Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 3

Jan Janssen, Michel Hop, Jasper Wouda en Judith Hollenberg

zeker weten

Wetenschappelijke verantwoording

Rekenen-Wiskunde 3.0 voor groep 3

Jan Janssen
Michel Hop
Jasper Wouda
Judith Hollenberg

Cito
Arnhem, 2015

Inhoud

Bijlagen 77

1 Inleiding

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

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

Criteriumvaliditeit is voor de toetsen Rekenen-Wiskunde 3.0 groep 3 van het Cito Volgsysteem niet van toepassing aangezien de toetsen geen voorspellende meetpretentie hebben.
Daarnaast is de verantwoording bestemd voor docenten die zich nader willen verdiepen in de achtergronden van de toetsen. Ook personen en instanties die in tweede instantie toetsscores van leerlingen in handen krijgen, vinden in deze verantwoording voldoende aanknopingspunten voor de interpretatie van deze scores.
De toetsen Rekenen-Wiskunde 3.0 bestaan – naast deze wetenschappelijke verantwoording – uit de volgende onderdelen:

- Handleiding, waarin opgenomen een inhoudsverantwoording
- De toetsboekjes M3, M3E3 en E3 voor de afnames op papier
- Afnamekaarten en nakijkkaarten voor de afnames op papier
- De digitale toetsen M3, M3E3 en E3 voor afnames achter de computer.

De reguliere toetsmomenten zijn medio groep 3 en eind groep 3. Naast een M3 toets en een E3 toets is er een tussentoets M3E3 ontwikkeld die ook halverwege het leerjaar of aan het eind van het leerjaar afgenomen kan worden. Medio groep 3 zou de tussentoets M3E3 ingezet kunnen worden voor leerlingen die al wat verder in rekenontwikkeling zijn en eind groep 3 kan de tussentoets M3E3 ingezet worden voor leerlingen waarbij de rekenvaardigheid zich minder snel ontwikkeld heeft.
Naast een rapportage van het A tot en met E niveau of I tot en met V niveau, worden ook functionerings-niveaus gerapporteerd die aangeven met welk gemiddeld niveau de vaardigheid van een leerling het best te vergelijken is.
De digitale en papieren varianten zijn op één schaal gebracht en zijn volledig vergelijkbaar.
Binnen het Cito Volgsysteem kan het toetsen op maat gerealiseerd worden.

De nieuwe toetsen rekenen-wiskunde voor groep 3 worden de toetsen Rekenen-Wiskunde 3.0 voor groep 3 uit het Cito Volgsysteem primair en speciaal onderwijs genoemd. De toetsen van de vorige uitgave worden de tweede generatie toetsen of LVS II genoemd. Om de leesbaarheid van dit document te bevorderen worden de toetsen Rekenen-Wiskunde 3.0 uit het Cito Volgsysteem ook met andere benamingen aangeduid. Deze kunnen zijn: derde generatie toetsen, LVS III – toetsen of Rekenen-Wiskunde 3.0.

2 Uitgangspunten van de toetsconstructie

2.1 Meetpretentie

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

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

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

2.2 Doelgroep

De toetsen zijn een onderdeel van het Cito Volgsysteem primair en speciaal onderwijs. De toetsen Rekenen-Wiskunde groep 3 zijn bedoeld voor leerlingen in groep 3 van het primair onderwijs en leerlingen in het speciaal (basis)onderwijs die functioneren op het niveau van groep 3 in het reguliere basisonderwijs. De toetsen zijn ook te gebruiken voor leerlingen in andere leerjaren die een rekenvaardigheid hebben op het niveau van groep 3. In de handleiding is toegelicht hoe dit toetsen op maat, voor wat betreft het selecteren van toetsen en interpreteren van de resultaten, werkt.
Voor leerlingen met een ontwikkelingsachterstand en/of extra onderwijsbehoeften zijn in de handleiding extra aanwijzingen opgenomen. Daartoe bevat de toetsmap een extra toetsboekje en zijn alternatieve rapportages ontwikkeld.

2.3 Gebruiksdoel en functie

De hoofddoelen van de toetsen Rekenen-Wiskunde zijn het in kaart brengen van het rekenvaardigheids-niveau en de ontwikkeling van de rekenvaardigheid van (groepen) leerlingen te 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.
- in kaart te brengen of er bij individuele leerlingen of bij de groep als geheel verschillen in vaardigheid zijn tussen het oplossen van kale opgaven en het oplossen van contextopgaven. Deze informatie is te vinden in een categorieënanalyse kaal-context die eveneens op te vragen is met het Computer-programma LOVS. Voor groep 3 worden voor deze analyse de bewerkingsopgaven van de categorie optellen en aftrekken gebruikt.

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

Hoe kunnen we de LVS-toetsen Rekenen-Wiskunde 3.0 inzetten om leerlingen te volgen in de tijd? Globaal kunnen we de toetsresultaten van leerlingen (of groepen leerlingen) op twee manieren interpreteren:

a. We kunnen het toetsresultaat van een leerling vergelijken met die van andere leerlingen op hetzelfde meettijdstip (afnamemoment).
b. We kunnen de toetsresultaten van dezelfde leerling vergelijken met diens eigen toetsresultaten op eerdere of latere meettijdstippen (afnamemomenten).

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

Ad a.
Bij de eerstgenoemde vergelijking worden de prestaties van een leerling vergeleken met de prestaties van de hele populatie op een gegeven afnamemoment. Hoe doet een leerling het, bijvoorbeeld, ten opzichte van de gemiddelde leerling? Voor dit doel is de populatie, op basis van de data die verzameld zijn in het kader van het normeringsonderzoek (zie hiervoor hoofdstuk 4 van deze wetenschappelijke verantwoording), ingedeeld in vaardigheidsniveaus (I-V, A-E). Vaardigheidsniveau I, bijvoorbeeld, bevat de 20% hoogst scorende leerlingen. Door de vaardigheidsscore van een leerling te vergelijken met deze vaardigheids-niveaus (die zijn afgebakend door percentielpunten die horen bij specifieke vaardigheidsscores), zijn uitspraken mogelijk zoals "Meriam heeft op afnamemoment einde leerjaar 3 vaardigheidsniveau IV 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 M3 had Meriam vaardigheidsniveau IV en op tijdstip E3 was het vaardigheidsniveau V". Bij de tweede methode worden de **vaardigheidsscores** op de twee verschillende momenten vergeleken: vaardigheidsscore 106, bijvoorbeeld, op tijdstip M3 en vaardigheidsscore 110 op tijdstip E3. Ook hier geldt, net als bij het vergelijken van prestaties met die van andere leerlingen, dat bij eventuele verdere acties van de leerkracht ook andere aspecten moeten worden betrokken.

Bij alle vergelijkingen die mogelijk zijn, zowel die ad a. als die ad b., dienen uitspraken over leerlingen te worden gerelativeerd. In de handleiding is meer informatie te vinden over de wijze waarop de gebruiker dit kan doen. Hieronder gaan we vooral in op het belang van de (on)betrouwbaarheid van de afgenomen toetsen hierbij.
Voor elke toets geldt dat de vaardigheidsscore die bij een toetsresultaat van een leerling hoort, behept is met een meetfout. Als we rekening houden met die meetfout dan zou het best zo kunnen zijn dat Meriam vaardigheidsniveau III heeft behaald op het eerste tijdstip en niet vaardigheidsniveau IV. Of dat we moeten concluderen dat het verschil in de prestaties op de twee tijdstippen statistisch niet significant is: Meriam is voor- noch achteruitgegaan. In alle gevallen speelt het betrouwbaarheidsinterval (BI) rondom de vaardigheidsscore een belangrijke rol. Dat geldt ook voor de indeling in vaardigheidsniveaus. Hoe het BI doorwerkt in de indeling in vaardigheidsniveaus en de verdere gevolgen daarvan wordt beschreven in hoofdstuk 5. Op deze plaats beperken we ons tot de vraag of we de uitspraak kunnen doen dat een leerling of groep (werkelijk) 'gegroeid' is. De eenvoudigste manier is om te kijken of de BI's voor de twee tijdstippen overlappen. Als deze twee BI's niet overlappen dan is er sprake van een significant verschil in vaardigheid tussen beide tijdstippen. Overlappen ze wel dan is er geen verschil in vaardigheid. Deze eenvoudige manier van vergelijken kan de leerkracht zelf uitvoeren en wordt ook in de handleiding beschreven.

We geven hier een voorbeeld. Bij de afname Rekenen-Wiskunde M3 behaalde Wout een vaardigheidsscore van 93 met een 67% betrouwbaarheidsinterval van 85-100. Bij de afname E3 behaalde Wout een vaardigheidsscore van 118; het bijbehorende betrouwbaarheidsinterval daarbij is 111-125. Aangezien de betrouwbaarheidsintervallen niet overlappen kunnen we zeggen dat Wouts vaardigheid is toegenomen.

Conclusie
De vaardigheidsgroei voor Rekenen-Wiskunde voltrekt zich langzaam in de tijd. De verschillen tussen vaardigheidsscores op achtereenvolgende meettijdstippen zijn betrekkelijk klein, al is de gemiddelde vaardigheidstoename in groep 3 en groep 4 wat groter dan in de hogere leerjaren. 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’.

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

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

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

Domeinbeschrijving groep 3 tot en met 8

Voor de uitwerking van de kerndoelen tot een domeinbeschrijving is gebruikgemaakt van de inhoud van de referentieniveaus (Expertgroep doorlopende leerlijnen taal en rekenen, 2008 en Expertgroep doorlopende leerlijnen, 2008a), de tussendoelen van de SLO (Buijs, 2008 en Buijs, 2008a), de publicaties van het TAL-team (TAL-team 1999, 2001, 2004, 2005 en 2007) 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.
De domeinbeschrijving is gemaakt voor groep 3 tot en met 8, om zo de doorgaande lijn in beeld te houden en bestaat uit een beschrijving van het leerstofgebied rekenen-wiskunde in de vorm van een lijst met leerdoelen.
De verschillende onderdelen van het domein rekenen-wiskunde vormen een samenhangend geheel van getalbegrip en rekenvaardigheid. Hierin staan inzicht in getallen, maatinzicht, ruimtelijk inzicht en het kunnen uitvoeren van operaties met getallen en het kunnen toepassen van die kennis en inzichten in uiteenlopende situaties centraal. We onderscheiden in de domeinbeschrijving voor het basisonderwijs, overeenkomstig de referentieniveaus, de volgende vier domeinen:

- Getallen;
- Verhoudingen;
- Meten en meetkunde;
- Verbanden

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

Uitwerking domeinbeschrijving

Getallen

1. Getalbegrip

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

2. Bewerkingen

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

Verhoudingen

3. Verhoudingen

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

Meten en meetkunde

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 veel voorkomende inhoudsmaten, kennis van maten en het toepassen van deze aspecten.

4.4. Meten: gewicht

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

4.5. Tijd en snelheid

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

4.6. Geld

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

5. Meetkunde

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

Verbanden

6. Tabellen, diagrammen en grafieken

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

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

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

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

Voor groep 3 zijn twee subdomeinen van belang, namelijk het domein Getallen en het domein Meten en meetkunde. Nadere informatie over de invulling van deze onderdelen in groep 3 staat in hoofdstuk 3.

Ontwikkeling van rekenen-wiskunde
De verschillende methodes bieden een gevarieerd aanbod van oefeningen en activiteiten aan om te zorgen dat leerlingen in staat zijn het systeem van hele getallen te construeren. Getallen, zowel het telgetal, het getal dat aantallen voorstelt, als het meetgetal komen aan de orde. Leerlingen ontwikkelen een mentaal beeld van hoeveelheden, orde van grootte, notie van de plaats van getallen in de telrij en de associatie van hoeveelheid en plaats in de telrij met de uitspraak en schrijfwijze van de getallen.
In primair onderwijs richt het onderwijs in rekenen-wiskunde zich op de ontwikkeling van handelings-gebonden, via contextgebonden, naar systeemgebonden rekenen.

- Bij *handelingsgebonden rekenen* gebruikt de leerling objecten of zijn vingers om een rekensituatie af te beelden. Getallen staan voor de leerlingen in eerste instantie los van elkaar, leerlingen ontdekken vervolgens dat aantallen samengevoegd kunnen worden en dus dat elk getal de som is van twee of meer andere getallen.
- Binnen het *contextgebonden rekenen* ontdekt de leerling de structuur van getallen, bijvoorbeeld de structuur in tientallen en eenheden. In deze fase ontwikkelen de leerlingen drie vormen van rekenen: rijgen, splitsen en afleiden. Bij rijgen maakt de leerling bijvoorbeeld sprongen van tien op een (denkbeeldige) getallenlijn. Bij splitsen rekent de leerling apart met tientallen en met eenheden en voegt ze later weer bij elkaar. Afleiden is een vorm van rekenen waarbij de leerling handig gebruik-maakt van rekenfeitjes. In de fase van het contextgebonden rekenen komen leerlingen erachter dat de verschillende operaties (optellen, aftrekken, vermenigvuldigen en delen) relaties met elkaar hebben. Deze informatie is waardevol om afleiden als rekenvorm te kunnen toepassen.
- Leerlingen kunnen *systeemgebonden rekenen* als zij gebruikmaken van rijg- en splitsalgoritmen en wanneer ze afleiden volgens meer abstracte regels zoals verdubbelen-halveren en de nulregels. De leerlingen leren in deze fase om bij het rijgen rekenhandelingen maximaal te verkorten. Ze maken wat splitsen betreft de stap van kolomsgewijs rekenen naar cijferend rekenen. Dit betekent dat leerlingen van rechts naar links in plaats van van links naar rechts rekenen (Kraemer, 2008).

2.4.2 Psychometrisch

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

2.4.2.1 Opgavenbanken en constructieprocedures
Bij de constructie van opgaven wordt in de regel een veelvoud van het aantal items dat uiteindelijk in de normeringstoets moet worden ingezet afgenomen in een proeftoets. Er moet immers rekening worden gehouden met uitval, bijvoorbeeld wegens meer of minder triviale fouten in de constructie of extreme moeilijkheid of gemakkelijkheid. Ook ontstaat er op deze manier een overschot aan kwalitatief goede opgaven, die aan de opgavenbank worden toegevoegd. Een nieuwe toets wordt samengesteld uit een aantal nieuw geproeftoetste opgaven en uit opgaven die al eerder in de opgavenbank waren opgenomen. Een belangrijk kenmerk van deze opgavenbanken is dat ze gekalibreerd zijn volgens de principes van de IRT: item respons theorie. Voor Rekenen-Wiskunde 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 toetsconstructeur min of meer naar willekeur een aantal items selecteert om een nieuwe toets te construeren. Hieronder wordt beschreven wat de vereisten zijn om van een deugdelijke en psychometrisch goed gefundeerde opgavenbank te kunnen spreken.

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

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

– *Kansmodel*
De ruwe omschrijving van de moeilijkheidsgraad die in de vorige alinea werd gehanteerd (de mate van vaardigheid die nodig is om het item goed te kunnen beantwoorden) behoeft verdere uitwerking. Men zou deze omschrijving kunnen opvatten als een drempel: heeft een leerling die mate van vaardigheid niet, dan is hij niet in staat het item juist beantwoorden; heeft hij die drempel wel gehaald, dan geeft hij (gegarandeerd) het juiste antwoord. Deze interpretatie weerspiegelt een deterministische kijk op het antwoordgedrag 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 meetmodellen).

– *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 moeilijkheids-graden te achterhalen. Dit gebeurt met een statistische schattingsmethode die wordt toegepast op de itemantwoorden die bij een steekproef van leerlingen zijn verzameld. Het hele proces van moeilijkheidsgraden schatten en verifiëren of de modelveronderstellingen houdbaar zijn, wordt kalibratie of ijking genoemd. De steekproef van leerlingen (in de boven al aangeduide proeftoets) die hiervoor wordt gebruikt heet kalibratiesteekproef.

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

– *Implicaties van gekalibreerde opgavenverzameling*
Als de kalibratie met succes uitgevoerd is, is het resultaat een zogenoemde gekalibreerde itembank. In het kalibratieproces worden de items die niet passen bij de verzameling uit de collectie verwijderd. De opgavenbank bevat voor elk item niet alleen zijn feitelijke inhoud, maar ook zijn psychometrische eigenschappen en de statistische zekerheid dat alle items dezelfde vaardigheid aanspreken (unidimensionaliteit). Dit houdt onder meer het volgende in:
1 In principe kunnen we met een willekeurige selectie items uit de bank de vaardigheid meten bij een willekeurige leerling. De meting is nauwkeuriger wanneer het niveau van de opgaven beter aansluit bij het niveau van de leerling.
Het voorgaande geldt tevens voor de digitale items. Ook deze items komen uit de itembank Rekenen-Wiskunde. Dus ook met een selectie van digitale items kan de vaardigheid van een leerling bepaald worden. Al hetgeen dat geldt voor de ‘papieren’ items uit de itembank, geldt daarom eveneens voor ‘digitale’ items uit dezelfde itembank.
2 We kunnen een schatting maken van de verdeling van de vaardigheid in een welomschreven populatie, door selecties van items voor te leggen aan aselecte steekproeven van leerlingen uit populaties die van belang zijn voor de normering. In het geval van het LVS zijn dat steekproeven van leerlingen op de verschillende normeringsmomenten vanaf medio groep 3 tot medio groep 8. Daarbij maakt het, behoudens wat bij 1 is vermeld over nauwkeurigheid, niet uit welke selectie van items aan een leerling binnen een normeringsgroep wordt afgenomen. Een van de eigenschappen van gekalibreerde itembanken is immers dat met elke selectie items de vaardigheid van leerlingen kan worden bepaald. In de praktijk komt dit meestal neer op het schatten van gemiddelde en standaardafwijking in de veronderstelling dat de vaardigheid normaal verdeeld is. Met deze schattingen kunnen dan ook schattingen gemaakt worden van de percentielen in de populatie.
3 Aan leerlingen die niet tot de betreffende referentiepopulatie behoren, kan dezelfde toets worden voorgelegd. De toetsscore wordt omgezet in een schatting van de vaardigheid en deze schatting kan geplaatst worden in de vaardigheidsverdeling van de populatie. Een leerling met achterstand in groep 8 kan een toets maken die normaliter aan groep 6 wordt voorgelegd, en zijn vaardigheids-schatting kan behalve met de populatie van groep 8 ook vergeleken worden met de percentielen in de populatie van groep 6, met bijvoorbeeld de uitspraak: “De vaardigheid van deze leerling komt overeen met de mediane vaardigheid in groep 6.”
4 De vergelijking die bij punt 3 gemaakt is, kan evengoed plaatsvinden als de (achterstands)leerling een andere toets (i.e. een selectie uit de opgavenbank) maakt dan de toets die normaliter aan groep 6 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 is te vinden in hoofdstuk 4 over de normering van de toets.

2.4.2.2 Het gehanteerde meetmodel
In het normeringsonderzoek is gebruikgemaakt van een op de itemresponstheorie (IRT) gebaseerd meetmodel zoals dat bij Cito gebruikelijk is. Dergelijke modellen verschillen in een aantal opzichten sterk van de klassieke testtheorie (Verhelst, 1993; Verhelst & Kleintjes, 1993a; Verhelst, Glas en Verstralen, 1995). Bij de klassieke testtheorie staan de toets en de toetsscore centraal. Het theoretisch belangrijkste begrip in deze theorie is de zogenaamde ware score, de gemiddelde score die de persoon zou behalen

indien de test een oneindig aantal keren onder dezelfde condities zou worden afgenomen. Deze klassieke testtheorie zou in dit onderzoek niet gebruikt kunnen worden, aangezien het normeringsonderzoek van de rekenwiskundetoetsen een onvolledig design betrof: niet alle leerlingen hadden alle opgaven gemaakt.
Het gebruik van het IRT-model heeft enkele belangrijke voordelen. Op de eerste plaats kunnen de 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 Xi de toevalsvariabele die het antwoord op item *i* voorstelt. Xi neemt de waarde 1 aan in geval van een correct antwoord en 0 in geval van een fout antwoord. Als symbool voor de vaardigheid kiezen we θ (theta). We wijzen erop dat θ niet rechtstreeks observeerbaar is. Dat zijn alleen de antwoorden op de opgaven. Dat is de reden waarom θ een 'latente' variabele wordt genoemd. De itemresponsfunctie *fi*(θ) is gedefinieerd als een conditionele kans:

$$f_i(\theta) = P(X_i = 1 | \theta) \tag{2.1}$$

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

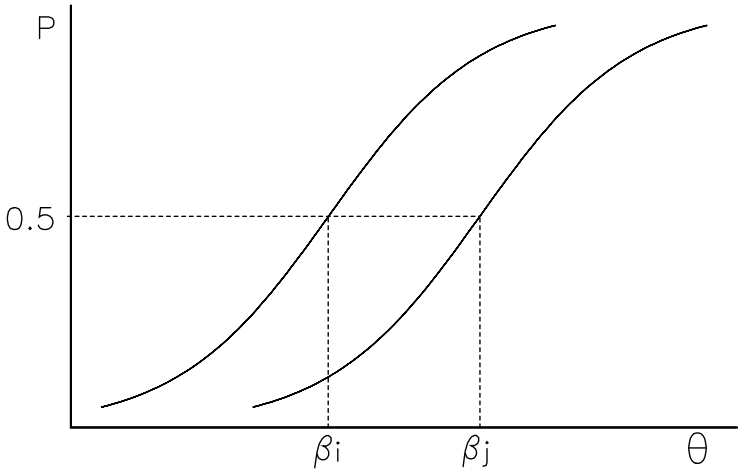
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{2.2}$$

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

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

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.

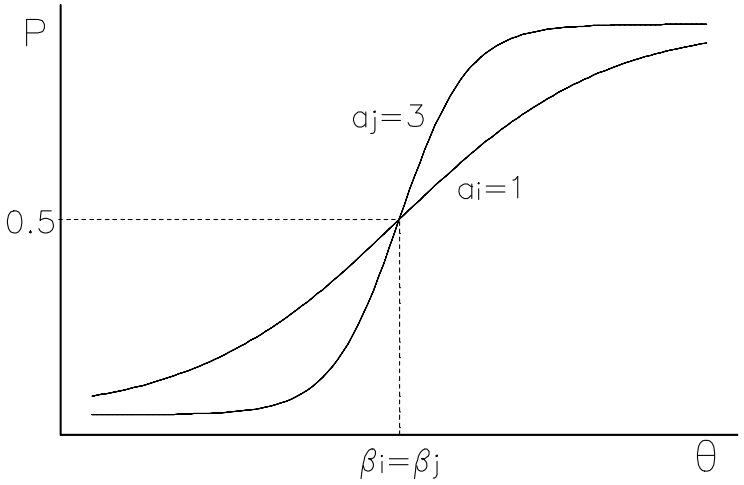
Alvorens het hier gebruikte model te introduceren, is eerst een kanttekening nodig bij het schatten van de moeilijkheidsparameters in het Raschmodel. Een vaak toegepaste schattingsmethode is de 'conditionele grootste aannemelijkheidsmethode' (in het Engels: Conditional Maximum Likelihood, verder aangeduid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

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

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

waarin *ai* 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 scoringsproblemen geven en bij doelen die op andere wijze moeilijk te toetsen zijn is gebruikgemaakt van de meerkeuzevorm. Daarnaast zijn de pure gokkansen bij de meerkeuze-opgaven in de reken-wiskundetoetsen niet zeer groot: bij het willekeurig invullen meestal .25. Hierdoor en door een verstandige dataverzamelingsprocedure toe te passen en met name niet te moeilijke opgaven te selecteren in de test, kan het OPLM toch toegepast worden op meerkeuzevragen, waarbij de overeenkomst tussen model en data de uiteindelijke doorslag over die geschiktheid moet geven. Indien het meetmodel op grond van de kalibratieresultaten aanvaard kan worden,

dat wil zeggen dat er na onderzoek geen praktische reden meer is om aan het meetmodel te twijfelen, dan kan men het meetmodel gebruiken om te gaan meten. Bij deze meetprocedure worden de itemparameters vastgezet op hun geschatte waarde uit de kalibratie.

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

Geldigheid van de normen
De toetsen van het Cito Volgsysteem primair en speciaal onderwijs worden elke acht tot tien jaar vernieuwd. Niet alleen de inhoud wordt volledig vernieuwd en aangepast aan de ontwikkelingen in het onderwijs, ook worden de normen opnieuw vastgesteld. Omdat er enige tijd verloopt tussen de data-verzameling in het normeringsonderzoek en het moment waarop een vernieuwde toets wordt uitgebracht, kan men voor de toetsen Rekenen-Wiskunde groep 3 een geldigheid aanhouden tot en met 2022. 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 voor groep 3 uit het Cito Volgsysteem primair en speciaal onderwijs bevat in totaal drie papieren toetsen: M3, M3E3 en E3 en drie digitale toetsen: M3, M3E3 en E3. De toetsen M3 en E3 zijn de reguliere toetsen, bedoeld voor afname op de reguliere afnamemomenten medio (M) en einde (E) schooljaar. Naast deze toetsen voor de reguliere afnamemomenten bevat het toetspakket ook de tussentoets M3E3. Door het toevoegen van de tussentoets M3E3 is er eind groep 3 ook een toets beschikbaar voor leerlingen waarbij de rekenvaardigheid zich minder snel ontwikkeld heeft. Voor deze leerlingen (vaak zal het gaan om speciale leerlingen) zijn er hierdoor meer toetsen beschikbaar voor het meten van hun rekenvaardigheid. Er is dus een toets voor het functioneringsniveau M3 (medio groep 3) en E3 (eind groep 3), maar ook voor het functioneringsniveau M3E3. Deze laatste toets is wat moeilijker dan de toets voor M3 en wat gemakkelijker dan de toets voor E3. Aan een leerling waarbij de rekenvaardigheid zich minder snel ontwikkelt kan aan het einde van groep 3 dus de toets M3E3 voorgelegd worden.
Deze leerling hoeft op deze manier geen te moeilijke toets (E3) te maken, maar ook niet twee keer dezelfde toets (M3).

Opbouw
In totaal bevat elke toets 52 opgaven. De reguliere toetsen M3 en E3 bestaan uit 2 taken van 26 opgaven. Ze kunnen voor speciale leerlingen opgedeeld worden in drie taken (van respectievelijk 18, 17 en 17 opgaven). In de handleiding zijn de verdelingen daarvoor aangegeven. De tussentoets M3E3 bestaat eveneens uit drie taken van respectievelijk 18, 17 en 17 opgaven. De taken dienen bij voorkeur te worden afgenomen op twee verschillende dagdelen zodat de leerlingen geconcentreerd aan alle taken kunnen werken.

Vorm
De toetsen van groep 3 bevatten naast een beperkt aantal meerkeuzeopgaven vooral korte antwoord-vragen. De opgaven bestaan uit een plaatje. Bij elk plaatje hoort een tekst die niet in het opgavenboekje of op het computerscherm staat. Bij een papieren afname leest de leerkracht de tekst voor, bij een digitale afname leest de computer de opgave voor. Bij de papieren afname beantwoorden de leerlingen de vragen door het antwoord in het hokje bij de afbeelding te schrijven of aan te kruisen. 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 E3-toets voor leerlingen eind groep 3) 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 E3-leerling maakt een toets M3) of een volgend afnamemoment (een M3-leerling maakt de toets E3). Voor leerlingen met een vertraagde ontwikkeling is, zoals eerder aangegeven, de extra toets M3E3 in de toetsmap opgenomen. Deze toets zit qua niveau tussen toets M3 en toets E3 in en kan voorgelegd worden aan leerlingen voor wie de toets E3 nog net te moeilijk is.
Voor zeer zwakke leerlingen of extreem vaardige leerlingen maakt de leerkracht, op basis van eigen observaties, resultaten op methodetoetsen en indien aanwezig eerdere resultaten op de LVS-toetsen een

inschatting van de best passende toets. Hierbij vormt het onderwijsaanbod een belangrijk uitgangspunt: de toets dient zoveel mogelijk aan te sluiten bij de lesstof waar de leerling op dat moment aan werkt.
Een leerling die eind groep 3 nog met de lesstof van medio groep 3 bezig is, kan aan het einde van groep 3 bijvoorbeeld de toets M3 maken.

Het afnemen van de toetsen
De papieren toetsen kunnen zowel klassikaal als individueel afgenomen worden. De leerkracht leest de opgave voor om te zorgen dat zwakke lezers evenveel kans hebben als goede lezers om de opdrachten te begrijpen en goed te maken. Bij elke opgave hoort een afbeelding die de opgavetekst ondersteunt.
Zo staan in de afbeelding veelal de genoemde getallen en/of aantallen en laat de afbeelding waar mogelijk zien welk soort bewerking uitgevoerd moet worden. De afbeeldingen staan in het opgavenboekje van de leerling. De leerling vult zijn of haar antwoord steeds in het vakje in dat onderaan de afbeelding staat.

De digitale versies maakt de leerling individueel. Elke opgave wordt automatisch door de computer voorgelezen. De leerling kan desgewenst door te klikken op het oortje in het scherm het geluidsfragment nogmaals beluisteren.

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

Correctie van de toetsen
De papieren toetsen Rekenen-Wiskunde 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 de Cito portal de vaardigheidsscore bij de behaalde toetsscore op te zoeken.
Bij de digitale versies van de toetsen worden de antwoorden van de leerlingen door de computer gescoord en hoeft de leerkracht de toetsen dus niet zelf na te kijken. Via het Computerprogramma LOVS worden de toetsscores omgezet naar de bijbehorende vaardigheidsscores.

Verwerking resultaten en interpretatie
De resultaten van de digitale afnames worden standaard met de computer verwerkt. De resultaten van de papieren afnames kan de leerkracht zowel met de computer als handmatig verwerken. Voor de handmatige verwerking zijn rapportageformulieren ontwikkeld die beschikbaar zijn via de 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 keuze om functioneringsniveaus op te vragen.

In onderstaand overzicht wordt de betekenis van de vaardigheidsniveaus I tot en met V en A tot en met E toegelicht.

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

Bij de indeling in I tot en met V worden op de overzichten de laagste groep en de hoogste groep nog onderverdeeld in twee groepen die ieder 10% van de leerlingen bevatten. Deze groepen worden van elkaar gescheiden door een stippellijn.
In de eerste versie van de LVS-toetsen werd alleen de indeling A tot en met E gebruikt. In de praktijk bleek deze enkele nadelen te hebben. Ten eerste is deze indeling niet symmetrisch. Bovendien zien sommige leerkrachten C als de gemiddelde groep. Het belangrijkste nadeel is echter dat er geen gemiddelde groep is. Daarom is bij de tweede versie van de toetsen een indeling toegevoegd met de niveaus I tot en met V. De indeling in de niveaus I tot en met V is symmetrisch opgebouwd en heeft als voordeel dat er een gemiddelde groep is. Deze indeling sluit aan bij de niveau-indeling van andere Cito-toetsinstrumenten zoals de Entreetoets.

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 E3 betekent bijvoorbeeld dat de vaardigheidsscore van de leerling heel dicht ligt bij de score van de gemiddelde leerling eind groep 3. De indeling in

functioneringsniveaus is oorspronkelijk ontwikkeld voor het speciaal (basis)onderwijs zodat de school op deze manier voor leerlingen met forse leerachterstanden meer inzicht kregen in hun niveau. Mede dankzij de komst van het ‘passend onderwijs’ ontstond ook bij regulier onderwijs de wens om functioneringsniveaus te gebruiken en zijn de functioneringsniveaus opgenomen in de rapportages.

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

In de hogere leerjaren zijn andere categorieën van toepassing dan in de toetsen van groep 3. Voor M3 bijvoorbeeld worden alleen de categorieën getallen (GET), optellen en aftrekken (O&A), vermenigvuldigen en delen (V&D) en meten (ME) gehanteerd. Het onderdeel breuken, verhoudingen en procenten komt in de hogere groepen wel aan bod, maar is voor groep 3 nog niet relevant. Niet elke categorie is met evenveel items vertegenwoordigd, dat zou immers geen recht doen aan de relatieve belangrijkheid van de categorieën in het onderwijs.
Naast een analyse op basis van rekendomeinen is er voor groep 3 binnen de categorie Optellen en Aftrekken ook een analyse mogelijk tussen kale (bijvoorbeeld ‘6+3=?’) en contextopgaven (bijvoorbeeld ‘5 knuffels liggen op tafel en 2 knuffels liggen op de grond. Hoeveel knuffels zijn dat samen?’). Op basis van de vaardigheidsscore van de leerling berekent het Computerprogramma LOVS een verwachting van het aantal goed beantwoorde kale opgaven rondom optellen en aftrekken en het aantal goed beantwoorde contextopgaven uit dit domein. 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 bij kale of contextopgaven beter of zwakker scoort dan verwacht en of het verschil als niet opvallend, opvallend of zeer opvallend moet worden geïnterpreteerd.

Naast een ‘categorieënanalyse leerling is er ook een zogenoemde ‘categorieënanalyse groep. In deze laatste rapportage staan eerst per leerling alle resultaten uit de ‘categorieënanalyse leerling’ overzichtelijk onder elkaar voor de hele groep. Vervolgens staat in een tabel aangegeven hoeveel leerlingen uit die groep per categorie beneden en hoeveel leerlingen boven verwachting scoren, inclusief de gemiddelde afwijking naar beneden/boven.Op basis van de gemiddelde verschillen wordt met een t-toets nagegaan of er sprake is van significantie (tweezijdig). Significantie op het 10% niveau maar niet op het 5% niveau levert de kwalificatie opvallend op. Significantie op het 5% niveau levert de kwalificatie zeer opvallend op. Wanneer een groot deel van de leerlingen uit de groep beneden verwachting scoort en/of wanneer de gemiddelde afwijking naar beneden groot is, wordt een signaal (zeer) opvallend gegeven. Dit geeft aan dat deze categorie op groepsniveau om extra aandacht vraagt. Het signaal (zeer) opvallend kan echter ook betekenen dat in de groep juist opvallend veel leerlingen zijn die boven verwachting scoren en dat het dus een categorie betreft die de groep makkelijk af gaat.
In de handleiding bij het Computerprogramma LOVS is voor de leerkrachten een uitvoerige beschrijving opgenomen van de categorieënanalyse en de interpretatie van de uitkomsten. Ook in hoofdstuk 4 van de

handleiding in de toetsmap staan aanwijzingen over de interpretatie en het gebruik van de ‘categorieën-analyse leerling’ en de ‘categorieënanalyse groep’.
Naar de categorieënanalyse is geen empirisch onderzoek verricht en deze moet dan ook puur gezien worden als een handreiking naar de leerkracht. De statistiek geeft aan hoe groot de verschillen zijn tussen verwacht en geobserveerd en of op basis van kansrekening aan de verschillen belang kan worden gehecht. Of er daadwerkelijk conclusies voor het onderwijs uit afgeleid kunnen worden hangt af van nadere analyse en interpretatie van de antwoorden van de leerling.
De toetsen en rapportagemogelijkheden maken deel uit van een systeem van leerlingenzorg waarbij een school werkt volgens de cyclus signaleren, analyseren, plannen en handelen. De rapportages richten zich hierbij op de eerste twee fases van deze cyclus.

3.2 Inhoudsverantwoording

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

Deze fases worden hieronder nader toegelicht.
Deze informatie vormt een aanvulling op de inhoudsverantwoording die opgenomen is in de handleiding van het toetspakket Rekenen-Wiskunde 3.0 voor groep 3. 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 onderscheiden onderwerpen. Met behulp van die overzichten kunnen de leerkrachten de scores van leerlingen inhoudelijk interpreteren. De paragrafen bestaan uit grafieken waarop de p50- en p80-kanspunten van de items in de toetsen, geordend op basis van p-waarde, zijn afgebeeld, alsmede de vaardigheidsverdelingen op een aantal afnamemomenten. 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 3. Deze onderwerpen komen aan bod in de meest gebruikte methodes voor rekenen-wiskunde groep 3. Bij het construeren van de opgaven en het samenstellen van de toets is gekeken naar de wijze waarop en de mate waarin deze onderwerpen in de methodes naar voren komen. Doordat de leerlijnen van de methodes op hoofdlijnen aan elkaar gelijk zijn, en bij het construeren van de opgaven is nagegaan of groepen leerlingen die met een andere reken-wiskundemethode werken de betreffende stof aangeboden hebben gekregen, kunnen de toetsen bij elke rekenen-wiskunde methode van groep 3 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 3 bij die onderwerpen globaal aan bod komt.
De toetsen voor groep 3 bevatten opgaven uit twee domeinen: 1. Getallen en 2. Meten en meetkunde.
Bi de verschillende onderwerpen binnen die domeinen zijn vervolgens doelen opgesteld die aansluiten bij de leerlijnen en leerstof van leerlingen in groep 3.
In onderstaand overzicht staat per onderwerp aangegeven welke leerstof in de toetsen voor groep 3 aan bod komt.

Uitwerking doelen voor toetsen groep 3

Getallen

1. Getalbegrip

Positiewaarde en positioneren

- Bepalen van de waarde van cijfers in getallen, bijvoorbeeld weten en begrijpen dat in het getal 15 de 1 niet 1 maar 10 voorstelt.
- Inzicht in de plaats van getallen in de telrij onder andere door het plaatsen van getallen op de getallenlijn.
- De plaats van getallen op de getallenlijn herkennen.
- Getallen plaatsen tussen andere getallen in de telrij.

Tellen en samenstellen

- Resultatief tellen van zowel geordende als ongeordende hoeveelheden.
- Structurerend tellen en samenstellen, gebruikmakend van verschillende groepen (onder andere van 2, 3, 4, 5 en 10).
- Verder- en terugtellen met sprongen van 1, 2 (vanaf een willekeurig getal tot 20) en 5 (vanaf een veelvoud van 5) en 10

Structureren in parten

- Hoeveelheden splitsen in twee of meer groepen die al of niet gelijk zijn.
- Splitsen op basis van de positiewaarde: 14 = 4 + __.

Vergelijken

- Vergelijken en ordenen van getallen en hoeveelheden met behulp van de begrippen meer, minder, evenveel, groter en kleiner, het dichtste bij.

2. Bewerkingen

Optellen in contextopgaven en kale opgaven

- Toepassen van het optellen waarin leerlingen hoeveelheden aan de hand van verschillende contexten moeten samennemen, toevoegen en vergelijken.
- Optellingen met getallen zonder context, waarbij werkwijzen gebruikt kunnen worden als hergroeperen, splitsen en doortellen met sprongen.
  Voor M3 in het getallengebied tot en met 10
  Voor E3 in het getallengebied tot en met 20

Aftrekken in contextopgaven en kale opgaven

- Toepassen van het aftrekken waarbij leerlingen met hoeveelheden bewerkingen uitvoeren in verschillende contexten zoals eraf halen, aanvullen en verschil bepalen .
- Aftrekkingen met getallen zonder context, waarbij werkwijzen gebruikt kunnen worden als aanvullen, terugtellen met sprongen, hergroeperen en splitsen.
  Voor M3 in het getallengebied tot en met 10
  Voor E3 in het getallengebied tot en met 20

Vermenigvuldigen in contextopgaven

- Informeel vermenigvuldigen in eenvoudige contexten waarbij werkwijzen gebruikt kunnen worden als verdubbelen en tellen met sprongen.

Delen in contextopgaven

- Informeel delen in eenvoudige contexten waarbij werkwijzen gebruikt kunnen worden als halveren en splitsen van een hoeveelheid in groepen.

Meten en meetkunde

4.1. Meten: lengte

In groep 3 heeft dit onderdeel met name betrekking op het meten met natuurlijke maten, maar in E3 komt ook het meten van lengte met een liniaal aan bod.

4.2. Meten: oppervlakte

In groep 3 bevat de toets voor deze categorie uitsluitend opgaven waarbij de leerling de oppervlakte moet bepalen aan de hand van natuurlijke maten, bijvoorbeeld door aan te geven hoeveel tegels op een afgebeelde oppervlakte passen.

4.3*. Meten: inhoud*

Voor groep 3 gaat het hier net als bij de eerder genoemde meetonderdelen om het werken met natuurlijke maten. Denk aan het bepalen hoeveel pakjes in een doos passen, het bepalen van het aantal blokken in een bouwwerk en het vergelijken van verschillende bouwwerken. .

4.4. Meten: gewicht

De opgaven rondom meten gaan in groep 3 over het aflezen van de weegschaal en om het hanteren van de begrippen zwaar(der/st) en licht(er/st)

4.5. Tijd

In groep 3 lezen de leerling in de toets met name tijdstippen op de klok af zoals hele en halve uren.

4.6. Geld

In groep 3 bevatten de opgaven munten van 1 en 2 euro. Ook het biljet van 5 euro komt een enkele keer aan bod.

Bij de verschillende doelen voor groep 3 op een afnamemoment zijn opgaven geconstrueerd die een operationalisering vormen van die doelen.
Voor een uitvoerige beschrijving van de inhoud van de toetsen M3, M3E3 en E3 verwijzen we naar de Inhoudsverantwoording in het toetspakket (Cito, 2013). Daar is per toets een uitgebreide inhoudsbeschrijving opgenomen die geïllustreerd wordt met voorbeeldopgaven uit de toetsen, alsmede een aanduiding van de moeilijkheidsgraad van die opgaven.

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

Tabel 3.1 Verdeling opgaven over categorieën in de toetsen van de uitgave Rekenen-Wiskunde groep 3

		M3		M3E3		E3	
Getallen en getalrelaties		16		16		16	
Optellen en aftrekken	**kaal**	20	10	20	10	20	10
	context		10		10		10
Vermenigvuldigen en delen		8		8		8	
Meten, tijd en geld		8		8		8	
Totaal		52		52		52	

Itemconstructie
De toetsen bestaan voornamelijk uit open opgaven waarbij de leerling een kort antwoord in de vorm van een getal geeft. Meerkeuzeopgaven komen beperkt voor. De meerkeuzevorm is in groep 3 alleen gebruikt bij opgaven rondom meetkunde. Bij deze opgaven geeft de leerling zijn antwoord door een kruisje bij één van de alternatieven te zetten (papier) of door op het betreffende alternatief te klikken (digitaal).
De opgaven, aansluitend bij de domeinbeschrijving, zijn geconstrueerd door itemschrijfcommissies die bestonden uit leerkrachten basisonderwijs, schoolbegeleiders en pabodocenten. De constructeurs kregen een opdracht, opgesteld door toetsdeskundigen van Cito. In deze opdracht stond omschreven voor welke categorieën opgaven geconstrueerd 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 2010 en 2011 halverwege en aan het einde van leerjaar 3 nieuw ontwikkelde opgaven voorgelegd aan leerlingen. Daarbij zijn per afnamemoment ongeveer 150 nieuwe opgaven geproeftoetst. Elke deelnemende school maakte één taak met 25 opgaven. Voor elke opgave zijn ongeveer 125 à 150 responsen verzameld. Na deze proeftoetsing 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 2012. Het afnamedesign werd zo ingericht dat a) de nieuw geconstrueerde opgaven bij de kalibratie konden worden gekoppeld aan de opgaven van de bestaande toetsen van het leerlingvolgsysteem (LVS-II) en b) de opgaven van het betreffende afnamemoment (per categorie) konden worden gekoppeld aan de opgaven van zowel een eerder als een later afnamemoment en c) alle nieuwe opgaven onderling konden worden gekoppeld.

Het afnamedesign voor het normeringsonderzoek voor afnamemoment medio groep 3 (M3) bestond uit in totaal negen opgavenboekjes. Alle leerlingen maakten de bestaande M3-toets van LVS-II (twee taken) en één taak met nieuw geconstrueerde opgaven. De nieuw geconstrueerde opgaven hadden deels betrekking op het betreffende afnamemoment en deels op het latere afnamemoment E3. Daarnaast werden ook opgaven meegenomen voor groep 2, met het oog op de voorgenomen constructie van een toets voor dat leerjaar. Deze opgaven zijn niet opgenomen in de toetsuitgave voor M3, maar zullen op een later tijdstip als ankeropgaven worden meegenomen in de normeringsonderzoeken voor de toetsen van groep 2.
Het normeringsonderzoek voor afnamemoment einde groep 3 (E3) kende een soortgelijk afnamedesign, eveneens bestaande uit negen opgavenboekjes, met dien verstande dat de toegevoegde ankeropgaven hier betrekking hadden op het eerdere afnamemoment M3 en het latere afnamemoment M4.
Voor beide afnamedesigns 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.

In het normeringsonderzoek M3 zijn op deze wijze in totaal 180 verschillende items voorgelegd aan 2270 leerlingen van groep 3 verdeeld over negen boekjes. Elk boekje bestond uit 73 of 74 opgaven verdeeld over drie taken. De 50 opgaven (verdeeld over twee taken) uit de bestaande LVS-II taken werden door alle 2270 leerlingen gemaakt. In de nieuwe taken kwamen de E2 anker opgaven en de E3 anker opgaven één keer voor. Elke nieuwe opgave voor medio groep 3 kwam in twee boekjes voor. De nieuwe opgaven werden gemiddeld door 411 leerlingen gemaakt.

Tabel 3.2 Design medio groep 3 met aantallen opgaven en aantallen leerlingen per boekje (B1-B9)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
E2 anker	2	2	2	2	2	2	2	2	2
M3 nieuwe opgaven	18	20	18	19	18	18	19	19	19
E3 anker	3	2	4	3	4	3	3	3	3
Nieuwe taak totaal	23	24	24	24	24	23	24	24	24
LVSM3 generatie II taak 1	25	25	25	25	25	25	25	25	25
LVSM3 generatie II taak 2	25	25	25	25	25	25	25	25	25
Totaal aantal opgaven	73	74	74	74	74	73	74	74	74
Aantal leerlingen	267	254	253	222	260	249	262	251	252
Aantal scholen	11	11	11	10	11	11	11	10	9

De taak met nieuw geconstrueerde opgaven voor M3 (aangeduid met ‘t1’) in elk van de negen boekjes bevatte opgaven uit alle vijf onderscheiden opgavencategorieën. Uit onderstaand overzicht is af te lezen hoe de verdeling van de opgaven over de categorieën is gerealiseerd.

Tabel 3.3 Aantallen opgaven per categorie in de nieuwe taken

	B1 t1	B2 t1	B3 t1	B4 t1	B5 t1	B6 t1	B7 t1	B8 t1	B9 t1
getallen	7	9	6	8	7	5	8	8	6
optellen / aftrekken	6	5	7	4	6	6	7	5	6
vermenigvuldigen / delen	3	3	2	6	7	5	2	3	3
meten, tijd en geld	3	5	5	4	2	3	5	4	5
kale opgaven	4	2	4	2	2	4	2	4	4
Totaal	23	24	24	24	24	23	24	24	24

In het normeringsonderzoek E3 zijn in totaal 190 verschillende items voorgelegd aan 2145 leerlingen van eind groep 3 verdeeld over negen boekjes. Elk boekje bestond uit 74, 75 of 76 opgaven verdeeld over drie taken. De 50 opgaven (verdeeld over twee taken) in de LVS-II taken werden door alle 2145 leerlingen gemaakt. In de nieuwe taken kwamen de M3-ankeropgaven en de M4-ankeropgaven één keer voor. De E3 nieuwe opgaven kwamen in twee boekjes voor en werden gemiddeld door 480 leerlingen gemaakt. De positie van de nieuwe taak in de boekjes varieerde.

Tabel 3.4 Design eind groep 3 met aantallen opgaven en aantallen leerlingen per boekje (B1-B9)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
M3 anker	3	3	3	3	2	3	5	3	3
E3 nieuwe opgaven	18	18	19	19	20	19	18	18	15
M4 anker	3	3	3	3	3	3	3	3	6
Nieuwe taak totaal	24	24	25	25	25	25	26	24	24
LVSE3 generatie II taak 1	25	25	25	25	25	25	25	25	25
LVSE3 generatie II taak 2	25	25	25	25	25	25	25	25	25
Totaal aantal opgaven	74	74	75	75	75	75	76	74	74
Aantal leerlingen	252	230	188	241	255	243	248	241	247
Aantal scholen	10	10	9	12	11	11	10	9	9

In het volgende overzicht staan de aantallen opgaven die in de nieuwe taken voor eind groep 3 voor de verschillende categorieën zijn opgenomen.

Tabel 3.5 Aantallen opgaven per categorie in de nieuwe taken

	B1 t1	**B2 t1**	**B3 t1**	**B4 t1**	**B5 t1**	**B6 t1**	**B7 t1**	**B8 t1**	**B9 t1**
getallen	5	8	7	8	7	7	10	8	4
optellen / aftrekken	7	4	4	7	4	7	4	7	8
vermenigvuldigen / delen	4	4	6	3	4	3	6	4	6
meten, tijd en geld	4	6	4	4	4	4	4	3	3
kale opgaven	4	2	4	3	6	4	2	2	3
Totaal	24	24	25	25	25	25	26	24	24

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. Itemparameters 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 groep 3 tot en met medio groep 4 een schaal geconstrueerd en is een selectie van opgaven gemaakt voor de papieren uitgave.

Het kalibratie-onderzoek digitaal en papier-digitaal
De geselecteerde opgaven voor de papieren uitgave van medio groep 3 en eind groep 3 (van LVS generatie III) zijn, met voor elke categorie enkele extra opgaven, gedigitaliseerd. De digitale opgaven zijn in een digitaal onderzoek en een (vergelijkend) papier-digitaal onderzoek, afgenomen in januari 2013 bij leerlingen van medio groep 3 en in juni 2013 bij leerlingen van eind groep 3. Het doel van deze afnames was om gegevens te verzamelen voor het schatten van itemparameters van de digitale itemvarianten en om de digitale items op dezelfde schaal onder te brengen als de papieren items.
Bij zowel leerlingen van medio groep 3 als eind groep 3 zijn digitale taken met nieuwe opgaven afgenomen in combinatie met de twee papieren taken van LVS generatie II en in combinatie met de twee digitale taken van LVS generatie II. Bij de afnames waren vier groepen leerlingen betrokken. In de tabellen hieronder is te zien welke taken die groepen leerlingen gemaakt hebben.

Tabel 3.6 Design medio groep 3 digitaal onderzoek (B2 en B4) en papier-digitaal onderzoek (B1 en B3)

	B1	B2	B3	B4
LVS M3 generatie III digitaal taak 1 (nieuw)				
LVS M3 generatie III digitaal taak 2 (nieuw)				
LVS M3 generatie II papier taak 1				
LVS M3 generatie II papier taak 2				
LVS M3 generatie II digitaal taak 1				
LVS M3 generatie II digitaal taak 2				
Aantal leerlingen	98	178	118	177
Aantal scholen	6	6	6	10

Tabel 3.7 Design eind groep 3 digitaal onderzoek (B2 en B4) en papier-digitaal onderzoek (B1 en B3)

	B1	B2	B3	B4
LVS E3 generatie III digitaal taak 1 (nieuw)				
LVS E3 generatie III digitaal taak 2 (nieuw)				
LVS E3 generatie II papier taak 1				
LVS E3 generatie II papier taak 2				
LVS E3 generatie II digitaal taak 1				
LVS E3 generatie II digitaal taak 2				
Aantal leerlingen	171	198	246	317
Aantal scholen	10	9	11	16

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

Samenstelling toetsen

De opgaven voor de toetsen zijn geselecteerd uit de verzameling opgaven van de gekalibreerde itembank. Die bank bevat opgaven van LVS generatie II en LVS generatie III voor papieren en digitale toetsen.
In totaal zijn zes toetsen samengesteld. Voor de papieren variant drie toetsen: M3, M3E3 en E3 en voor de digitale variant drie toetsen: M3, M3E3 en E3.
De toetsen voor M3 en E3 van de papieren variant zijn samengesteld uit opgaven die in de nieuwe taken bij het normeringsonderzoek zaten. De toetsen voor M3 en E3 van de digitale variant zijn samengesteld uit opgaven die in de nieuwe taken zaten van het digitale en papier-digitale onderzoek. Daarnaast is er voor groep 3 (zowel voor de papieren variant als de digitale variant) een tussentoets M3E3 samengesteld. De tussentoets M3E3 bevat niet alleen nieuwe opgaven, maar ook opgaven uit LVS-II en enkele opgaven uit de M3 en E3 toetsen van LVS-III (met name kale opgaven).

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. In de matrijs is te zien dat we gestreefd hebben naar een goede balans tussen kale opgaven en contextopgaven. Daarnaast is bij het selecteren van opgaven ook gekeken 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 M3, M3E3 en E3 in relatie tot de vaardigheidsverdelingen van M3, E3 en M4. 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 M3, M3E3 en E3 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 M3 ligt het merendeel van de balkjes op en rond de gemiddelde vaardigheidsscore behorende bij medio groep 3. Bij de toets E3 liggen de opgaven veelal hoger op de vaardigheidsschaal, rond de gemiddelde vaardigheidsscore eind groep 3. Bij de tussentoets M3E3 is gezorgd voor een variatie aan niveaus. Deze toets bevat zowel opgaven rond de gemiddelde vaardigheidsscore behorende bij het gemiddelde medio groep 3 als bij het gemiddelde eind groep 3. Zie ook de alinea in het begin van dit hoofdstuk ‘Keuze van een passende toets: Toetsen op maat’. In de figuren is zichtbaar dat bij alle toetsen geldt dat zij relatief veel ‘gemakkelijke’ opgaven bevatten, oftewel opgaven behorend bij lagere vaardigheidsscores dan de gemiddelde vaardigheidsscore op het betreffende toetsmoment. Deze keuze is gemaakt om te zorgen dat het merendeel van de leerlingen een succeservaring heeft bij het maken van de toets. Er is gezocht naar een optimale balans tussen nauwkeurig uiteenlopende vaardigheidsniveaus in beeld brengen en zorgen voor een prettige toetservaring voor leerling en leerkrachten.

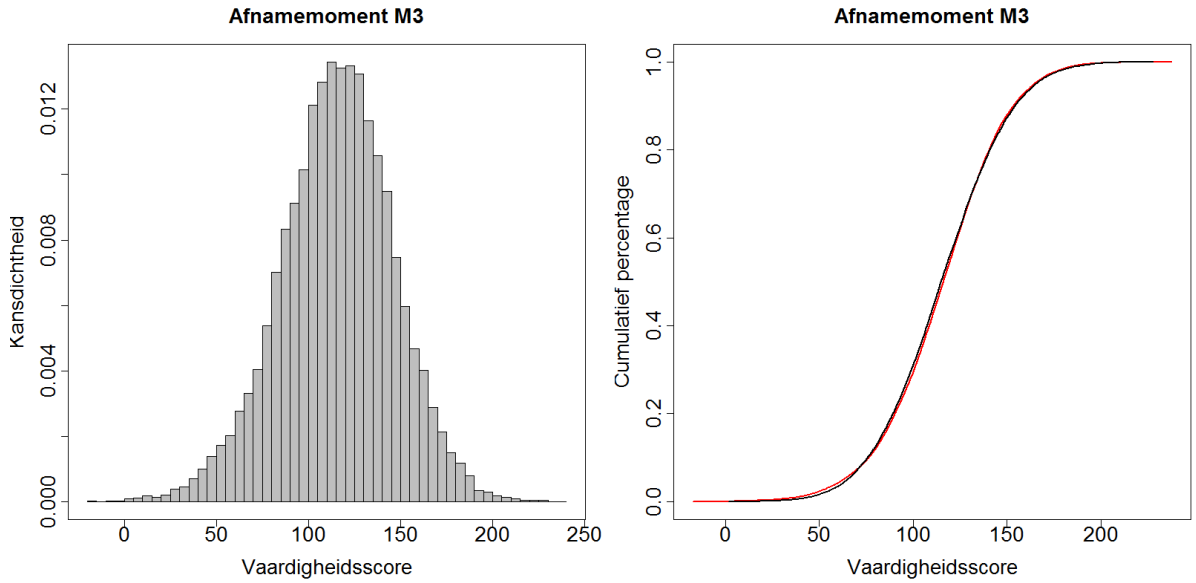
3.3 Statistische beschrijving

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 M3, M3E3 en E3 toetsen, zowel op de ruwe scoreschaal als op de vaardigheidsschaal.
De gegevens zijn gebaseerd op de prestaties van 4512 leerlingen voor M3 en op de prestaties van 4319 leerlingen voor E3. Deze aantallen leerlingen zijn groter dan de aantallen leerlingen die in de normeringssteekproef zaten. Naast de data van leerlingen uit de normeringssteekproef zijn namelijk ook data gebruikt van leerlingen uit dataretour. Cito dataretour is een exporttool die basisscholen in staat stelt om jaarlijks hun LVS-resultaten naar Cito te sturen voor (interne) onderzoeksdoeleinden. Het opsturen van resultaten vindt geautomatiseerd plaats via het computerprogramma LOVS (Keuning, 2014). Daarover rapporteren we in hoofdstuk 4.
De waarden in tabel 3.8 en de figuren 3.1 en 3.2 laten zien dat de vaardigheidsverdeling bij benadering normaal is.

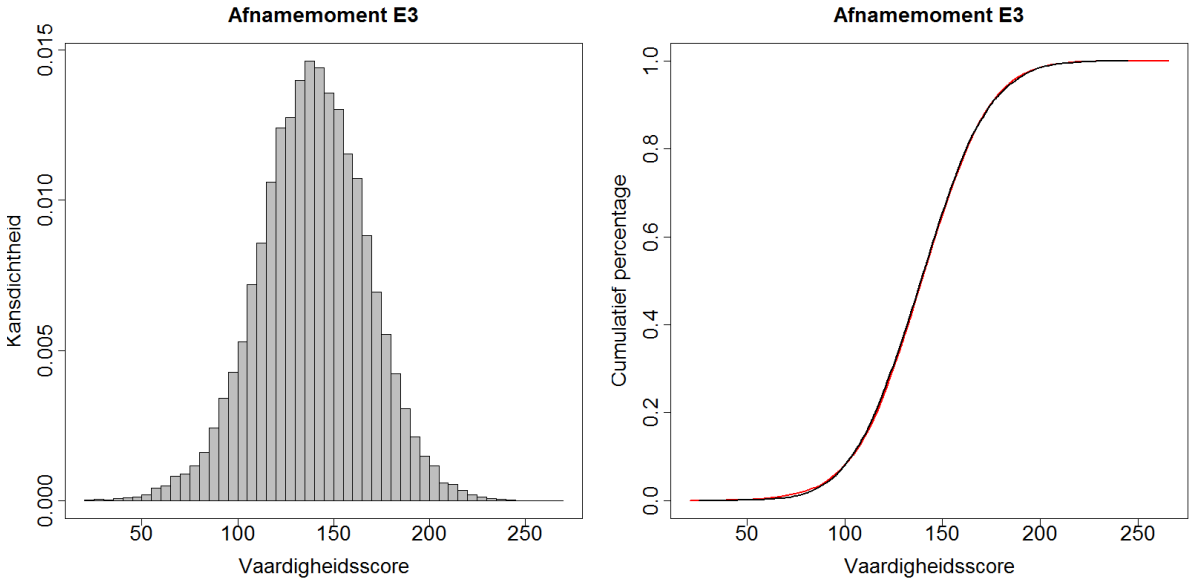
Tabel 3.8 Beschrijvende gegevens toetsen M3, M3E3 en E3 op ruwe scoreschaal en vaardigheidsschaal

	Gemiddelde	Standaarddeviatie	Kurtosis	Scheefheid
M3 Ruwe score	36,95	10,36	0,27	-0,91
M3 Vaardigheid	115,31	30,72	0,32	-0,18
M3E3 Ruwe score	38,76	9,74	0,31	-0,93
M3E3 Vaardigheid	139,26	28,17	0,27	-0,10
E3 Ruwe score	37,97	10,11	0,36	-0,92
E3 Vaardigheid	139,26	28,17	0,27	-0,10
M3 digitaal Ruwe score	31,80	11,74	-0,79	-0,37
M3 digitaal Vaardigheid	115,31	30,72	0,32	-0,18
M3E3 digitaal Ruwe score	38,70	9,81	0,26	-0,92
M3E3 digitaal Vaardigheid	139,26	28,17	0,27	-0,10
E3 digitaal Ruwe score	33,20	11,13	-0,60	-0,48
E3 digitaal Vaardigheid	139,26	28,17	0,27	-0,10

Figuur 3.1 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling M3



Figuur 3.2 Weergave van behaalde vaardigheidsscores en cumulatieve verdeling E3



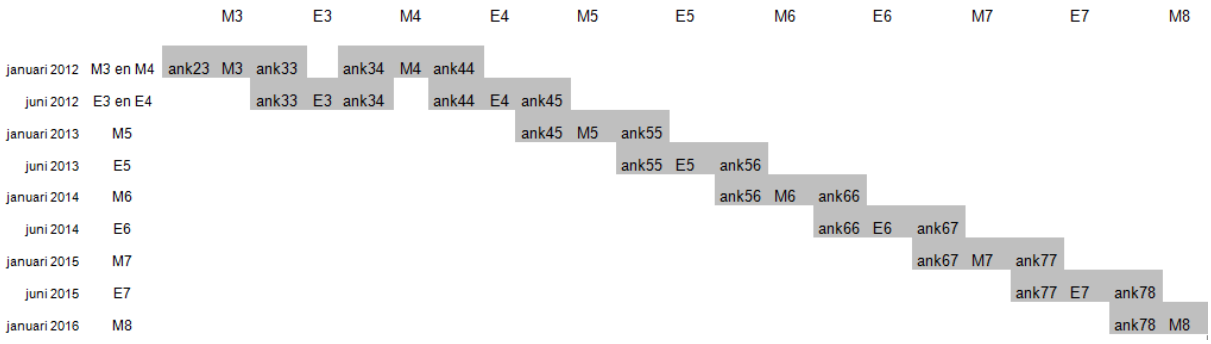
4 Kalibratie en normering

4.1 Opzet normeringsonderzoeken LVS: het macrodesign

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 M5 moeten vergelijkbaar zijn met die van andere populaties, bijvoorbeeld E4 en E5, oftewel het dataverzamelingsdesign, dient verbonden te zijn. Hiertoe dient een longitudinale opzet gebruikt te worden.

De verbondenheid van het design
Het idee van een gemeenschappelijke schaal impliceert strikt genomen dat men iemands vaardigheid zou kunnen schatten aan de hand van een willekeurig samengestelde toets. Het spreekt echter vanzelf dat het een zinloze onderneming is een toets die geconstrueerd is voor groep 7 voor te leggen aan leerlingen van groep 3, omdat zo'n toets ongetwijfeld opgaven zal bevatten die een beroep doen op kennis van leerstof die in groep 3 niet is onderwezen. Dit betekent dat we door de algemene kenmerken van het curriculum in het rekenonderwijs tamelijk beperkt zijn in het voorleggen van itemmateriaal aan leerlingen voor wie het niet specifiek is geconstrueerd. Daarom is er besloten dat het overlapmateriaal dat aan een bepaalde (half-) jaargroep kan worden voorgelegd alleen itemmateriaal mag bevatten dat specifiek voor die halfjaargroep is geconstrueerd en voor de twee belendende halfjaargroepen. Voor M5 betekent dit dat de leerlingen in het normeringsonderzoek items voorgelegd krijgen die specifiek voor M5 zijn geconstrueerd, en (een minderheid aan) items die geconstrueerd zijn voor E4 en E5.
Het macro-design is weergegeven in figuur 4.1.

Figuur 4.1 Het macrodesign voor de normeringsafnames



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

Longitudinale opzet
Een volledig longitudinaal design impliceert dat een cohort leerlingen gevolgd wordt van M3 tot en met M8. Een dergelijk design heeft een aantal zwaarwegende nadelen. Het is onvermijdelijk dat er uitval plaats zal vinden. Bij een hoog percentage uitval wordt het steeds ingewikkelder, zo niet onmogelijk, om betrouwbare normen op te stellen. Bovendien is een longitudinale studie belastend voor de deelnemende scholen en leerlingen. Dit brengt het risico mee van ongewenste en moeilijk controleerbare neveneffecten.
Daarom is ervoor gekozen het longitudinale karakter van het onderzoek in te perken, en aan de deelnemende scholen te vragen deel te nemen op maximaal drie opeenvolgende meetmomenten, waarbij het startmoment verspreid is voor verschillende scholen. Bijvoorbeeld: school A start met groep 3 op het mediomoment van schooljaar x en zal eveneens deelnemen aan de opvolgende momenten E3 (schooljaar x) en M4 (schooljaar x+1). School B zal starten op moment E3 (schooljaar x) en zal eveneens deelnemen aan de opvolgende momenten M4 (schooljaar x+1) en E4 (schooljaar x+1). Op deze manier wordt rekening gehouden met de belasting voor scholen en worden toch de benodigde longitudinale data verkregen.
Aansluitend bij de verbondenheid van het design via opeenvolgende toetsmomenten en de longitudinale opzet zal de kalibratie per leerjaar worden uitgevoerd op een beperkt deel van de gemeenschappelijke schaal. De kalibratie zal plaatsvinden op basis van de verzamelde data voor dat leerjaar op de afnamemomenten, aangevuld met de gegevens van het voorgaande en het opvolgende afnamemoment. Voor groep 3 vindt de kalibratie plaats op basis van de gegevens die op de afnamemomenten M3, E3 en M4 verzameld zijn. In het geval van leerjaar 4 vindt de kalibratie plaats op basis van de verzamelde gegevens op de afnamemomenten E3, M4, E4 en M5. Dit sluit aan bij de inhoudelijke kenmerken van de aangeboden opgaven, een sterke leerling in groep 4 zal wel opgaven uit groep 5 kunnen maken, maar geen opgaven uit groep 8 omdat deze qua inhoud nog niet allemaal zijn behandeld. 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 3 betreft het dus het gedeelte uit figuur 4.1, dat in figuur 4.2 is weergegeven.

Figuur 4.2 Design groep 3

Januari 2012	M3	ank23	M3	ank33		
Juni 2012	E3			ank33	E3	ank34

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.

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 toets LVS-II toets en een derde 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 groep 3 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 3 en eind groep 3 (tabel 3.2 en tabel 3.4). In het *embedded field* onderzoek bij de papieren afnames zijn in totaal negen boekjes voor medio groep 3 en negen boekjes voor eind groep 3 afgenomen. Naast de twee taken van de reguliere uitgave van LVS-II maakte elke leerling één taak met nieuwe opgaven. Elke nieuwe taak bevatte naast nieuwe opgaven ook ankeropgaven uit het voorafgaande en het volgende afnamemoment. In bijlage 2: Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M3 en bijlage 3: Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek E3 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 daar het OPLM behoort tot de exponentiële familie, met de gewogen somscore:

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

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

voor de diverse statistische toetsen in OPLM. De toetsingsgrootheid voor de veronderstelde discriminatie-indices is gegeven door

$$M \quad f_{s\,H}(p(\ \ |s) \quad prop(\ \ |s)) \quad f_{s\,L}(prop(\ \ |s) \quad p(\ \ |s)). \tag{4.2}$$

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

$$S \quad f(p(\ \ |s) \quad prop(\ \ |s)).$$

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:

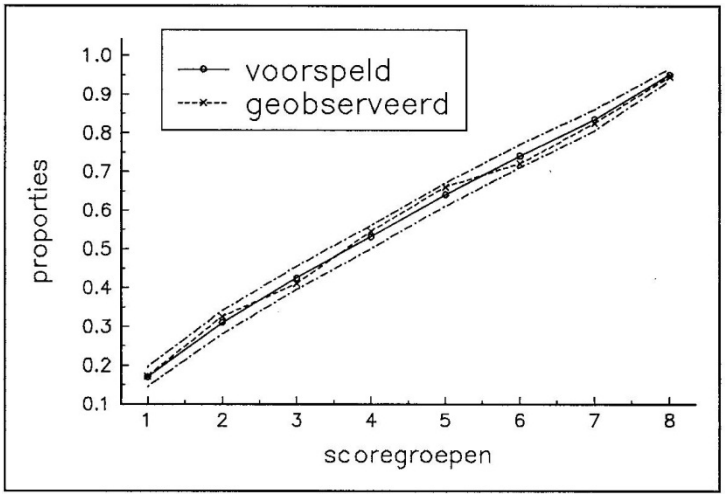
1. Met behulp van het programma OPCAT stellen we de discriminatie-indices in OPLM in en hercoderen we indien noodzakelijk de antwoordcategorieën in de data.
2. Vervolgens schatten we de itemparameters met behulp van de CML-methode.
3. Met behulp van de M-toetsen controleren we of de discriminatie-indices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

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

4.2.3 Toetsing van het IRT-model

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

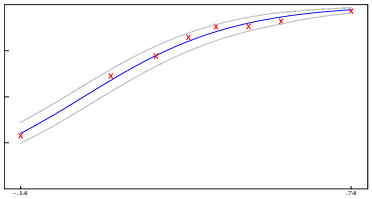
Figuur 4.3 Grafische voorstelling van een S_i–toets



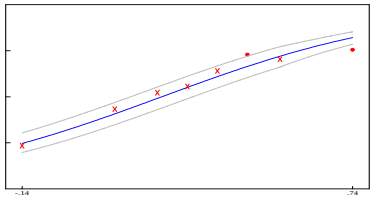
Het is ondoenlijk om voor alle opgaven dergelijke grafische voorstellingen in deze verantwoording op te nemen. Daarom beperken we ons steeds per toetsversie tot het item met de slechtste en de beste S-passing, aangevuld met een qua S-toetsingsresultaat gemiddelde (dat wil zeggen, meest representatieve) passing. De voorbeelden in figuur 4.4 illustreren dat voor de toetsen M3 en E3 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 Rekenen-Wiskunde-toetsen 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 toeten Rekenen-Wiskunde M3, M3E3 en E3 met de best passende, de slechtst passende en een qua passing representatieve opgave

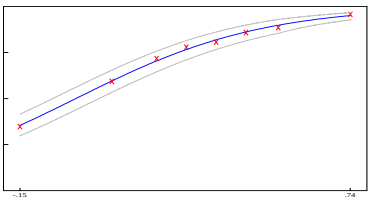
M3



Best passend

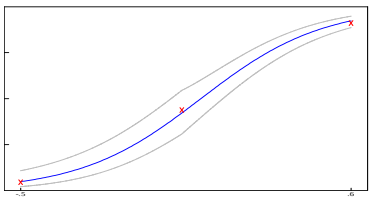


Slechtst passend

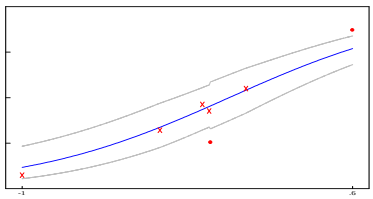


Representatieve passing

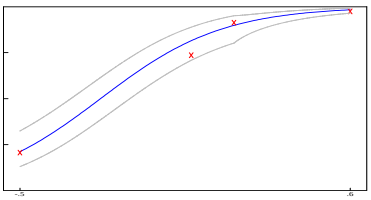
M3E3



Best passend

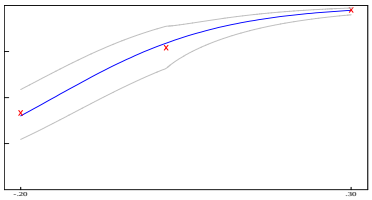


Slechtst passend

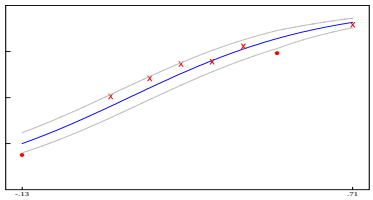


Representatieve passing

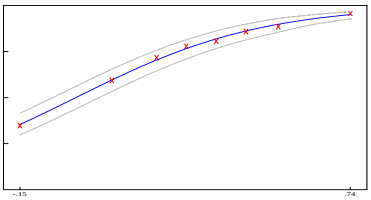
E3



Best passend

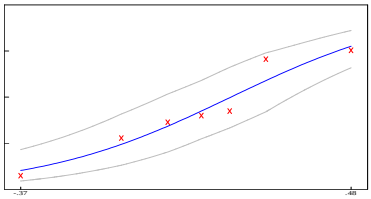


Slechtst passend

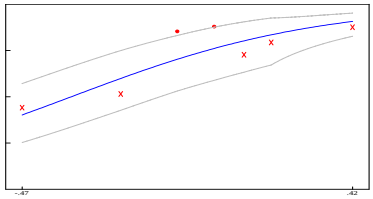


Representatieve passing

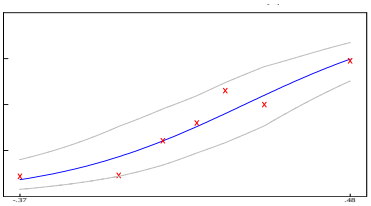
M3 digitaal



Best passend

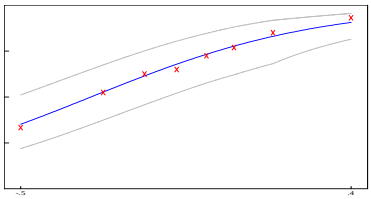


Slechtst passend

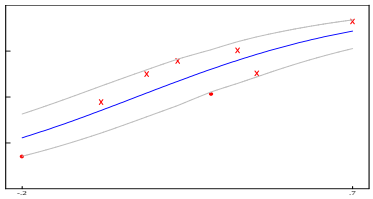


Representatieve passing

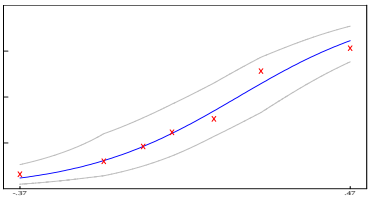
M3E3 digitaal



Best passend

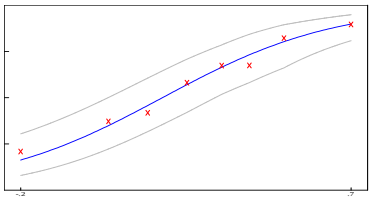


Slechtst passend

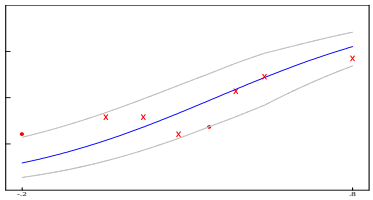


Representatieve passing

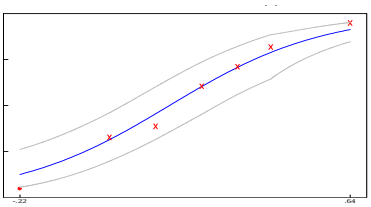
E3 digitaal



Best passend



Slechtst passend



Representatieve passing

In feite kan men bij de kalibratie beter varen op deze grafische weergaven dan op toetsingsresultaten in termen van exacte getallen en de significantie daarvan. Niettemin zijn er bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij zijn we vooral geïnteresseerd in de distributie van de overschrijdingskansen van deze verzameling toetsingsresultaten. Als we de S-toetsen opvatten als onafhankelijk, dan zouden de overschrijdingskansen uniform verdeeld moeten zijn binnen het (0,1) interval, uiteraard met zo weinig mogelijk significante resultaten. Tabel 4.1 waarin het (0,1) interval is opgedeeld in tien gelijke stukken, geeft een beeld van de uitkomsten bij een kalibratie van alle opgaven van de toets LVS-III Rekenen-Wiskunde M3-E3. Daarnaast is aangegeven in hoeveel gevallen de overschrijdingskans kleiner was dan 0,01, respectievelijk 0,05. Het is duidelijk dat voor de toets de verdeling redelijk gelijkmatig is over het gehele interval van overschrijdingskansen. Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat met uitzondering van enkele opgaven, sprake is van niet-significante S-toetsen. Hierbij valt op te merken dat het aantal opgaven waarbij de S-toets significant was op het niveau lag dat te verwachten valt onder een nul-model, zoals bij een significantie niveau van 5% te verwachten valt dat 5% van de resultaten significant is, zonder dat dit betekenis heeft. De resultaten zoals die hier gevonden worden passen in dat beeld. Al met al vormen zij een kwantitatieve ondersteuning van de conclusie dat de opgaven een unidimensionaal construct representeren.

Tabel 4.1 Verdeling van overschrijdingskansen bij S-toetsen voor toetsen M3, M3E3 en E3

	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3	3	1	1	6	2	9	5	4	2	7	5	7
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3E3	1	5	3	3	5	5	4	5	6	6	7	2
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
E3	1	4	7	4	0	4	4	8	7	5	5	3
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3 digitaal	0	2	1	6	3	3	3	6	4	3	12	9
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
M3E3 digitaal	1	0	4	2	3	5	5	1	3	10	6	12
	0.-/---/---------.1----------.2----------.3----------.4----------.5----------.6----------.7----------.8----------.9---------1.											
E3 digitaal	2	1	1	3	5	4	7	4	4	7	9	5

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 M3, M3E3 en E3 geldt zowel voor de papieren versie als de digitale versie dat de R1c minder dan anderhalf maal het aantal vrijheidsgraden bedraagt. De toetsingsgrootheid is significant bij de papieren toetsen. Aan dit laatste moet bij steekproeven met een dergelijke omvang niet te veel waarde worden gehecht.

Tabel 4.2 R1c-waarden voor M3-E3

Toetsversie	**R1c**	**df**	**p**
M3	393,6	324	<0,005
M3E3	624,3	509	<0,005
E3	401,6	321	<0,005
M3 digitaal	330,1	353	0,804
M3E3 digitaal	371,6	354	0,251
E3 digitaal	397,4	400	0,260

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 ruim onder de 0,20 wat een oordeel goed inhoudt. Op één opgave na bij M3 digitaal (die een waarde van 0,21 had) hadden alle opgaven bij alle toetsen een waarde onder de 0,20. De nauwkeurigheid van de schatting kan daarmee als (zeer) goed beoordeeld worden.

Tabel 4.3 Nauwkeurigheid van de itemparameterschattingen (constante ‘c’)

	Constante ‘c’	
Toetsmoment	**Range**	**Gemiddelde**
M3	0,057 – 0,178	0,098
E3	0,059 – 0,161	0,092
M3 digitaal	0,077 – 0,210	0,118
E3 digitaal	0,062 – 0,183	0,102

Op basis van de hierboven beschreven resultaten kan de conclusie luiden dat voor de toetsen LVS-III Rekenen-Wiskunde groep 3 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 validiteits-vraagstuk 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 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 leerling-volgsysteemtoetsen 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 dekkingsgraad van de LVS toetsen Rekenen-Wiskunde is bijzonder hoog: de toetsen worden door 90% tot 95% van de scholen toegepast. Voor de deelnemende scholen aan het normeringsonderzoek is nagegaan of zij als groep afweken van wat men voor de totale populatie van scholen zou mogen verwachten. De gemiddelde score op de Cito Eindtoets Basisonderwijs bleek voor de scholen in het normeringsonderzoek niet af te wijken van het populatiegemiddelde voor deze toets. Voor het bepalen van de normering worden deze gegevens aangevuld met gegevens uit Cito dataretour.

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 die volledig in de feitelijke toetssituatie verzameld zijn af. Dit is een gemiste kans. Juist het combineren van het *embedded field* normeringsonderzoek met Cito dataretour biedt grote voordelen ten opzichte van alternatieve onderzoeksdesigns. Enerzijds wordt er op

deze manier voor gezorgd dat de toetsresultaten die gebruikt worden bij het bepalen van de normen zoveel mogelijk in de feitelijke toetssituatie verzameld zijn. Anderzijds is het mogelijk om via Cito dataretour de "kwaliteit" van het *embedded field* normeringsonderzoek te checken. Een belangrijke randvoorwaarde is wel dat de uiteindelijke normeringsteekproef representatief is voor de landelijke populatie van scholen en leerlingen. Representativiteit van de normeringssteekproef zoals die samengesteld wordt op basis 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 er voor gekozen om de indeling naar vijf niveaus die gebruikelijk is bij het CBS te reduceren tot een tweedeling in enerzijds niet tot matig verstedelijkt (platteland) en anderzijds sterk tot zeer sterk verstedelijkt (stad). Een dergelijke tweedeling blijkt in de praktijk goed te volstaan (cf. Van Boxtel & Hemker, 2009).

- ***Schooltype***. Bij de definitie van de variabele *schooltype* is gebruikgemaakt van de formatiegewichten van de leerlingen binnen een school volgens de meest recente regeling van OCW. Daarin worden drie niveaus onderscheiden die gebaseerd zijn op het opleidingsniveau van de ouders:
  0.0 één van de ouders of beide ouders heeft of hebben een opleiding gehad uit categorie 3
  0.3 beide ouders of de ouder die belast is met de dagelijkse verzorging heeft of hebben een opleiding uit categorie 2 gehad
  1.2 één van de ouders heeft een opleiding gehad uit categorie 1 en de ander een opleiding uit categorie 1 óf 2

  In deze indeling wordt verwezen naar de volgende categorieën in het opleidingsniveau van de ouders: 1 = maximaal basisonderwijs of (V)SO-ZMLK, 2 = maximaal LBO/VBO, praktijkonderwijs of VMBO basis- of kaderberoepsgerichte leerweg, en 3 = overig VO en hoger. Leerlingen met een formatiegewicht van 0.3 of 1.2 zijn te definiëren als achterstandsleerlingen. Scholen zijn ingedeeld naar het percentage 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 × 2 × 4 × 2 = 64 verschillende categorieën representeren. De variabelen *regio*, *urbanisatiegraad* en *schooltype* zijn op het niveau van de school gedefinieerd. De variabele *sekse* is op het niveau van de leerling gedefinieerd. Het is niet goed mogelijk om bij het selecteren van data tegelijkertijd rekening te houden met school- én leerlingvariabelen. Daarom vindt de dataselectie in twee stappen plaats. In de

eerste stap worden iteratief scholen uit Cito dataretour toegevoegd aan de dataset met normerings-gegevens. Niet elke school heeft daarbij evenveel kans om geselecteerd te worden. Bij de selectie wordt namelijk rekening gehouden met de regio en de urbanisatiegraad van de school en het aantal achterstandsleerlingen. De kans w_{ijk} dat een school met regio i, urbanisatiegraad j en schooltype k geselecteerd wordt, hangt af van het reeds geselecteerde aantal leerlingen N_{S}, het gewenste aantal leerlingen N_{T}, en het beschikbare aantal leerlingen in Cito dataretour N_{D}:

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

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

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

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

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

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.

Samenvattend gaat het algoritme voor het genereren van een representatieve normeringssteekproef op basis van een normeringsonderzoek (S) en Cito dataretour (D) dus als volgt te werk:

Voorbereiding data normeringsonderzoek

bereken w_{ijk} voor S
indien $w_{ijk} < 0$
  herhaal
    trek aselect een school y en verwijder deze uit S
    bereken w_{ijk}
  totdat $w_{ijk} \geq 0$
retourneer S

Toevoegen data uit Cito dataretour

bereken w_{ijk} voor S
herhaal
  trek een school y uit D gegeven w_{ijk} en voeg deze toe aan S
  bereken w_{ijk}
  indien $w_{ijk} < 0$
    verwijder school y uit S
    bereken w_{ijk}
totdat $N_{\mathrm{S}} \geq N_{\mathrm{T}}$
retourneer S

Check leerlingvariabele sekse

bereken w_q voor S
voor elke school y
  herhaal
    trek een leerling uit S_y gegeven $w_{y,q}$ en voeg deze toe aan $\tilde{\mathrm{S}}_y$
  totdat $N_{\tilde{\mathrm{S}}_y} = N_{\mathrm{S}_y}$
retourneer $\tilde{\mathrm{S}}$

Het algoritme is toegepast bij de ontwikkeling van LVS-III Rekenen-Wiskunde. Het uitgangspunt was om de data die tijdens het *embedded field* normeringsonderzoek verzameld zijn te verdubbelen met behulp van data uit Cito dataretour. Hieronder wordt weergegeven hoe het selectiealgoritme functioneerde voor de eerste afnamemomenten: medio groep 3 en eind groep 3.

Aan het normeringsonderzoek op afnamemoment medio groep 3 hebben 2253 leerlingen van 94 verschillende basisscholen deelgenomen. Het gewenste aantal leerlingen is dus ingesteld op $N_{\mathrm{T}} \quad 2 \quad 2253 \quad 4506$. Constante C is ingesteld op 1.05. Sommige categorieën bleken geheel niet vertegenwoordigd te zijn in het databestand met normeringsgegevens, terwijl andere categorieën in het licht van de gekozen $N_{\mathrm{T}}C$ oververtegenwoordigd waren. Scholen met meer dan 40 procent achterstandsleerlingen op het platteland waren bijvoorbeeld niet vertegenwoordigd in het databestand. In de verstedelijkte gebieden was er juist sprake van een oververtegenwoordiging van achterstandsscholen. Om deze reden zijn aselect 9 scholen uit het databestand verwijderd. Op deze scholen zaten 248 leerlingen. Na verwijdering van deze leerlingen bevatte het databestand met normeringsgegevens dus 85 scholen en 2005 leerlingen. Dit bestand is iteratief aangevuld met data uit Cito dataretour totdat $N_{\mathrm{S}} \geq N_{\mathrm{T}}$. In totaal zijn er door het selectiealgoritme 179 basisscholen geselecteerd. Van de geselecteerde scholen

bleken er 67 niet geschikt te zijn, omdat een bepaalde categorie dan oververtegenwoordigd raakte in het licht van de gekozen $N_{\mathrm{T}}C$. Dit betekent dat er uiteindelijk 112 scholen met in totaal 2501 leerlingen vanuit Cito dataretour toegevoegd zijn aan het databestand met normeringsgegevens. De uiteindelijke normeringssteekproef voor de LVS-toets Rekenen-Wiskunde op afnamemoment medio groep 3 bevatte dus 85 + 112 = 197 scholen (43 procent normeringsonderzoek en 57 procent Cito dataretour) en 2005 + 2501 = 4506 leerlingen (44 procent normeringsonderzoek en 56 procent Cito dataretour).

Het selectiealgoritme leverde bij afnamemoment eind groep 3 vergelijkbare resultaten op. Aan het normeringsonderzoek op afnamemoment eind groep 3 hebben 2157 leerlingen van 92 verschillende basisscholen deelgenomen. Het gewenste aantal leerlingen is dus ingesteld op $N_{\mathrm{T}} = 2 \times 2157 = 4314$. Voor constante C is dezelfde waarde gekozen als eerder bij afnamemoment medio groep 3. Wederom bleken bepaalde categorieën niet vertegenwoordigd te zijn in het databestand met normeringsgegevens. Tevens was er sprake van een oververtegenwoordiging van enkele categorieën. Daarom zijn voorafgaand aan de toevoeging van data uit Cito dataretour aselect 11 scholen verwijderd uit het databestand met normeringsgegevens. Op deze scholen zaten in totaal 300 leerlingen. Na verwijdering van deze leerlingen bevatte het databestand met normeringsgegevens voor afnamemoment eind groep 3 dus 81 scholen en 1857 leerlingen. Dit bestand is aangevuld met data uit Cito dataretour. In totaal zijn er door het selectiealgoritme 222 basisscholen geselecteerd. Van de geselecteerde scholen bleken er 121 niet geschikt te zijn, omdat een bepaalde categorie dan oververtegenwoordigd raakte in het licht van de gekozen $N_{\mathrm{T}}C$. Dit betekent dat er uiteindelijk 101 scholen met in totaal 2458 leerlingen vanuit Cito dataretour toegevoegd zijn aan het databestand met normeringsgegevens. De uiteindelijke normeringssteekproef voor de LVS-toets Rekenen-Wiskunde op afnamemoment eind groep 3 bevatte dus 81 + 101 = 182 scholen (44 procent normeringsonderzoek en 56 procent Cito dataretour) en 1857 + 2458 = 4315 leerlingen (43 procent normeringsonderzoek en 57 procent Cito dataretour).

Zowel voor afnamemoment medio groep 3 als eind groep 3 heeft het selectiealgoritme tot de gewenste oplossing geleid. Wel valt op dat het selectiealgoritme relatief veel scholen ongeschikt verklaart. Dat komt doordat een erg kleine waarde voor constante C is gekozen. Het gevolg is dat het selectiealgoritme weinig ruimte heeft gekregen om af te wijken van de gewenste aantallen in elke categorie. Vooral in de laatste iteraties waarin het geobserveerde leerlingaantal al dicht bij het gewenste leerlingaantal ligt, kan toevoeging van een school leiden tot een oververtegenwoordiging van bepaalde categorieën. In beginsel is het geen probleem dat de selectie van relatief veel scholen na de berekening van w_{ijk} ongedaan gemaakt wordt. Wel is het de vraag in hoeverre het zinvol is om te streven naar steekproeven die volledig representatief zijn voor de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse*. Ook in aselecte steekproeven kan de verdeling van leerlingen over de verschillende categorieën immers afwijken van de verdeling in de populatie. In een aselecte steekproef is deze afwijking per definitie het gevolg van toeval. Statistische weging is in een aselecte steekproef dan ook niet op zijn plaats. Door bij de normering van LVS-III de benodigde data representatief te trekken, zijn de afwijkingen die we vinden in relatie tot de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse* in zekere zin ook toe te schrijven aan toeval. Afwijkingen tussen de steekproef en de populatie kunnen in dat geval verdedigbaar zijn. Niettemin wordt in een vervolgstap de landelijke representativiteit van de normeringssteekproef ter controle onderzocht.

In tabel 4.4 wordt weergegeven welke aantallen van de steekproef en van dataretour uiteindelijk zijn meegenomen in de normering.

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

Afnamemoment	Aantal leerlingen			Aantal scholen
	Steekproef	Dataretour	Normering	Normering
M3	1947	2565	4512	190
E3	1864	2455	4319	184

Nagegaan is ook of de groep scholen die geselecteerd zijn met de gegevens uit Cito dataretour een goede afspiegeling vormen van de landelijke populatie. Dat de scholen een goede afspiegeling vormen van de landelijke populatie blijkt wel uit het feit dat de gemiddelde score op de Cito Eindtoets Basisonderwijs voor deze groep niet afwijkt van het populatiegemiddelde.

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 4.4 worden de resultaten van de representativiteitsanalyses getoond. De steekproef is geanalyseerd in relatie tot de variabelen regio, urbanisatiegraad, schooltype en sekse.

Formule 4.1 Berekening van de effectgrootte ϕ

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

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

	Populatie		Steekproef		
stratum	**%**	**M3**	**%**	**E3**	**%**
0 – 10%	60,6	2788	61,9	2676	62,0
10 – 25%	26,4	1199	26,6	1131	26,2
25 – 40%	6,6	266	5,9	261	6,0
> 40%	6,4	251	5,6	251	5,8

M3 $\chi^2(3, N = 4504) = 9{,}653$; $p = 0{,}022$; $\phi = 0.046$

E3 $\chi^2(3, N = 4319) = 5{,}920$; $p = 0{,}116$; $\phi = 0.037$

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

	Populatie		Steekproef		
regio	**%**	**M3**	**%**	**E3**	**%**
Noord	10,2	448	10,0	412	9,5
Oost	22,7	1013	22,5	984	22,8
West	47,1	2145	47,6	2050	47,5
Zuid	20,0	898	19,9	873	20,2

M3 $\chi^2(3, N = 4504) = 0{,}700$; $p = 0{,}873$; $\phi = 0{,}012$

E3 $\chi^2(3, N = 4319) = 2{,}070$; $p = 0{,}558$; $\phi = 0{,}022$

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

	Populatie			Steekproef	
Urbanisatie	**%**	**M3**	**%**	**E3**	**%**
Stad	56,3	2557	56,8	2461	57,0
Land	43,7	1947	43,2	1858	43,0

M3 $\chi^2(1, N = 4504) =$ 0,477; p = 0,490; ϕ = 0,010

E3 $\chi^2(1, N = 4319) =$ 0,909; p = 0,340; ϕ = 0,015

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

	Populatie			Steekproef	
Geslacht	**%**	**M3**	**%**	**E3**	**%**
jongen	50,4	2243	51,0	2175	51,2
meisje	49,6	2159	49,1	2077	48,9

M3 $\chi^2(1, N = 4402) =$ 0,499; p = 0,480; ϕ = 0,011

E3 $\chi^2(1, N = 4252) =$ 0,908; p = 0,341; ϕ = 0,015

De χ^2 -waarden zijn laag en in slechts één geval significant. Bij grotere steekproeven zegt significantie echter weinig. Het is beter om de effectgrootte ϕ als uitgangspunt te nemen. We zien dat de effectgroottes ver onder de .10 liggen en daarmee zeer klein zijn (cf. Cohen, 1988). Ze zijn met .046 nog het hoogst voor de variabele *stratum* bij M3. De conclusie is niettemin dat de normeringssteekproeven een zeer goede afspiegeling vormen van de populatie.

4.3.3 Normeringsresultaten

Na de normeringssteekproef te hebben samengesteld, konden de normen worden bepaald. Naast het gemiddelde werden de percentielen berekend. Dat gebeurde op basis van de verdeling van scores in de normeringssteekproef zoals die is samengesteld op basis van het *embedded field* normeringsonderzoek en Cito dataretour. Om de scores die leerlingen behalen te kunnen vergelijken over de tijd worden vaardigheidsscores gebruikt. Uit de ruwe scores van de leerlingen uit het *embedded field* normerings-onderzoek 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. In paragraaf 3.3 is de verdeling van *plausible values* voor afnamemomenten M3-E3 te zien. De *plausible values* voor dit afnamemoment vormen 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. Daarbij wordt uitgegaan van de empirische cumulatieve verdelingsfunctie. Tabel 4.9 geeft de normgegevens voor LVS-III Rekenen-Wiskunde M3 en E3.
De tussentoets M3E3 kan indien gewenst afgenomen worden medio groep 3 of eind groep 3. M3E3 is dus zelf niet genormeerd. Bij die toets wordt de normering gebruikt die past bij het afnamemoment (medio groep 3 ofwel eind groep3).

Tabel 4.9 Normtabel op leerlingniveau voor LVS-III Rekenen-Wiskunde M3-E3

Tijdstip	*M*	*SD*	*Kurt.*	*Skew.*	*P*10	*P*20	*P*25	*P*40	*P*50	*P*60	*P*75	*P*80
M3	115,3	30,7	0,319	-0,179	76,5	90,4	95,8	108,7	116,1	123,6	135,8	140,6
E3	139,3	29,9	0,272	-0,101	103,7	116,6	120,9	132,6	139,5	146,5	158,2	162,6

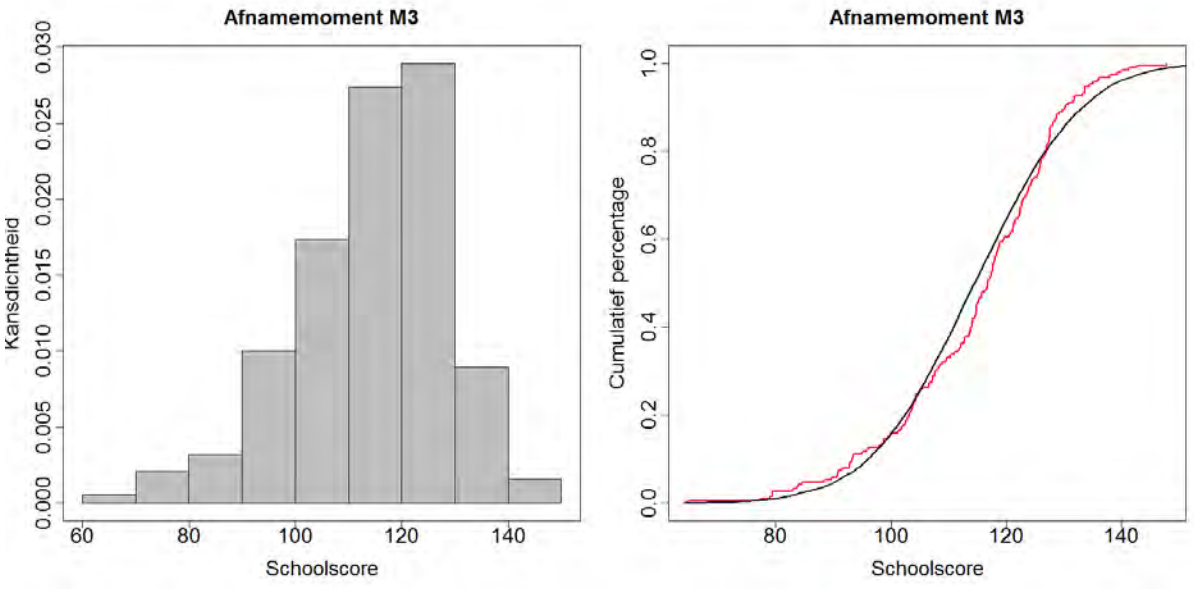
Naast een normering op leerlingniveau kent Cito ook een normering op schoolniveau. Om de schoolverdeling te bepalen wordt het intercept-only multilevel model gebruikt met een gemiddelde per school en een variantie op school- en leerlingniveau. De schatting van het model verloopt via een bootstrap procedure. 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 bootstrap-procedure 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 M3-E3

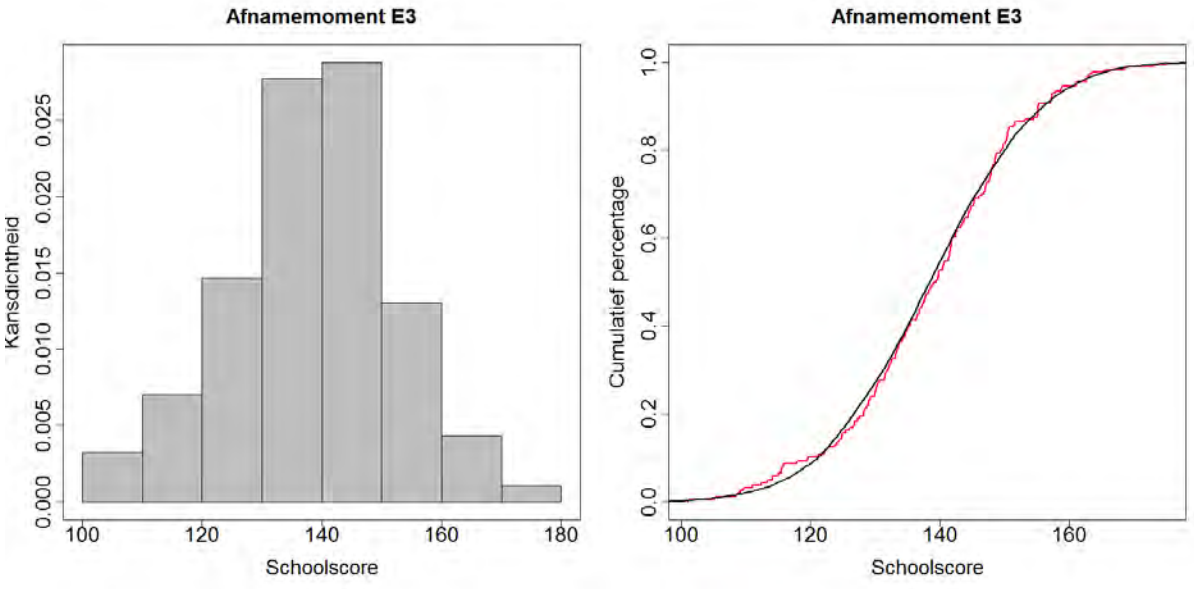
Moment	Aantal replicaties	Aantal scholen	Gemiddelde	SD School	SD Leerling	ICC
M3	20	190	114,7	13,6	29,1	0,18
E3	20	184	138,3	13,0	27,1	0,19

Figuur 4.5 laat de verdeling van schoolgemiddelden zien. Het is lastig te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 161. Op het eerste gezicht lijkt de verdeling redelijk normaal verdeeld. Op basis van het eindresultaat uit de bootstrap procedure zijn de percentielen voor de vaardigheidsverdeling A tot en met E en I tot en met V berekend. Tabel 4.11 geeft de normgegevens op schoolniveau. De percentielen komen dichter bij elkaar te liggen dan in de leerlingverdeling. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in de verschillende niveaus.

Figuur 4.5 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde M3



Figuur 4.6 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde E3



Tabel 4.11 Normtabel op schoolniveau voor LVS-III Rekenen-Wiskunde M3-E3

Tijd	***M***	***SD***	***P*10**	***P*20**	***P*25**	***P*40**	***P*50**	***P*60**	***P*75**	***P*80**
M3	114,7	13,6	97,3	103,2	105,5	111,2	114,7	118,1	123,9	126,1
E3	138,3	13,0	121,6	127,4	129,5	135,0	138,3	141,6	147,1	149,3

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 opgenomen zijn. De betrouwbaarheid van de uitgegeven toetsen in klassieke zin is dan ook niet rechtstreeks te bepalen. Van de toetsen in het proefonderzoek –met deels opgaven die niet in de uiteindelijke toetsen terecht zijn gekomen- zijn wel de klassieke betrouwbaarheden bekend. Deze toetsen waren niet alle van gelijke lengte. Met de Spearman-Brown formule is berekend wat de betrouwbaarheden van deze toetsen zijn, als deze de uiteindelijke lengte van 52 opgaven zouden hebben. De gemiddelde alpha van deze toetsen, herberekend naar een toetslengte van 52 opgaven bleek iets boven de 0,90 te liggen.
Het is echter ook mogelijk om de betrouwbaarheid van elke uiteindelijke toets te schatten door gebruik te maken van het feit dat alle items die zijn opgenomen in de toetsen OPLM-geschaald zijn. Ook andere beschrijvende gegevens, zoals de gemiddelde score en de standaardmeetfout, zijn te schatten op grond van het feit dat de toetsen volledig bestaan uit OPLM-gekalibreerde items. Om relevante beschrijvende gegevens bij de verschillende toetsen te genereren, is gebruikgemaakt van het programma OPLAT (Verhelst, Glas en Verstralen, 1995).
In OPLAT wordt een door Verhelst, Glas en Verstralen (1995, pp. 99-100) ontwikkelde coëfficiënt berekend die qua interpretatie een grote overeenkomst vertoont met betrouwbaarheidscoëfficiënten uit de klassieke testtheorie. Het begrip ware score is wat meer geëxpliciteerd, namelijk als de verwachte score op een (vaste) toets, maar dan gezien als functie van de latente variabele θ. Deze verwachte waarde wordt aangeduid met τ(θ). Als bovendien bekend is hoe θ in de populatie verdeeld is, kunnen ook het gemiddelde en de variantie van de ware scores in de populatie bepaald worden. De variantie van de ware scores in de populatie worden aangegeven met het symbool Var(τ). Tussen θ en τ(θ) bestaat een een-op-een relatie, immers de een kan uit de andere berekend worden. Het is echter niet zo dat een persoon met vaardigheid θ per se de toetsscore τ(θ) moet behalen (dat is alleen zo als de toets oneindig lang wordt). De geobserveerde score bij een eenmalige afname zal dan ook een afwijking vertonen van de verwachte score, waardoor met een eenmalige toetsafname niet meer zonder fout de waarde van θ bepaald kan worden. De variantie van de geobserveerde toetsscore wordt aangegeven met Var(t|τ(θ)), en door weer gebruik te maken van de distributie van θ in de populatie kan ook de gemiddelde variantie van de geobserveerde toetsscores berekend gaan worden.

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

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

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

waarin MAcc staat voor 'Accuracy of Measurement'.

5.2 Betrouwbaarheid: resultaten

Tabel 5.1 bevat informatie over de meeteigenschappen van de vaardigheidsschaal Rekenen-Wiskunde. In de tweede kolom staat de maximumscore, voor iedere toets is deze gelijk aan het aantal opgaven dat deel uitmaakt van de totale toets. De derde kolom geeft de geschatte gemiddelde scores van de leerlingen op de verschillende toetsen. De vierde kolom bevat informatie over de geschatte standaardmeetfout op de ruwe

score van iedere toets. De vijfde kolom laat zien wat de geschatte betrouwbaarheidscoëfficiënt (MAcc) van de verschillende toetsen (of toetsonderdelen) is. De schattingen van de gemiddeldes, de standaard-meetfouten 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. Merk op dat deze betrouwbaarheden iets hoger liggen dan de waarden van geobserveerde toetsen herschaald naar 52 opgaven. Daar is een tweetal redenen voor aan te voeren. Ten eerste waren die betrouwbaarheden bepaald als coëfficiënt alpha, wat een onderschatting is van de daadwerkelijke betrouwbaarheid. Een andere belangrijke reden is dat de opgaven die in de uiteindelijke toets terecht zijn gekomen ook mede geselecteerd zijn op hun onderscheidend vermogen: het zijn de betere opgaven die zodoende ook een betrouwbaardere toets opleveren.

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

Toets	**Maximum score**	**Gemiddelde**	**Standaardmeetfout**	**MAcc**	**Test-hertest (simulatie)**
M3	52	36,9	2,82	0,92	0,92
M3E3	52	39,2	2,67	0,90	0,90
E3	52	38,0	2,80	0,92	0,92
M3 digitaal	52	31,8	3,00	0,94	0,94
M3E3 digitaal	52	38,7	2,72	0,92	0,92
E3 digitaal	52	33,2	3,03	0,93	0,93

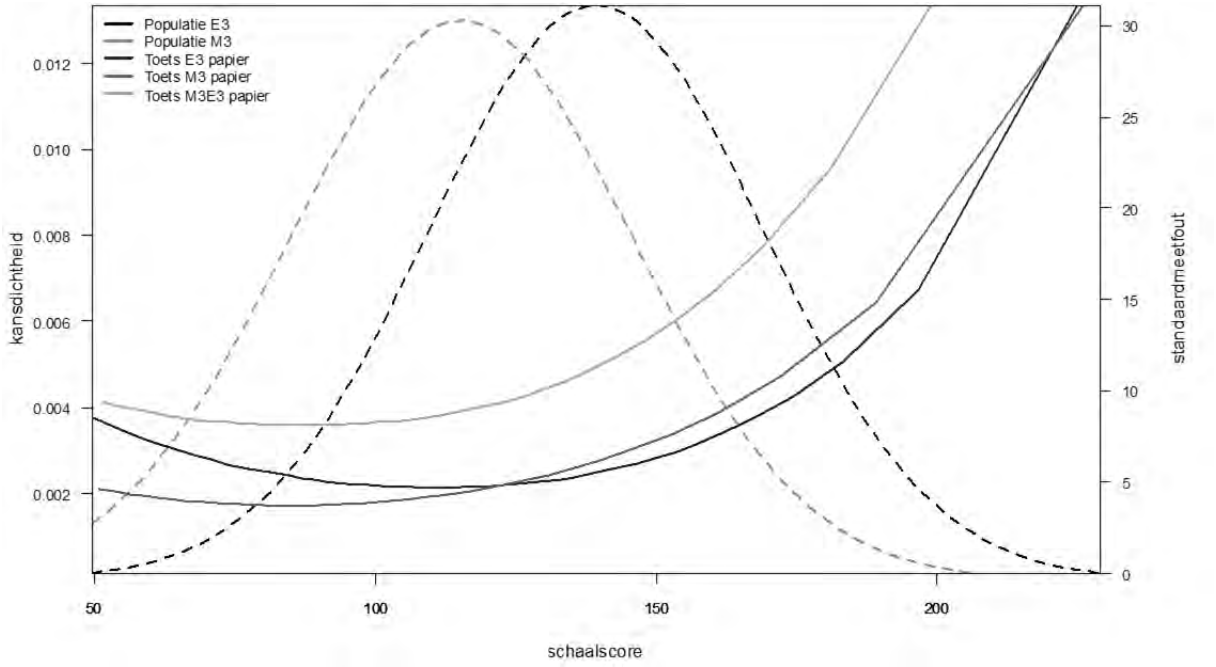
Het feit dat alle items OPLM-gekalibreerd zijn, maakt het mogelijk een hertest te simuleren. We hebben een dubbele afname gesimuleerd van 1 000 000 leerlingen. Daarbij hebben we enerzijds de vaardigheids-verdeling van alle leerlingen, anderzijds alle itemparameters als uitgangspunt genomen. Steeds is een bepaalde vaardigheid aselect uit de verdeling genomen en zijn twee bij deze vaardigheid horende afnames gesimuleerd. Uiteindelijk is de correlatie tussen deze 1 000 000 dubbele (virtuele) afnames berekend. Men kan deze simulatie beschouwen als een test-hertestonderzoek onder ideale condities. De tweede toets-afname 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 exact overeen met eerder berekende coëfficiënten en leiden dan ook tot dezelfde conclusies met betrekking tot de betrouwbaarheid van de Rekenen-Wiskunde-toetsen.

5.3 Lokale betrouwbaarheid en meetnauwkeurigheid

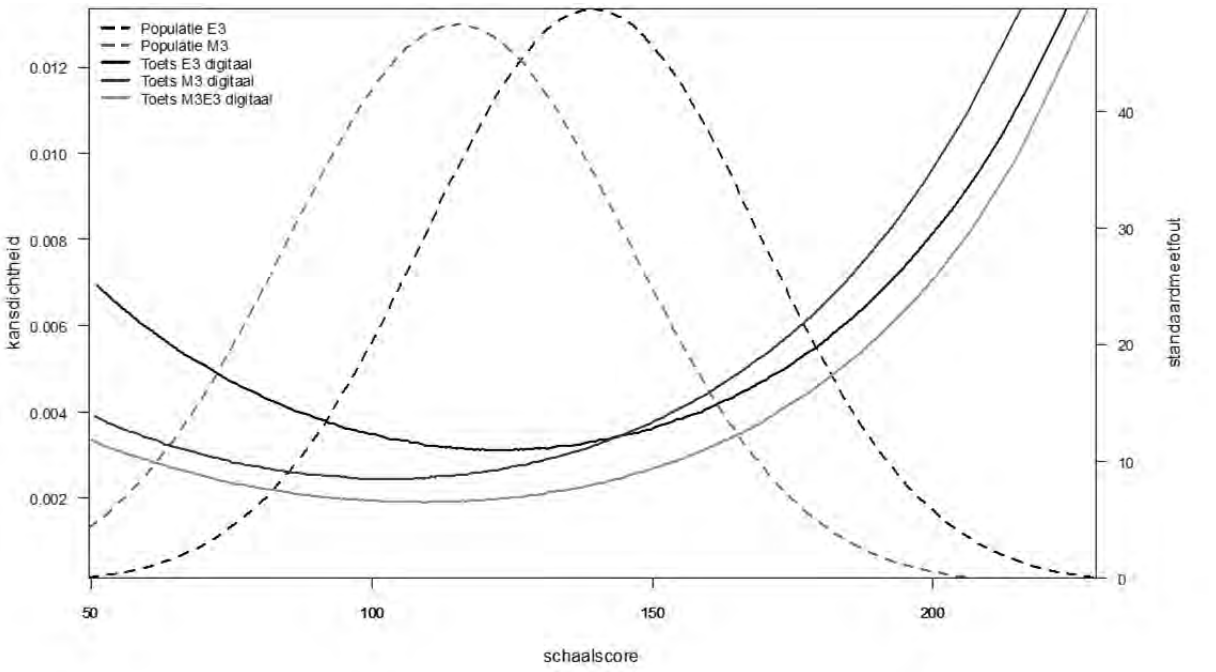
De hiervoor vermelde betrouwbaarheidscoëfficiënten hebben alleen betrekking op de globale meetnauwkeurigheid van de toetsen en geven geen beeld van de lokale meetnauwkeurigheid van de verschillende toetsen Rekenen-Wiskunde. De figuren 5.1 en 5.2 geven grafisch weer hoe het gesteld is met de lokale meetnauwkeurigheid bij de papieren en digitale toetsen M3, M3E3 en E3. 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 toetsen Rekenen-Wiskunde M3, M3E3 en E3 en de kansdichtheidsfuncties voor de M3 en E3-populaties



Figuur 5.2 Grootte van de meetfouten voor de digitale toetsen Rekenen-Wiskunde M3, M3E3 en E3 en de kansdichtheidsfuncties voor de M3 en E3-populaties



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 5.7. De betrouwbaarheidstabellen laten het effect van de lokale meetnauwkeurigheid zien.
Zo laat tabel 5.2 bijvoorbeeld zien dat 81,8 procent van de leerlingen die bij de M3-toets in scoregroep A vallen met hun geschatte vaardigheidsscore ook met hun werkelijke vaardigheidsscore in deze scoregroep vallen. Anders gezegd: de kans dat een A-leerling terecht als een A-leerling wordt bestempeld is ongeveer 81,3 procent. Verder laat de tabel zien dat 18,3 procent van de leerlingen in niveaugroep A een vaardigheidsscore heeft die in werkelijkheid in scoregroep B valt.
Verdere gedetailleerde informatie over de meetnauwkeurigheid van de toetsen is te vinden in de handleiding van het toetspakket (Cito, 2013). Bij de portalbestanden is de tabel van toetsscore naar vaardigheidsscore en niveau opgenomen. In deze tabel 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 betrouwbaarheidstabellen vaak samengevat in één of meerdere indices. Wij gebruiken de *plus/minus 1 niveau-index* en de *Marginal Classification Accuracy*. De eerste maat is bedacht door Pilliner (1969). Hij stelt als ambitieniveau dat 95 procent van de leerlingen in een scoregroep in werkelijkheid ook in die scoregroep moet scoren, ***of*** één scoregroep daarboven ***of*** één 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, maar in de praktijk niet te realiseren, toetsafname lijkt de *Marginal Classification Accuracy* rond 0,75 - 0,80 uit te komen (Keuning & Béguin, in voorbereiding). In de praktijk liggen de waarden vaak tussen 0,60 en 0,70.

De samenvattende indices voor de afnamemomenten medio groep 3 en einde groep 3 zijn te vinden in de tabellen 5.8 (voor de papieren toetsen) en 5.9 (voor de digitale toetsen). In deze tabellen laten de Marginal Classification Accuracy waarden zien dat de meeste leerlingen op basis van hun geschatte vaardigheidsscore geplaatst worden in de niveaugroep waar ze werkelijk thuishoren en de Accuracy plus/minus 1 niveau waarden maken aannemelijk dat de uitkomsten duidelijk in lijn liggen met het ambitieniveau zoals dat geformuleerd is door Pilliner (1969).

Tabel 5.2 Betrouwbaarheidstabel Toets M3 papier voor afnamemoment medio 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	89,3	10,7	0,1	0,0	0,0	E	88,0	11,9	0,0	0,0	0,0
IV	14,5	66,9	18,2	0,5	0,0	D	12,2	69,5	18,2	0,0	0,0
III	0,3	21,5	56,3	21,3	0,7	C	0,0	13,3	71,3	15,2	0,1
II	0,0	1,8	23,1	54,2	20,9	B	0,0	0,1	19,1	62,4	18,3
I	0,0	0,1	2,4	18,5	79,0	A	0,0	0,0	1,0	17,9	81,1

Tabel 5.3 Betrouwbaarheidstabel Toets M3E3 papier voor afnamemoment medio 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	87,0	12,6	0,4	0,0	0,0	E	85,4	14,4	0,3	0,0	0,0
IV	20,7	56,1	21,1	2,0	0,0	D	17,3	61,3	21,0	0,3	0,0
III	2,1	25,9	45,7	23,9	2,5	C	0,6	18,4	60,4	20,0	0,7
II	0,2	5,5	24,2	43,4	26,6	B	0,0	1,3	23,5	52,2	23,0
I	0,1	1,1	5,6	19,7	73,5	A	0,0	0,2	3,7	21,1	75,0

Tabel 5.4 Betrouwbaarheidstabel Toets E3 papier voor afnamemoment einde 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	89,0	10,9	0,1	0,0	0,0	E	86,6	13,3	0,1	0,0	0,0
IV	15,8	65,5	18,2	0,5	0,0	D	13,7	68,4	17,9	0,0	0,0
III	0,4	22,1	56,5	20,5	0,6	C	0,1	14,0	69,8	16,1	0,1
II	0,0	1,7	22,7	54,4	21,1	B	0,0	0,2	18,8	63,3	17,7
I	0,0	0,1	2,5	19,1	78,3	A	0,0	0,0	1,0	19,1	79,8

Tabel 5.5 Betrouwbaarheidstabel Toets M3 digitaal voor afnamemoment medio 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	88,9	11,0	0,1	0,0	0,0	E	86,4	13,5	0,1	0,0	0,0
IV	13,1	69,7	17,0	0,2	0,0	D	12,4	69,4	18,2	0,0	0,0
III	0,1	18,2	62,9	18,6	0,2	C	0,0	11,5	75,1	13,3	0,0
II	0,0	0,5	20,0	62,7	16,7	B	0,0	0,0	15,3	69,9	14,8
I	0,0	0,0	0,7	16,1	83,2	A	0,0	0,0	0,2	14,7	85,1

Tabel 5.6 Betrouwbaarheidstabel Toets M3E3 digitaal voor afnamemoment medio 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	89,4	10,6	0,1	0,0	0,0	E	87,0	12,9	0,1	0,0	0,0
IV	15,5	66,0	18,0	0,5	0,0	D	13,2	69,3	17,4	0,0	0,0
III	0,4	22,2	56,3	20,5	0,6	C	0,1	13,8	70,1	16,0	0,1
II	0,0	2,0	23,1	53,4	21,6	B	0,0	0,2	19,2	62,5	18,0
I	0,0	0,2	3,0	19,5	77,3	A	0,0	0,0	1,3	19,7	79,0

Tabel 5.7 Betrouwbaarheidstabel Toets E3 digitaal voor afnamemoment einde 3

	Scoregroep waarin de ware score valt						Scoregroep waarin de ware score valt				
Score-groepen V t/m I	V	IV	III	II	I	Score-groepen E t/m A	E	D	C	B	A
V	88,4	11,4	0,1	0,0	0,0	E	85,2	14,7	0,2	0,0	0,0
IV	15,4	66,3	17,9	0,4	0,0	D	14,3	67,3	18,4	0,0	0,0
III	0,2	20,5	59,7	19,2	0,3	C	0,1	13,3	71,3	15,3	0,0
II	0,0	0,9	20,9	59,6	18,6	B	0,0	0,1	16,7	67,5	15,7
I	0,0	0,0	1,1	17,2	81,7	A	0,0	0,0	0,3	16,7	83,0

Tabel 5.8 Samenvattende indices toetsen M3, M3E3 en E3 op afnamemomenten groep 3 Papier

	Toets M3, afnamemoment M3		Toets M3E3, afnamemoment E3		Toets E3, afnamemoment E3	
	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	69,1	72,9	60,1	64,3	68,5	71,9
Accuracy plus/minus 1 niveau	98,8	99,7	96,1	98,6	98,8	99,7

Tabel 5.9 Samenvattende indices toetsen M3, M3E3 en E3 op afnamemomenten groep 3 Digitaal

	Toets M3, afnamemoment M3		Toets M3E3, afnamemoment E3		Toets E3, afnamemoment E3	
	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E	scoregroep I t/m V	scoregroep A t/m E
Marginal classification accuracy	73,5	76,6	68,2	71,8	70,1	73,8
Accuracy plus/minus 1 niveau	99,7	99,9	98,7	99,7	99,4	99,9

Gemiddeld gezien scoort, afhankelijk van het afnamemoment en de gekozen indeling in scoregroepen, 96 tot 99 procent van de leerlingen in een scoregroep ook in werkelijkheid in die scoregroep, ***of*** één scoregroep daarboven ***of*** één scoregroep daaronder. De *Marginal Classification Accuracy* loopt uiteen van 60 tot 77 procent. Dit betekent dat de classificatie op basis van de geschatte vaardigheidsscore bij beide afnamemomenten gemiddeld gezien in ruim 60 procent van de gevallen overeenstemt met de classificatie op basis van de (gesimuleerde) werkelijke vaardigheidsscore. De resultaten zijn hiermee positief te noemen: het percentage misclassificaties is beperkt. De laagste waarden zien we bij toets M3/E3, en dan met name bij de hoogste scoregroep. Dat is conform verwachting, aangezien deze toets – die wat makkelijker is dan de toets E3 – expliciet bedoeld is voor de minst vaardige leerlingen aan het eind van groep 3. De (boven)gemiddeld vaardige leerlingen zullen deze toets in de praktijk ook niet maken.

Op basis van bovenstaande gegevens concluderen we dat op basis van de toetsen Rekenen Wiskunde 3.0 groep 3 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 zal 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 leervorderingstoetsen, zoals deze toets Rekenen-Wiskunde voor groep 3, 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 soortgenoot onderzoek 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 een 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 3.

6.2 Unidimensionaliteit, respectievelijk structuur

Zoals in hoofdstuk 4 al aangegeven, zijn bij de kalibratie S-toetsen uitgevoerd die een indicatie geven van de kwaliteit van de kalibratie. Daarbij is duidelijk geworden dat voor de toets de verdeling gelijkmatig is over het gehele interval van overschrijdingskansen. Dit resultaat geeft een bevestiging van het eerder geschetste beeld, dat sprake is van niet-significante S-toetsen. Het aantal significante S-toetsen was te verwaarlozen onder het 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 de momenten M3 en E3 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 3. De toetsopgaven richten zich op inhoudelijk relevante onderwerpen als *Getallen en getalrelaties, Optellen en aftrekken, Vermenigvuldigen en delen en Meten.* De schaal zou wellicht een nog fraaiere passing vertonen als slechts één van deze onderdelen gekozen zou zijn – dit ligt dan dichter tegen het theoretische ideaal van unidimensionaliteit aan -, maar zou inhoudelijk te eenzijdig zijn om praktisch zinvolle uitspraken over rekenvaardigheid te doen.
Ook de grote onderlinge samenhang tussen de vier verschillende inhoudelijke subvaardigheden geeft aan dat in de praktijk deze vier subvaardigheden samengenomen kunnen worden om een unidimensionele uitspraak te doen. De correlaties tussen de subschalen, zoals gegeven in tabel 6.1, variërend van 0,85 tot 0,94 (gemiddeld 0,885) 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. De correlatie van de subschalen met de totaalscores van gemiddeld 0,98 geeft al aan dat een enkele rekenvaardigheidsscore geoorloofd is.
Dat de correlatie niet geheel gelijk aan 1 is, heeft voor de schaal betekent dat er nog wel enige informatie beschikbaar is in de data die niet geheel gevangen wordt door de unidimensionele score. Dat is gezien de hoge correlaties niet veel, maar die wordt ondervangen door ook gebruik te maken van profielscores uit de categorieënanalyses. Voor het zeer ruime merendeel van de leerlingen hebben deze geen betekenis gezien de hoge mate van unidimensionaliteit, maar voor een klein percentage leerlingen waar dit nog wel een rol speelt kan dat hiermee ondervangen worden.

Correlatie scores deeltaken met elkaar en totaal

Tabel 6.1 Latente correlaties tussen score op deeltaken en totaalscore van de toetsen M3 en E3

M3 (N=2250)

	Getallen en getalsrelaties (cat. 12)	Optellen en aftrekken (cat. 13)	Vermenigvuldigen en delen (cat. 14)	Meten (cat21)
Optellen en aftrekken (cat. 13)	0,89			
Vermenigvuldigen en delen (cat. 14)	0,86	0,90		
Meten (cat. 21)	0,90	0,88	0,89	
Totaal	0,99	0,98	0,98	0,99

E3 (N=1891)

	Getallen en getalsrelaties (cat. 12)	Optellen en aftrekken (cat. 13)	Vermenigvuldigen en delen (cat. 14)	Meten (cat21)
Optellen en aftrekken (cat. 13)	0,89			
Vermenigvuldigen en delen (cat. 14)	0,94	0,87		
Meten (cat. 21)	0,88	0,85	0,87	
Totaal	0,99	0,97	0,99	0,99

De gevonden hoge correlaties geven ook al aan dat een factor analyse hier weinig meer uit zou halen dan een enkele (reken)vaardigheid die gemeten zou worden, aangezien die voor de overgrote meerderheid van de leerlingen het gedrag op de toets bepaald. In dit geval is het doen van een factoranalyse binnen een onvolledig design praktisch niet aan te raden.
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 0,10 uitstekend te noemen. De hoogst gevonden waarde is 0,21 bij één van de 52 opgaven bij M3 digitaal. Bij alle andere opgaven is de c-waarde (veel) lager dan 0,20. De conclusie mag luiden dat we ook op basis van deze analyse de kalibratie geslaagd kunnen noemen. Deze nauwkeurigheid van de parameters hangt samen met het hoge aantal observaties per opgave: iedere opgave is door gemiddeld 411 M3-leerlingen en 480 E3-leerlingen gemaakt, wat boven het beoogde aantal van 400 ligt dat noodzakelijk is voor schattingen van een 2-parameter model. Daarbij kan nog opgemerkt worden dat het gebruikte model een hybride is tussen een één- en een twee-parameter model.

6.3 Itemkwaliteit

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

	P-waarden		R_{it}-waarden		N items
	Range	Gemiddelde	Range	Gemiddeld	
M3	0,41 - 0,88	0,71	0,25 - 0,58	0,42	52
M3E3	0,41 - 0,94	0,75	0,29 - 0,58	0,40	52
E3	0,51 - 0,91	0,73	0,30 - 0,57	0,45	52
M3 dig	0,42 - 0,82	0,61	0,26 - 0,59	0,45	52
M3E3 dig	0,50 - 0,91	0,74	0,24 - 0,58	0,42	52
E3 dig	0,43 - 0,81	0,64	0,23 - 0,56	0,43	52

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 M3, M3E3 en E3. Bij alle toetsen is te zien dat de p-waarden liggen tussen de 0,41 en 0,94. In het algemeen geldt dat, enkele uitzonderingen daargelaten, de p-waarden van de items tussen de 0,40 en de 0,90 moeten liggen. Er is gezorgd voor een goede spreiding van moeilijkheid over de items. De gemiddelde moeilijkheid van de M3-, M3E3- en E3-toets ligt tussen de 0,61 en 0,75. Er wordt in het algemeen voor groep 3 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. Bij geen enkele toets ligt de R_{it}-waarde onder de 0,23. De gemiddelde R_{it}-waarden is voor alle drie de toetsen 0,40 of hoger. Door de COTAN wordt een R_{it}-waarde van boven de 0,30 gekwalificeerd als goed. Met een gemiddelde van 0,40 of hoger is de itemkwaliteit van de toetsen uitstekend te noemen. Bijlage 4 bevat een volledig overzicht van de p-waarden en de R_{it}-waarden van de items van de toetsen.
In tabel 6.3 zijn de verdelingskarakteristieken gegeven van de ruwe scores op de verschillende toetsmomenten. De gemiddelden komen uiteraard overeen met wat men bij een gegeven aantal items mag verwachten bij de gekozen (gemiddelde) moeilijkheidsgraad. Omdat deze gemiddelde moeilijkheidsgraad

voor alle onderdelen rond de 0,70 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 vijf van de zes schalen gematigd linksscheef. De uitzondering is de toets M3E3 digitaal, die met een waarde van 0,26 bij benadering als symmetrisch gezien kan worden (met een rechtsscheve tendens). De verdelingen zijn ééntoppig. De toppen van de verdelingen van de papieren toetsen lijken vrij sterk op elkaar en datzelfde geldt voor de toppen van de verdelingen van de digitale toetsen. Opvallend is dat de toppen van de digitale toetsen wat platter zijn terwijl de papieren toetsen wat spitser zijn.

Tabel 6.3 Verdelingskenmerken van de toetsen Rekenen-Wiskunde groep 3

Meetmoment	Aantal opgaven	Gemiddelde	SD	Skewness	Kurtosis
M3	52	36,9	10,2	-0,81	0,01
M3E3	52	39,2	8,5	-0,90	0,47
E3	52	38,0	10,1	-0,85	0,10
M3 dig	52	31,8	11,7	-0,79	-0,37
M3E3 dig	52	38,7	9,8	0,26	-0,92
E3 dig	52	33,2	11,1	-0,60	-0,48

6.4 Itembias

Er is onderzoek uitgevoerd naar differentieel itemfunctioneren (*Differential Item Functioning,* DIF) met betrekking tot sekse. 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 slechts één item van de selectie van E3 (S = 35,262; DF =15; P = 0,002) differentieel functioneren zien (bij α = 0,01). Bij E3 is bij geen van de items deze toets significant. Voor de toetsen Rekenen-Wiskunde van leerjaar 3 is dus nauwelijks sprake van DIF met betrekking tot sekse.

6.5 Soortgenootonderzoek

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 correlaties tussen gewogen scores gebaseerd op items uit LVS-II en die van LVS-III waren hoog voor M3 (r=0,94; N=2239; p<0,005) en E3 (r=0,95; N=2131; p<0.005). 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 vijf scholen naast de 3de generatie rekenen wiskunde toets M3 en E3, ook de Schoolvaardigheidstoets rekenen-wiskunde van Boomtestuitgevers afgenomen.
De scholen werden benaderd op basis van het feit dat zij de nieuwe toetsen Rekenen-Wiskunde voor groep 3 (generatie 3) hadden aangeschaft. Er werden alleen scholen aangeschreven die meer dan één pakket Rekenen-Wiskunde van het leerlingvolgsysteem hadden aangeschaft zodat van een school meerdere klassen of locaties zouden kunnen deelnemen. Dit is gedaan om het onderzoek zo efficiënt mogelijk op te zetten. In totaal hebben vijf scholen met zeven groepen deelgenomen aan het onderzoek.
Er is daarbij sprake van een redelijke landelijke spreiding, die volstaat voor het doel van het onderzoek. De scholen zijn afkomstig uit de provincies Utrecht, Gelderland en Overijssel en zijn gelegen in de plaatsen

Staphorst (1 locatie, 19 leerlingen), Amersfoort (1 locatie, 32 leerlingen), Putten (1 locatie, 15 leerlingen), Dieren (2 locaties, 20 leerlingen) en Apeldoorn (2 locaties, 50 leerlingen). In totaal hebben 136 leerlingen de schoolvaardigheidstoets rekenen-wiskunde van Boom gemaakt. Er zijn bij de afname geen leerlingen van een groep uitgesloten van deelname. Alle leerlingen van de groepen die deelnamen hebben de toets gemaakt.

Tabel 6.4 Correlatie van de Cito toetsen Rekenen-Wiskunde M3 en E3 met diverse toetsen op het gebied van rekenen wiskunde

	Cito Rekenen-Wiskunde M3 generatie 3			Cito Rekenen-Wiskunde E3 generatie 3	
	correlatie	N		correlatie	N
Schoolvaardigheidstoets Rekenen wiskunde (Boom)	0,70	105		0,78	106
Cito Rekenen-Wiskunde kleuters M1	0,70	35		0,62	35
Cito Rekenen-Wiskunde kleuters E1	0,68	43		0,64	43
Cito Rekenen-Wiskunde kleuters M2	0,64	67		0,65	67
Cito Rekenen-Wiskunde kleuters E2	0,66	80		0,65	80
Cito Rekenen-Wiskunde M3	1,00	105		0,85	105
Cito Rekenen-Wiskunde E3	0,85	105		1,00	106
Cito Rekenen-Wiskunde M3 Gen2				0,80	19
Cito Rekenen-Wiskunde E3 Gen2				0,79	19
Tempotest rekenen (Boom)	0,39	49		0,43	50

In tabel 6.4 staan de correlaties van de nieuwe toetsen Rekenen-Wiskunde van groep 3 weergeven met relevante toetsen rekenen wiskunde. Alle in de tabel vermelde toetsen zijn positief beoordeeld door de COTAN op de Tempotest rekenen na. De correlatie met de Schoolvaardigheidstoets rekenen-wiskunde van Boom testuitgevers is 0,7 voor de M3 toets en 0,78 voor de E3 toets. De correlatie met de andere rekentoets van Boom, namelijk de Tempotest rekenen, ligt een stuk lager namelijk 0,39 voor de M3 rekenen-wiskunde toets van Cito en 0,43 voor de E3 toets. Hier lijkt dan ook een net iets andere vaardigheid te worden gemeten dan bij de andere toetsen Rekenen-Wiskunde. Aangezien de validiteit van de Tempotest rekenen als onvoldoende is beoordeeld kan aan deze uitkomst ook niet teveel waarde worden gehecht. De correlatie van de Schoolvaardigheidstoets rekenen-wiskunde van Boom met de tempotoets van Boom is eveneens lager dan de correlatie tussen de andere rekentoetsen, namelijk 0,29 op basis van 50 leerlingen (staat niet in het overzicht hierboven). Ook hier geldt dat de gemeten vaardigheid een wat ander karakter heeft, wellicht dat het specifieke karakter van de afnamevorm, namelijk snelheid / tempo, bij de tempotoets voor de lagere correlaties zorgt dan wel het gebrek aan validiteit de oorzaak is van de lagere correlatie.
De hoogste correlatie (0,85) wordt gevonden tussen de M3 toets en de E3 toets van de 3de generatie LVS rekenen wiskunde. Dit ligt in de lijn der verwachting aangezien de toetsen qua opzet, samenstelling en afnamevorm identiek zijn en alleen in moeilijkheid verschillen. De correlatie van de 3de generatie toetsen rekenen-wiskunde met de 2de generatie toetsen rekenen-wiskunde zijn eveneens hoog te noemen, namelijk rond de 0,8. Daarbij moet wel de opmerking worden gemaakt dat deze correlatie slechts op 19 leerlingen is gebaseerd (1 groep uit Staphorst).
De correlaties van de 3de generatie toetsen LVS rekenen-wiskunde met de kleutertoetsen rekenen-wiskunde lopen van 0,62 tot 0,70. Omdat de kleutertoetsen Rekenen-Wiskunde en de 3de generatie toetsen Rekenen-Wiskunde LVS qua opzet verschillen en er meer tijd tussen de afnames heeft gezeten, is dat niet verwonderlijk. De correlatie is nog steeds hoog te noemen.

De correlatie met de verschillende leestoetsen loopt van 0,14 bij Woordenschat E3 tot 0,56 bij begrijpend lezen (zie tabel 6.5). In het algemeen kan gesteld worden dat de correlatie van de 3de generatie rekenen wiskunde toetsen met de leestoetsen (Spelling, Begrijpend lezen, DMT en Woordenschat) rond de 0,40 liggen. Daarmee is de correlatie beduidend lager dan de correlatie van de 3de generatie Rekenen-Wiskunde toetsen met de verschillende rekentoetsen die in het algemeen rond de 0,70 liggen.

Tabel 6.5 Correlatie van de derde generatie Cito LVS toetsen Rekenen-Wiskunde M3 en E3 met diverse toetsen op het gebied van leervorderingen

	Cito Rekenen-Wiskunde M3			Cito Rekenen-Wiskunde E3	
	correlatie	N		correlatie	N
Cito Spelling M3	0,47	105		0,42	105
Cito Spelling E3	0,45	105		0,47	106
Cito DMT M3	0,36	63		0,46	63
Cito DMT E3	0,35	63		0,46	64
Cito Begrijpend lezen E3	0,52	49		0,56	50
Cito Woordenschat M3	0,44	49		0,32	49
Cito Woordenschat E3	0,32	49		0,14	50

Aan de scholen die hebben deelgenomen aan het normeringsonderzoek is gevraagd of dataretour van de betreffende leerlingen van andere LVS-onderdelen gebruikt mocht worden. Met de dataretour functie zijn van de leerlingen van het normeringsonderzoek ook de scores op andere LVS-toetsen beschikbaar. In de tabel hieronder is de correlatie tussen de toets Rekenen-Wiskunde E3 en de toetsen Spelling, Begrijpend lezen en Technisch lezen weergegeven.

Tabel 6.6 Correlaties tussen Rekenen-Wiskunde E3 en verschillende andere LVS-onderdelen

	Rekenen-Wiskunde*	Aantal leerlingen
Cito Spelling E3	0,52	425
Cito Begrijpend lezen E3	0,58	436
Cito Technisch lezen – DMT E3	0,40	467
Cito Technisch lezen – Leestechniek E3	0,59	150
Cito Technisch lezen – Leestempo E3	0,34	97

*Deze correlaties zijn gecorrigeerd voor attenuatie

Bij de leesvakken is ook bij de data van het normeringsonderzoek de correlatie met de toets Rekenen-Wiskunde beduidend lager dan de correlatie van de toets Rekenen-Wiskunde met de andere toetsen Rekenen-Wiskunde. Bij onderdelen waar snelheid een belangrijke rol speelt is de correlatie over het algemeen lager (Leestempo).

Tabel 6.7 Correlatie van de toets begrijpend lezen met de schoolvaardigheidstoets rekenen-wiskunde, Rekenen-Wiskunde M3 en Rekenen-Wiskunde E3.

	SVT RW	N	RW M3	N	RW E3	N
Begrijpend lezen E3	0,55	69	0,52	49	0,56	50

Opvallend is dat de correlatie van de toetsen Rekenen-Wiskunde met Begrijpend lezen allemaal rond de 0,55 liggen (zie tabel 6.7). In het veld wordt regelmatig aangegeven dat de toetsen van Cito eigenlijk Begrijpend lezen meten omdat de toetsen opgaven met contexten bevatten. De reken-wiskunde toets van Boom (SVT RW) is opgezet zonder contextopgaven en dus zou men op basis daarvan verwachten dat de correlatie met begrijpend lezen lager zou liggen dan voor de toetsen met contextopgaven van Cito. Echter uit de overeenkomende correlaties met begrijpend lezen kan dat niet worden bevestigd. De correlatie van de “kale” toets SVT RW met de Begrijpend lezen toets is even hoog als de correlatie van de toets Rekenen-Wiskunde van Cito met Begrijpend lezen. We hebben deze gegevens niet kunnen bevestigen op basis van de wetenschappelijke verantwoording van de schoolvaardigheidstoets rekenen-wiskunde van Boom. Alhoewel bij de uitgangspunten en opzet van die toets uitgebreid wordt stilgestaan op de invloed van de taligheid op de toetsresultaten, worden er in de wetenschappelijke verantwoording geen correlatie gegevens gepresenteerd over divergente validiteit zoals bijvoorbeeld de correlatie met Begrijpend lezen. Mogelijk doet het kleine verschil in correlatie tussen begrijpend lezen en de schoolvaardigheidstoets rekenen-wiskunde enerzijds en begrijpend lezen en de toetsen Rekenen-Wiskunde van Cito anderzijds zich uitsluitend voor bij de lagere groepen doordat de vaardigheden bij jongere leerlingen homogener van aard zijn. Immers het onderwijs op de verschillende leergebieden is nog maar van korte duur geweest. Samenvattend kan dus gesteld worden dat de correlaties van de 3de generatie toetsen Rekenen-Wiskunde LVS conform verwachting zijn. De correlatie met Cotan goedgekeurde toetsen rekenen-wiskunde zijn hoger dan de correlatie met leestoetsen en vormen daarmee een ondersteuning voor de validiteit van de toetsen. De data geven aan dat er gemeten wordt wat men beoogt te meten, namelijk rekenen-wiskunde.

6.6 Verschillen tussen relevante subgroepen

In de onderstaande tabellen worden per afnamemoment de gemiddelde scores van de leerlingen per lesmethode weergegeven. Opvallend daarbij is het hoge gemiddelde van de methode Rekenrijk in leerjaar 3. De andere methoden liggen in leerjaar 3 dicht bij elkaar. Er is daarbij sprake van geen tot een klein effect. Het verschil van Rekenrijk met de andere methode in leerjaar 3 varieert van klein tot groot. Van de methode Rekenrijk is bekend dat deze het formele rekenen eerder aanbiedt dan de andere methoden.

Tabel 6.8 Gemiddelde score per lesmethode

Moment	Lesmethode	M	SD	Aantal
M3	alles telt	102,4	31,5	279
	pluspunt	112,1	29,4	707
	rekenrijk	126,0	28,5	231
	wig	114,5	31,6	700
	wis en reken	107,4	31,9	74
	onbekend	112,6	26,9	132

Moment	Lesmethode	M	SD	Aantal
E3	alles telt	129,59	27,06	266
	pluspunt	134,13	28,55	674
	rekenrijk	150,7	27,07	106
	wig	137,72	28,33	593
	wis en reken	135,74	32,96	75
	onbekend	135,57	30,62	343

De volgende tabellen geven de gemiddelde score weer van de verschillende halfjaargroepen.

Tabel 6.9 Gemiddelde score per halfjaargroep

M3

Halfjaargroep	Aantal	M	SD
6	126	114,0	28,4
6,5	821	110,7	31,4
7	867	115,8	31,8
7,5	261	110,8	27,3
8	61	107,1	27,5

E3

Halfjaargroep	Aantal	M	SD
6,5	122	140,5	27,9
7	805	134,2	29,1
7,5	830	138,7	29,4
8	251	129,7	27,0
8,5	58	128,6	27,4

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 2 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. Van de leerlingen die in de eigen jaargroep zitten is bekend dat de oudere leerlingen een hoger gemiddelde scoren dan de jongere halfjaargroep. Dit patroon is bij beide afnamemomenten terug te vinden.

Tabel 6.10 gemiddelde score jongen-meisje

Geslacht

Moment	Geslacht	Aantal	M	SD
M3	jongen	1054	115,5	31,5
	meisje	1050	110,2	29,7

Moment	Geslacht	Aantal	M	SD
E3	jongen	1051	137,7	30,0
	meisje	1003	133,3	27,8

Per afnamemoment is in de bovenstaande tabellen de gemiddelde score van jongens en meisjes weergegeven. Uit onderzoek is bekend dat de effectgrootte rond de 0,3 is halverwege de basisschool (Hop, M. (red.) (2012). Balans van het reken-wiskunde onderwijs halverwege de basisschool 5, Arnhem, Cito). Op alle afnamemomenten scoren jongens hoger dan meisjes. In termen van effectgrootte is er sprake van geen effect (m3 0,17 en e3 0,15).

7 Samenvatting

In dit hoofdstuk geven we kort weer wat in de voorafgaande hoofdstukken is besproken.
De toetsen Rekenen-Wiskunde 3.0 voor groep 3 vormen een hulpmiddel om vast te stellen in hoeverre leerlingen rekenvaardig zijn en hoe deze rekenvaardigheid zich ontwikkelt door leerlingen te volgen. Met behulp van categorieënanalyses kan in kaart worden gebracht op welk domein leerlingen ten opzichte van hun algemene rekenvaardigheid het relatief beter of zwakker doen. We beschrijven in hoofdstuk 2 dat de inhoud van de toetsen aansluit bij de kerndoelen primair onderwijs en bij de referentieniveaus. In de domeinbeschrijving onderscheiden we voor groep 3 de onderwerpen getallen, optellen en aftrekken, vermenigvuldigen en delen en meten. Dat zijn ook de domeinen waarop bij de categorieënanalyses bij M3 en E3 gerapporteerd wordt. We geven in het 2e 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 3 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 vaardigheidsschaal rekenen-wiskunde passen kunnen we zeggen dat de papieren en digitale opgaven dezelfde vaardigheid meten.
In hoofdstuk 4 rapporteerden we over de kalibratie en normering. We beschreven de opzet, de gevolgde stappen bij de kalibratie en de toetsing van het IRT-model dat gebruikt is bij de analyses. Uit de S-toetsing kan geconcludeerd worden dat het meetinstrument en het meetmodel adequaat is om het gedrag van leerlingen te verklaren. Bovendien blijkt dat verschillen in gedrag tussen de leerlingen zijn te verklaren door een unidimensioneel concept. Uit de resultaten van de analyses met betrekking tot R1c- waarden en de constante ‘c’ trekken we de conclusie dat de kalibratie geslaagd is.
In paragraaf 4.3.2 wordt aangetoond dat de normeringssteekproef op basis van de variabelen regio, 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 0,90 en hoger, goed te noemen. Verder zijn in dit hoofdstuk betrouwbaarheidstabellen opgenomen die de betekenis van de meetnauwkeurigheid voor de beslissingen die met de toetsen genomen worden, laten zien. In het laatste hoofdstuk, hoofdstuk 6, wordt de validiteit van de toetsen behandeld. Zowel de begripsvaliditeit als de inhoudsvaliditeit komt aan bod. De inhoudsvaliditeit wordt aangetoond door te verwijzen naar de verschillende bronnen die in het Nederlands onderwijs richtinggevend zijn voor de inhoud van het rekenen-wiskunde domein. Vervolgens komt in hoofdstuk 6 de begripsvaliditeit aan bod. In eerste instantie door te verwijzen naar het unidimensionele karakter van de toets zoals dat in hoofdstuk 4 is aangetoond met de s-toetsen. Een goede modelfit wordt vervolgens aangetoond door te verwijzen naar de verhouding van de R1c ten opzichte van de vrijheidsgraden (df). Ook de gepresenteerde correlaties van de verschillende domeinen met elkaar en de verschillende domeinen met de totaalscore bieden ondersteuning voor het unidimensionale karakter van de toetsen. Als laatste in het kader van bewijsvoering van de structuur van de toets wordt in hoofdstuk 6 verwezen naar de constante c, die op een enkele uitzondering 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 rit-waarden. In hoofdstuk 6 worden gegevens gepresenteerd over DIF onderzoek. Voor wat betreft sekse is vastgesteld dat er nauwelijks sprake is van DIF.
In hoofdstuk 6 wordt uitgebreid aandacht besteed aan de soortgenootvaliditeit. Als eerste bewijsvoering voor de validiteit van de toetsen wordt de hoge correlatie van de tweede generatie toetsen (LVS toetsen) met de toetsen Rekenen-Wiskunde 3.0 uit het Cito Volgsysteem opgevoerd. Daarmee is een belangrijk bewijsstuk geleverd voor de validiteit van de toetsen Rekenen-Wiskunde 3.0. Er wordt vervolgd met de

presentatie van de correlatiegegevens van een onderzoek waarbij bij een groep leerlingen naast de toets Rekenen-Wiskunde M3 en E3 een (Cotan positief beoordeelde) toets van een andere uitgever heeft gemaakt. Ook uit dat onderzoek blijkt dat de correlatie tussen deze toets met de toetsen Rekenen-Wiskunde 3.0 toetsen M3 en E3 groep 3 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 waarbij gesteld kan worden dat de leerlingen van de methode Rekenrijk het beter doen dan de andere methoden. 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. Alhoewel jongens, zoals verwacht, hoger scoren dan meisjes, is dit verschil in termen van effectgrootte er niet.
Al met al kunnen we concluderen dat de validiteit van de toetsen Rekenen-Wiskunde 3.0 goed te noemen is.

Literatuur

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

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

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

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

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

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

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

Cito (2005). *Leerling- en onderwijsvolgsysteem, Rekenen-Wiskunde, groep 3.* Arnhem: Cito*.*

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

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

Cito (2002). *Rekenen hulpboek groep 3 medio. Arnhem*: Cito.

Cito (2002). *Rekenen hulpboek groep 3 eind.* 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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.* Copenhage: Danish Institute for Educational Research.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

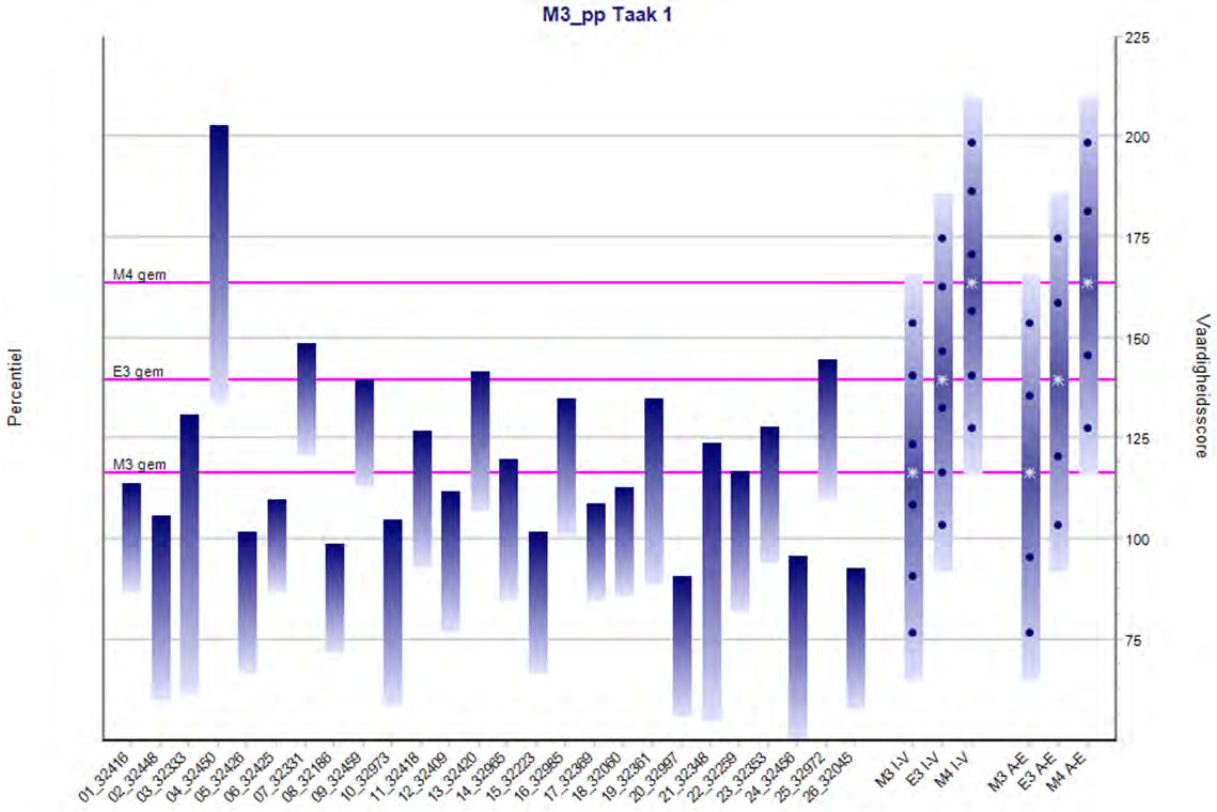
Verhelst, N.D., Verstralen, H.H.F.M., & Eggen, T.H.J.M. (1991). *Finding starting values for the item parameters and suitable discrimination indices in the one-parameter logistic model.* Measurement and Research Department Reports 91-10. Arnhem: Cito.

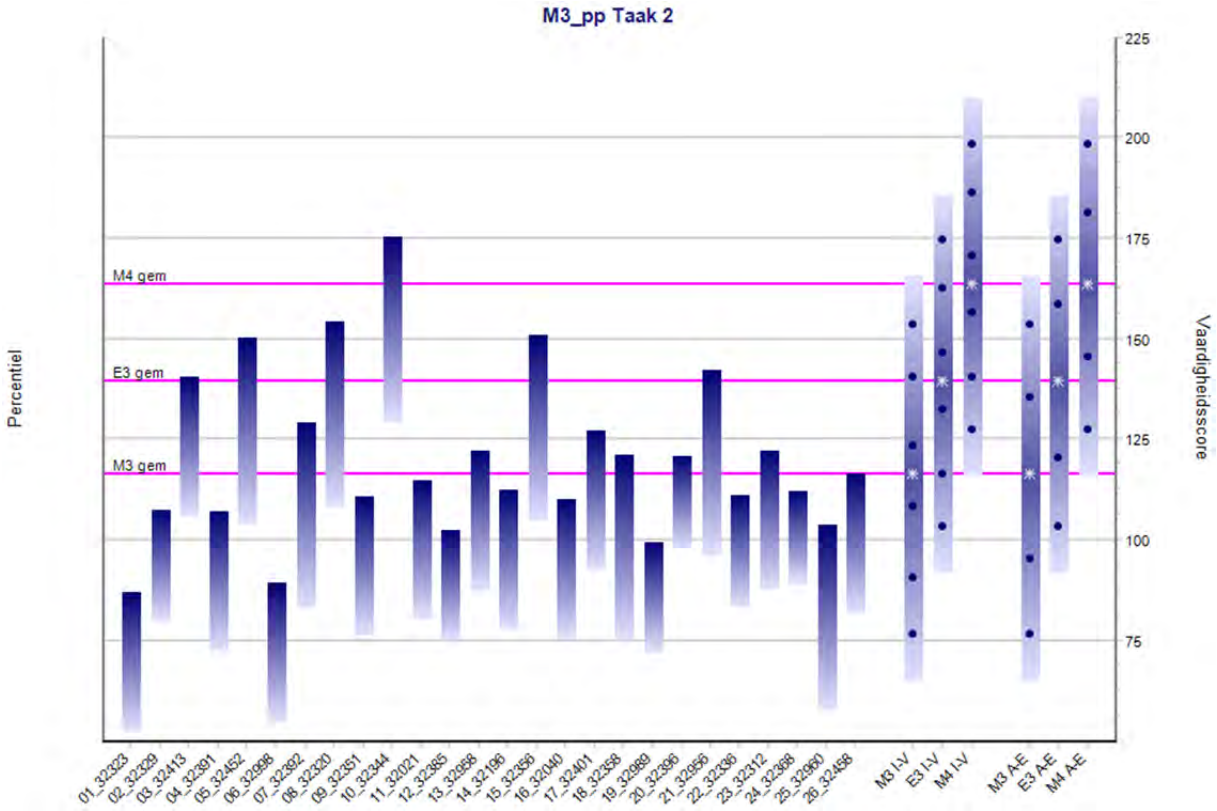
Verhelst, N., Eggen, T. (1989). *Psychometrische aspecten van peilingsonderzoek* (PPON-rapport, nr 4). Arnhem: Cito.

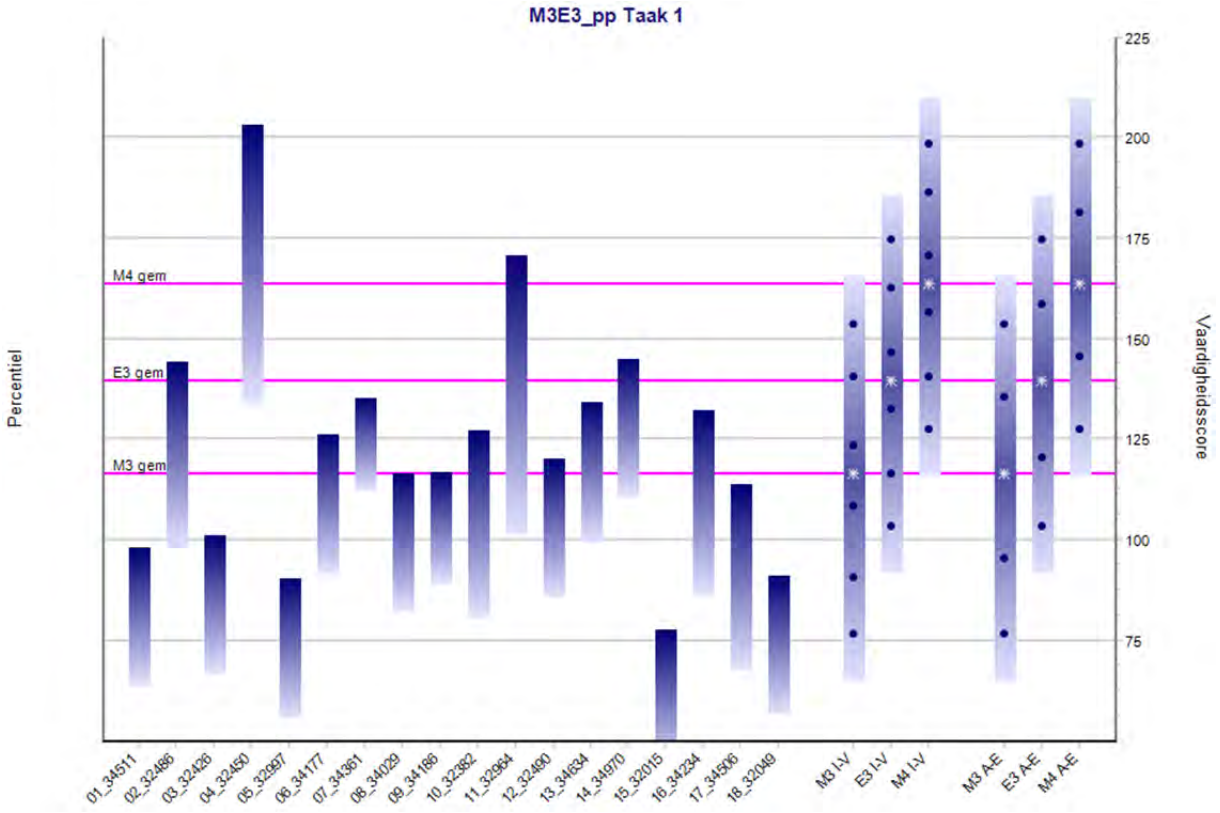
Verstralen, H.H.F.M. (1997*). OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem; Cito

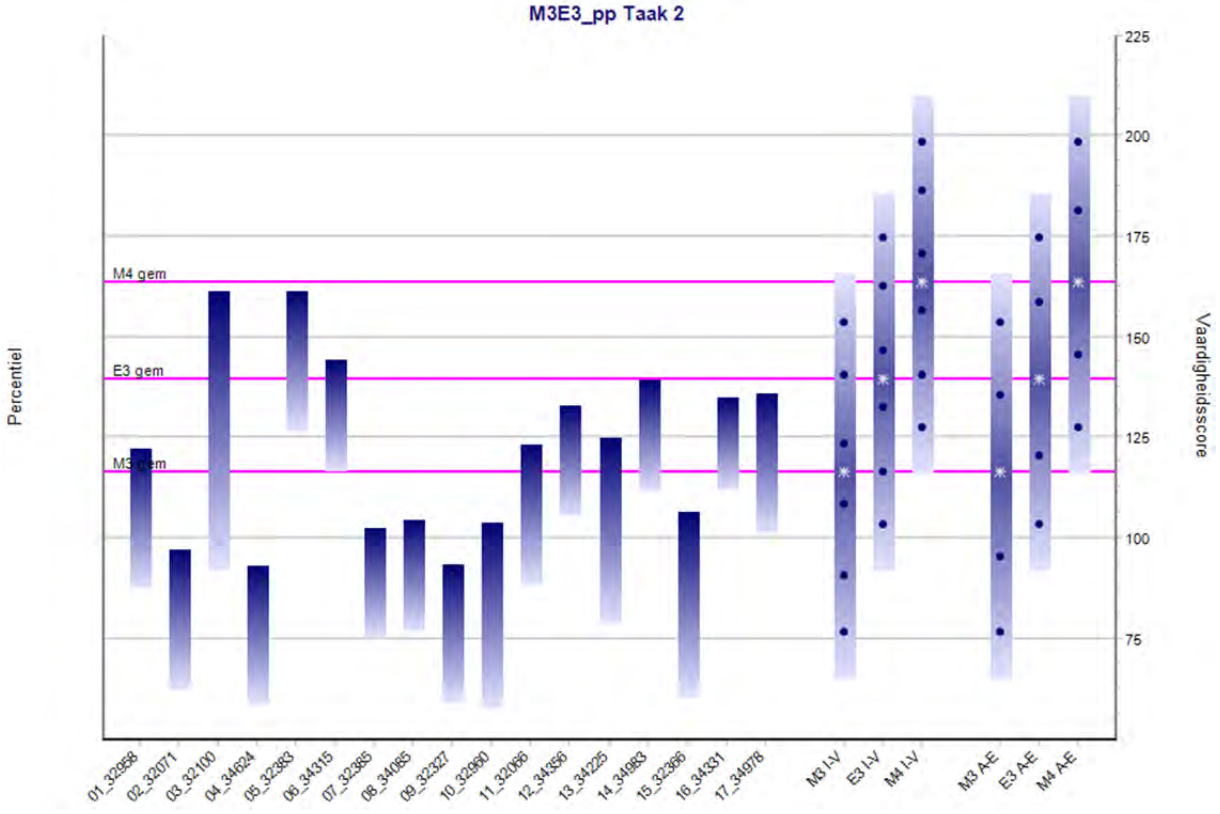
Bijlagen

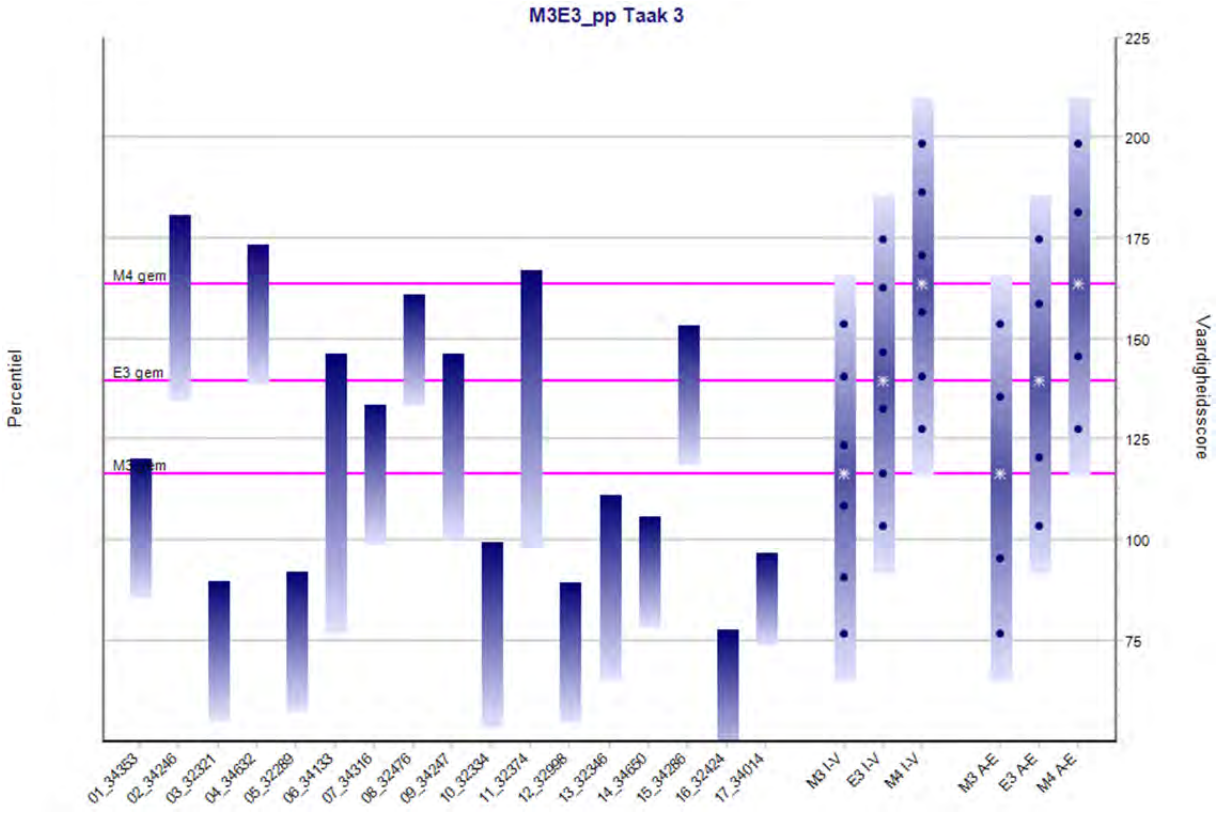
Bijlage 1 p50 en p80-kanspunten van de opgaven in de papieren toetsen en digitale toetsen M3, M3E3 en E3 in relatie tot de vaardigheidsverdelingen van M3, E3 en M4

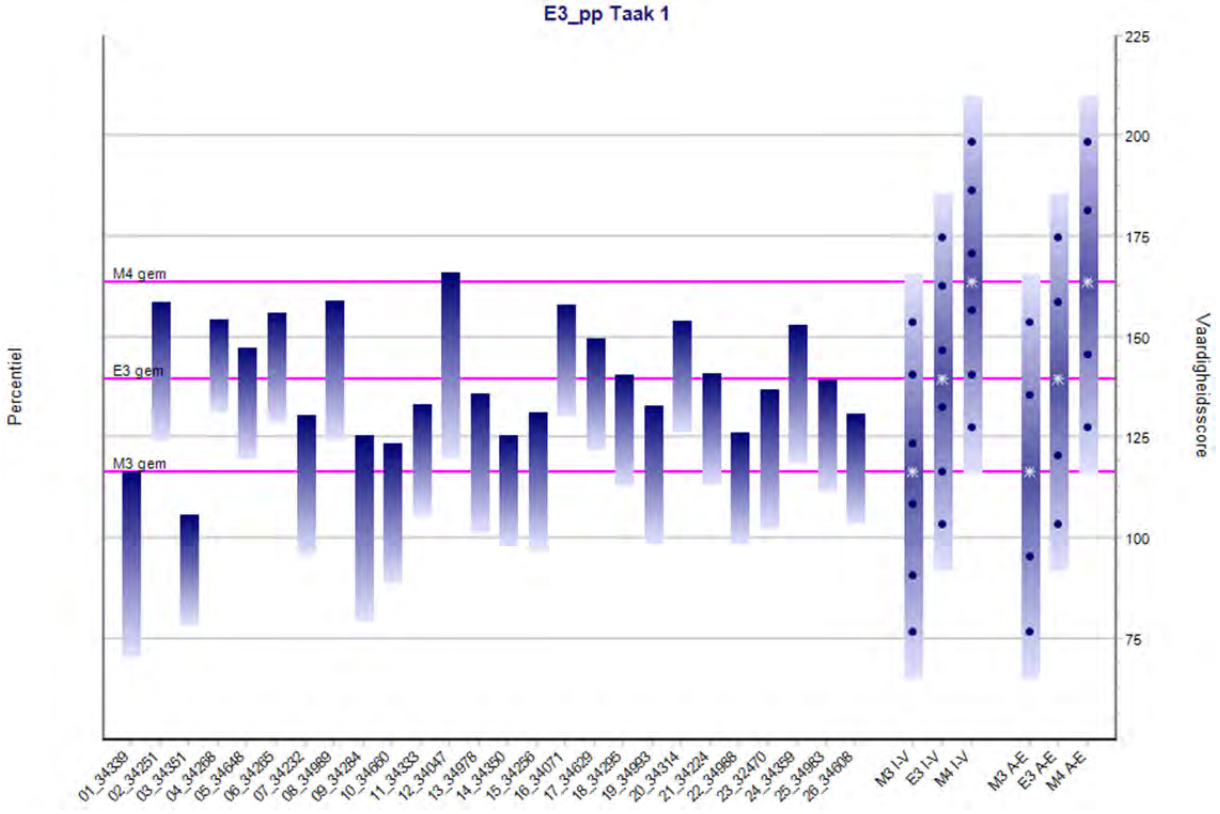


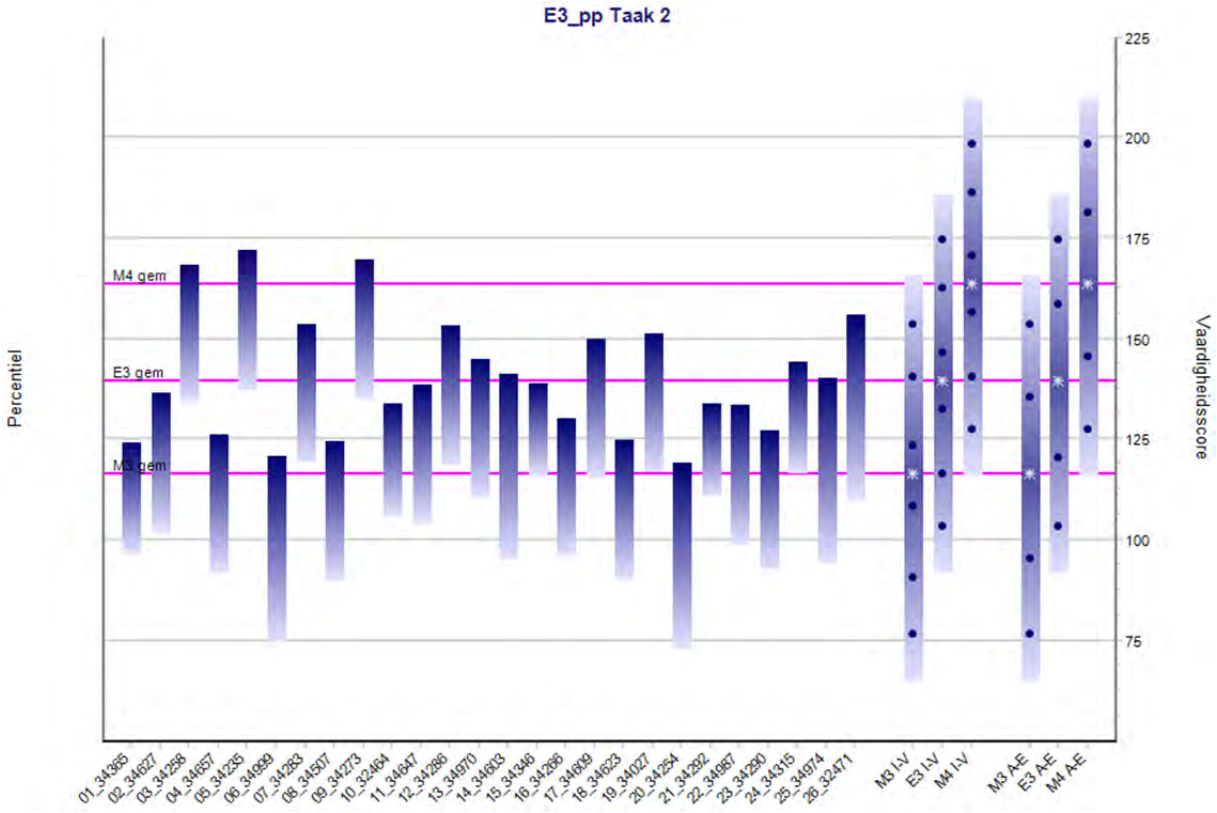


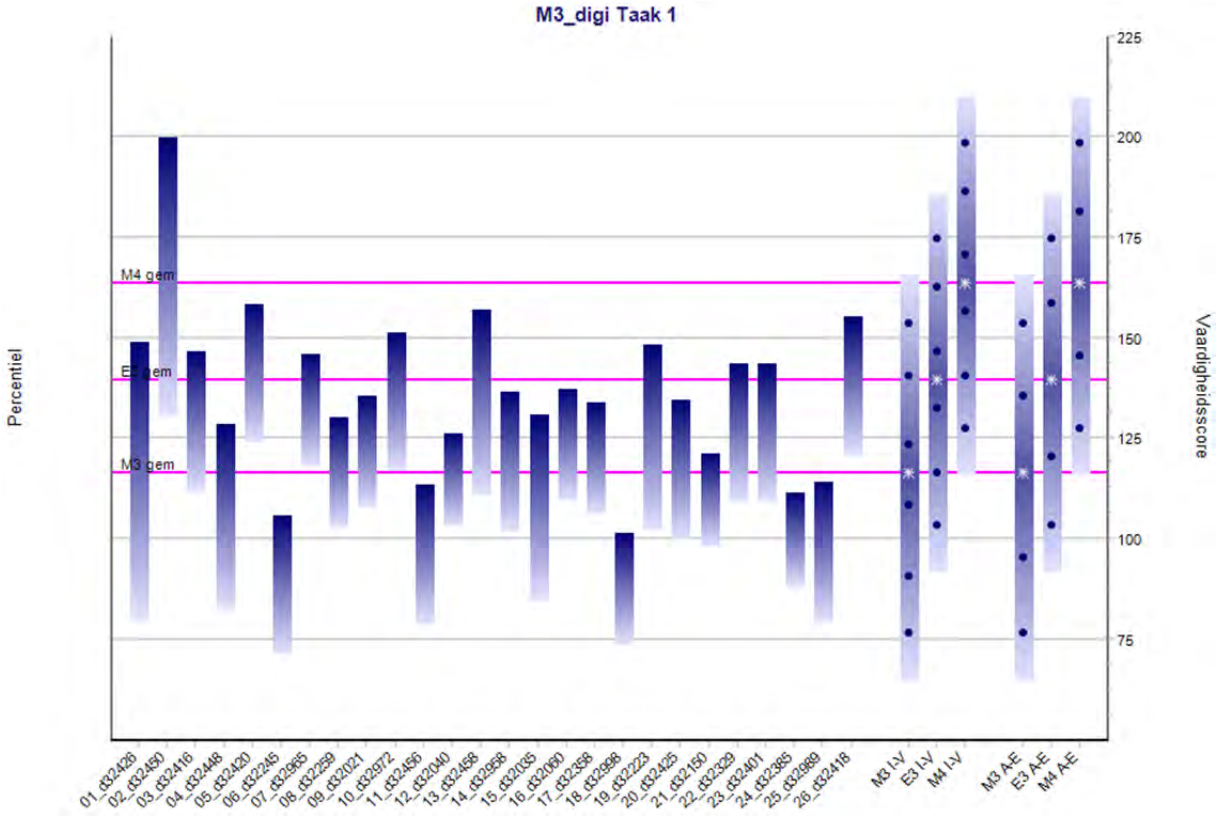

M3E3_pp Taak 1
Percentiel
Vaardigheidsscore
225
200
175
150
125
100
75
M4 gem
E3 gem
M3 gem
01_34511
02_32486
03_32426
04_32450
05_32997
06_34177
07_34361
08_34029
09_34186
10_32382
11_32964
12_32490
13_34634
14_34970
15_32015
16_34234
17_34506
18_32049
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

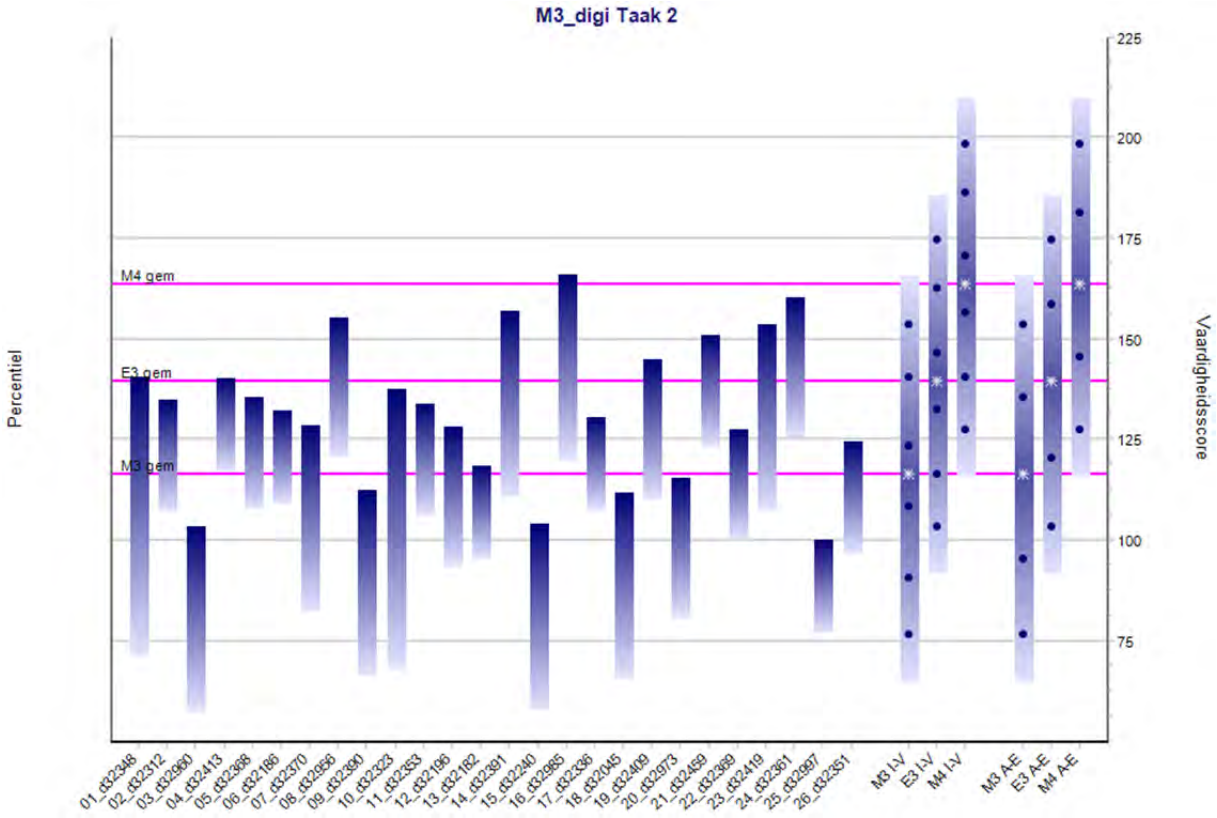

M3E3_pp Taak 2
Percentiel
Vaardigheidsscore
225
200
175
150
125
100
75
M4 gem
E3 gem
M3 gem
01_32868
02_32071
03_32100
04_34624
05_32383
06_34315
07_32385
08_34085
09_32327
10_32960
11_32066
12_34356
13_34225
14_34983
15_32366
16_34331
17_34978
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

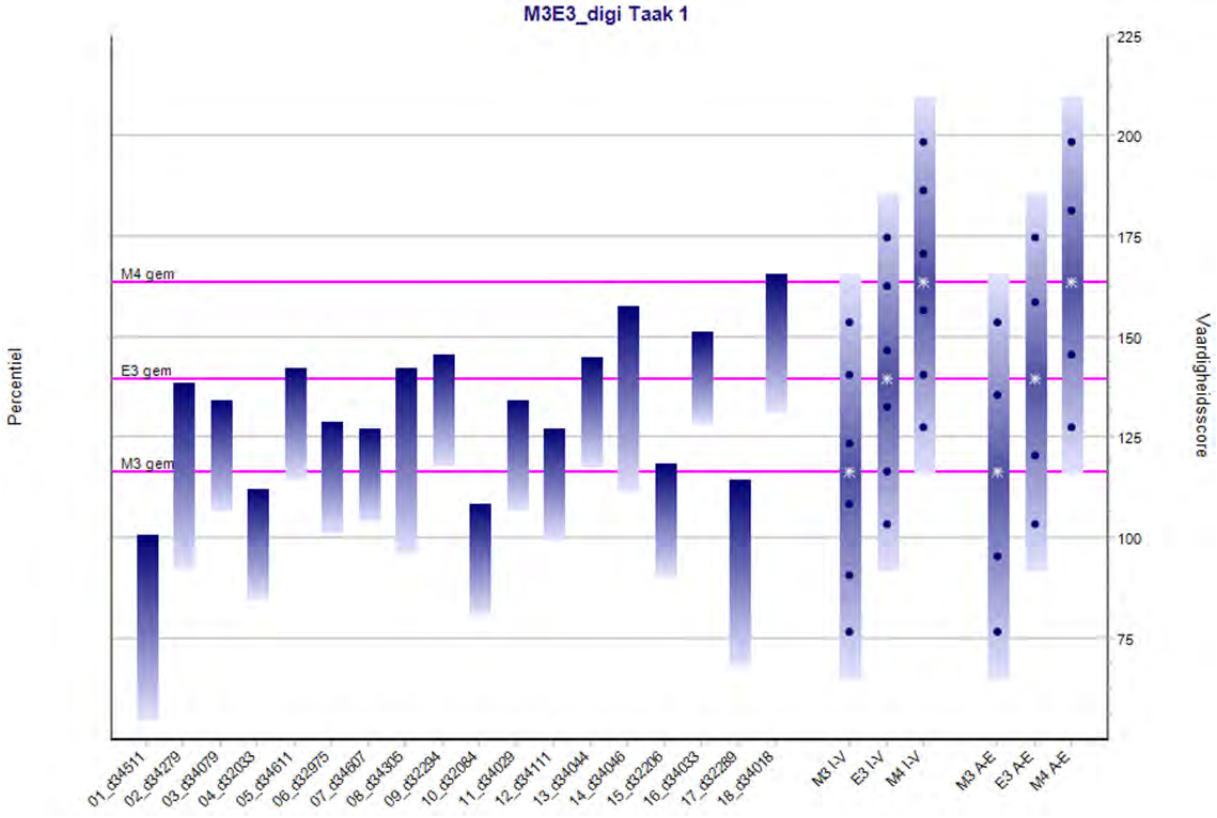

M3E3_pp Taak 3
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_34353
02_34246
03_32321
04_34632
05_32289
06_34133
07_34316
08_32476
09_34247
10_32334
11_32374
12_32998
13_32346
14_34650
15_34286
16_32424
17_34014
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

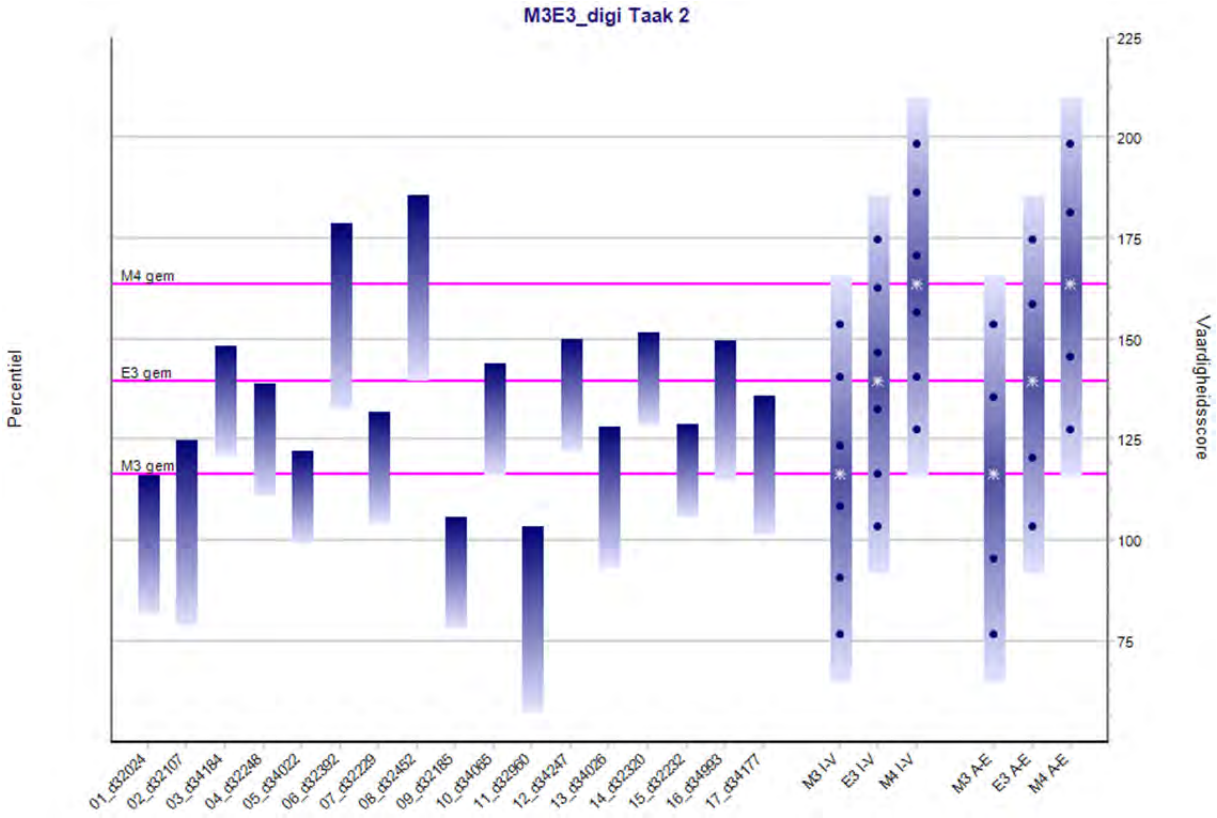

E3_pp Taak 1
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_34339
02_34251
03_34351
04_34288
05_34648
06_34285
07_34232
08_34989
09_34284
10_34660
11_34333
12_34047
13_34978
14_34350
15_34256
16_34071
17_34629
18_34295
19_34983
20_34314
21_34224
22_34988
23_32470
24_34359
25_34983
26_34608
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

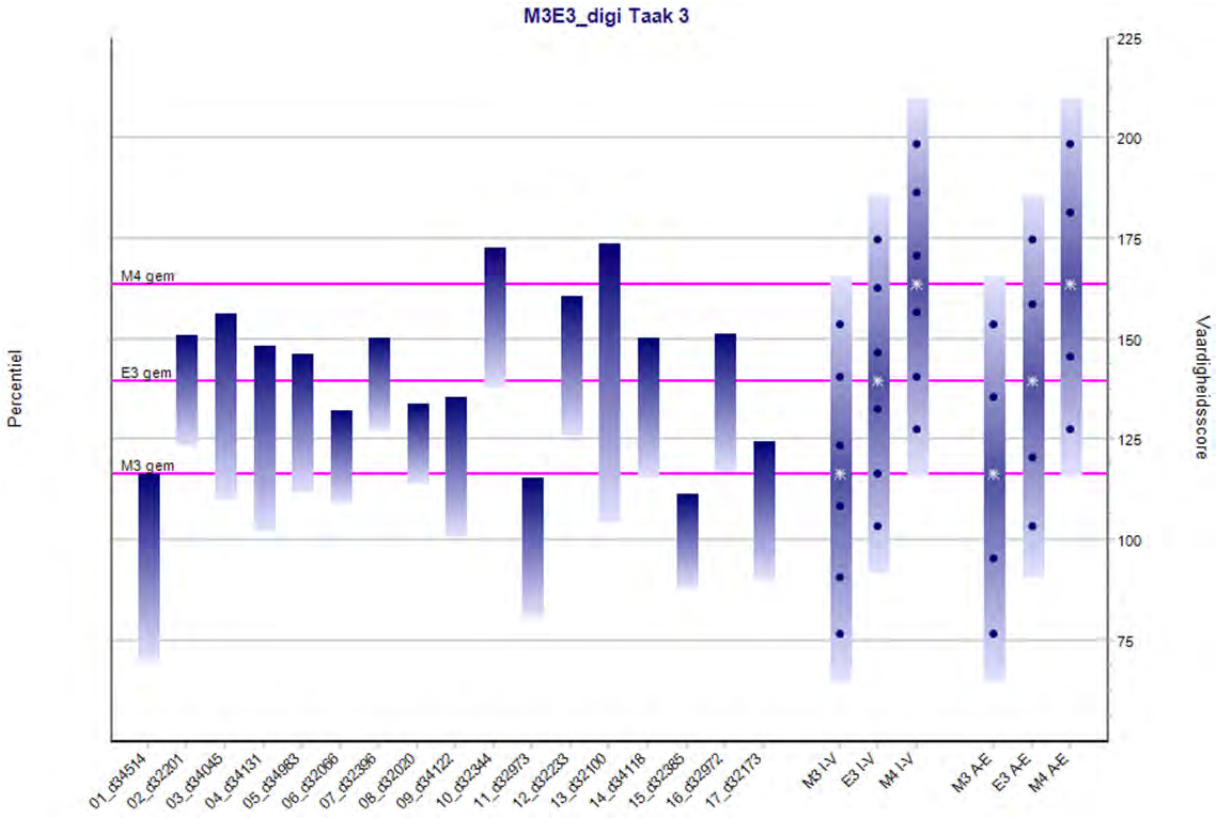

E3_pp Taak 2
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_34365
02_34627
03_34258
04_34657
05_34235
06_34999
07_34283
08_34507
09_34273
10_32464
11_34647
12_34286
13_34970
14_34603
15_34346
16_34286
17_34909
18_34623
19_34027
20_34254
21_34292
22_34987
23_34290
24_34315
25_34974
26_32471
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

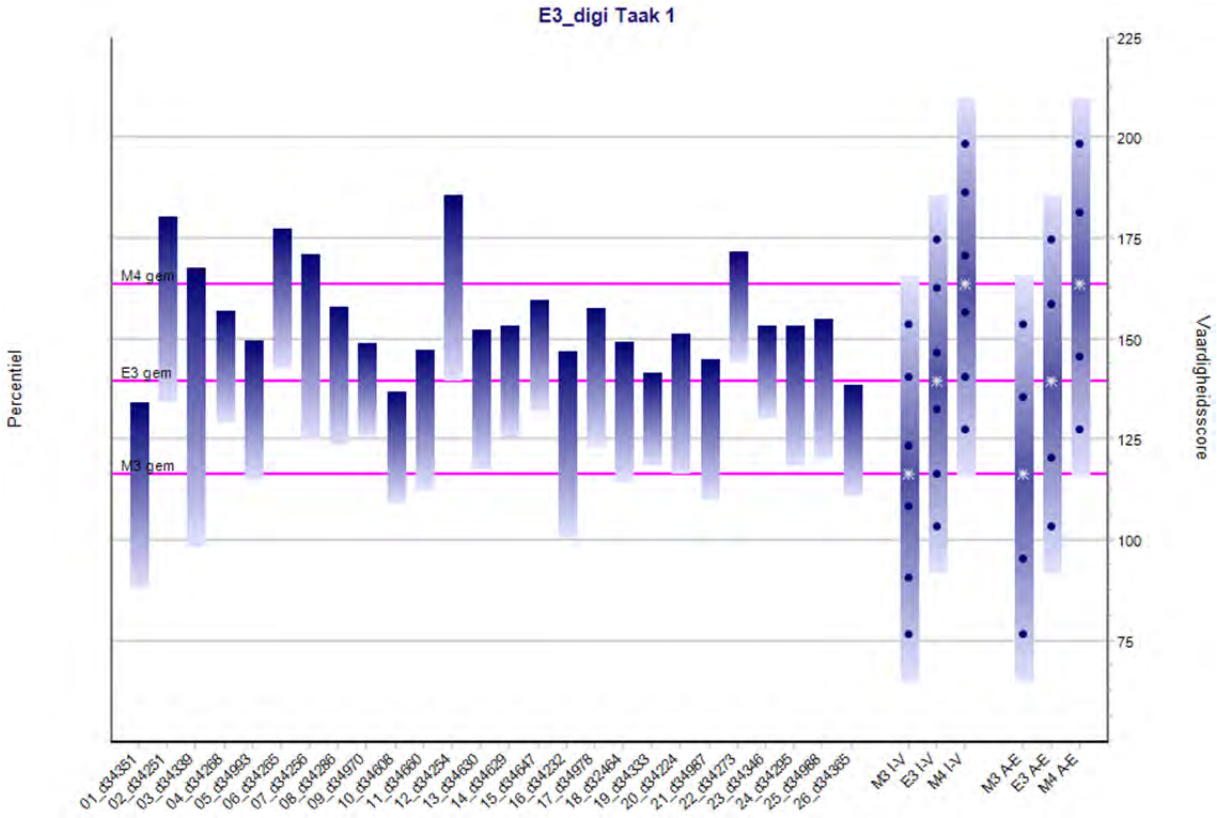

M3_digi Taak 1
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_d32426
02_d32450
03_d32416
04_d32448
05_d32420
06_d32245
07_d32965
08_d32259
09_d32021
10_d32972
11_d32456
12_d32040
13_d32458
14_d32958
15_d32035
16_d32060
17_d32358
18_d32998
19_d32223
20_d32425
21_d32150
22_d32329
23_d32401
24_d32385
25_d32989
26_d32418
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

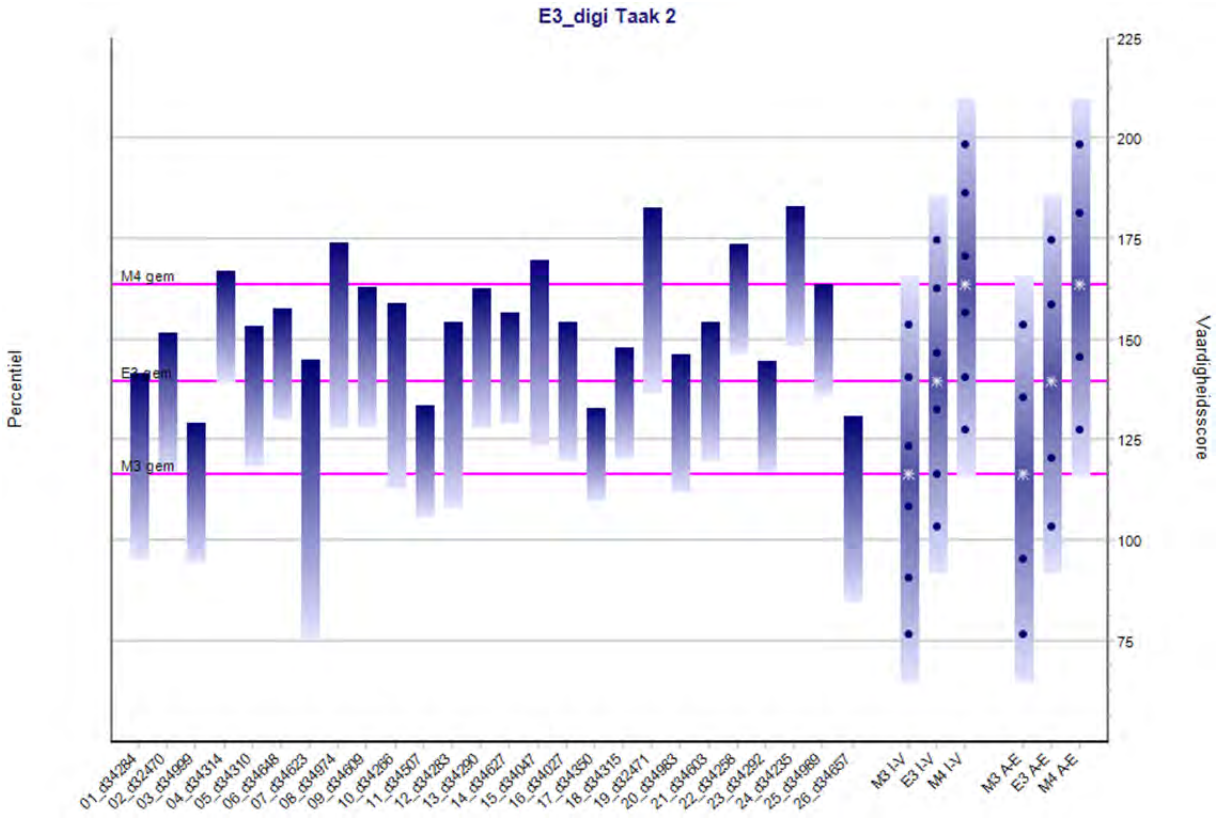

M3_digi Taak 2
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_d32348
02_d32312
03_d32660
04_d32413
05_d32368
06_d32186
07_d32370
08_d32656
09_d32390
10_d32323
11_d32353
12_d32196
13_d32182
14_d32391
15_d32240
16_d32685
17_d32336
18_d32045
19_d32409
20_d32973
21_d32459
22_d32369
23_d32419
24_d32361
25_d32997
26_d32351
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E


M3E3_digi Taak 1
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_d34511
02_d34279
03_d34079
04_d32033
05_d34611
06_d32975
07_d34607
08_d34305
09_d32294
10_d32064
11_d34029
12_d34111
13_d34044
14_d34046
15_d32206
16_d34033
17_d32289
18_d34018
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E


M3E3_digi Taak 2
Percentiel
Vaardigheidsscore
225
200
175
150
125
100
75
M4 gem
E3 gem
M3 gem
01_d32024
02_d32107
03_d34184
04_d32248
05_d34022
06_d32392
07_d32229
08_d32452
09_d32185
10_d34065
11_d32960
12_d34247
13_d34026
14_d32320
15_d32232
16_d34993
17_d34177
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E


M3E3_digi Taak 3
Percentiel
Vaardigheidsscore
225
200
175
150
125
100
75
M4 gem
E3 gem
M3 gem
01_d34514
02_d32201
03_d34045
04_d34131
05_d34983
06_d32066
07_d32396
08_d32020
09_d34122
10_d32344
11_d32973
12_d32233
13_d32100
14_d34118
15_d32365
16_d32972
17_d32173
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E


E3_digi Taak 1
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_d34351
02_d34251
03_d34339
04_d34268
05_d34993
06_d34265
07_d34256
08_d34286
09_d34970
10_d34606
11_d34660
12_d34254
13_d34630
14_d34629
15_d34647
16_d34232
17_d34978
18_d32464
19_d34333
20_d34224
21_d34987
22_d34273
23_d34346
24_d34295
25_d34988
26_d34365
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E


E3_digi Taak 2
Percentiel
Vaardigheidsscore
M4 gem
E3 gem
M3 gem
225
200
175
150
125
100
75
01_d34284
02_d32470
03_d34999
04_d34314
05_d34310
06_d34648
07_d34623
08_d34974
09_d34609
10_d34266
11_d34507
12_d34283
13_d34290
14_d34627
15_d34047
16_d34027
17_d34350
18_d34315
19_d32471
20_d34983
21_d34603
22_d34258
23_d34292
24_d34235
25_d34989
26_d34657
M3 I-V
E3 I-V
M4 I-V
M3 A-E
E3 A-E
M4 A-E

Bijlage 2 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek M3

M3			Taak1	Taak2	Taak3	Taak4	Taak5	Taak6	Taak7	Taak8	Taak9
	Aantal items	N	23	24	24	24	24	23	24	24	24
e2	get	2	2								
	get	2		2							
	get	2			2						
	opaf	2				2					
	opaf	2					2				
	opaf	2						2			
	mtg	2							2		
	mtg	2								2	
	mtg	2									2
m3	get	2	2	2							
	get	3					3	3			
	get	2							2		2
	get	3	3						3		
	get	2			2			2			
	get	2		2	2						
	get	2				2					2
	get	3		3						3	
	get	3				3			3		
	get	2					2			2	
	get	2					2				2
	opaf	2		2				2			
	opaf	2						2		2	
	opaf	2	2			2					
	opaf	3			3					3	
	opaf	2			2				2		
	opaf	2	2		2						
	opaf	2					2		2		
	opaf	2	2				2				
	opaf	3		3							3
	verdel	3	3							3	
	verdel	3		3		3					
	verdel	2						2	2		
	verdel	3						3			3
	verdel	3				3	3				
	mtg	2			2	2					
	mtg	2		2			2				
	mtg	2				2				2	
	mtg	3			3						3
	mtg	3		3					3		
	kaal	2	2								2
	kaal	2	2					2			
	kaal	2								2	2
	kaal	2				2		2			
	kaal	2							2	2	
	kaal	2			2		2				
e3	get	3				3					
	get	3								3	
	opaf	3							3		
	opaf	3									3
	verdel	2			2						
	verdel	2					2				
	verdel	2					2				
	mtg	3	3								
	mtg	3						3			
	kaal	2		2							
	kaal	2			2						

Legenda:
get: getallen
opaf: optellen en aftrekken
verdel: vermenigvuldigen en delen
mtg: meten, tijd en geld
kaal: sommen zonder context als 7+8=

Bijlage 3 Overzicht samenstelling nieuwe taken voor kalibratie en normeringsonderzoek E3

E3			Taak1	Taak2	Taak3	Taak4	Taak5	Taak6	Taak7	Taak8	Taak9
	Aantal items	N	24	24	25	25	25	25	26	24	24
E3	get	4								4	4
	get	4		4				4			
	get	4				4	4				
	get	4			4				4		
	get	4		4		4					
	get	4							4	4	
	get	2	2						2		
	opaf	4	4							4	
	opaf	4		4							4
	opaf	4			4	4					
	opaf	4					4	4			
	opaf	4							4		4
	verdel	4		4						4	
	verdel	4	4				4				
	verdel	3			3						3
	verdel	3				3		3			
	mtg	4						4	4		
	mtg	4				4	4				
	mtg	4	4		4						
	kaal	4	4				4				
	kaal	4			4			4			
	kaal	2		2						2	
m3	get	3	3								
	get	3			3						
	opaf	3						3			
	opaf	3				3					
	verdel	3							3		
	verdel	3									3
	mtg	3		3							
	mtg	3								3	
	kaal	2							2		
	kaal	2					2				
m4	get	3						3			
	get	3					3				
	opaf	3								3	
	opaf	3	3								
	verdel	3							3		
	verdel	3			3						
	mtg	3		3							
	mtg	3									3
	kaal	3									3
	kaal	3				3					

Legenda:
get: getallen
opaf: optellen en aftrekken
verdel: vermenigvuldigen en delen
mtg: meten, tijd en geld
kaal: sommen zonder context als 7+8=

Bijlage 4 Klassieke en IRT-indices van de opgaven in de M3, M3E3, E3 papieren en digitale toetsen

M3 papier

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,73	0,51	5	-0,13	3,55
2	0,81	0,33	3	-0,40	1,25
3	0,73	0,26	2	-0,38	0,74
4	0,41	0,27	2	0,34	0,90
5	0,82	0,41	4	-0,33	1,93
6	0,75	0,56	6	-0,13	4,51
7	0,45	0,50	5	0,21	4,33
8	0,83	0,47	5	-0,28	2,70
9	0,52	0,52	5	0,13	4,36
10	0,81	0,33	3	-0,41	1,22
11	0,67	0,46	4	-0,08	2,77
12	0,77	0,43	4	-0,23	2,30
13	0,56	0,46	4	0,07	3,04
14	0,72	0,45	4	-0,15	2,57
15	0,82	0,41	4	-0,33	1,94
16	0,61	0,46	4	0,01	2,95
17	0,76	0,55	6	-0,15	4,38
18	0,74	0,51	5	-0,15	3,49
19	0,66	0,37	3	-0,11	1,72
20	0,87	0,37	4	-0,44	1,54
21	0,75	0,25	2	-0,45	0,70
22	0,74	0,44	4	-0,18	2,46
23	0,66	0,46	4	-0,07	2,80
24	0,84	0,31	3	-0,50	1,06
25	0,54	0,46	4	0,10	3,06
26	0,86	0,38	4	-0,42	1,61
27	0,88	0,36	4	-0,48	1,42
28	0,78	0,50	5	-0,20	3,19
29	0,57	0,46	4	0,06	3,03
30	0,79	0,42	4	-0,27	2,14
31	0,57	0,38	3	0,04	1,87
32	0,87	0,37	4	-0,45	1,51
33	0,69	0,37	3	-0,17	1,65
34	0,55	0,38	3	0,08	1,89
35	0,77	0,43	4	-0,24	2,28
36	0,41	0,36	3	0,30	1,85
37	0,75	0,44	4	-0,20	2,41
38	0,81	0,48	5	-0,25	2,90
39	0,70	0,45	4	-0,12	2,64
40	0,76	0,44	4	-0,22	2,33
41	0,57	0,38	3	0,05	1,88
42	0,78	0,43	4	-0,24	2,25
43	0,67	0,46	4	-0,07	2,78
44	0,74	0,36	3	-0,25	1,52
45	0,83	0,47	5	-0,28	2,73
46	0,66	0,58	6	-0,02	5,23
47	0,62	0,38	3	-0,04	1,81
48	0,76	0,50	5	-0,17	3,38
49	0,70	0,45	4	-0,12	2,65
50	0,73	0,56	6	-0,11	4,67
51	0,82	0,33	3	-0,42	1,20
52	0,74	0,44	4	-0,18	2,46

M3E3 papier

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,93	0,29	4	-0,36	0,99
2	0,75	0,33	3	-0,02	1,51
3	0,92	0,30	4	-0,33	1,08
4	0,52	0,25	2	0,34	0,93
5	0,94	0,27	4	-0,44	0,79
6	0,82	0,38	4	-0,08	1,97
7	0,75	0,53	6	0,12	4,72
8	0,87	0,35	4	-0,18	1,60
9	0,87	0,42	5	-0,11	2,30
10	0,82	0,30	3	-0,19	1,19
11	0,67	0,25	2	0,02	0,83
12	0,85	0,36	4	-0,14	1,74
13	0,78	0,40	4	0,00	2,27
14	0,72	0,42	4	0,11	2,66
15	0,96	0,22	4	-0,57	0,52
16	0,80	0,31	3	-0,14	1,29
17	0,87	0,28	3	-0,32	0,94
18	0,94	0,27	4	-0,43	0,80
19	0,84	0,37	4	-0,12	1,80
20	0,93	0,29	4	-0,37	0,96
21	0,71	0,24	2	-0,08	0,78
22	0,94	0,28	4	-0,41	0,86
23	0,60	0,42	4	0,27	3,07
24	0,70	0,48	5	0,17	3,94
25	0,92	0,36	5	-0,25	1,52
26	0,92	0,37	5	-0,23	1,61
27	0,94	0,28	4	-0,41	0,87
28	0,90	0,25	3	-0,42	0,77
29	0,84	0,37	4	-0,11	1,84
30	0,78	0,47	5	0,06	3,32
31	0,83	0,30	3	-0,21	1,15
32	0,73	0,48	5	0,12	3,68
33	0,89	0,26	3	-0,40	0,81
34	0,75	0,53	6	0,12	4,70
35	0,77	0,40	4	0,01	2,33
36	0,85	0,36	4	-0,14	1,73
37	0,53	0,34	3	0,35	1,94
38	0,95	0,26	4	-0,45	0,77
39	0,50	0,41	4	0,39	3,18
40	0,94	0,27	4	-0,43	0,83
41	0,76	0,23	2	-0,23	0,69
42	0,79	0,40	4	-0,01	2,24
43	0,55	0,48	5	0,34	4,50
44	0,74	0,33	3	0,00	1,55
45	0,91	0,24	3	-0,46	0,70
46	0,68	0,25	2	-0,02	0,81
47	0,95	0,26	4	-0,45	0,77
48	0,88	0,27	3	-0,35	0,89
49	0,91	0,37	5	-0,22	1,68
50	0,66	0,42	4	0,19	2,90
51	0,96	0,22	4	-0,57	0,53
52	0,95	0,37	6	-0,26	1,55

E3 papier

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,86	0,28	3	-0,29	0,99
2	0,62	0,43	4	0,24	3,03
3	0,91	0,36	5	-0,22	1,69
4	0,58	0,55	6	0,32	5,88
5	0,67	0,50	5	0,20	4,09
6	0,60	0,50	5	0,29	4,41
7	0,80	0,39	4	-0,04	2,13
8	0,62	0,44	4	0,25	3,03
9	0,83	0,30	3	-0,20	1,16
10	0,84	0,37	4	-0,11	1,86
11	0,78	0,47	5	0,06	3,33
12	0,62	0,35	3	0,20	1,84
13	0,77	0,40	4	0,01	2,33
14	0,82	0,45	5	-0,02	2,85
15	0,80	0,39	4	-0,03	2,16
16	0,58	0,50	5	0,31	4,45
17	0,65	0,50	5	0,22	4,19
18	0,72	0,49	5	0,13	3,76
19	0,79	0,40	4	-0,01	2,23
20	0,62	0,50	5	0,26	4,34
21	0,72	0,49	5	0,13	3,78
22	0,82	0,45	5	-0,01	2,90
23	0,77	0,41	4	0,02	2,37
24	0,66	0,43	4	0,19	2,90
25	0,73	0,48	5	0,12	3,68
26	0,79	0,46	5	0,04	3,20
27	0,83	0,44	5	-0,03	2,77
28	0,77	0,41	4	0,02	2,36
29	0,54	0,43	4	0,34	3,16
30	0,82	0,38	4	-0,08	1,97
31	0,51	0,43	4	0,38	3,18
32	0,85	0,29	3	-0,25	1,07
33	0,65	0,43	4	0,19	2,92
34	0,83	0,37	4	-0,10	1,90
35	0,53	0,43	4	0,35	3,17
36	0,77	0,47	5	0,06	3,36
37	0,76	0,41	4	0,04	2,44
38	0,66	0,43	4	0,19	2,90
39	0,72	0,42	4	0,11	2,66
40	0,76	0,33	3	-0,05	1,46
41	0,72	0,54	6	0,16	5,02
42	0,80	0,39	4	-0,04	2,12
43	0,68	0,43	4	0,16	2,82
44	0,83	0,38	4	-0,10	1,91
45	0,67	0,43	4	0,17	2,85
46	0,85	0,29	3	-0,27	1,04
47	0,76	0,53	6	0,11	4,62
48	0,79	0,40	4	-0,01	2,24
49	0,82	0,38	4	-0,07	2,01
50	0,70	0,49	5	0,17	3,95
51	0,77	0,33	3	-0,06	1,44
52	0,68	0,35	3	0,10	1,71

M3 digitaal

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,62	0,47	4	0,00	2,94
2	0,53	0,39	3	0,11	1,90
3	0,69	0,27	2	-0,28	0,79
4	0,54	0,47	4	0,11	3,07
5	0,60	0,47	4	0,02	2,97
6	0,47	0,53	5	0,19	4,36
7	0,55	0,47	4	0,09	3,06
8	0,61	0,59	6	0,04	5,49
9	0,56	0,54	5	0,08	4,31
10	0,63	0,53	5	0,00	4,12
11	0,42	0,46	4	0,26	3,01
12	0,48	0,58	6	0,18	5,71
13	0,82	0,51	6	-0,23	3,74
14	0,80	0,41	4	-0,28	2,10
15	0,56	0,54	5	0,08	4,31
16	0,55	0,39	3	0,08	1,89
17	0,55	0,54	5	0,10	4,34
18	0,43	0,46	4	0,24	3,03
19	0,74	0,43	4	-0,19	2,43
20	0,82	0,32	3	-0,43	1,20
21	0,57	0,54	5	0,07	4,30
22	0,66	0,53	5	-0,03	4,02
23	0,55	0,47	4	0,09	3,06
24	0,58	0,38	3	0,03	1,86
25	0,58	0,54	5	0,07	4,28
26	0,78	0,34	3	-0,34	1,35
27	0,70	0,37	3	-0,17	1,64
28	0,53	0,39	3	0,11	1,90
29	0,66	0,46	4	-0,06	2,81
30	0,82	0,46	5	-0,26	2,85
31	0,56	0,59	6	0,10	5,66
32	0,58	0,54	5	0,06	4,28
33	0,61	0,53	5	0,03	4,20
34	0,69	0,37	3	-0,15	1,67
35	0,74	0,55	6	-0,12	4,61
36	0,66	0,27	2	-0,20	0,83
37	0,81	0,32	3	-0,42	1,21
38	0,70	0,26	2	-0,32	0,78
39	0,78	0,34	3	-0,34	1,36
40	0,43	0,28	2	0,31	0,90
41	0,52	0,47	4	0,12	3,08
42	0,46	0,46	4	0,21	3,06
43	0,76	0,43	4	-0,21	2,36
44	0,46	0,46	4	0,21	3,06
45	0,47	0,38	3	0,20	1,90
46	0,68	0,57	6	-0,04	5,10
47	0,57	0,59	6	0,08	5,62
48	0,70	0,37	3	-0,17	1,64
49	0,66	0,58	6	-0,02	5,25
50	0,75	0,43	4	-0,20	2,39
51	0,49	0,47	4	0,17	3,08
52	0,43	0,52	5	0,24	4,30

M3E3 digitaal

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,51	0,43	4	0,38	2,77
2	0,71	0,49	5	0,15	4,37
3	0,77	0,47	5	0,07	4,29
4	0,80	0,46	5	0,01	4,16
5	0,71	0,42	4	0,12	3,07
6	0,50	0,35	3	0,40	1,75
7	0,77	0,53	6	0,10	5,66
8	0,69	0,49	5	0,18	4,36
9	0,66	0,25	2	0,05	0,91
10	0,84	0,49	6	-0,01	5,31
11	0,82	0,39	4	-0,06	2,80
12	0,77	0,32	3	-0,07	1,77
13	0,66	0,50	5	0,21	4,33
14	0,78	0,47	5	0,05	4,24
15	0,77	0,41	4	0,01	2,96
16	0,87	0,28	3	-0,32	1,40
17	0,67	0,43	4	0,17	3,08
18	0,72	0,34	3	0,03	1,86
19	0,91	0,37	5	-0,22	3,10
20	0,62	0,55	6	0,28	5,47
21	0,61	0,55	6	0,29	5,43
22	0,86	0,42	5	-0,09	3,75
23	0,67	0,35	3	0,12	1,90
24	0,56	0,43	4	0,31	2,92
25	0,81	0,51	6	0,05	5,52
26	0,89	0,40	5	-0,15	3,45
27	0,65	0,50	5	0,23	4,31
28	0,74	0,48	5	0,11	4,35
29	0,69	0,49	5	0,18	4,36
30	0,91	0,24	3	-0,45	1,15
31	0,68	0,35	3	0,11	1,90
32	0,90	0,25	3	-0,43	1,20
33	0,87	0,35	4	-0,18	2,46
34	0,54	0,35	3	0,33	1,82
35	0,70	0,49	5	0,16	4,37
36	0,76	0,33	3	-0,04	1,81
37	0,68	0,43	4	0,16	3,08
38	0,86	0,28	3	-0,30	1,43
39	0,90	0,44	6	-0,12	4,61
40	0,60	0,43	4	0,26	3,01
41	0,87	0,35	4	-0,19	2,43
42	0,90	0,38	5	-0,19	3,24
43	0,77	0,47	5	0,07	4,29
44	0,83	0,38	4	-0,10	2,71
45	0,75	0,58	7	0,14	7,10
46	0,83	0,30	3	-0,21	1,58
47	0,80	0,52	6	0,06	5,57
48	0,81	0,45	5	0,00	4,11
49	0,77	0,41	4	0,02	2,97
50	0,64	0,50	5	0,24	4,30
51	0,68	0,43	4	0,15	3,08
52	0,60	0,55	6	0,29	5,41

E3 digitaal

InBk	P-Val	RIT	Dsc	Beta	Info
1	0,71	0,42	4	0,12	3,07
2	0,60	0,36	3	0,24	1,88
3	0,53	0,51	5	0,36	3,91
4	0,69	0,49	5	0,17	4,37
5	0,65	0,44	4	0,20	3,07
6	0,57	0,51	5	0,32	4,06
7	0,77	0,23	2	-0,24	0,82
8	0,81	0,31	3	-0,15	1,67
9	0,59	0,44	4	0,29	2,97
10	0,47	0,44	4	0,43	2,64
11	0,50	0,50	5	0,40	3,76
12	0,58	0,51	5	0,30	4,13
13	0,67	0,43	4	0,18	3,08
14	0,52	0,36	3	0,37	1,78
15	0,53	0,36	3	0,35	1,81
16	0,57	0,36	3	0,28	1,86
17	0,75	0,47	5	0,09	4,33
18	0,68	0,43	4	0,15	3,08
19	0,44	0,49	5	0,46	3,42
20	0,43	0,43	4	0,49	2,46
21	0,74	0,48	5	0,11	4,35
22	0,65	0,44	4	0,20	3,07
23	0,59	0,36	3	0,25	1,88
24	0,67	0,43	4	0,17	3,08
25	0,65	0,44	4	0,21	3,06
26	0,59	0,51	5	0,30	4,15
27	0,76	0,33	3	-0,04	1,80
28	0,63	0,56	6	0,26	5,53
29	0,59	0,44	4	0,28	2,98
30	0,72	0,42	4	0,11	3,07
31	0,45	0,50	5	0,44	3,52
32	0,69	0,35	3	0,09	1,89
33	0,66	0,35	3	0,13	1,90
34	0,59	0,56	6	0,30	5,34
35	0,67	0,43	4	0,17	3,08
36	0,73	0,33	3	0,01	1,85
37	0,68	0,25	2	-0,01	0,90
38	0,66	0,43	4	0,19	3,07
39	0,70	0,54	6	0,19	5,70
40	0,59	0,51	5	0,29	4,16
41	0,62	0,44	4	0,24	3,04
42	0,63	0,44	4	0,23	3,04
43	0,69	0,43	4	0,15	3,08
44	0,79	0,31	3	-0,12	1,72
45	0,81	0,38	4	-0,05	2,83
46	0,77	0,46	5	0,06	4,28
47	0,50	0,36	3	0,40	1,75
48	0,62	0,51	5	0,26	4,25
49	0,66	0,43	4	0,19	3,07
50	0,77	0,52	6	0,10	5,67
51	0,67	0,50	5	0,21	4,34
52	0,70	0,42	4	0,13	3,08

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

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

Fotografie: Ron Steemers