,9	13,6	0,0	C	0,0	8,3	77,8	13,8	0,0
II	0,0	0,0	13,2	72,9	13,9	B	0,0	0,0	12,1	75,7	12,2
I	0,0	0,0	0,1	15,5	84,4	A	0,0	0,0	0,0	13,3	86,7

Tabel 5.2a Samenvattende indices toets B8 papier voor afnamemoment begin 8

Toets B8 papier, afnamemoment B8		
	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	79,3	80,4
Accuracy plus/minus 1 niveau	100	100

Tabel 5.3 Betrouwbaarheidstabel toets M8 papier voor afnamemoment medio 8

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	91,7	8,3	0,0	0,0	0,0	E	88,7	11,3	0,0	0,0	0,0
IV	10,4	75,1	14,4	0,0	0,0	D	10,5	75,0	14,5	0,0	0,0
III	0,0	15,4	69,1	15,5	0,0	C	0,0	9,7	78,1	12,1	0,0
II	0,0	0,1	16,6	69,0	14,3	B	0,0	0,0	13,1	74,8	12,1
I	0,0	0,0	0,2	14,4	85,4	A	0,0	0,0	0,0	12,6	87,4

Tabel 5.3a Samenvattende indices toets M8 papier voor afnamemoment medio 8

Toets B8 papier, afnamemoment B8		
	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	77,8	80,3
Accuracy plus/minus 1 niveau	99,9	100

Tabel 5.4 Betrouwbaarheidstabel toets B8 papier met rekenmachine voor afnamemoment B8

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	91,4	8,6	0,0	0,0	0,0	E	89,2	10,8	0,0	0,0	0,0
IV	9,5	76,6	13,9	0,0	0,0	D	10,0	78,0	12,0	0,0	0,0
III	0,0	12,3	75,2	12,5	0,0	C	0,0	7,7	79,4	13,0	0,0
II	0,0	0,0	12,1	75,1	12,8	B	0,0	0,0	11,2	77,4	11,4
I	0,0	0,0	0,0	14,5	85,5	A	0,0	0,0	0,0	12,4	87,6

Tabel 5.4a Samenvattende indices toets B8 papier met rekenmachine voor afnamemoment begin 8

Toets B8 papier, afnamemoment B8		
	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	80,9	81,8
Accuracy plus/minus 1 niveau	100	100

Tabel 5.5 Betrouwbaarheidstabel toets M8 papier met rekenmachine voor afnamemoment medio 8

Score- groepen V t/m I	Scoregroep waarin de ware score valt					Score- groepen E t/m A	Scoregroep waarin de ware score valt				
	V	IV	III	II	I		E	D	C	B	A
V	92,3	7,7	0,0	0,0	0,0	E	89,5	10,5	0,0	0,0	0,0
IV	9,5	77,0	13,4	0,0	0,0	D	9,6	76,9	13,4	0,0	0,0
III	0,0	14,2	71,3	14,4	0,0	C	0,0	8,9	79,8	11,3	0,0
II	0,0	0,1	15,4	71,3	13,3	B	0,0	0,0	12,0	76,7	11,2
I	0,0	0,0	0,1	13,5	86,5	A	0,0	0,0	0,0	11,7	88,3

Tabel 5.5a Samenvattende indices toets M8 papier met rekenmachine voor afnamemoment medio 8

	M8 papier met rekenmachine, afnamemoment medio 8	
	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	79,5	81,8
Accuracy plus/minus 1 niveau	100	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 77,8 tot 81,8 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 8 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 8, 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 8.

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. 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 nul-model. 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 alle toetsen 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 8. De toetsopgaven richten zich op inhoudelijk relevante domeinen als Getallen, Verhoudingen, Meten en meetkunde, Verbanden en er is een optionele taak met rekenmachine. De schaal zou wellicht een nog fraaiere passing vertonen als slechts één van deze onderdelen gekozen zou zijn – dit ligt dan dicht 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 vijf verschillende inhoudelijke domeinen geeft aan dat in de praktijk deze vijf domeinen samengenomen kunnen worden om een unidimensionele uitspraak te doen. De correlaties tussen de subschalen, zoals gegeven in tabel 6.1 variërend van 0,90 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 M8 op basis van dataretour

M8

	Getallen	Verhoudingen	Meten	Verbanden	Rekenmachine
Getallen		12858	12858	12858	3637
Verhoudingen	0,97		12858	12858	3637
Meten	0,96	0,98		12858	3637
Verbanden	0,97	0,99	0,98		3637
Rekenmachine	0,91	0,90	0,90	0,92	

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 lager en ook deze is prima met een gemiddelde van rond de 0,05. Bij de papieren toetsen is er slechts een enkele opgave van de 192 met een c-waarde boven de 0,30. Bij de digitale toetsen geen 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_{it} -waarden naar toets

	p-waarden			R_{it} -waarden	
	N-items	Range	Gemiddelde	Range	Gemiddeld
B8-3	96	0,37-0,90	0,64	0,27-0,60	0,44
B8-4	116	0,37-0,90	0,64	0,26-0,60	0,43
B8-3 dig	96	0,36-0,83	0,62	0,24-0,60	0,41
B8-4 dig	116	0,36-0,83	0,61	0,24-0,61	0,42
M8-3	96	0,40-0,91	0,67	0,24-0,59	0,41
M8-4	116	0,40-0,91	0,67	0,24-0,59	0,41
M8-3 dig	96	0,40-0,85	0,65	0,23-0,58	0,39
M8-4 dig	116	0,40-0,86	0,65	0,23-0,58	0,40

In tabel 6.2 zijn de ranges en de gemiddelden weergegeven voor de p-waarden en de R_{it} -waarden van de items van de papieren en digitale toetsen B8/M8 op de momenten B8 en M8. Bij de toetsen is te zien dat de p-waarden liggen tussen de 0,36 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 M8 en B8 ligt tussen de 0,61 en 0,67. Er wordt in het algemeen voor groep 8 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_{it} 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 0,39 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_{it} , die hier gegeven is, ligt altijd iets lager dan de R_{it} . Er is in de toetsen B8/M8 papier en B8/M8 digitaal geen enkel item met een R_{it} van lager dan 0,20. Bijlage 3 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 0,70 ligt, zijn de verdelingen linksscheef (vergelijk de negatieve waarden in de kolom 'skewness'), de ene wat meer dan de andere. De gevonden scheefheid is bij alle verdelingen gematigd linksscheef. De verdelingen zijn ééntoppig.

Tabel 6.3 Verdelingskenmerken van de B8/M8-toetsen op de afnamemomenten Rekenen-Wiskunde groep 8

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
B8 papier	60,6	20,39	-0,64	-0,45
B8 papier met rekenmachinetaak	73,8	24,25	-0,61	-0,46
B8 digitaal	59,2	19,83	-0,69	-0,38
B8 digitaal met rekenmachinetaak	71,1	23,91	-0,70	-0,37
M8 papier	63,7	19,04	-0,42	-0,56
M8 papier met rekenmachinetaak	77,7	22,61	-0,38	-0,58
M8 digitaal	62,2	18,59	-0,50	-0,50
M8 digitaal met rekenmachinetaak	74,8	22,43	-0,49	-0,51

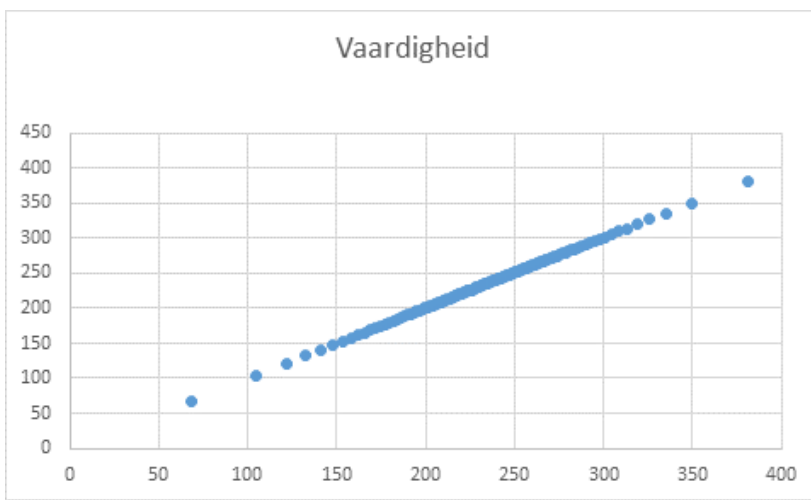
6.4 Itembias

Er is een analyse uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning*, DIF) met betrekking tot sekse. Hiervoor is de papieren kalibratie als uitgangspunt genomen voor M8. 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 de papieren toets van M8 bij 12 van de 116 items DIF zien. 7 daarvan waren in het voordeel van jongens, de andere 5 in het voordeel van meisjes. Bij de digitale toets M8 waren er 10 items van de 116 met DIF. 6 in het voordeel van de jongens 4 in het voordeel van de meisjes.

Om aan te tonen dat deze DIF-items inderdaad geen schade berokkenen berekenen we de vaardigheidsschattingen (voor elk ruwe score): een die gebaseerd is op de data van de jongens en een die gebaseerd is op de data van de meisjes, waarbij de itemparameters van de items met DIF verschillend zijn voor de jongens en de meisjes. In figuur 6.1 is te zien dat de verschillen in itemparameters voor de DIF-items voor jongens en meisjes niet tot een andere vaardigheidsschatting leiden.

Figuur 6.1 Vaardigheidsverdeling jongens vs meisjes.



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 8.

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 M8 papier ($r=0,95$; $N=2207$) en M8 digitaal ($r=0,93$; $N=118$). 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 drie 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 derde generatietoetsen Rekenen-Wiskunde LVS van groep 8. De scholen waren gelegen in Oss, Apeldoorn en Zevenaar.

Tabel 6.5 Correlatie van Citotoetsen Rekenen-Wiskunde met de Schoolvaardigheidstoets (Boom)

	Cito Rekenen-Wiskunde M8 generatie 3	
	correlatie	N
Schoolvaardigheidstoets Rekenen-Wiskunde (Boom)	0,81	113

*gecorrigeerd voor attenuatie

In tabel 6.5 staat de correlatie van de nieuwe toets Rekenen-Wiskunde M8 weergegeven 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 0,81 voor M8. De gevonden correlatie ligt in lijn met de correlatie die eerder gevonden is tussen de derde generatietoetsen LVS Rekenen-Wiskunde 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. De correlatie van de toetsen van groep 7 van LVS met de Boom toetsen was met 0,91 hoger dan de ander correlaties van LVS met de Boom toetsen rekenen.

Tabel 6.6 Correlatie van de derde generatie Cito LVS toets Rekenen-Wiskunde B8/M8 papier met toetsen op het gebied van leervorderingen

	Rekenen-Wiskunde*	Aantal leerlingen
Cito Spelling 2017 M8	0,53	91388
Cito Begrijpend Lezen 2017 M8	0,78	91038
Cito Woordenschat 2017 M8	0,61	17640
Cito Spelling WW 2017 M8	0,64	84538

*Deze correlaties zijn gecorrigeerd voor attenuatie.

Dataretourgegevens zijn gebruikt om de correlatie met scores op andere LVS-toetsen te berekenen. In de tabel hierboven is de correlatie tussen de toets Rekenen-Wiskunde B8/M8 3.0 papier en de toetsen Spelling, Spelling Werkwoorden (WW), Begrijpend lezen en Woordenschat weergegeven. De correlaties zijn in lijn met de eerder gevonden correlaties in de wetenschappelijke verantwoordingen van LVS Rekenen 3.0 groep 3 tot en met groep 7.

6.6 Verschillen tussen relevante subgroepen

In de tabel 6.7 wordt voor afnamemoment M8 de gemiddelde scores van de leerlingen per lesmethode weergegeven. De data zijn afkomstig van het normeringsonderzoek.

Er is nauwelijks sprake van verschillen tussen de gemiddelden van de methoden. Alleen bij Wizwijs is er sprake van een klein effect. Het aantal scholen is daar echter beperkt. In de andere gevallen is er geen sprake van een effect.

Tabel 6.7 Gemiddelde score per lesmethode

Moment	Lesmethode	M	SD	Aantal	Effect
M8	WIG4	271,5	24,5	899	
	Pluspunt	275,2	22,1	413	0,15
	Alles telt	269,4	27,8	222	-0,08
	Zeker	273,1	33,2	103	0,07
	Rekenrijk	275,0	22,1	56	0,14
	Wizwijs	264,5	25,0	28	-0,28

In tabel 6.8 wordt de score per halfjaar groep gepresenteerd voor M8.

De data zijn afkomstig van het normeringsonderzoek M8.

Gebruikelijk is een patroon te vinden waarbij de jongste leerlingen het hoogst scoren en de oudste leerlingen het laagst. (Janssen, 2015; Hop, 2016)

Tabel 6.8 Gemiddelde score per halfjaargroep

Moment	Halfjaargroep	N	M	SD	Effect
M8	11,5	132	289,2	22,1	
	12	816	273,0	22,3	-0,69
	12,5	559	273,8	25,5	-0,65
	13	269	261,5	25,2	-1,15
	13,5	53	253,6	22,5	-1,47

Moment	Halfjaargroep	N	M	SD	Effect
B8	11	173	275,5	23,5	
	11,5	483	271,2	23,3	-0,18
	12	384	268,2	26,1	-0,29
	12,5	95	255,1	27,1	-0,80
	13	28	244,3	25,0	-1,28

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 6.9 Gemiddelde score jongen-meisje

Moment	Sekse	N	M	SD	Effect
B8	jongen	579	273,6	24,5	
	meisje	584	264,2	25,4	-0,38

Moment	Sekse	N	M	SD	Effect
M8	jongen	908	274,8	25,3	
	meisje	882	269,9	24,2	-0,25

Per afnamemoment is in de bovenstaande tabellen de gemiddelde score van jongens en meisjes weergegeven. De data zijn afkomstig van het normeringsonderzoek. 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 (B8 0,38 en M8 0,25).

Tabel 6.10 Gemiddelde score leerlinggewicht

Moment	LG	N	M	SD	Effect
M8	0.00	1501	273,9	24,4	
	0.30	78	261,6	25,9	-0,35
M8	1.20	95	263,9	27,1	-0,27

Bij de toets B8/M8 is een matig effect te zien tussen leerlingen met een gewicht van 0 en leerlingen met een gewicht van 0.30 of 1.20. Bij B8/M8 is 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). De data zijn afkomstig van het normeringsonderzoek.

7 Samenvatting

In dit hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken.

De toetsen Rekenen-Wiskunde 3.0 voor groep 8 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. Na de inleiding in hoofdstuk 1 beschrijven we 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 8 zijn alle domeinen van belang.

We geven in dit 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 8 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 R_{1c} -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, urbanisatiegraad, 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 domein rekenen-wiskunde.

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 R_{1c} 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.

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 Volgstelsel 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 correlatiegegevens van een onderzoek waarbij een groep leerlingen naast de toets Rekenen-Wiskunde B8/M8 een (COTAN positief-beoordeelde) toets van een andere uitgever heeft gemaakt. Ook uit dat onderzoek blijkt dat de correlatie tussen deze toets met de toets Rekenen-Wiskunde 3.0 B8/M8 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. Daaruit blijkt onder andere dat de leerlingen over de verschillende methoden heen ongeveer gelijk scoren. De effecten die daar gevonden worden zijn klein te noemen. 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 (B8 0,38 en M8 0,25).

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, & Hemker, B.T. (2009). *Wetenschappelijke verantwoording van de Intelligentietest Eindtoets Basisonderwijs*. Arnhem: Cito.
- Buijs, K., Klep, J., & Noteboom, A. (2008). *Tule – Rekenen/Wiskunde*. Enschede: SLO.
- Buijs, K., Scherpenzeel, P. van, Voorde, M. ten, & Zwaard, P. van der (2008a). *Werken aan de doorlopende leerlijn rekenen wiskunde van po naar vo*. Enschede. Enschede: SLO.
- Cito (2014). Primair en speciaal onderwijs. *Cito Volgstelsel. Rekenen-Wiskunde 3.0. Groep 4*. Arnhem: Cito.
- Cito (2013). Primair en speciaal onderwijs. *Cito Volgstelsel. Rekenen-Wiskunde 3.0. Groep 3*. Arnhem: Cito.
- Cito (2006). *Leerling- en onderwijsvolgstelsel, 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 & Treffers, A. (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., Kordes, J. & Van Weerden, J.J. (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-wiskundeonderwijs halverwege 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., Van der Schoot, F. en Hemker, B. (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., Van der Schoot, F. & Van Rijn, P. (2009). *Balans van het reken-wiskundeonderwijs in het speciaal basisonderwijs. Uitkomsten van de derde peiling in 2006*. PPON-reeks nummer 39. Arnhem: Cito.

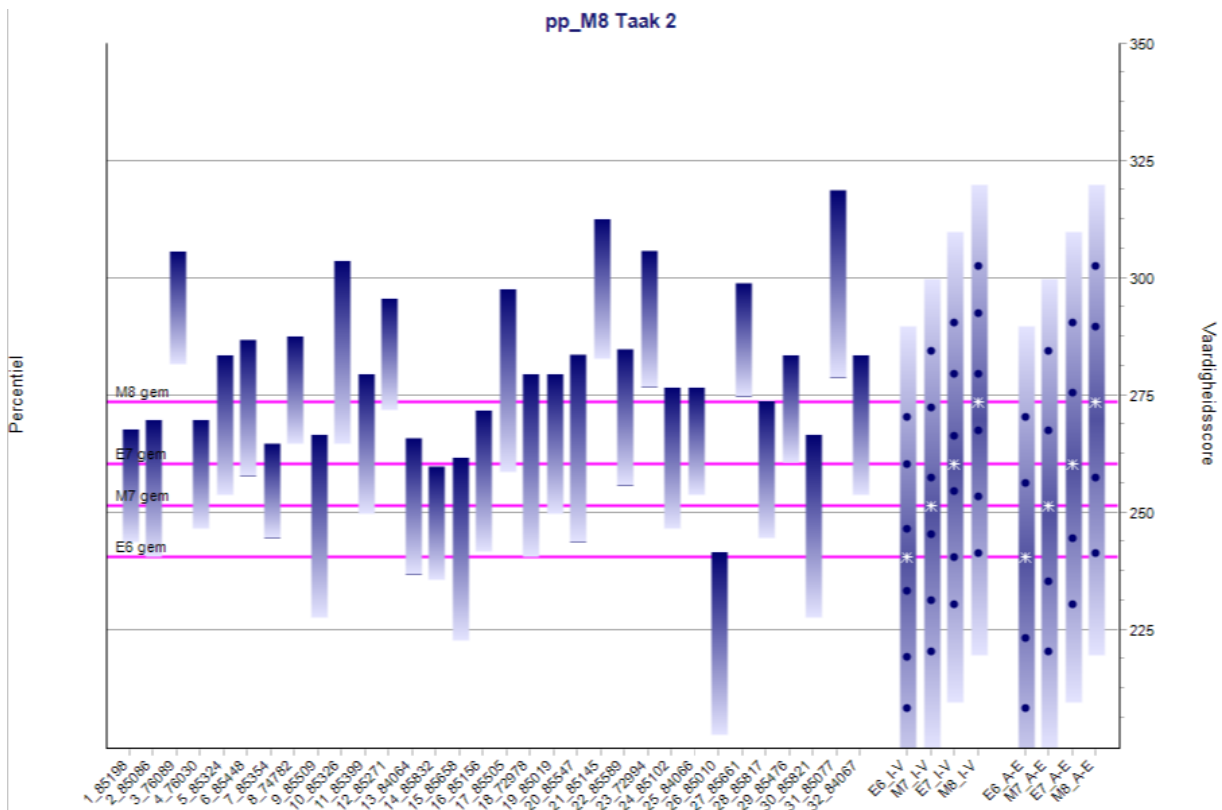
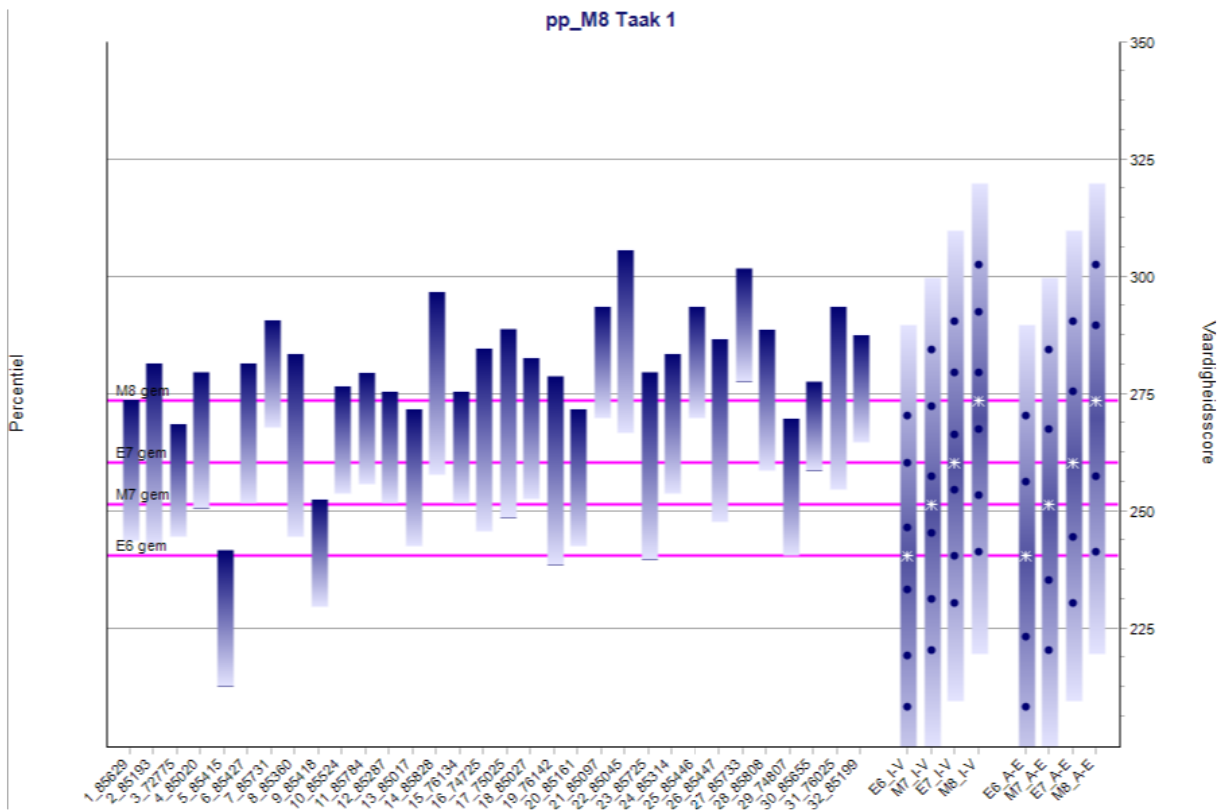
Kraemer, J-M. (2009a). *Balans over de strategieën en procedures bij het hoofdrekenen halverwege de basisschool. Uitkomsten van de peiling in 2005*. PPON-reeks nummer 40. Arnhem: Cito.

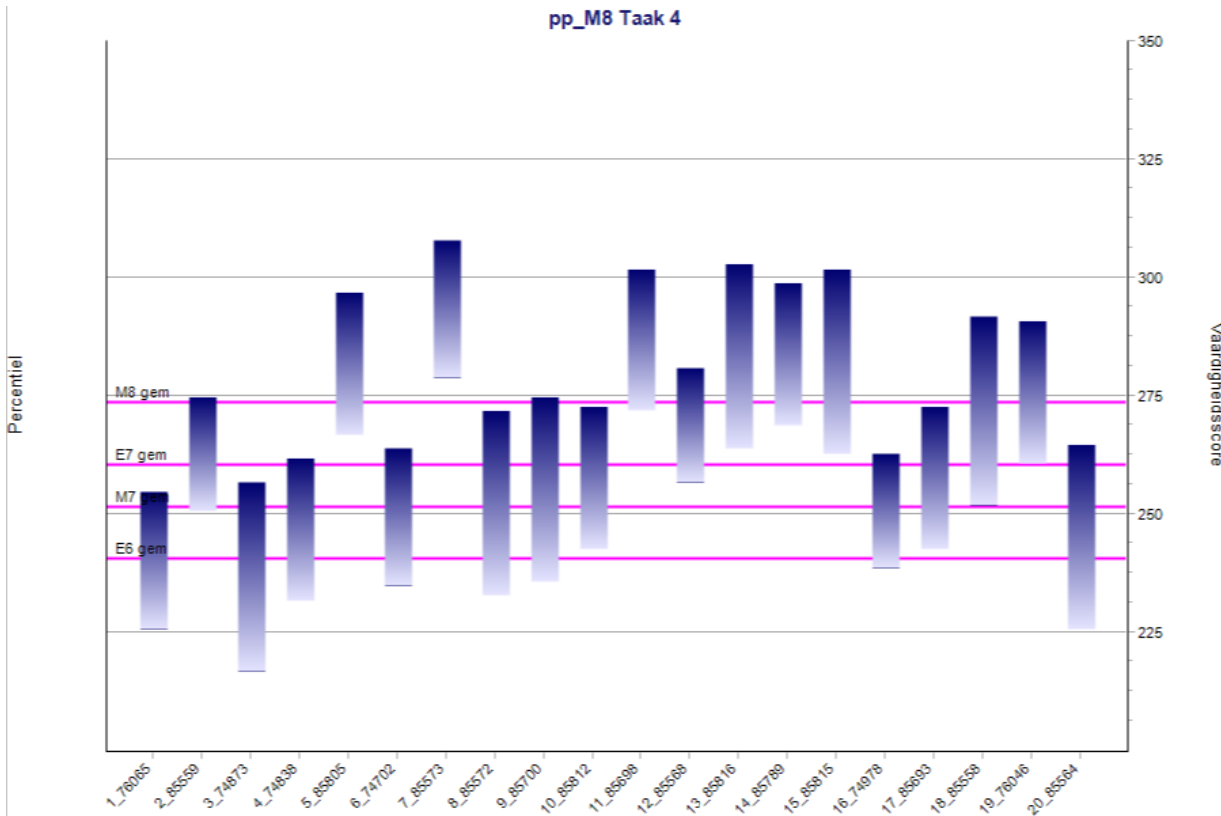
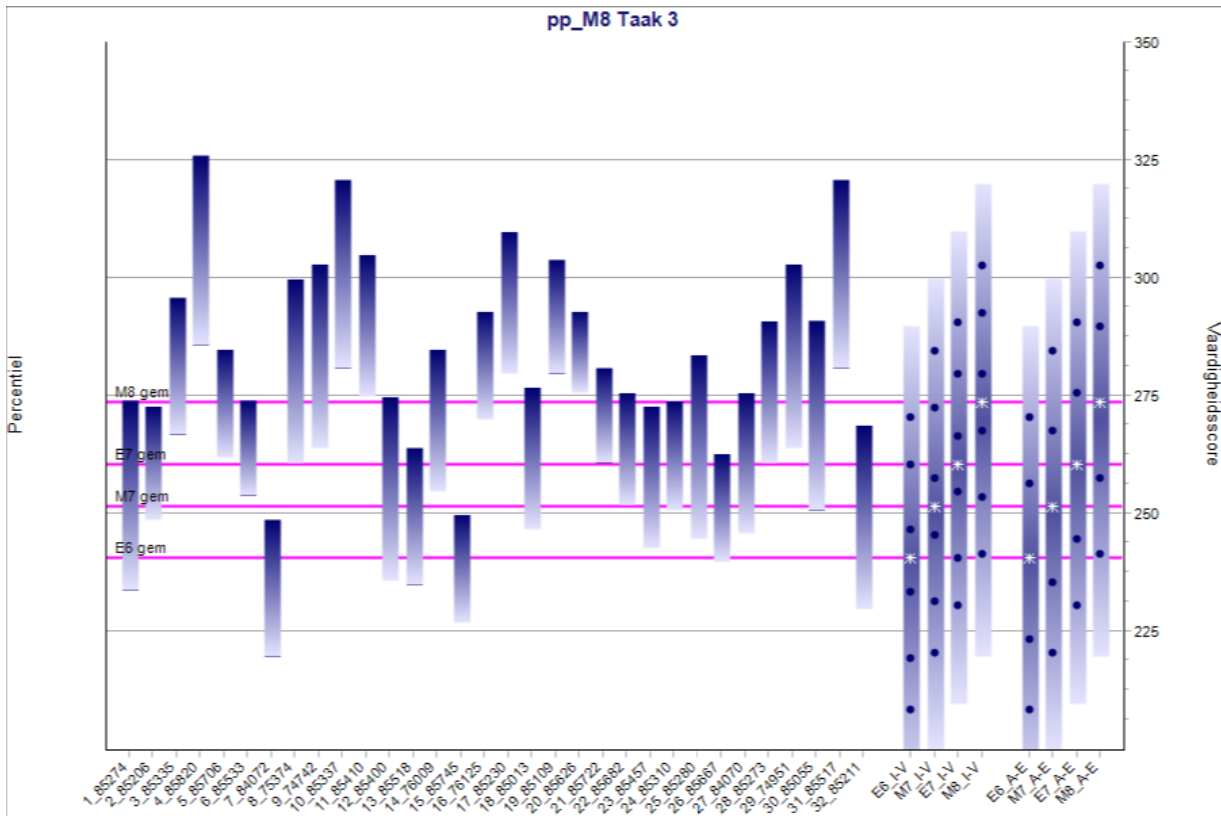
- Kraemer, J-M (2008). *Diagnosticeren en plannen in de onderbouw*. Arnhem: Cito.
- Kraemer, J-M., Janssen, J., Van der Schoot, F. & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs halverwege de basisschool 4. Uitkomsten van de vierde peiling in 2003*. PPON-reeks nr. 31. Arnhem: Cito.
- Koninklijke 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* 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*. Enschede: 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 examination 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*. Copenhagen: Danish Institute for Educational Research.
- Sanders, P. & Verstralen, H. (2010). *Het beoordelen van toetscores*. In P. Sanders (ed.), *Toetsen op school* (pp. 143-155). Arnhem: Cito.
- Scheltens, F., Hemker, B., Vermeulen, J. (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*. Enschede: Universiteit Twente.
- TAL-team (2007). *Metten 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.

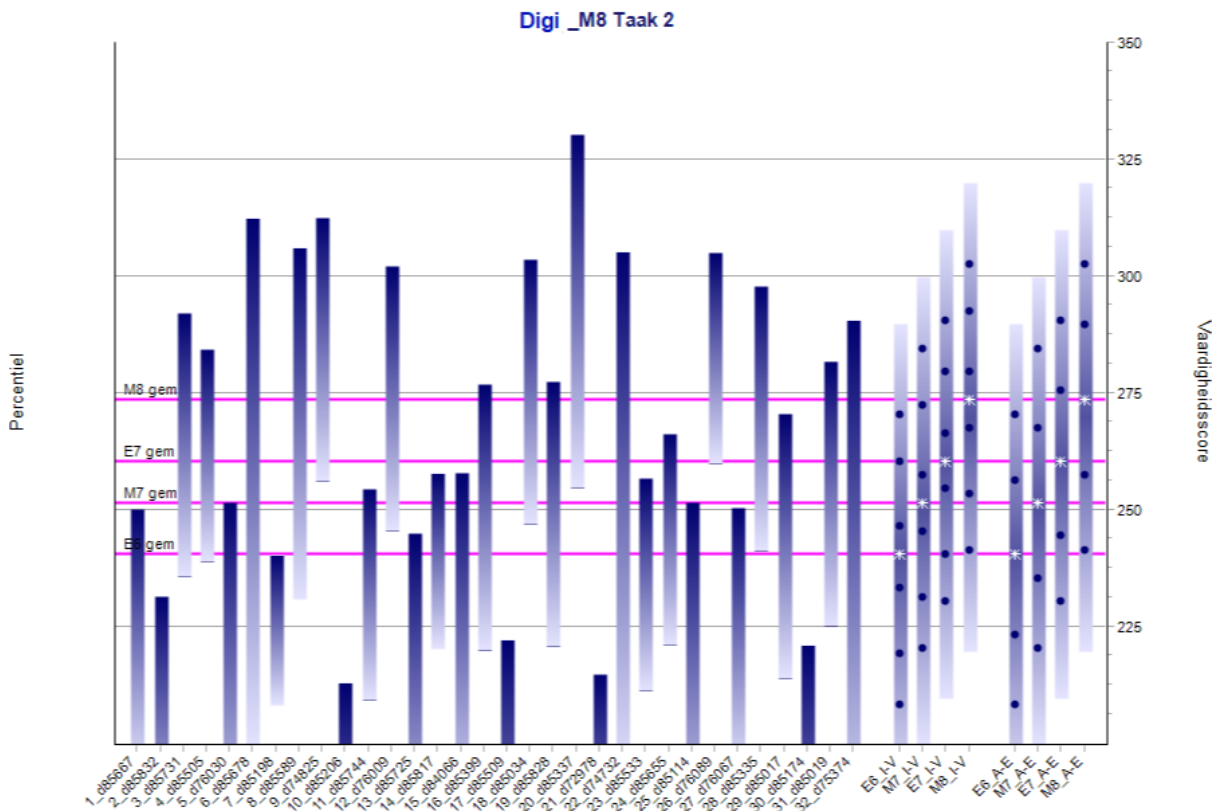
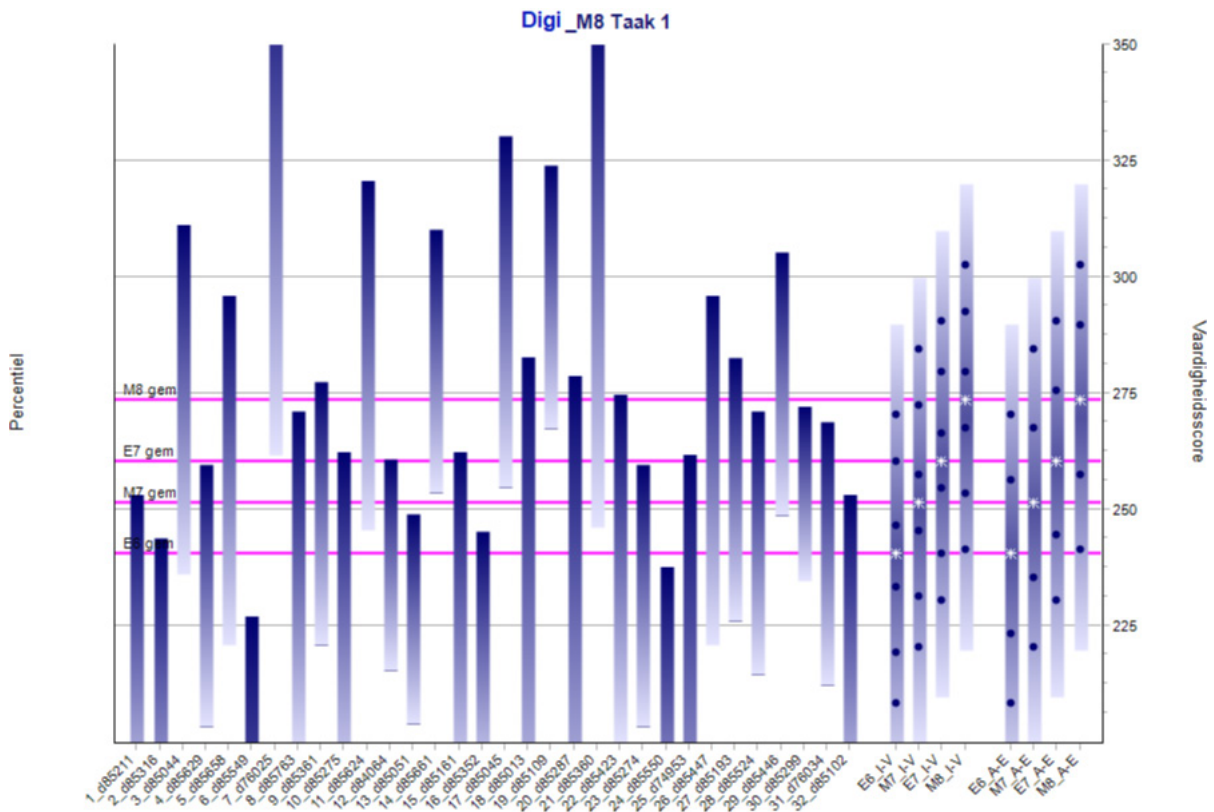
- TAL-team (1999). *Jonge kinderen leren rekenen: hele getallen*. Groningen: Wolters Noordhoff.
- Van der Linden, W.J., & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.
- Verhelst, N. (2007). *Profielanalyse met Item Respons Theorie*. Arnhem: Cito.
- Verhelst, 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 nummer 4. Arnhem: Cito.
- Verstralen, H.H.F.M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem: Cito.
- Wheadon, C., & Stockford, I. (2011). Classification accuracy and consistency in GSCE and A level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. Coventry, UK: Office of Qualifications and Examinations Regulation.

Bijlagen

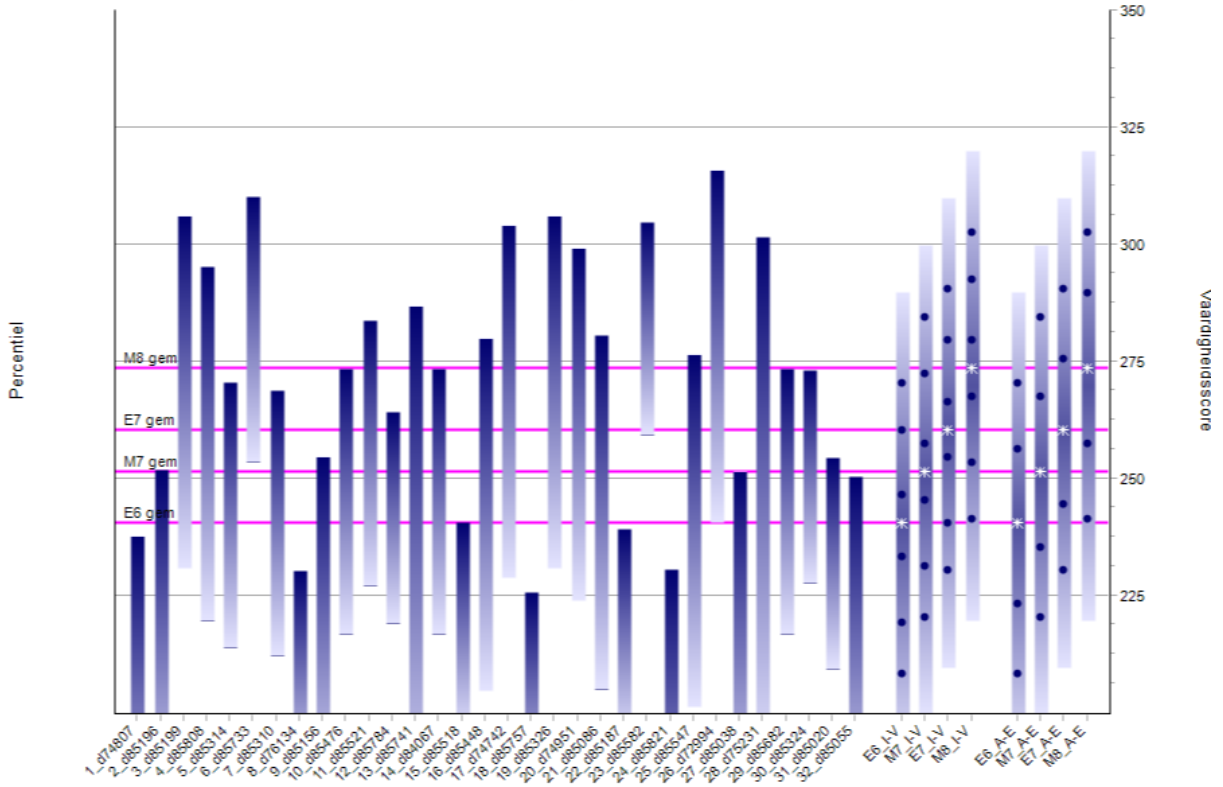
Bijlage 1 p50- en p80-kanspunten van de opgaven in de papieren toetsen en digitale toetsen B8/M8 in relatie tot de vaardigheidsverdelingen van E6, M7, E7 en M8



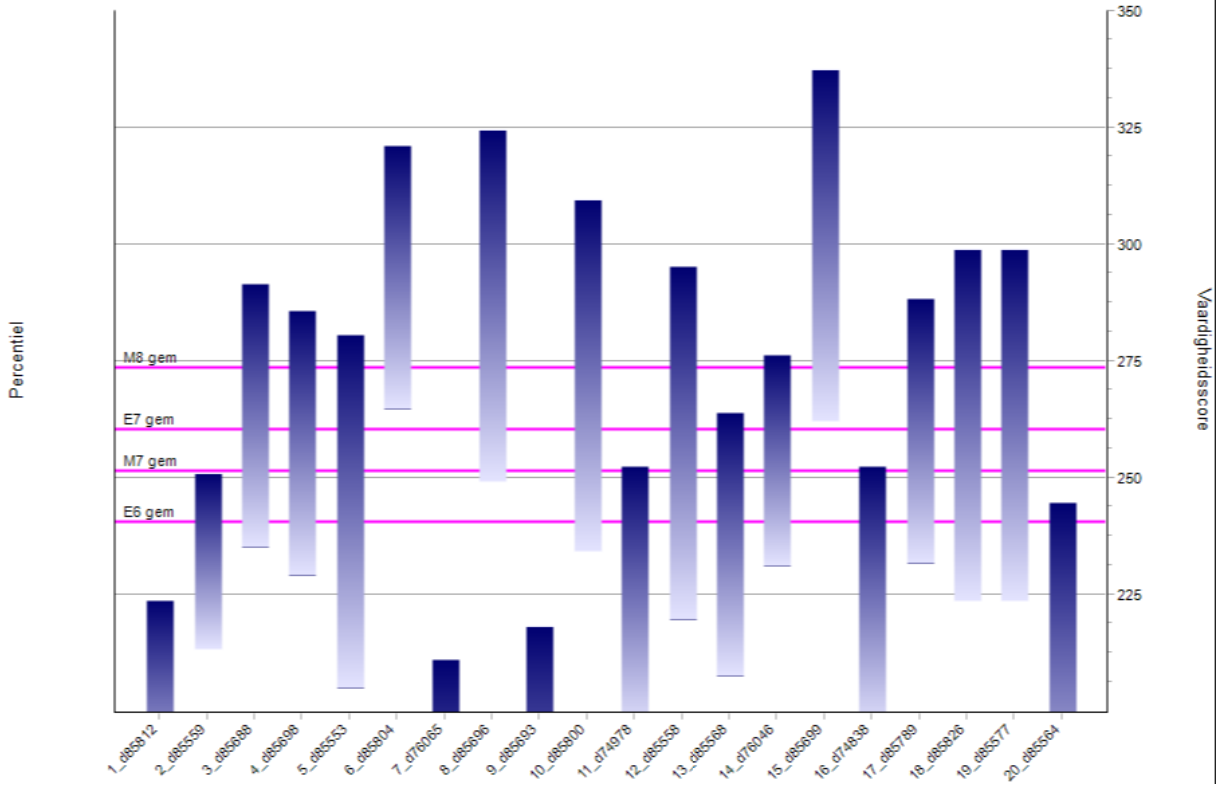




Digi_M8 Taak 3



Digi_M8 Taak 4



Bijlage 2 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M8

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

type	type	blok
M8(E7)	ME	43
M8(E7)	ME	44
M8(E7)	ME	45
E7	ME	46
E7	ME	47
RSA	ME	48
RSA	ME	49
M8	VB	50
M8	VB	51
M8	VB	52
M8	VB	53
M8	VB	54
M8	VB	55
M8(E7)	VB	56
E7	VB	57
RSA	VB	58
RSA	VB	59
M8	KL	60
M8	KL	61
M8	KL	62
M8	KL	63
M8	KL	64
M8(E7)	KL	65
E7	KL	66

1	2	3	4	5	6	7	8	9	10	11	12
			3					3			
				3							3
					3					3	
								4			
					4						
			3								
				3							
4			4								
	4			4							
		4			4						
3								3			
	3								3		
						3				3	
							3				3
								3			
									3		
										3	3
											4
							3				
								3			
3	3										
		3	3								
				3	3						
						3	3				
								3	3		
										3	3
									4		
32	31	30	31	31	32	31	31	32	32	32	31

Taak 13 en 14 zijn rekenmachinetaken. Deze zijn onder te verdelen in de domeinen getallen, verhoudingen, meten en verbanden. 16 opgaven van taak 13 en 14 komen zowel in taak 13 als in taak 14 voor. De samenstelling van taak 13 en 14:

taak	getallen	verhoudingen	meten
13	23	7	2
14	23	4	5

RS = Referentieset
 GB = Getallen
 ME = Meten en meetkunde
 VH = Verhoudingen
 VB = Verbanden
 KL = Kaal

Bijlage 3 Klassieke en IRT-indices van de opgaven in de papieren en digitale toetsen B8/M8

dig_M8_3

Item	InBk	Label	Dsc	P-Val	RIT	RIR
1214	1	d85017	4	0,646	0,453	0,432
1394	2	d85667	4	0,732	0,431	0,412
1403	3	d85661	5	0,460	0,515	0,496
1429	4	d85682	5	0,630	0,520	0,501
1452	5	d85044	3	0,537	0,379	0,355
1384	6	d85476	5	0,635	0,519	0,500
1436	7	d74951	3	0,583	0,377	0,354
1420	8	d85102	3	0,744	0,347	0,327
1238	9	d74953	2	0,750	0,250	0,227
1213	10	d74807	3	0,789	0,330	0,310
1375	11	d85725	3	0,770	0,338	0,318
1385	12	d72978	4	0,848	0,370	0,353
1225	13	d75231	2	0,656	0,269	0,245
1398	14	d85114	3	0,751	0,345	0,324
1445	15	d85211	4	0,740	0,429	0,409
1248	16	d84064	5	0,658	0,516	0,497
1423	17	d85524	5	0,639	0,519	0,500
1379	18	d76030	3	0,751	0,345	0,324
1392	19	d85518	5	0,752	0,488	0,470
1232	20	d74825	4	0,448	0,451	0,430
1218	21	d85034	4	0,492	0,457	0,435
1442	22	d85521	4	0,586	0,459	0,437
1437	23	d85020	5	0,688	0,509	0,491
1391	24	d72994	3	0,518	0,378	0,355
1227	25	d75374	2	0,684	0,265	0,241
1241	26	d85013	2	0,703	0,261	0,238
1401	27	d85629	4	0,693	0,443	0,423
1223	28	d85156	3	0,740	0,349	0,328
1231	29	d85399	4	0,618	0,456	0,435
1388	30	d85199	3	0,557	0,378	0,355
1454	31	d85763	3	0,687	0,363	0,341
1396	32	d76034	5	0,654	0,516	0,497
1239	33	d85731	4	0,545	0,460	0,438
1446	34	d85352	5	0,747	0,490	0,473
1236	35	d85198	7	0,723	0,592	0,576
1425	36	d85187	5	0,756	0,486	0,469
1400	37	d85447	3	0,596	0,376	0,353
1216	38	d85448	3	0,655	0,369	0,347
1433	39	d85533	5	0,678	0,512	0,493
1386	40	d85733	4	0,461	0,453	0,432
1405	41	d85361	4	0,616	0,456	0,435
1438	42	d85817	6	0,646	0,569	0,552
1215	43	d85808	4	0,591	0,458	0,437
1406	44	d85423	4	0,674	0,447	0,427
1383	45	d85086	3	0,654	0,370	0,347
1226	46	d85206	3	0,849	0,296	0,278
1390	47	d85589	3	0,557	0,378	0,355
1448	48	d85161	3	0,715	0,356	0,335
1221	49	d76134	5	0,793	0,468	0,451
1402	50	d85274	5	0,693	0,508	0,489
1431	51	d85582	5	0,424	0,508	0,489
1434	52	d85505	5	0,535	0,523	0,504
1230	53	d85193	4	0,591	0,458	0,437
1426	54	d76067	4	0,730	0,432	0,413
1377	55	d85757	5	0,810	0,457	0,441
1240	56	d85299	6	0,563	0,575	0,557
1381	57	d85821	3	0,809	0,320	0,301
1247	58	d85549	3	0,809	0,320	0,301
1447	59	d85045	3	0,463	0,375	0,351

1397	60	d85828	4	0,616	0,456	0,435
1451	61	d85275	3	0,715	0,356	0,335
1407	62	d85360	3	0,499	0,377	0,354
1243	63	d85051	5	0,714	0,502	0,484
1424	64	d85055	3	0,754	0,344	0,323
1234	65	d84066	4	0,722	0,435	0,415
1219	66	d85038	3	0,750	0,345	0,325
1430	67	d85655	5	0,630	0,520	0,501
1441	68	d85019	4	0,596	0,458	0,437
1389	69	d85310	4	0,654	0,451	0,430
1244	70	d85109	4	0,395	0,441	0,420
1435	71	d85832	5	0,796	0,466	0,449
1222	72	d85314	4	0,646	0,453	0,432
1380	73	d76009	4	0,499	0,457	0,436
1229	74	d85509	4	0,821	0,389	0,371
1422	75	d74742	3	0,564	0,378	0,355
1395	76	d76089	5	0,422	0,508	0,488
1224	77	d85784	5	0,639	0,519	0,500
1376	78	d85196	3	0,751	0,345	0,324
1237	79	d85335	4	0,518	0,459	0,437
1443	80	d74732	2	0,647	0,271	0,247
1393	81	d84067	4	0,635	0,454	0,433
1453	82	d76025	2	0,454	0,276	0,251
1409	83	d85624	3	0,499	0,377	0,354
1449	84	d85316	3	0,773	0,337	0,317
1217	85	d85678	2	0,628	0,273	0,249
1387	86	d85550	3	0,790	0,329	0,310
1440	87	d85547	3	0,668	0,367	0,345
1404	88	d85658	3	0,596	0,376	0,353
1228	89	d85446	4	0,483	0,456	0,434
1421	90	d85174	3	0,831	0,308	0,289
1220	91	d85741	2	0,693	0,263	0,240
1378	92	d85326	3	0,557	0,378	0,355
1245	93	d85287	3	0,702	0,360	0,338
1444	94	d85744	5	0,688	0,509	0,491
1427	95	d85337	3	0,463	0,375	0,351
1235	96	d85324	5	0,594	0,523	0,504

dig_M8_4

Item	InBk	Label	Dsc	P-Val	RIT	RIR
1214	1	d85017	4	0,646	0,451	0,434
1394	2	d85667	4	0,732	0,430	0,413
1403	3	d85661	5	0,460	0,515	0,498
1429	4	d85682	5	0,630	0,519	0,503
1452	5	d85044	3	0,537	0,376	0,357
1384	6	d85476	5	0,635	0,518	0,502
1436	7	d74951	3	0,583	0,375	0,356
1420	8	d85102	3	0,744	0,345	0,328
1238	9	d74953	2	0,750	0,247	0,228
1213	10	d74807	3	0,789	0,328	0,311
1375	11	d85725	3	0,770	0,336	0,319
1385	12	d72978	4	0,848	0,368	0,354
1225	13	d75231	2	0,656	0,266	0,247
1398	14	d85114	3	0,751	0,343	0,326
1445	15	d85211	4	0,740	0,427	0,411
1248	16	d84064	5	0,658	0,515	0,499
1423	17	d85524	5	0,639	0,518	0,502
1379	18	d76030	3	0,751	0,343	0,326
1392	19	d85518	5	0,752	0,487	0,472
1232	20	d74825	4	0,448	0,450	0,432
1218	21	d85034	4	0,492	0,455	0,438
1442	22	d85521	4	0,586	0,457	0,439
1437	23	d85020	5	0,688	0,508	0,493
1391	24	d72994	3	0,518	0,376	0,357
1227	25	d75374	2	0,684	0,262	0,242
1241	26	d85013	2	0,703	0,258	0,239
1401	27	d85629	4	0,693	0,441	0,424
1223	28	d85156	3	0,740	0,346	0,329
1231	29	d85399	4	0,618	0,455	0,437
1388	30	d85199	3	0,557	0,376	0,357
1454	31	d85763	3	0,687	0,361	0,343
1396	32	d76034	5	0,654	0,515	0,499
1239	33	d85731	4	0,545	0,458	0,440
1446	34	d85352	5	0,747	0,489	0,474
1236	35	d85198	7	0,723	0,591	0,578
1425	36	d85187	5	0,756	0,485	0,470
1400	37	d85447	3	0,596	0,374	0,355
1216	38	d85448	3	0,655	0,367	0,349
1433	39	d85533	5	0,678	0,511	0,495
1386	40	d85733	4	0,461	0,452	0,434
1405	41	d85361	4	0,616	0,455	0,437
1438	42	d85817	6	0,646	0,569	0,554
1215	43	d85808	4	0,591	0,457	0,439
1406	44	d85423	4	0,674	0,446	0,429
1383	45	d85086	3	0,654	0,367	0,349
1226	46	d85206	3	0,849	0,294	0,279
1390	47	d85589	3	0,557	0,376	0,357
1448	48	d85161	3	0,715	0,354	0,336
1221	49	d76134	5	0,793	0,467	0,453
1402	50	d85274	5	0,693	0,507	0,491
1431	51	d85582	5	0,424	0,507	0,491
1434	52	d85505	5	0,535	0,522	0,506
1230	53	d85193	4	0,591	0,457	0,439
1426	54	d76067	4	0,730	0,430	0,414
1377	55	d85757	5	0,810	0,456	0,442
1240	56	d85299	6	0,563	0,574	0,559
1381	57	d85821	3	0,809	0,318	0,302
1247	58	d85549	3	0,809	0,318	0,302
1447	59	d85045	3	0,463	0,372	0,353
1397	60	d85828	4	0,616	0,455	0,437
1451	61	d85275	3	0,715	0,354	0,336
1407	62	d85360	3	0,499	0,375	0,356

1243	63	d85051	5	0,714	0,501	0,486
1424	64	d85055	3	0,754	0,342	0,325
1234	65	d84066	4	0,722	0,433	0,417
1219	66	d85038	3	0,750	0,343	0,326
1430	67	d85655	5	0,630	0,519	0,503
1441	68	d85019	4	0,596	0,456	0,439
1389	69	d85310	4	0,654	0,449	0,432
1244	70	d85109	4	0,395	0,440	0,422
1435	71	d85832	5	0,796	0,465	0,451
1222	72	d85314	4	0,646	0,451	0,434
1380	73	d76009	4	0,499	0,456	0,438
1229	74	d85509	4	0,821	0,387	0,373
1422	75	d74742	3	0,564	0,376	0,357
1395	76	d76089	5	0,422	0,507	0,491
1224	77	d85784	5	0,639	0,518	0,502
1376	78	d85196	3	0,751	0,343	0,326
1237	79	d85335	4	0,518	0,457	0,439
1443	80	d74732	2	0,647	0,268	0,248
1393	81	d84067	4	0,635	0,452	0,435
1453	82	d76025	2	0,454	0,273	0,252
1409	83	d85624	3	0,499	0,375	0,356
1449	84	d85316	3	0,773	0,335	0,318
1217	85	d85678	2	0,628	0,270	0,250
1387	86	d85550	3	0,790	0,327	0,311
1440	87	d85547	3	0,668	0,365	0,346
1404	88	d85658	3	0,596	0,374	0,355
1228	89	d85446	4	0,483	0,454	0,437
1421	90	d85174	3	0,831	0,306	0,291
1220	91	d85741	2	0,693	0,260	0,241
1378	92	d85326	3	0,557	0,376	0,357
1245	93	d85287	3	0,702	0,357	0,339
1444	94	d85744	5	0,688	0,508	0,493
1427	95	d85337	3	0,463	0,372	0,353
1235	96	d85324	5	0,594	0,522	0,506
1419	97	d85688	4	0,548	0,458	0,440
1410	98	d74838	4	0,720	0,434	0,417
1374	99	d85812	5	0,819	0,451	0,437
1415	100	d85804	4	0,408	0,443	0,425
1206	101	d85800	3	0,544	0,376	0,357
1414	102	d85564	3	0,769	0,336	0,319
1369	103	d85568	4	0,676	0,445	0,428
1367	104	d85693	3	0,838	0,301	0,286
1373	105	d74978	4	0,724	0,433	0,416
1416	106	d76046	5	0,575	0,523	0,507
1371	107	d76065	3	0,855	0,291	0,276
1209	108	d85577	4	0,577	0,458	0,440
1211	109	d85698	4	0,577	0,458	0,440
1368	110	d85789	4	0,565	0,458	0,440
1207	111	d85826	4	0,577	0,458	0,440
1411	112	d85696	3	0,484	0,374	0,355
1205	113	d85559	6	0,683	0,562	0,547
1366	114	d85558	3	0,600	0,374	0,355
1412	115	d85553	3	0,652	0,368	0,349
1417	116	d85699	3	0,433	0,369	0,350

pap_M8_4

Item	InBk	Label	Dsc	P-Val	RIT	RIR
1089	1	85017	4	0,756	0,423	0,407
755	2	85667	5	0,811	0,460	0,446
1084	3	85661	5	0,474	0,514	0,498
786	4	85682	5	0,706	0,505	0,490
805	5	85077	3	0,450	0,370	0,350
990	6	85097	5	0,524	0,520	0,504
1153	7	85476	5	0,627	0,520	0,504
1124	8	74951	3	0,567	0,375	0,356
1151	9	85102	4	0,724	0,434	0,418
1121	10	74807	4	0,771	0,416	0,401
761	11	85725	3	0,730	0,351	0,333
1108	12	85457	4	0,752	0,424	0,409
1066	13	72978	3	0,726	0,352	0,334
1004	14	75025	3	0,672	0,365	0,346
817	15	84064	4	0,796	0,404	0,389
1000	16	85211	3	0,790	0,329	0,312
815	17	85524	5	0,692	0,509	0,494
1146	18	85518	4	0,811	0,396	0,382
767	19	74782	5	0,586	0,522	0,506
875	20	76030	5	0,756	0,488	0,474
829	21	85020	4	0,697	0,442	0,425
869	22	75374	3	0,589	0,374	0,355
1100	23	72994	4	0,460	0,450	0,432
795	24	85013	4	0,722	0,435	0,419
979	25	85629	4	0,746	0,427	0,411
1095	26	76142	3	0,737	0,349	0,331
826	27	84070	4	0,729	0,433	0,416
1061	28	85145	4	0,404	0,439	0,422
779	29	85156	4	0,760	0,421	0,406
983	30	85722	6	0,636	0,571	0,557
1087	31	85199	5	0,585	0,522	0,506
1081	32	85399	4	0,701	0,441	0,424
872	33	76125	5	0,534	0,521	0,505
789	34	85731	5	0,554	0,522	0,506
748	35	85198	5	0,776	0,479	0,465
1097	36	85410	4	0,475	0,452	0,434
1134	37	85447	3	0,683	0,362	0,344
1143	38	85448	4	0,636	0,453	0,436
995	39	85533	6	0,709	0,557	0,543
1056	40	85733	5	0,440	0,508	0,492
1055	41	85817	4	0,740	0,429	0,413
1107	42	85206	5	0,733	0,497	0,482
1060	43	85161	4	0,755	0,423	0,408
1050	44	85589	4	0,654	0,450	0,433
1142	45	85273	4	0,607	0,456	0,438
1117	46	76134	5	0,706	0,505	0,490
1145	47	85505	3	0,606	0,373	0,354
1064	48	85274	3	0,766	0,339	0,322
999	49	85418	5	0,871	0,412	0,400
705	50	85193	3	0,712	0,356	0,338
1005	51	85230	4	0,432	0,445	0,427
1140	52	85271	5	0,507	0,519	0,503
1130	53	85354	6	0,789	0,524	0,511
713	54	85400	3	0,756	0,342	0,325
994	55	85808	4	0,622	0,454	0,437
1115	56	85086	4	0,771	0,417	0,401
737	57	84072	4	0,886	0,338	0,325
1137	58	85045	3	0,545	0,376	0,356
1154	59	85821	3	0,801	0,323	0,307
1114	60	85820	3	0,398	0,361	0,342
1059	61	85828	3	0,612	0,373	0,354
1003	62	85360	3	0,702	0,358	0,340

1072	63	84066	5	0,695	0,508	0,493
854	64	85055	3	0,657	0,367	0,349
845	65	72775	5	0,767	0,484	0,469
1129	66	85310	5	0,723	0,500	0,485
1152	67	85019	4	0,701	0,441	0,424
734	68	85655	6	0,660	0,568	0,553
998	69	85109	5	0,423	0,504	0,488
848	70	85832	5	0,832	0,445	0,432
1043	71	85314	4	0,668	0,448	0,431
1067	72	76009	4	0,658	0,450	0,433
1141	73	85509	3	0,799	0,325	0,309
1148	74	74742	3	0,567	0,375	0,356
840	75	74725	3	0,696	0,360	0,342
851	76	85784	5	0,673	0,513	0,498
1099	77	76089	5	0,403	0,499	0,483
993	78	85335	4	0,556	0,457	0,440
855	79	85415	4	0,913	0,306	0,295
1096	80	85517	3	0,434	0,367	0,348
1123	81	84067	4	0,667	0,448	0,431
1139	82	76025	3	0,633	0,371	0,352
1040	83	85280	3	0,703	0,358	0,340
1063	84	85010	3	0,899	0,255	0,243
1062	85	85547	3	0,705	0,357	0,340
1058	86	85626	7	0,458	0,599	0,585
1002	87	85446	5	0,530	0,521	0,505
858	88	85658	3	0,825	0,311	0,296
1128	89	85427	4	0,684	0,445	0,428
866	90	85745	5	0,886	0,396	0,384
1054	91	85326	3	0,562	0,375	0,356
1051	92	85287	5	0,708	0,505	0,490
1049	93	85337	3	0,434	0,367	0,348
1045	94	85027	4	0,677	0,446	0,429
1077	95	85324	4	0,665	0,448	0,431
1125	96	85706	5	0,614	0,521	0,505
1171	97	85572	3	0,772	0,336	0,320
1157	98	85693	4	0,751	0,425	0,409
1160	99	85568	5	0,658	0,516	0,500
1188	100	85564	3	0,809	0,320	0,304
1198	101	74873	3	0,848	0,297	0,282
1161	102	74838	4	0,824	0,388	0,374
1166	103	74702	4	0,809	0,397	0,383
1178	104	85812	4	0,751	0,425	0,409
1159	105	74978	5	0,812	0,459	0,446
1164	106	76046	4	0,606	0,456	0,438
1162	107	76065	4	0,858	0,363	0,350
1173	108	85573	4	0,444	0,447	0,430
1170	109	85698	4	0,505	0,455	0,437
1175	110	85789	4	0,533	0,457	0,439
1179	111	85815	3	0,577	0,375	0,356
1156	112	85558	3	0,651	0,368	0,350
1158	113	85559	5	0,716	0,503	0,488
1163	114	85805	4	0,551	0,457	0,440
1189	115	85700	3	0,758	0,342	0,325
1169	116	85816	3	0,568	0,375	0,356

pap_M8_3

Item	InBk	Label	Dsc	P-Val	RIT	RIR
979	1	85629	4	0,746	0,428	0,409
705	2	85193	3	0,712	0,358	0,337
845	3	72775	5	0,767	0,484	0,467
829	4	85020	4	0,697	0,443	0,424
855	5	85415	4	0,913	0,306	0,293
1128	6	85427	4	0,684	0,446	0,426
789	7	85731	5	0,554	0,524	0,505
1003	8	85360	3	0,702	0,360	0,339
999	9	85418	5	0,871	0,412	0,398
815	10	85524	5	0,692	0,510	0,492
851	11	85784	5	0,673	0,514	0,496
1051	12	85287	5	0,708	0,506	0,488
1089	13	85017	4	0,756	0,424	0,406
1059	14	85828	3	0,612	0,375	0,353
1117	15	76134	5	0,706	0,506	0,488
840	16	74725	3	0,696	0,362	0,340
1004	17	75025	3	0,672	0,367	0,345
1045	18	85027	4	0,677	0,448	0,428
1095	19	76142	3	0,737	0,351	0,330
1060	20	85161	4	0,755	0,425	0,406
990	21	85097	5	0,524	0,522	0,503
1137	22	85045	3	0,545	0,378	0,355
761	23	85725	3	0,730	0,353	0,332
1043	24	85314	4	0,668	0,450	0,430
1002	25	85446	5	0,530	0,523	0,503
1134	26	85447	3	0,683	0,365	0,343
1056	27	85733	5	0,440	0,510	0,491
994	28	85808	4	0,622	0,456	0,436
1121	29	74807	4	0,771	0,418	0,399
734	30	85655	6	0,660	0,569	0,552
1139	31	76025	3	0,633	0,373	0,351
1087	32	85199	5	0,585	0,524	0,505
748	33	85198	5	0,776	0,480	0,463
1115	34	85086	4	0,771	0,418	0,400
1099	35	76089	5	0,403	0,501	0,482
875	36	76030	5	0,756	0,489	0,472
1077	37	85324	4	0,665	0,450	0,430
1143	38	85448	4	0,636	0,455	0,434
1130	39	85354	6	0,789	0,524	0,509
767	40	74782	5	0,586	0,524	0,505
1141	41	85509	3	0,799	0,326	0,307
1054	42	85326	3	0,562	0,378	0,355
1081	43	85399	4	0,701	0,442	0,422
1140	44	85271	5	0,507	0,521	0,501
817	45	84064	4	0,796	0,406	0,388
848	46	85832	5	0,832	0,445	0,430
858	47	85658	3	0,825	0,313	0,294
779	48	85156	4	0,760	0,422	0,404
1145	49	85505	3	0,606	0,376	0,353
1066	50	72978	3	0,726	0,354	0,333
1152	51	85019	4	0,701	0,442	0,422
1062	52	85547	3	0,705	0,360	0,338
1061	53	85145	4	0,404	0,441	0,421
1050	54	85589	4	0,654	0,452	0,432
1100	55	72994	4	0,460	0,452	0,431
1151	56	85102	4	0,724	0,435	0,416
1072	57	84066	5	0,695	0,509	0,491
1063	58	85010	3	0,899	0,257	0,242
1084	59	85661	5	0,474	0,516	0,497
1055	60	85817	4	0,740	0,430	0,411
1153	61	85476	5	0,627	0,521	0,502
1154	62	85821	3	0,801	0,325	0,306

805	63	85077	3	0,450	0,372	0,349
1123	64	84067	4	0,667	0,450	0,430
1064	65	85274	3	0,766	0,341	0,321
1107	66	85206	5	0,733	0,498	0,480
993	67	85335	4	0,556	0,459	0,438
1114	68	85820	3	0,398	0,364	0,341
1125	69	85706	5	0,614	0,522	0,503
995	70	85533	6	0,709	0,558	0,541
737	71	84072	4	0,886	0,338	0,323
869	72	75374	3	0,589	0,377	0,354
1148	73	74742	3	0,567	0,378	0,355
1049	74	85337	3	0,434	0,370	0,347
1097	75	85410	4	0,475	0,454	0,433
713	76	85400	3	0,756	0,344	0,324
1146	77	85518	4	0,811	0,397	0,380
1067	78	76009	4	0,658	0,451	0,431
866	79	85745	5	0,886	0,396	0,382
872	80	76125	5	0,534	0,523	0,504
1005	81	85230	4	0,432	0,447	0,426
795	82	85013	4	0,722	0,436	0,417
998	83	85109	5	0,423	0,506	0,487
1058	84	85626	7	0,458	0,600	0,584
983	85	85722	6	0,636	0,572	0,555
786	86	85682	5	0,706	0,506	0,488
1108	87	85457	4	0,752	0,426	0,407
1129	88	85310	5	0,723	0,501	0,484
1040	89	85280	3	0,703	0,360	0,339
755	90	85667	5	0,811	0,460	0,444
826	91	84070	4	0,729	0,434	0,415
1142	92	85273	4	0,607	0,458	0,437
1124	93	74951	3	0,567	0,378	0,355
854	94	85055	3	0,657	0,369	0,348
1096	95	85517	3	0,434	0,370	0,347
1000	96	85211	3	0,790	0,330	0,311

Cito

Amsterdamseweg 13
6814 CM Arnhem
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11

Fotografie: Gijs Versteeg

