

Actualiteit en kwaliteit van normen

Een werkwijze voor het normeren van een leerlingvolgsysteem

Jos Keuning, Herman van Boxtel, Nienke Lansink, Jacqueline Visser,
Anke Weekers en Ronald Engelen



Actualiteit en kwaliteit van normen

Een werkwijze voor het normeren van een leerlingvolgsysteem

Jos Keuning
Herman van Boxtel
Nienke Lansink
Jacqueline Visser
Anke Weekers
Ronald Engelen

© Cito B.V. Arnhem (2015)

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

Inhoud

1	Inleiding	5
2	Normeren van een toets	7
2.1	Inleiding	7
2.2	Kwaliteit en actualiteit	7
2.3	Praktijkvoorbeeld	9
3	Werkwijze bij het normeren van LVS-III	13
3.1	Inleiding	13
3.2	Onderzoeksopzet	13
3.3	Normeringssteekproef	15
3.4	Check op landelijke representativiteit	20
3.5	Normering op leerlingniveau	22
3.6	Normering op schoolniveau	25
4	Werkwijze bij het hernormeren van LVS-II en LVS-III	29
4.1	Inleiding	29
4.2	Hernormeringssteekproef	29
4.3	Check op landelijke representativiteit	30
4.4	Praktijkvoorbeeld	32
4.5	Perspectief	39
	Literatuur	45

1 Inleiding

Eén van de criteria aan de hand waarvan toetsen beoordeeld worden, heeft betrekking op de normen. De actualiteit van de normen is daarbij een belangrijk onderdeel. Het is bekend dat normen aan slijtage onderhevig zijn. Ze zijn gevoelig voor maatschappelijke veranderingen, veranderingen in het onderwijs en veranderingen in het gebruiksdoel, waarbij de mate waarin slijtage optreedt, kan verschillen per toets. Daarnaast vormen de afnamecondities tijdens het normeringsonderzoek een invloedrijke factor om rekening mee te houden. Het is van belang dat de afnamecondities overeenkomen met de toepassings-situatie waarvoor de toets bedoeld is. Maar ook als de toetsinhoud en de afname-instructie tijdens het normeringsonderzoek hetzelfde zijn als tijdens de échte afname na uitgave, kan de afnamesituatie in de loop van de tijd gaan afwijken van de afnamesituatie zoals die was ten tijde van het normeringsonderzoek. In het onderwijs zien we bijvoorbeeld dat toetsresultaten in korte tijd een steeds grotere rol zijn gaan spelen in de interne en externe verantwoording die scholen afleggen. Bij scholen neemt hierdoor de druk toe om hun onderwijsprestaties te verbeteren. Langzaam schuiven leervorderingentoetsen voor de leerling, de leerkracht en/of de school dan ook op van *low stake* naar *high stake*.

Voor Cito is de veronderstelde verschuiving van *low stake* naar *high stake* aanleiding geweest om na te gaan in hoeverre de normen van de LVS-toetsen in het *Cito Volgsysteem primair en speciaal onderwijs* actueel zijn. Analyses op systematisch geregistreerde afnamegegevens laten zien dat (a) de resultaten bij het daadwerkelijk gebruik van de toetsen in de dagelijkse onderwijspraktijk snel of zelfs direct afwijken van de initiële normering, en (b) normeringen sneller slijten in de tijd dan gedacht. Dit was voor Cito reden om een nieuwe werkwijze te ontwikkelen voor het normeren van leerlingvolgsysteemtoetsen. Om iets te doen aan de eerste constatering, proberen we in de nieuwe werkwijze de situatie waarin de toets tijdens het normeringsonderzoek wordt afgenomen zoveel mogelijk te laten lijken op de feitelijke situatie waarin de toets na uitgave wordt toegepast. Daarnaast wordt de actualiteit van elke normering in de nieuw ontwikkelde werkwijze jaarlijks gemonitord om snelle slijtage tijdig te kunnen signaleren. Het monitoren van de normering is inmiddels toegepast bij LVS-II *Rekenen-Wiskunde*, *Begrijpend lezen*, en *Spelling*. In schooljaar 2013/2014 zijn voor deze toetsen nieuwe normen beschikbaar gekomen. Daarnaast is de werkwijze gebruikt bij de normering van de eerste LVS-toetsen van de derde generatie (LVS-III). Het gaat om *Rekenen-Wiskunde 3.0*, *Begrijpend lezen 3.0* en *Spelling 3.0* voor de groepen 3 en 4.

Dit rapport is als volgt opgebouwd. In Hoofdstuk 2 gaan we in op factoren die van invloed zijn op de actualiteit en kwaliteit van normgegevens. Tevens laten we in Hoofdstuk 2 aan de hand van een praktijkvoorbeeld (de toets *Woordenschat*) zien in hoeverre deze factoren van toepassing zijn op de leerlingvolgsysteemtoetsen van Cito. In Hoofdstuk 3 beschrijven we de werkwijze die Cito volgt bij de ontwikkeling van de derde generatie leerlingvolgsysteemtoetsen. We geven onder andere aan hoe de normeringsonderzoeken worden opgezet, hoe de normeringssteekproeven gecheckt worden op landelijke representativiteit, en hoe de normen op leerling- en schoolniveau vastgesteld worden. In Hoofdstuk 4 gaan we in op de vraag hoe de normgegevens van bestaande leerlingvolgsysteemtoetsen gemonitord en, indien nodig, geüpdatet worden. We illustreren de werkwijze aan de hand van een concreet voorbeeld. In het voorbeeld gaan we uit van LVS-II *Rekenen-Wiskunde*. We sluiten af met het schetsen van een perspectief. We geven in het perspectief aan tegen welke problemen en beslissingen we aanlopen bij het monitoren en tussentijds bijstellen van normeringen.

Dit rapport is op te vatten als een algemene wetenschappelijke verantwoording van een normerings-methodiek. De concrete gegevens die horen bij de feitelijke normering van een toetsuitgave zijn te vinden in de betreffende wetenschappelijke verantwoording of in addenda die per leerstofdomein worden uitgebracht als er sprake is van een hernormering.

2 Normeren van een toets

2.1 Inleiding

Een toetsscore krijgt pas betekenis door deze te vergelijken met een norm. De toetsscore kunnen we interpreteren door: (1) de toetsscore te vergelijken met de toetsscores van andere leerlingen uit de doelgroep, (2) de toetsscore te vergelijken met toetsscores die de leerling bij eerdere afnamemomenten behaald heeft, of (3) de toetsscore te vergelijken met een standaard die ook wel als cesuur of norm aangeduid wordt (Sanders & Verstralen, 2010). In de eerste twee gevallen spreken we van relatief normeren; in het derde geval van absoluut normeren. In het Cito Volgsysteem primair en speciaal onderwijs worden toetsresultaten relatief beoordeeld. Relatieve normen worden opgesteld via een procedure die uit drie stappen bestaat. In de eerste stap wordt beschreven welke leerlingen tot de doelgroep behoren. Er wordt bijvoorbeeld vastgelegd of doubleurs wel of niet tot de doelgroep gerekend worden. In de tweede stap wordt de toets afgenomen bij een aselechte steekproef van ten minste 400 leerlingen die binnen de doelgroep vallen (Evers, Lucassen, Meijer & Sijtsma, 2010). In de derde en laatste stap wordt de normschaal geconstrueerd. Ook als deze stappen strikt gevolgd worden, is het niet zeker dat de normering aan alle kwaliteitseisen voldoet en dat de normering voor een lange periode actueel is. In paragraaf 2.2 gaan we in op factoren die van invloed zijn op de actualiteit en kwaliteit van een (relatieve) normering. Vervolgens laten we in paragraaf 2.3 aan de hand van een praktijkvoorbeeld zien in hoeverre deze factoren van toepassing zijn op de leerlingvolgsysteemtoetsen (LVS-toetsen) van Cito.

2.2 Kwaliteit en actualiteit

Bij de normering van een toets zijn er veel factoren die onder controle gehouden moeten worden om ervoor te zorgen dat de normering adequaat is. Het beoordelingssysteem van de COTAN (Evers, et al., 2010) legt de nadruk op steekproefomvang en representativiteit ten opzichte van de populatie waarvoor de toets bedoeld is. Het is ook belangrijk dat de situatie waarin de toets in het normeringsonderzoek wordt afgenomen, overeenkomt met de situatie waarin de toets na uitgave feitelijk wordt toegepast. Het beoordelingssysteem van de COTAN schenkt hieraan weinig aandacht. Veel meer dan de eis dat er na de normering geen wijzigingen in de iteminhoud en de instructie mogen worden aangebracht, wordt er in feite niet over vermeld. Toch kunnen de afnamecondities in een normeringsonderzoek aanzienlijk afwijken van de afnamecondities tijdens échte afnamen. Eén van de belangrijkste factoren bij leervorderingen-toetsen is de mate waarin degene die getoetst wordt, waarneemt dat er iets voor hem op het spel staat. Daarnaast kan het belang dat een leerkracht heeft bij een toetsafname van invloed zijn op de uitkomsten. Als hier wijzigingen in optreden, kan er een discrepantie ontstaan tussen de normeringssituatie en de feitelijke afname.

Als we ons beperken tot de leerling, lijkt het vooral te gaan om wat in de literatuur wordt aangeduid met *test taking motivation*, ofwel "... *a student's engagement and expenditure of energy toward the goal of attaining the highest possible score on the test*" (Wise & DeMars, 2005, p. 2). Over dit type motivatie is veel geschreven. In de Verenigde Staten blijkt het voor onderwijsinstellingen vaak lastig te zijn om leerlingen gemotiveerd te laten deelnemen aan gestandaardiseerde toetsafnamen. Dat ongemotiveerde leerlingen lager scoren dan gemotiveerde leerlingen is uitgebreid empirisch onderbouwd (zie bijvoorbeeld Wise & DeMars, 2005). Het meeste onderzoek heeft weliswaar betrekking op het hoger onderwijs, maar motivatieproblemen worden ook elders gemeld (Brophy & Ames, 2005). Volgens het *expectancy-value* model wordt motivatie beïnvloed door (a) de waarde die aan de toets wordt gegeven – wat levert een goede score mij op?, en (b) de verwachte moeilijkheid van de taak – kan ik de taak aan?

Het verdient aanbeveling om onderscheid te maken tussen enerzijds *low stake* toetssituaties, waarin er zowel voor de leerling als voor de leerkracht weinig op het spel lijkt te staan, en anderzijds *high stake* toetssituaties, waarin zowel de leerling als de leerkracht veel waarde hechten aan het toetsresultaat. Van een *low stake* toetssituatie is bijvoorbeeld sprake tijdens landelijke peilingsonderzoeken en normeringsonderzoeken. Tijdens de afname van de (verplichte) eindtoets basisonderwijs en de centrale eindexamens is er sprake van een *high stake* toetssituatie. Andere toetssituaties kunnen ergens op een continuüm tussen deze twee uiteinden worden ondergebracht. We mogen verwachten dat de prestaties van leerlingen in *low stake* toetssituaties soms fors lager liggen dan in *high stake* toetssituaties (zie bijvoorbeeld Brophy & Ames, 2005; Wise & DeMars, 2005; Hemker, 2012). In recent onderzoek van Hemker (2012) zijn de resultaten van de *low stake* periodieke peiling van het onderwijsniveau (PPON) vergeleken met de resultaten van de *high stake* Cito Eindtoets Basisonderwijs. In de *high stake* conditie bleken leerlingen aanzienlijk hoger te scoren dan in de *low stake* conditie (effectgrootte > .36). Bij jongens was het prestatieverschil groter dan bij meisjes. In vergelijking met de *high stake* conditie hadden leerlingen in de *low stake* conditie sterker de neiging om items over te slaan. In proeftoets- en normeringsonderzoeken is het niet mogelijk om volledig te controleren voor afnameconditie. Als leerlingen (en scholen) tijdens de proeftoetsing en het normeringsonderzoek minder goed hun best doen dan tijdens de echte afnamen die na uitgave plaatsvinden, zou de normering van bepaalde toetsen *gebiased* kunnen zijn. Dat kan bijvoorbeeld gelden voor leerlingvolgsysteemtoetsen.

Slijtage is een andere factor die van invloed kan zijn op de adequaatheid – of beter gezegd: actualiteit – van een normering. Afwegende wat praktisch haalbaar en wenselijk is, houdt het beoordelingssysteem van de COTAN de volgende regel aan bij het beoordelen van de actualiteit van een normering: “Om de gebruiker te attenderen op mogelijk versleten normen, wordt aan de beoordeling van tests waarvan hernormerings- of ijkingsonderzoek sinds vijftien jaar na het afsluiten van het normeringsonderzoek niet heeft plaatsgevonden, de kwalificatie ‘*De normen zijn verouderd*’ toegevoegd. Na nog eens vijf jaar zonder dergelijk onderzoek wordt deze kwalificatie gewijzigd in: ‘*Wegens veroudering zijn de normen niet meer bruikbaar en wordt de beoordeling onvoldoende*’ (p.21).” De mate van veroudering of slijtage is afhankelijk van allerlei factoren. Er zijn signalen dat de normen van leervorderingentoetsen mogelijk sneller slijten dan voorheen. Voor de snellere slijtage zijn verschillende mogelijke oorzaken aan te dragen (zie bijvoorbeeld Evers, et al., 2010; Onderwijsinspectie, 2011; Koretz & Béguin, 2010; Popham, 1999; Shepard, 1990):

- De leerlingpopulatie kan over de tijd veranderen. Als er bijvoorbeeld minder leerlingen vanuit het reguliere basisonderwijs worden doorgeleid naar het speciaal basisonderwijs kan het (gemiddelde) prestatieniveau veranderen.
- Het onderwijsaanbod kan veranderen. Sommige (onderdelen van) leergebieden kunnen (weer) belangrijker worden en andere (onderdelen van) leergebieden kunnen naar de achtergrond verschuiven.
- Er kan een verandering optreden in de manier waarop leerkrachten en scholen leerlingvolgsystemen gebruiken. Nu basisscholen verplicht zijn om een leerlingvolgsysteem te gebruiken, worden leerkrachten zich mogelijk bewuster van hiaten bij (groepen) leerlingen. Het gevolg kan zijn dat probleemleerlingen beter geremedieerd worden en de onderwijskundig-didactische aanpak van leerkrachten effectiever wordt.
- Als toetsen vaker gebruikt worden, kan de toetsinhoud geleidelijk bekend raken. Leerkrachten passen het onderwijsaanbod mogelijk bewust of onbewust aan op die kennis. Soms vinden er ook oneigenlijke trainingen plaats op de toetsinhoud. In dat geval worden de prestaties van een leerling op de specifieke toets die wordt afgenomen beter, maar gaat de werkelijke vaardigheid niet vooruit.
- Het belang dat door de leerling, de leerkracht of de school gehecht wordt aan een toetsresultaat kan over de tijd groter worden. In het onderwijs zien we bijvoorbeeld dat toetsresultaten een steeds grotere rol gaan spelen in de interne en externe verantwoording die scholen afleggen. Bij scholen neemt daardoor de druk toe om hun onderwijsprestaties te verbeteren. Langzaam lijken leervorderingentoetsen steeds meer op te schuiven van *low stake* naar *high stake*. Het gevolg hiervan kan zijn dat de resultaten op een specifieke toets omhoog gaan, maar dat de leerlingen niet daadwerkelijk beter worden in bijvoorbeeld rekenen of spellen.

2.3 Praktijkvoorbeeld

Er is nagegaan in hoeverre de normen van LVS-II na uitgave verschuiven en/of slijten. In deze paragraaf presenteren we ter illustratie de resultaten van de analyses die verricht zijn voor LVS-II *Woordenschat* op afnamemoment eind groep 3. De normen voor deze toets zijn gebaseerd op een databestand uit schooljaar 2007/2008 met gegevens van 33 verschillende basisscholen met in totaal 803 leerlingen. In Van Berkel et al. (2010) wordt uitgebreid ingegaan op de representativiteit van de normeringssteekproef en de werkwijze die gevolgd is bij het normeren van LVS-II *Woordenschat*. In de schooljaren na uitgave zijn via Cito dataretour op grote schaal toetsgegevens verzameld. Cito dataretour is een exporttool die basisscholen in staat stelt om jaarlijks op vrijwillige basis hun LVS-resultaten naar Cito te sturen voor (interne) onderzoeksdoeleinden. Het opsturen van resultaten vindt geautomatiseerd plaats via het Computerprogramma LOVS. Verreweg de meeste basisscholen geven gehoor aan de oproep die Cito jaarlijks doet. Het grote voordeel van Cito dataretour is dat er longitudinale toetsgegevens beschikbaar komen van honderdduizenden leerlingen op verschillende toetsen. Vanaf schooljaar 2008/2009 tot schooljaar 2012/2013 zijn in totaal van 149475 leerlingen gegevens verzameld op afnamemoment eind groep 3. Na opschoning van Cito dataretour (zie paragraaf 3.3 en 4.2) bleven er toetsgegevens over van 132762 leerlingen. De leerlingen waren als volgt verspreid over de schooljaren: 2008/2009 = 16816, 2009/2010 = 33023, 2010/2011 = 31805, 2011/2012 = 30118, 2012/2013 = 21000.

Tabel 2.1 laat zien hoe verschillende leerlingcohorten op afnamemoment eind groep 3 gepresteerd hebben op LVS-II *Woordenschat*. De eerste regel in Tabel 2.1 (jaar 2007/2008) geeft weer hoe de leerlingen tijdens het normeringsonderzoek presteerden. Er vallen twee dingen op. In de eerste plaats zien we een relatief grote verschuiving in prestatie direct na uitgave van de toets. Leerlingen scoorden in het eerste jaar gemiddeld $50.03 - 47.51 = 2.52$ punten hoger dan tijdens het normeringsonderzoek. In de tweede plaats zien we relatief kleine prestatieverschuivingen over de tijd. In schooljaar 2012/2013 scoorden leerlingen bijvoorbeeld $52.65 - 52.05 = .60$ punten hoger dan in schooljaar 2011/2012. Gemiddeld bedraagt de prestatieverschuiving bij de overgang van het ene naar het andere schooljaar .66 punten. De standaarddeviatie blijkt nauwelijks over de tijd te veranderen.

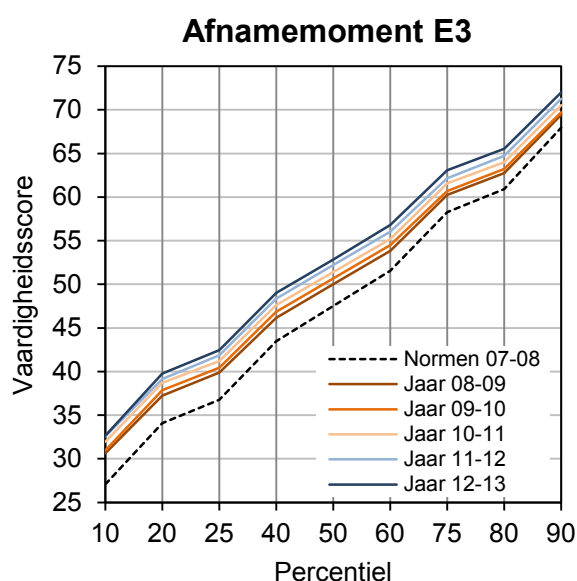
Tabel 2.1 Prestaties van leerlingen bij LVS-II *Woordenschat* einde 3 in de periode 07/08-12/13

Jaar	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
07/08	47.51	15.94	----	----	27.1	34.1	36.8	43.5	47.5	51.6	58.3	60.9
08/09	50.03	14.93	-.16	.02	30.6	37.2	39.9	46.2	50.0	53.8	60.3	62.7
09/10	50.58	14.83	-.22	.02	31.0	37.9	40.4	46.9	50.7	54.5	60.7	63.2
10/11	51.38	14.81	-.19	.03	32.0	38.7	41.2	47.6	51.4	55.2	61.6	64.0
11/12	52.05	14.82	-.23	.00	32.6	39.2	41.9	48.4	52.2	56.1	62.2	64.7
12/13	52.65	15.06	-.25	-.03	32.7	39.8	42.5	49.0	52.9	56.8	63.1	65.6

Noot: K = Kurtosis, S = Skewness

In Figuur 2.1 worden de onderzoeksresultaten grafisch weergegeven. Op de x-as staan enkele percentielen die voor de gebruikelijke Cito niveau-indelingen (A-E; I-V) relevant zijn. Op de y-as staat de vaardigheidscore die bij het betreffende percentiel hoort. De grafische weergave laat zien dat de percentielen opschuiven naarmate de toets langer op de markt is. De verschuiving is direct na uitgave relatief groot. Het lijkt er dus op dat de *low stake* condities tijdens het normeringsonderzoek effect hebben gehad op de prestaties van leerlingen. Als leerlingen tijdens het normeringsonderzoek minder goed hun best doen dan tijdens de échte afname zijn de normen te soepel en wordt het relatieve niveau van leerlingen bij de échte afname overschat. In de jaren na het eerste jaar na uitgave zijn de verschuivingen relatief klein, maar wel consistent. Dit wijst erop dat de normering aan slijtage onderhevig is. Het gevolg is dat we de relatieve prestaties van leerlingen over de tijd steeds meer overschatten.

Figuur 2.1 Visuele weergave van de prestaties van leerlingen bij LVS-II Woordenschat einde 3



Figuur 2.1 maakt duidelijk dat de prestaties van leerlingen veranderen in de tijd. We weten niet precies hoe we de veranderingen moeten duiden. Daarom zijn effectgroottes d uitgerekend. Een effectgrootte geeft weer hoe relevant of betekenisvol een verschil is. Een effectgrootte d van .20 kan beschouwd worden als een klein effect, een effectgrootte d van .50 als een gemiddeld effect, en een effectgrootte d van .80 als een groot effect (Cohen, 1988). In Tabel 2.2 is de effectgrootte voor alle mogelijk combinaties van schooljaren te zien. In termen van de richtlijnen van Cohen (1988) is het prestatieverschil tussen het normeringsonderzoek en het eerste afnamejaar als klein aan te merken ($d = .168$). In schooljaar 2012/2013 is het effect toegenomen tot .341. De effectgroottes voor de prestatieverschillen tussen opeenvolgende schooljaren variëren van .037 tot .054. Deze verschuivingen zijn zeer klein. De toename in prestatie over de tijd is opmerkelijk constant.

Tabel 2.2 Effectgroottes tussen afnamejaren bij LVS-II Woordenschat einde 3

	Normen 07/08	Jaar 08/09	Jaar 09/10	Jaar 10/11	Jaar 11/12	Jaar 12/13
Normen 07/08	----					
Jaar 08/09	.168	----				
Jaar 09/10	.206	.037	----			
Jaar 10/11	.261	.091	.054	----		
Jaar 11/12	.305	.136	.099	.045	----	
Jaar 12/13	.341	.175	.139	.085	.041	----

Ten slotte is nagegaan wat de praktische implicaties zijn als een normering aan slijtage onderhevig is. Op basis van het databestand voor *LVS-II Woordenschat* is uitgerekend voor hoeveel leerlingen het kwartiel verandert als niet het normeringsonderzoek dat in schooljaar 2007/2008 heeft plaatsgevonden als referentie fungeert, maar het schooljaar waar de leerling op dat moment daadwerkelijk in zit. In Tabel 2.3 zien we dat 84.9 procent van de leerlingen die in schooljaar 2008/2009 in het vierde kwartiel vallen, ook op basis van de geldende normering in het vierde kwartiel geclassificeerd zouden worden. Voor 15.1 procent van de leerlingen verandert de classificatie. Waar zij op basis van de geldende normering in het hoogste, vierde, kwartiel geclassificeerd zouden worden, vallen zij op basis van de hypothetische 2008/2009-normering in het lagere, derde, kwartiel. De overige percentages in Tabel 2.3 zijn op dezelfde manier te

interpreteren. Als de slijtage gering is, zal de diagonaal in alle schooljaren steeds dezelfde, hoge, waarden bevatten. Tabel 2.3 laat zien dat hier bij *LVS-II Woordenschat* geen sprake van is. We zien juist dat de waarden op de diagonaal steeds iets lager worden. In schooljaar 2008/2009 maakt het voor ruim 84 procent van de leerlingen niet uit of de geldende normering of de hypothetische 2008/2009-normering gebruikt wordt. In schooljaar 2012/2013 is dat percentage afgenomen tot ruim 68 procent.

Tabel 2.3 Percentage leerlingen per kwartiel bij *LVS-II Woordenschat* einde 3 in de periode 07/08-12/13

Jaar	Kwartiel van het betreffende schooljaar	Kwartiel op basis van het normeringsonderzoek			
		4	3	2	1
08/09	4	84.9	0.0	0.0	0.0
	3	15.1	75.6	0.0	0.0
	2	0.0	24.4	75.5	0.0
	1	0.0	0.0	24.5	100.0
09/10	4	82.2	0.0	0.0	0.0
	3	17.8	70.1	0.0	0.0
	2	0.0	29.9	70.4	0.0
	1	0.0	0.0	29.6	100.0
10/11	4	77.3	0.0	0.0	0.0
	3	22.7	63.1	0.0	0.0
	2	0.0	36.9	63.5	0.0
	1	0.0	0.0	36.5	100.0
11/12	4	72.7	0.0	0.0	0.0
	3	27.3	56.3	0.0	0.0
	2	0.0	43.7	58.5	0.0
	1	0.0	0.0	41.5	100.0
12/13	4	69.0	0.0	0.0	0.0
	3	31.0	50.4	0.0	0.0
	2	0.0	49.6	54.2	0.0
	1	0.0	0.0	45.8	100.0

Samenvattend laten de analyses zien dat de *low stake* afnameconditie in het normeringsonderzoek effect heeft gehad en dat de normen van *LVS-II Woordenschat* voor afnamemoment eind groep 3 aan slijtage onderhevig zijn. Ook bij andere *LVS-II* toetsen spelen deze twee fenomenen, maar niet altijd in even sterke mate en op dezelfde manier. Bij de *LVS-II* toetsen die tot nog toe geanalyseerd zijn, was het effect van de afnameconditie in sommige gevallen bijvoorbeeld iets kleiner en was de slijtage in de normen minder groot. Ook kwam het voor dat de normen over de tijd nauwelijks onderhevig waren aan slijtage. Er is geen duidelijk antwoord te geven op de vraag waarom er verschillen bestaan tussen leergebieden en jaargroepen. Elementen die een rol zouden kunnen spelen bij deze verschillen in slijtage, zijn:

- De mate waarin een vaardigheid gemakkelijk leerbaar is en/of toetsinhoud (on)bewust gemakkelijk geoefend kan worden.
- Het belang dat aan een toets of vaardigheid gehecht wordt in het kader van opbrengstgericht werken en het afleggen van verantwoording.

Om, los van de oorzaken, met beide gevonden fenomenen rekening te kunnen houden, stellen we in Hoofdstuk 3 en 4 een nieuwe werkwijze voor om toetsen te normeren. In die werkwijze proberen we de situatie waarin de toets tijdens het normeringsonderzoek wordt afgenomen zoveel mogelijk te laten lijken op de feitelijke situatie waarin de toets na uitgave wordt toegepast. Daarnaast wordt elke normering in de nieuw ontwikkelde werkwijze gemonitord op basis van Cito dataretour.

3 Werkwijze bij het normeren van LVS-III

3.1 Inleiding

In Hoofdstuk 2 is duidelijk geworden dat de prestaties van leerlingen soms al vrij snel na uitgave van een toets verschuiven. Deze verschuiving wordt mogelijk veroorzaakt door een discrepantie tussen de afnamesituatie tijdens het normeringsonderzoek en de feitelijke afnamesituatie na uitgave. Vermoedelijk doen leerlingen tijdens het normeringsonderzoek minder goed hun best dan tijdens de echte afnamen die na uitgave plaatsvinden. De *low stake* condities tijdens het normeringsonderzoek leiden in dat geval tot *bias* in de normen. Bij de ontwikkeling van LVS-III wordt geprobeerd om *bias* in de normen te vermijden door de situatie waarin de toets tijdens het normeringsonderzoek wordt afgenomen zoveel mogelijk te laten lijken op de feitelijke situatie waarin de toets na uitgave wordt toegepast. Dit betekent dat er bij de ontwikkeling van LVS-III een andere werkwijze wordt gevolgd dan bij eerdere generaties van het LVS van Cito. In paragraaf 3.2 gaan we in op de onderzoeksopzet die gehanteerd wordt voor het normeringsonderzoek. Vervolgens beschrijven we in paragraaf 3.3 hoe de normeringssteekproef tot stand komt. In paragraaf 3.4 geven we aan hoe de normeringssteekproef gecheckt wordt op landelijke representativiteit. In paragrafen 3.5 en 3.6, ten slotte, beschrijven we hoe we de normen op leerling- en schoolniveau bepalen.

3.2 Onderzoeksopzet

Bij de ontwikkeling van LVS-I en LVS-II werd een verzameling items tijdens het normeringsonderzoek in een onvolledig *standalone* design voorgelegd aan leerlingen. Figuur 3.1 geeft een voorbeeld. We zien dat er in dit voorbeeld 9 verschillende taken (A – I), in combinaties van 2, aan leerlingen in leerjaren 3 en 4 worden voorgelegd en dat een groep leerlingen steeds een taak gemeenschappelijk heeft met een andere groep leerlingen. Vanwege deze zogenaamde ankertaken is het mogelijk om de items in de verschillende taken bij de analyse met behulp van modellen uit de item respons theorie op één onderliggende meetschaal te plaatsen (Kolen & Brennan, 2004; Eggen, 1993, 2004; Hambleton, Swaminathan & Rogers 1991). Als de items op één en dezelfde meetschaal liggen, kunnen de resultaten van leerlingen die verschillende taken maken gemakkelijk onderling vergelijkbaar gemaakt worden. Bovendien kunnen de normen en de eigenschappen van een itemverzameling in een populatie berekend worden onafhankelijk van de taken die zijn voorgelegd tijdens het normeringsonderzoek. Het is dus geen probleem dat de leerlingen tijdens het normeringsonderzoek niet volledig de nieuw uit te geven toetsen maken.

Figuur 3.1 Voorbeeld design LVS-I en LVS-II

Toets	Leerjaar	A	B	C	D	E	F	G	H	I
1	medio 3									
2	medio 3									
3	medio 3									
4	medio 3									
5	eind 3									
6	eind 3									
7	eind 3									
8	eind 3									
9	eind 3									
10	medio 4									
11	medio 4									
12	medio 4									
13	medio 4									

Hoewel het *standalone* design zoals dat is weergegeven in Figuur 3.1 wereldwijd wordt toegepast bij de ontwikkeling en normering van toetsen, garandeert het design niet dat de leerlingen in de steekproef de afname ook daadwerkelijk ervaren als een echte afname die meetelt (zie Schmeiser & Welch, 2006). Over het algemeen maken leerlingen de items namelijk buiten de gebruikelijke toetscyclus om en spelen de toetsresultaten geen rol in het onderwijs of de interne of externe verantwoording die een school aflegt. Daarom wordt bij de ontwikkeling en normering van LVS-III niet langer gebruikgemaakt van dit type design. In plaats daarvan worden zogenaamde *embedded field* normeringsonderzoeken georganiseerd waarin nieuw ontwikkelde items meedraaien in de bestaande toetscyclus. Dit betekent dat aan een reguliere LVS-II toetsafname eenmalig één of meer taken worden toegevoegd met nieuw materiaal. De leerlingen weten op voorhand niet om welk materiaal het gaat. Tevens zijn zij niet op de hoogte van het feit dat de toets óók voor onderzoeksdoeleinden ingezet wordt. Op deze manier wordt de discrepantie tussen de afnamesituatie tijdens het normeringsonderzoek en de feitelijke afnamesituatie na uitgave tot een minimum gereduceerd. Motivatie-effecten zijn zo goed als uit te sluiten.

Bij de ontwikkeling van LVS-III *Rekenen-Wiskunde* is voor het eerst gebruikgemaakt van een *embedded field* design. Het precieze design is te vinden in de wetenschappelijke verantwoording van LVS-III *Rekenen-Wiskunde* (Janssen, Wouda & Verbruggen, in voorbereiding). Figuur 3.2 laat zien hoe de designs die gebruikt worden bij de normering van LVS-III over het algemeen worden opgebouwd.

Figuur 3.2 Voorbeeld design LVS-III

Toets	Leerjaar	medio 3					einde 3					medio 4					einde 4		
		m3 LVS-II	m3 deel 1	m3 deel 2	m3e3 deel 1	m3e3 deel 2	e3 LVS-II	e3 deel 1	e3 deel 2	e3m4 deel 1	e3m4 deel 2	m4 LVS-II	m4 deel 1	m4 deel 2	m4e4 deel 1	m4e4 deel 2	e4 LVS-II	e4 deel 1	e4 deel 2
1	medio 3																		
2	medio 3																		
3	medio 3																		
4	medio 3																		
5	medio 3																		
6	medio 3																		
7	medio 3																		
10	medio 3																		
11	einde 3																		
12	einde 3																		
13	einde 3																		
14	einde 3																		
15	einde 3																		
16	einde 3																		
17	einde 3																		
18	einde 3																		
19	einde 3																		
20	einde 3																		
21	medio 4																		
22	medio 4																		
23	medio 4																		
24	medio 4																		
25	medio 4																		
26	medio 4																		
27	medio 4																		
28	medio 4																		
29	medio 4																		
30	medio 4																		

We zien dat alle leerlingen op afnamemoment medio groep 3 volgens het design van Figuur 3.2 de *m3-toets* uit LVS-II maken. De leerlingen op afnamemoment eind groep 3 maken allemaal de *e3-toets* uit LVS-II en op afnamemoment medio groep 4 maken de leerlingen de *m4-toets* uit LVS-II. Daarnaast maken de leerlingen twee andere taken met nieuw materiaal. De taken met dit materiaal vormen samen de uiteindelijk uit te geven toets voor een bepaald afnamemoment. Op basis van de taken met labels *m3 deel 1* en *m3 deel 2* ontstaat dus de uitgave voor afnamemoment medio groep 3. In LVS-III worden voor het eerst ook zogeheten tussentoetsen meegenomen in de onderzoeken. Deze toetsen kunnen ingezet worden bij leerlingen voor wie de reguliere toets naar verwachting te moeilijk (bijvoorbeeld afname *m3e3-toets* op afnamemoment einde groep 3) of te gemakkelijk (bijvoorbeeld afname *e3m4-toets* op afnamemoment einde groep 3) is. Elk taak bestaat voor ongeveer 10 procent uit reservemateriaal, zodat eventuele calamiteiten opgevangen kunnen worden. Het is de verwachting dat het reservemateriaal slechts incidenteel ingezet hoeft te worden, omdat alle items voorafgaand aan het normeringsonderzoek geproeftoetst zijn bij steekproeven met 150 tot 200 leerlingen.

In tegenstelling tot LVS-II gaat het normeringsonderzoek in LVS-III niet langer uit van een onvolledig *standalone* design, maar van een volledig *embedded field* design-plus. Zoals we in Figuur 3.2 kunnen zien fungeren de items uit LVS-II als een groot vast anker dat voor alle leerlingen hetzelfde is. Het plus-deel bevat het materiaal voor de nieuwe uitgave en is variabel voor de leerlingen. Zowel binnen een afnamemoment als over de afnamemomenten heen wordt geankerd. Het belangrijkste voordeel van het design in Figuur 3.2 is dat het nieuwe materiaal voor LVS-III in de bestaande toetscyclus meedraait en dat de normen voor de nieuw uit te geven toetsen op basis van de feitelijke toetssituatie bepaald kunnen worden. Daarnaast is er sprake van enkele bijkomende voordelen. Ten eerste kunnen de normen zoals die bepaald worden voor de nieuw uit te geven toetsen vergeleken worden met de normen van LVS-II, ten tweede kan de constructvaliditeit beter onderbouwd worden. De scores van de leerlingen op het oude en het nieuwe materiaal kunnen immers op hun onderlinge samenhang worden onderzocht. Daarmee kan de continuïteit tussen verschillende generatie toetsen empirisch in kaart worden gebracht. Ten slotte bestaat de mogelijkheid om Cito dataretour in te zetten bij het normeren van de toetsen. In paragraaf 3.3 wordt uitgebreid ingegaan op het gebruik van Cito dataretour bij het normeren van LVS-III.

3.3 Normeringssteekproef

Met het oog op de normering van LVS-III worden de data die tijdens de *embedded field* normeringsonderzoeken verzameld zijn, aangevuld met gegevens uit Cito dataretour (zie paragraaf 2.3). Vanzelfsprekend worden de data die via Cito dataretour binnenkomen opgeschoond voordat ze gebruikt worden. In de eerste plaats worden de scholen verwijderd die ook aan de *embedded field* normeringsonderzoeken deelnamen. Daarnaast worden de volgende categorieën leerlingen uit de bestanden verwijderd:

- Leerlingen uit het speciaal onderwijs.
Deze leerlingen worden niet meegenomen bij de normering, omdat het speciaal onderwijs de prestaties van hun leerlingen graag wil vergelijken met het reguliere onderwijs. Voor deze leerlingen gelden namelijk dezelfde kerndoelen. Leerkrachten in het speciaal onderwijs zijn onder meer geïnteresseerd in het functioneringsniveau van de leerlingen. Een referentiegroep met alleen leerlingen uit het reguliere onderwijs is vanuit dit gezichtspunt eenduidig en optimaal. Ten tweede is het problematisch om een referentiegroep samen te stellen met reguliere én speciale leerlingen. Cito kiest er namelijk voor om te normeren op basis van de jaargroep. In het speciaal onderwijs worden leerlingen in de regel ingedeeld naar vaardigheidsniveau; een indeling naar jaargroep is ongebruikelijk. Het is om deze reden onduidelijk van welke referentiegroep de leerlingen uit het speciaal onderwijs deel uit zouden moeten maken.
- Leerlingen van scholen die het LVS selectief inzetten.
In de hogere leerjaren blijken sommige scholen het LVS alleen in te zetten bij zwakkere leerlingen (zie Keuning, 2011). Om bias in de normering te voorkomen, laten we deze leerlingen buiten beschouwing.

- Leerlingen die op hetzelfde afnamemoment meerdere toetsen van dezelfde vaardigheid maken. Alleen de gegevens van de toets die bij het afnamemoment hoort, worden behouden.
We mogen verwachten dat leerlingen bij wie meerdere toetsen zijn afgenomen zeer zwak of zeer goed zijn in vergelijking met de andere leerlingen in de klas. Alleen bij deze leerlingen is het immers zinvol om ‘door’ te toetsen. Het alsnog afnemen van een toets van een hoger of lager niveau maakt de niveaubepaling namelijk betrouwbaarder. Door van dit soort leerlingen meer dan één score te handhaven, zouden zeer zwakke en zeer goede leerlingen oververtegenwoordigd raken in de normeringssteekproef. Dit is onwenselijk. We zouden de variantie bijvoorbeeld kunnen overschatten. Om deze reden nemen we deze leerlingen in het bestand mee met slechts één toetsafname.

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 .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 $[\cdot10, \cdot25)$, (3) percentage achterstandsleerlingen $[\cdot25, \cdot40)$ en (4) percentage achterstandsleerlingen $[\cdot40, 1]$.

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

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

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

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

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

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

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

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

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:

Vorbereiding data normeringsonderzoek

```

bereken  $w_{ijk}$  voor S
indien  $w_{ijk} < 0$ 
  herhaal
    trek aselekt 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  $N_S \geq N_T$ 
retourneer S

```

Check leerlingvariabele seks

```

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  $\tilde{S}_y$ 
  totdat  $N_{\tilde{S}_y} = N_{S_y}$ 
retourneer  $\tilde{S}$ 

```

Het algoritme is toegepast bij de ontwikkeling van de nieuwe LVS-III toetsen, bijvoorbeeld de toetsen *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. Ter illustratie wordt hieronder weergegeven hoe het selectiealgoritme functioneerde voor de eerste twee 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 $N_T = 2 \times 2253 = 4506$. 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 $N_T C$ 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 $N_S \geq N_T$. 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 $N_T C$. 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 $N_T = 2 \times 2157 = 4314$. 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 $N_T C$. 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 *seks*. 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 aselechte steekproef dan ook niet op zijn plaats. Door bij de normering van LVS-III de benodigde data representatief te trekken, zijn de afwijkingen die we vinden in relatie tot de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *seks*e 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.

3.4 Check op landelijke representativiteit

Het beoordelingssysteem van de COTAN (Evers, et al., 2010) geeft aan dat een normeringssteekproef moet worden beschreven in termen van *leeftijd*, *etniciteit*, *seks*e en *regio*. De variabele *leeftijd* is bij de ontwikkeling van LVS-III niet aan de orde, omdat de normering plaatsvindt naar leerjaar. Zowel tijdens de *embedded field* normeringsonderzoeken als in Cito dataretour worden steeds hele klassen geselecteerd. Daarom mag worden aangenomen dat de verdeling naar leeftijd over leerjaar een afspiegeling vormt van de populatie. Het beschrijven van een normeringssteekproef naar etnische herkomst is problematisch. Enerzijds omdat de etnische herkomst van leerlingen op scholen niet bekend is en anderzijds omdat er geen betrouwbare (en eenduidige) populatiegegevens voorhanden zijn. Daarom wordt met het oog op de ontwikkeling van LVS-III op schoolniveau informatie verzameld over het aantal gewichtenleerlingen en de urbanisatiegraad. Hoewel in de huidige gewichtenregeling van OCW de herkomst van de leerling geen rol meer speelt, mag aangenomen worden dat de combinatie van urbanisatiegraad en leerlinggewicht een goede benadering biedt voor etnische herkomst. Eerdere analyses op data uit landelijke peilingsonderzoeken hebben dit laten zien (Hemker, Kordes en Van Weerden, 2011). Gegevens over *seks*e en *regio* worden standaard verzameld tijdens de normeringsonderzoeken. Ook in Cito dataretour wordt deze informatie standaard opgevraagd.

Door toepassing van het selectiealgoritme uit paragraaf 3.3 is de representativiteit van de normeringssteekproeven voor LVS-III in principe gegarandeerd. Niettemin wordt er een controle uitgevoerd op de representativiteit door de populatieverdelingen te vergelijken met de steekproefverdelingen. In Tabel 3.1 wordt ter illustratie de representativiteitsanalyse voor de eerste twee afnamemomenten van LVS-III *Rekenen-Wiskunde* weergegeven. In de representativiteitsanalyse wordt de normeringsteekproef geanalyseerd in relatie tot de variabelen *regio*, *urbanisatiegraad*, *schooltype*, en *seks*e. In paragraaf 3.3 is te lezen hoe deze variabelen gedefinieerd zijn. De populatiegegevens zijn afkomstig van DUO (www.data.duo.nl), CBS (statline.cbs.nl) en de Cendris postcode-relatietabel.

Tabel 3.1 Representativiteitsanalyse LVS-III Rekenen-Wiskunde medio 3 en einde 3

Variabele	Categorie	Populatie	Steekproef medio 3		Steekproef einde 3	
			N	%	N	%
Regio	Noord	10.2	458	10.2	420	9.7
	Oost	22.7	1025	22.8	970	22.5
	West	47.1	2167	48.1	2077	48.1
	Zuid	20.0	856	19.0	848	19.7
Urbanisatiegraad	Platteland	56.3	2551	56.6	2466	57.2
	Stad	43.7	1955	43.4	1849	42.8
Schooltype	[0, .10)	60.6	2802	62.2	2672	61.9
	[.10, .25)	26.4	1181	26.2	1131	26.2
	[.25, .40)	6.6	262	5.8	261	6.1
	[.40, 1]	6.4	261	5.8	251	5.8
Sekse	Jongen	50.4	2220	50.5	2117	50.2
	Meisje	49.6	2177	49.5	2100	49.8

Voor beide afnamemomenten is goed te zien dat de steekproefverdelingen nauwelijks afwijken van de populatieverdelingen. Dit is ook statistisch getoetst. De resultaten staan in Tabel 3.2. De χ^2 -waarden zijn laag en in slechts één geval significant. Bij grotere steekproeven zegt significantie echter niet zoveel. 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 .044 en .036 nog het hoogst voor de variabele *schooltype*. De conclusie is niettemin dat de normeringssteekproef voor beide afnamemomenten een zeer goede afspiegeling vormt van de populatie.

Tabel 3.2 Statistische toetsing representativiteitsanalyse LVS-III Rekenen-Wiskunde medio 3 en einde 3

Variabele	Steekproef medio 3					Steekproef einde 3				
	χ^2	df	N	p	ϕ	χ^2	df	N	p	ϕ
Regio	3.499	3	4506	.32	.028	2.402	3	4315	.49	.024
Urbanisatiegraad	.227	1	4506	.63	.007	1.384	1	4315	.24	.018
Schooltype	8.903	3	4506	.03	.044	5.746	3	4315	.13	.036
Sekse	.008	1	4397	.93	.001	.082	1	4217	.78	.004

3.5 Normering op leerlingniveau

Na afname van een LVS-toets volgt een indicatie van het niveau van een leerling ten opzichte van een representatieve landelijke vergelijkingsgroep. Het relatieve niveau van een leerling wordt in het LVS van Cito aangeduid door middel van de letters A tot en met E of de Romeinse cijfers I tot en met V. De niveaus zijn als volgt gedefinieerd:

Niveau	Grens	Omschrijving
A	score $\geq P75$	De 25% hoogst scorende leerlingen
B	$P50 \leq \text{score} < P75$	De 25% leerlingen die boven de landelijke mediaan scoren
C	$P25 \leq \text{score} < P50$	De 25% leerlingen die onder de landelijke mediaan scoren
D	$P10 \leq \text{score} < P25$	De 15% leerlingen die ruim onder de landelijke mediaan scoren
E	score $< P10$	De 10% laagst scorende leerlingen

Niveau	Grens	Omschrijving
I	score $\geq P80$	De 20% hoogst scorende leerlingen
II	$P60 \leq \text{score} < P80$	De 20% leerlingen die boven de landelijke mediaan scoren
III	$P40 \leq \text{score} < P60$	De 20% leerlingen die rond de landelijke mediaan scoren
IV	$P20 \leq \text{score} < P40$	De 20% leerlingen die onder de landelijke mediaan scoren
V	score $< P20$	De 20% laagst scorende leerlingen

In de communicatie met scholen spreken we over het gemiddelde, omdat zij het begrip ‘mediaan’ meestal niet kennen. Daarnaast wordt ook het 90^{ste} percentiel in het LVS van Cito onderscheiden. Leerlingen met percentielscores $\geq P90$ zijn de zogenaamde A+ of I+ leerlingen. De percentielen worden bepaald op basis van de verdeling van scores die we vinden in de normeringssteekproef zoals die is samengesteld op basis het *embedded field* normeringsonderzoek en Cito dataretour. In veel toepassingen kunnen de percentielen gebaseerd worden op de ruwe scores die de leerlingen tijdens het normeringsonderzoek behalen op de toets die is voorgelegd. Bij een leerlingvolgsysteem is het niet zinvol om de percentielen te definiëren op de ruwe scoreschaal, omdat de ruwe scores die leerlingen behalen op verschillende toetsen in de tijd (in absolute zin) niet met elkaar te vergelijken zijn. Bij voorkeur wordt er in een leerlingvolgsysteem gebruik gemaakt van meetschalen waarmee de vaardigheid van leerlingen gevolgd kan worden onafhankelijk van de itemset die bij een toetsafname aan de leerlingen wordt voorgelegd. Dergelijke meetschalen kunnen onder andere geconstrueerd worden binnen het raamwerk van de item respons theorie (zie bijvoorbeeld, Embretson & Reise, 2000; Kolen & Brennan, 2004; Van der Linden & Hambleton, 1997). Als een item respons model geldt voor een verzameling items kunnen de ruwe scores van de leerlingen via het onderliggende model gecorrigeerd worden voor de moeilijkheidsgraad en de lengte van de toets. De gecorrigeerde ruwe scores worden in de praktijk meestal schaal- of vaardigheidsscores genoemd.

Aan de meeste toetsen in het LVS van Cito ligt het *One-Parameter Logistic Model* (OPLM) van Verhelst en Glas (1995) ten grondslag. Gegeven dit model zijn er verschillende manieren om de vaardigheidsverdeling van de normeringssteekproef te bepalen. We kunnen er bijvoorbeeld voor kiezen om de vaardigheidsscores van de individuele leerlingen (bijvoorbeeld de ML- of WML-schatting) in de normeringssteekproef te gebruiken. Deze aanpak is echter niet zonder risico, omdat door de schattingsfouten in de individuele vaardigheidsscores de kans bestaat dat we de variantie overschatten (Wu, 2005). Het gevolg is dat we de percentielen te laag of te hoog kiezen en we leerlingen niet optimaal indelen in niveau groepen. Een alternatief is om de vaardigheidsverdeling te schatten op basis van de itemantwoorden en de itemparameters uit het item respons model zonder de omweg van de individuele vaardigheidsscores. Verhelst en Verstralen (2002) hebben het programma SAUL ontwikkeld dat op deze manier te werk gaat. In essentie schat het programma conditionele gemiddelden volgens een lineair structureel model. Het basismodel wordt gegeven door de volgende vergelijking:

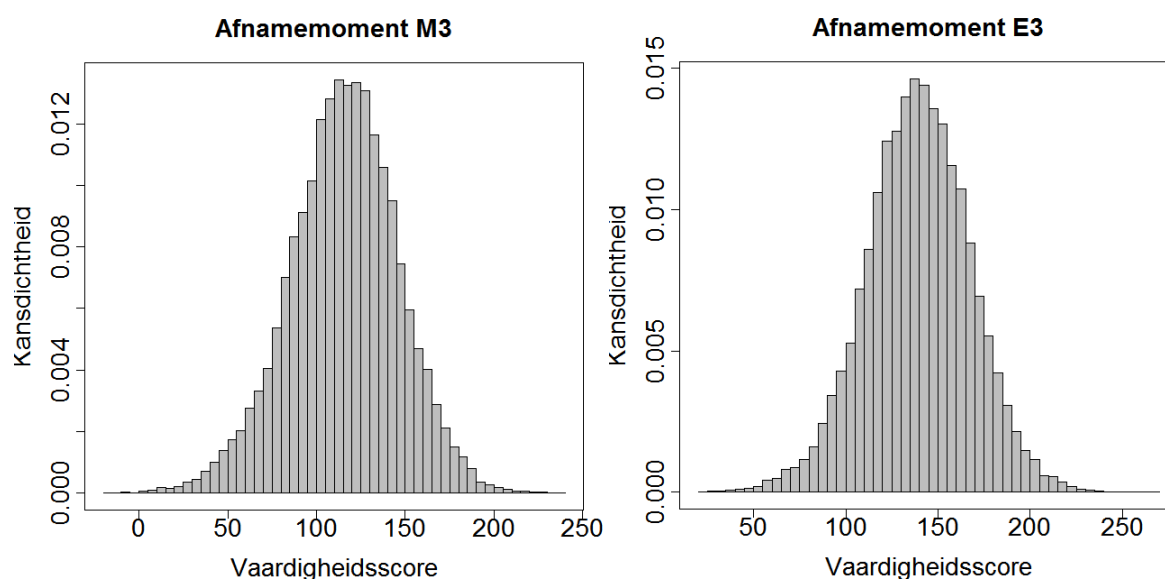
$$\theta = \beta_0 + \beta_1 X_1 + \dots + \beta_q X_q + \varepsilon,$$

waarin θ de latente vaardigheid van de leerling weergeeft, β_0 het algemene gemiddelde is, en β_1 tot en met β_q de effecten weergeven van achtergrondvariabelen X_1 tot en met X_q . Het model veronderstelt dat het residu ε normaal verdeeld is met gemiddelde 0 en variantie σ^2 . Het programma SAUL is vooral nuttig als de normeringssteekproef niet representatief is en de vaardigheidsverdeling hiervoor gecorrigeerd moet worden. Als er geen statistische weging hoeft plaats te vinden, kan via toepassing van *Marginal Maximum Likelihood* (met gefixeerde itemparameters), ofwel MML, hetzelfde resultaat bereikt worden. Hoewel de theoretische vaardigheidsverdeling goed geschat kan worden met behulp van SAUL of MML zijn de methoden niet direct toe te passen bij de normering van LVS-III. De methoden maken namelijk gebruik van de itemantwoorden van de leerlingen en die zijn niet altijd beschikbaar via Cito dataretour. De ruwe totaalscores zijn wel van alle leerlingen beschikbaar.

Met het oog op de normering van LVS-III wordt er op basis van de ruwe totaalscores voor elke leerling een *plausible value* gegenereerd op de nieuw ontwikkelde meetschaal. In feite representeren *plausible values* de volledige range aan vaardigheidsscores die een leerling zou kunnen hebben, gegeven zijn itemantwoorden. Een *plausible value* is dus niet gelijk aan de θ -parameter (de vaardigheidsscore) zoals die gedefinieerd is in bijvoorbeeld het OPLM. Er wordt namelijk niet één enkele puntschatting voor θ bepaald, maar er wordt een reeks van mogelijke waarden voor θ geschat die elk een bepaalde kans hebben om geobserveerd te worden. *Plausible values* zijn random trekkingen uit deze (geschatte) verdeling voor de vaardigheid θ van een leerling. Daardoor geven *plausible values* niet alleen informatie over de geschatte vaardigheid van een leerling, maar ook over de onzekerheid die bij die schatting hoort. Vanwege deze eigenschap kunnen *plausible values* zonder problemen met standaardprogrammatuur als SPSS en SAS geanalyseerd worden. Het gebruik van de puntschattingen voor de latente vaardigheid θ zou tot bias in sommige populatieparameters leiden. Zoals al eerder is opgemerkt zou de variantie bij gebruik van de ML- of WML-schatting van θ bijvoorbeeld overschat worden.

Plausible values kunnen op verschillende manieren gegenereerd worden (zie bijvoorbeeld, Albert, 1992; Béguin & Glas, 2001; Maris & Bechger, 2005; Mislevy, 1991). Bij de normering van de derde generatie LVS-toetsen wordt gebruikgemaakt van het algoritme zoals dat recent is voorgesteld door Marsman, Maris, Bechger en Glas (2011). In essentie genereert dit algoritme op basis van het OPLM en een getrokken θ^* uit de populatieverdeling een mogelijk responspatroon $\mathbf{y}$ voor een leerling. In geval het responspatroon $\mathbf{y}$ tot dezelfde ruwe score optelt als het responspatroon $\mathbf{x}$ dat geobserveerd is voor een leerling, dus in geval $\mathbf{y}_+ = \mathbf{x}_+$, dan weten we dat de getrokken θ^* afkomstig is uit de posterior verdeling voor die leerling en kunnen we de getrokken θ^* als *plausible value* toewijzen aan de leerling. De procedure wordt op een efficiënte manier herhaald totdat we aan alle leerlingen in de normeringssteekproef zoals die is samengesteld op basis van het *embedded field* normeringsonderzoek en Cito dataretour een *plausible value* hebben toegekend. De normering wordt gebaseerd op de *plausible values* van de leerlingen in de normeringssteekproef. Figuur 3.3 laat ter illustratie de verdeling van *plausible values* zien voor de eerste twee afnamemomenten van LVS-III *Rekenen-Wiskunde*. We zien dat de *plausible values* in dit voorbeeld op beide afnamemomenten een normale verdeling volgen. Dat hoeft niet zo te zijn. Hoewel binnen de item respons theorie uitgegaan wordt van vaardigheidsverdelingen met een specifieke (normaal verdeelde) vorm, kan in de empirie een andere vorm optreden. Bij de normering houden we hier rekening mee.

Figuur 3.3 Scoreverdelingen op basis van plausible values voor LVS-III Rekenen-Wiskunde



Op basis van de scoreverdelingen in Figuur 3.3 worden de percentielen berekend die horen bij de vaardigheidsindelingen A tot en met E enerzijds en I tot en met V anderzijds. We gaan daarbij uit van de **empirische** cumulatieve verdelingsfunctie. Als de vaardigheid van leerlingen niet normaal verdeeld is, zien we dat dus terug in de percentielen die we bepalen. Tabel 3.3 geeft de normgegevens voor LVS-III Rekenen-Wiskunde. Zoals verwacht mag worden, scoren de leerlingen bij afnamemoment eind groep 3 hoger dan bij afnamemoment medio groep 3. De percentielen liggen bovendien behoorlijk ver uit elkaar. Dit betekent dat het zonder meer mogelijk is om leerlingen op basis van de ontwikkelde toetsen voor medio 3 en einde 3 in te delen in de onderscheiden niveaugroepen.

Tabel 3.3 Normtabel op leerlingniveau voor LVS-III Rekenen-Wiskunde medio 3 en einde 3

Tijd	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
medio 3	115.3	30.7	.319	-.179	76.5	90.4	95.8	108.7	116.1	123.6	135.8	140.6
einde 3	139.3	28.2	.272	-.101	103.7	116.6	120.9	132.6	139.5	146.5	158.2	162.6

Noot: K = Kurtosis, S = Skewness

De normering is gecheckt door deze toe te passen op de normeringssteekproef. Bij de check is uitgegaan van de vaardigheidsindeling I tot V. We verwachten dat op grond van deze indeling 20 procent van de leerlingen niveau I scoort, 20 procent niveau II, 20 procent niveau III, 20 procent niveau IV en 20 procent niveau V. Tabel 3.4 laat zien dat deze verdeling in voldoende mate teruggevonden wordt in de normeringsteekproeven voor afnamemomenten medio groep 3 en eind groep 3. Er kan dan ook geconcludeerd worden dat de normtabel zoals die in dit voorbeeld voor LVS-III Rekenen-Wiskunde op basis van de *plausible values* is bepaald adequaat is. Kleine afwijkingen ten opzichte van de theoretische verdeling zijn ook heel normaal. Dit komt doordat *plausible values* elke waarde kunnen aannemen binnen een bepaalde range. Voor de vaardigheidsscores die we in de empirie waarnemen na afname van een specifieke toets geldt dat niet.

Tabel 3.4 Percentage leerlingen in de niveaucategorieën I tot en met V

Tijd	Niveau				
	V	IV	III	II	I
medio 3	20.32	20.35	21.50	18.62	19.22
einde 3	21.42	18.78	21.14	19.59	19.08

3.6 Normering op schoolniveau

Het Cito Volgsysteem kent naast een normering op leerlingniveau ook een normering op schoolniveau. Het relatieve niveau van een school wordt net als eerder bij de leerlingen aangeduid door middel van de letters A tot en met E of de Romeinse cijfers I tot en met V. In paragraaf 3.5 is de betekenis van de niveaucategorieën uiteengezet. De schoolnormering kan op verschillende manieren bepaald worden. Het meest eenvoudig is om op basis van de resultaten van de individuele leerlingen binnen een school, schoolgemiddelden uit te rekenen en de verdeling van schoolgemiddelden te gebruiken om de percentielen te bepalen die horen bij de niveaucategorieën A tot en met E en I tot en met V. Hoewel deze methode eenvoudig is toe te passen, kleeft er ook een belangrijk nadeel aan. De methode veronderstelt namelijk dat alle schoolgemiddelden even nauwkeurig te bepalen zijn, terwijl dat in de praktijk meestal niet zo is. De standaardfout van een schoolgemiddelde $\bar{X}$ wordt gegeven door:

$$\sigma(\bar{X}) = \frac{1}{\sqrt{n}} \sigma_X,$$

waarbij σ_X de standaarddeviatie van het schoolgemiddelde weergeeft en n gelijk is aan het aantal leerlingen waarop het schoolgemiddelde gebaseerd is. Zoals we kunnen zien, is de standaardfout direct afhankelijk van het aantal leerlingen binnen een school. De standaardfout is relatief groot als het schoolgemiddelde gebaseerd wordt op relatief weinig leerlingen. Naarmate het schoolgemiddelde gebaseerd kan worden op meer leerlingen wordt de standaardfout kleiner en wordt het schoolgemiddelde nauwkeuriger. Als dit verschijnsel bij het bepalen van de schoolnormering genegeerd wordt, bestaat het risico dat we de variantie op schoolniveau (sterk) overschatten. Bij de ontwikkeling van LVS-III wordt deze methode dan ook niet toegepast.

Multilevel modellen houden rekening met het aantal observaties op elk niveau. Het meest eenvoudige model bevat geen verklarende variabelen. Het model scheidt alleen de variantie van de afhankelijke variabele in een deel dat aan de scholen y is toe te schrijven en in een deel dat aan de leerlingen p is toe te schrijven. Bij de ontwikkeling van LVS-III maken we gebruik van dit zogeheten *intercept-only* model. Het model wordt gegeven door de volgende vergelijking:

$$\text{Niveau 1 (leerlingniveau): } X_{py} = \beta_{0y} + \varepsilon_{py}, \quad \varepsilon_{py} \sim N(0, \sigma^2)$$

$$\text{Niveau 2 (schoolniveau): } \beta_{0y} = \gamma_{00} + \mu_{0y}, \quad \mu_{0y} \sim N(0, \tau^2)$$

$$\text{Gecombineerd: } X_{py} = \gamma_{00} + \varepsilon_{py} + \mu_{0y}$$

Zoals we kunnen zien, beschrijft dit model op niveau 1 dat de scores X van leerlingen afhangen van twee bronnen: de schoolgemiddelden β_{0y} en de *random* individuele variatie ε_{py} rond het schoolgemiddelde. Het schoolgemiddelde voor een bepaalde school is opgebouwd uit een algemeen gemiddelde γ_{00} plus een afwijking van dat gemiddelde die specifiek is voor de betreffende school μ_{0y} . Er wordt verondersteld dat de afwijkingen ε_{py} ten opzichte van de schoolgemiddelden β_{0y} normaal verdeeld zijn met een gemiddelde

van 0 en een variantie gelijk aan σ^2 . De afwijkingen μ_{0y} ten opzichte van het algemene gemiddelde γ_{00} zijn eveneens normaal verdeeld. Het gemiddelde is gelijk aan 0 en een variantie is gelijk aan τ^2 . Als het model eenmaal geschat is, kunnen de percentielen die horen bij de vaardigheidsindelingen A tot en met E en I tot en met V gemakkelijk bepaald worden door de inverse van de cumulatieve **normale** verdeling te berekenen met gemiddelde γ_{00} en standaarddeviatie τ . We kunnen tevens bepalen in hoeverre de scholen bijdragen aan de variantie van de afhankelijke variabele. Hiertoe berekenen we de intraklasse-correlatie ρ die aangeeft hoeveel van de totale variantie is toe te schrijven aan de scholen:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

Als de leerlingen binnen de scholen veel op elkaar lijken is de intraklassecorrelatie groot. Als de leerlingen binnen de scholen minder op elkaar lijken is de intraklassecorrelatie kleiner. In onderwijskundig onderzoek liggen de intraklassecorrelaties doorgaans tussen de .05 en .25 (Snijders & Bosker, 1993; Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Schochet, 2008). Als vuistregel wordt vaak aangehouden dat een multilevel analyse zinvol is als de intraklasse-correlatie .04 of meer bedraagt. Anderen stellen dat het niet zozeer gaat om de intraklassecorrelatie, maar meer om het effect van de wijze waarop de steekproef is samengesteld (zie bijvoorbeeld Snijders & Bosker, 1999; Hox & Maas, 2001; Hox, 2002). In dit zogeheten design effect wordt dan ook niet alleen rekening gehouden met de intraklassecorrelatie, maar ook met de gemiddelde groepsgrootte:

$$\text{design effect} = 1 + \left(\frac{N_p}{N_y} - 1 \right) \times \rho,$$

waarin N_p het aantal leerlingen weergeeft en N_y het aantal scholen. Als het design effect groter is dan 2, is het wenselijk om in de analyse rekening te houden met het effect van de clustering (cf. Muthén & Satorra; 1995).

Bij het bepalen van de schoolnormen wordt het eerder beschreven multilevel model geschat in 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. De reden om te kiezen voor een bootstrap procedure is tweeledig. Ten eerste maakt de procedure het mogelijk om zicht te krijgen op de stabiliteit van de resultaten. Ten tweede wordt het mogelijk om in de analyse statistisch te wegen. Vanwege de werkwijze die gehanteerd wordt bij het samenstellen van de normeringssteekproef – waarbij we ervoor zorgen dat de normeringssteekproef representatief is – is dat laatste voordeel bij de ontwikkeling en normering van LVS-III niet van toepassing. Tabellen 3.5 en 3.6 laten ter illustratie zien tot welke resultaten de bootstrap procedure heeft geleid bij de eerste twee afnamemomenten van LVS-III *Rekenen-Wiskunde*. We zien dat de uitkomsten behoorlijk stabiel zijn met een scholenaantal >180 en een leerlingaantal > 4000. We zien tevens dat een multilevel analyse wenselijk is. De intraklassecorrelatie ligt zowel op medio 3 als einde 3 boven de .04. Het design effect is bij beide afnamemomenten duidelijk groter dan 2.

Tabel 3.5 *Uitkomsten multilevel analyse voor LVS-III Rekenen-Wiskunde medio 3*

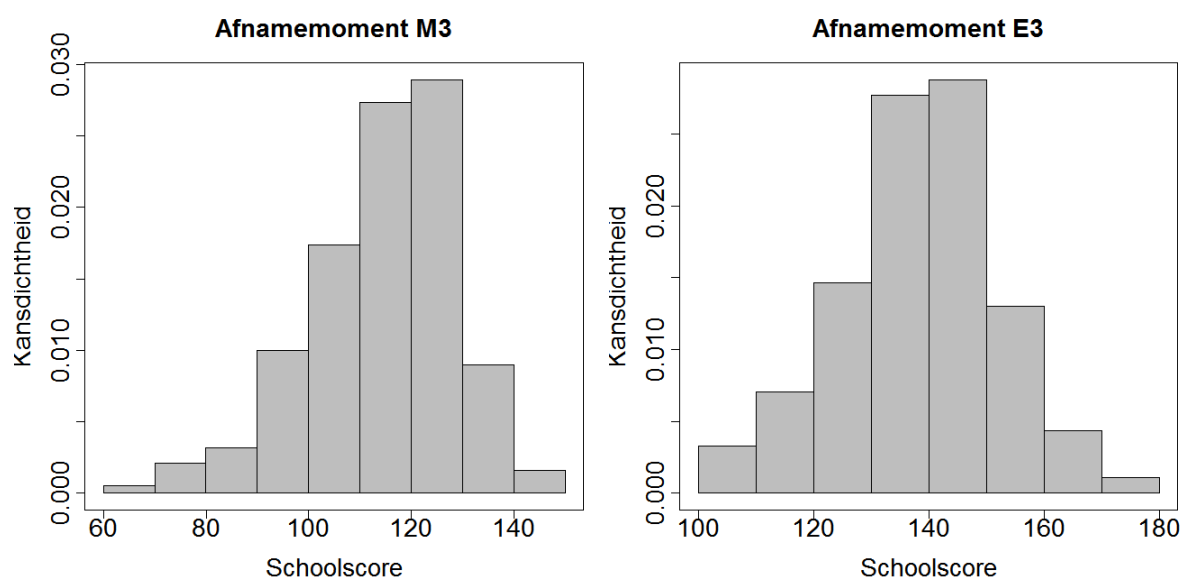
Replicatie	N_y	N_p	γ_{00}	τ	σ	ρ	Design eff.
1	190	4596	113.44	15.52	29.13	.22	6.12
2	190	4328	115.25	14.09	29.17	.19	5.12
3	190	4454	114.01	13.25	28.66	.18	4.95
4	190	4692	115.20	13.35	29.48	.17	5.03
5	190	4593	113.84	15.05	29.34	.21	5.82
...	...	...	...	...	...	...	...
15	190	4597	115.48	12.14	29.81	.14	4.29
16	190	4469	114.77	12.21	28.66	.15	4.47
17	190	4436	112.57	14.72	27.82	.22	5.89
18	190	4711	115.41	13.04	29.78	.16	4.83
19	190	4592	113.25	15.83	29.31	.23	6.24
20	190	4071	115.09	14.38	29.12	.20	5.00
Eindresultaat			114.69	13.60	29.07	.18	5.11

Tabel 3.6 *Uitkomsten multilevel analyse voor LVS-III Rekenen-Wiskunde einde 3*

Replicatie	N_y	N_p	γ_{00}	τ	σ	ρ	Design eff.
1	184	4556	138.74	12.88	26.75	.19	5.47
2	184	4573	138.59	12.18	26.53	.17	5.15
3	184	4012	138.64	12.14	26.59	.17	4.60
4	184	4386	137.19	11.93	27.98	.15	4.52
5	184	4069	137.78	14.18	27.22	.21	5.50
...	...	...	...	...	...	...	...
15	184	3975	138.57	13.07	27.00	.19	4.91
16	184	4012	137.71	13.83	27.74	.20	5.14
17	184	4314	138.42	13.74	26.96	.21	5.62
18	184	4348	135.22	14.27	26.74	.22	6.02
19	184	4586	137.64	14.52	26.48	.23	6.53
20	184	4267	138.52	12.45	27.66	.17	4.73
Eindresultaat			138.32	13.01	27.06	.19	5.15

Figuur 3.4 laat ter illustratie de verdeling van schoolgemiddelden zien voor de eerste twee afname-momenten van LVS-III *Rekenen-Wiskunde*. Het is lastig om te bepalen of de schoolgemiddelden een normale verdeling volgen met een scholenaantal van 190 bij medio 3 en 184 bij einde 3. Op het eerste gezicht lijkt de verdeling bij afnamemoment medio groep 3 iets linksscheef te zijn. Bij afnamemoment eind groep 3 lijkt een normale verdeling beter te verdedigen. De ervaring op basis van Cito dataretour bij LVS-II heeft geleerd dat de aanname van een normale verdeling op het niveau van de scholen eigenlijk altijd houdbaar is. Het lijkt dus geen probleem te zijn dat we bij het schatten van het multilevel model en het bepalen van de percentielen een normale verdeling veronderstellen.

Figuur 3.4 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde



Op basis van de rij *Eindresultaat* in Tabel 3.5 en 3.6 zijn de percentielen berekend die horen bij de vaardigheidsindelingen A tot en met E, en I tot en met V. Tabel 3.7 geeft de normgegevens op schoolniveau voor LVS-III *Rekenen-Wiskunde*. Zoals we mochten verwachten is de spreiding op schoolniveau kleiner dan op leerlingniveau. De percentielen komen dus dichterbij elkaar te liggen. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in niveaus A tot en met E of I tot en met V.

Tabel 3.7 Normtabel op schoolniveau voor LVS-III Rekenen-Wiskunde medio 3 en einde 3

Tijd	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
medio 3	114.7	13.6	----	----	97.3	103.2	105.5	111.2	114.7	118.1	123.9	126.1
einde 3	138.3	13.0	----	----	121.6	127.4	129.5	135.0	138.3	141.6	147.1	149.3

Noot: K = Kurtosis, S = Skewness

4 Werkwijze bij het hernormeren van LVS-II en LVS-III

4.1 Inleiding

Er zijn twee factoren die van invloed kunnen zijn op de actualiteit en kwaliteit van normen. Ten eerste spelen de *afnamecondities* tijdens het normeringsonderzoek een rol. In Hoofdstuk 3 hebben we beschreven hoe we bij de ontwikkeling van LVS-III proberen om bias in de normen te vermijden door de situatie waarin de toets tijdens het normeringsonderzoek wordt afgenomen zoveel mogelijk te laten lijken op de feitelijke situatie waarin de toets na uitgave wordt toegepast. Ten tweede kunnen normen onderhevig zijn aan *slijtage*. Met ingang van schooljaar 2013/2014 worden de normen van LVS-II en LVS-III jaarlijks gecheckt op basis van Cito dataretour. Als de onderzoeksresultaten aanleiding geven om de normen bij te stellen, wordt onderzocht hoe dit gerealiseerd kan worden. In dit hoofdstuk gaan we in op de werkwijze die gevolgd wordt bij het hernormeren van toetsen. De werkwijze is in veel opzichten vergelijkbaar met de reeds beschreven normeringsprocedure die we volgen bij de ontwikkeling van nieuwe LVS-III toetsen (zie Hoofdstuk 3). In paragraaf 4.2 beschrijven we hoe de hernormeringssteekproef tot stand komt. In paragraaf 4.3 geven we aan hoe de hernormeringssteekproef gecheckt wordt op landelijke representativiteit. In paragraaf 4.4 illustreren we de nieuw ontwikkelde procedure *hernormeren* aan de hand van LVS-II *Rekenen-wiskunde*. We sluiten dit hoofdstuk af met een perspectief. In het perspectief geven we aan hoe we in de toekomst om zouden kunnen gaan met veranderingen in de prestaties van leerlingen.

4.2 Hernormeringssteekproef

Bij het hernormeren van LVS-II (en in de toekomst LVS-III) wordt gebruikgemaakt van data uit Cito dataretour. We gebruiken indien mogelijk de gegevens van de laatste drie schooljaren. Dit betekent dat we in schooljaar 2013/2014 de data van schooljaren 2010/2011, 2011/2012 en 2012/2013 gebruiken. Deze werkwijze wijkt af van de werkwijze die gevolgd wordt als we een toets voor de eerste keer normeren. In dat geval gebruiken we, naast de data die tijdens het normeringsonderzoek verzameld zijn, namelijk alleen de data van het meest recente schooljaar (zie paragraaf 3.3). Bij het hernormeren gebruiken we data van meerdere schooljaren, omdat we dan beter kunnen controleren voor incidentele schommelingen in de prestaties van leerlingen. Via statistische weging wordt ervoor gezorgd dat de drie schooljaren een gelijk aandeel hebben in de normering.

In principe gebruiken we bij het hernormeren van een bepaald afnamemoment de data van *alle* leerlingen die tot de doelgroep behoren. Bij het hernormeren van afnamemoment eind groep 3 selecteren we dus de gegevens van *alle* leerlingen die een toets hebben gemaakt op dat afnamemoment. Hiermee zorgen we ervoor dat niet alleen de leerlingen worden meegenomen die de toets van afnamemoment eind groep 3 hebben gemaakt, maar ook de zwakkere leerlingen die een toets van een lager niveau hebben gemaakt (medio 3) en de betere leerlingen die een toets van een hoger niveau hebben gemaakt (medio 4). Omgekeerd worden de gegevens van leerlingen die *wél* de toets van eind groep 3 hebben gemaakt, maar op dat moment niet in groep 3 zaten, niet meegenomen. Voor de analyse is het geen probleem dat niet alle leerlingen dezelfde toetsversie gemaakt hebben. De ruwe scores van de leerlingen zijn op basis van het onderliggende meetmodel namelijk te vertalen naar zo genoemde vaardigheidsscores die over versies van toetsen vergelijkbaar zijn. De normen van de Cito LVS-toetsen worden uitgedrukt in vaardigheidsscores. De data worden opgeschoond voordat ze gebruikt worden. Net als eerder (zie paragraaf 3.3) worden de volgende categorieën leerlingen verwijderd:

- Leerlingen uit het speciaal onderwijs en leerlingen voor wie het onderwijstype onbekend is.
- Leerlingen van scholen die de LVS-toetsen selectief inzetten. In de hogere leerjaren blijken sommige scholen de toetsen 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.

Vervolgens worden de leerlingen verwijderd die extreem hoog of laag scoren in vergelijking met andere leerlingen. We verwijderen de leerlingen met een ruwe score die groter of kleiner is dan 1.5 keer de interkwartielafstand. Ook de leerlingen met een ruwe score groter dan percentiel 99.75 of kleiner dan percentiel .25 worden verwijderd. Er zijn twee belangrijke redenen om een deel van de scores voorafgaand aan de analyse als *outlier* aan te merken. Ten eerste worden de scores van de meeste toetsen handmatig door de leerkracht ingevoerd in het Computerprogramma LOVS. Bij het invoeren worden af en toe fouten gemaakt. De invoerfouten vallen door de gekozen systematiek buiten de hernormering. Ten tweede zijn er leerlingen die, om onduidelijke redenen, onder kansniveau presteren. Door toeval kan dit af en toe gebeuren, maar het is zeer de vraag of de leerlingen die ver onder kansniveau scoren ook meegenomen moeten worden bij de (her)normering. De leerling kan door omstandigheden weinig gemotiveerd zijn geweest om de toets te maken of er kan sprake zijn geweest van een (technisch) probleem. Soms komen deze scores onbedoeld toch in Cito dataretour terecht.

De scores die om de genoemde redenen al als ‘verdacht’ zijn aangemerkt, zouden ook in praktisch opzicht tot een probleem geleid hebben. Het is namelijk problematisch om *plausible values* te genereren voor leerlingen die in het licht van de prestaties van de totale (norm)groep zeer afwijkend presteren. Bij het genereren van *plausible values* worden random getallen getrokken uit een verdeling en op basis van bepaalde criteria toegewezen aan leerlingen (zie Marsman, Maris, Bechger & Glas, 2011). De verdeling wordt niet willekeurig gekozen, maar in een aantal iteraties gebaseerd op de prestaties van de normgroep. Als de score van een leerling op een andere plaats ligt dan de verdeling waaruit getrokken wordt, duurt het ‘oneindig’ lang voordat het algoritme voor het genereren van *plausible values* iets oplevert. Het algoritme kan, met andere woorden, niet tot een oplossing komen. Het aanmerken van deze gevallen als *outlier* heeft dus ook een belangrijk praktisch voordeel.

Nadat de *outliers* verwijderd zijn, wordt aan alle leerlingen op basis van de behaalde ruwe score en de itemparameters een *plausible value* toegekend. In paragraaf 3.5 is te lezen wat *plausible values* zijn, hoe die gegenereerd worden, en waarom ze gebruikt worden bij het (her)normeren van een toets. Na het opschonen van het databestand en het toekennen van *plausible values* is de hernormeringssteekproef gereed. Er wordt dus niet, zoals eerder in paragraaf 3.3, een steekproef getrokken; we gebruiken de gegevens van alle leerlingen bij het controleren (en mogelijk herzien) van de normen van een toets.

4.3 Check op landelijke representativiteit

Het is niet zeker of de hernormeringssteekproef voor een bepaald afnamemoment representatief is voor de landelijke populatie van scholen en leerlingen. Ook zou de hernormeringssteekproef beschouwd kunnen worden als een *sample of convenience*. Cito beschikt immers min of meer “toevallig” over de toetsgegevens van de leerlingen die in de hernormeringssteekproef zitten. De toetsgegevens zijn niet afkomstig uit een aselechte steekproef die getrokken is uit een populatie van scholen en leerlingen die allemaal een gelijke kans hadden om in de steekproef terecht te komen. Om deze reden vindt er een uitgebreide representativiteitsanalyse plaats en wordt er indien nodig gebruikgemaakt van een statistische wegings-procedure. De representativiteit van de hernormeringssteekproef wordt geanalyseerd in relatie tot de achtergrondvariabelen *regio*, *urbanisatiegraad*, *schooltype*, en *sekse*. In paragraaf 3.3 is te lezen hoe deze variabelen gedefinieerd zijn.

In de eerste stap wordt de steekproefverdeling voor elke achtergrondvariabele vergeleken met de populatieverdeling. De vergelijking vindt voor elk van de drie schooljaren afzonderlijk plaats. Via een χ^2 -toetsing wordt nagegaan of eventuele afwijkingen significant zijn. Bij grotere steekproeven zegt significantie echter niet zoveel. Daarom wordt bij de interpretatie vooral gekeken naar de effectgrootte ϕ van de χ^2 -toetsing. Een effectgrootte ϕ van .10 kan aangemerkt worden als een klein effect, een effectgrootte ϕ van .30 als een gemiddeld effect, en een effectgrootte ϕ van .50 als een groot effect (cf. Cohen, 1988). Als de effectgrootte ϕ in ten minste één van de schooljaren groter is dan .10, wordt geconcludeerd dat de hernormeringssteekproef onvoldoende representatief is met betrekking tot de achtergrondvariabele die

geanalyseerd is. In dat geval kan er aanleiding zijn om de onder- of overrepresentatie van een bepaalde groep leerlingen via statistische weging op te lossen.

In de tweede stap wordt nagegaan of de prestaties van groepen leerlingen betekenisvol van elkaar verschillen. Bij de achtergrondvariabele *regio* maken we bijvoorbeeld onderscheid tussen leerlingen die in het noorden, oosten, westen en zuiden van Nederland wonen. Als de noordelijke provincies beter presteren dan de zuidelijke provincies, en overgerepresenteerd zijn in de hernormeringssteekproef, zullen we de gemiddelde (landelijke) prestatie overschatten als we geen statistische weging toepassen. De normering zal in dat geval *gebiased* zijn. Als de verschillende regio's statistisch vergelijkbaar presteren, maakt het niet uit hoeveel leerlingen van elke regio er in de hernormeringssteekproef zitten. De gemiddelde (landelijke) prestatie kan op basis van elke regio correct geschat worden en de normering zal niet *gebiased* zijn. Statistische weging is in dat geval niet noodzakelijk.

Op basis van een regressieanalyse bepalen we of een achtergrondvariabele effect heeft op de prestaties van leerlingen. In de analyse zijn de *plausible values* de afhankelijke variabele en *schooljaar*, *regio*, *urbanisatiegraad*, *schooltype*, en *seks* de verklarende variabelen. De regressieparameters worden geïnterpreteerd in termen van significantie (*z*) en relevantie (*d*). De *z*-scores bepalen we door de regressieparameters te delen door de standaardfout. De effectgroottes *d* bepalen we door de regressieparameters te delen door de residuele standaarddeviatie. We kunnen spreken van een klein effect als $d = .20$, van een gemiddeld effect als $d = .50$, en van een groot effect als $d = .80$ (cf. Cohen, 1988). In geval de effectgrootte *d* van een bepaalde achtergrondvariabele groter is dan $.20$ wordt geconcludeerd dat het waargenomen effect relevant is. Om *bias* in de normen te voorkomen is het bij $d > .20$ nodig dat de steekproefverdeling in voldoende mate overeenkomt met de populatieverdeling. Anders is statistische weging gewenst.

De beslissing om de hernormeringssteekproef al dan niet statistisch te wegen, wordt genomen op basis van de uitkomsten van de representativiteitsanalyse en de regressieanalyse. Als de hernormeringssteekproef representatief is met betrekking tot een bepaalde achtergrondvariabele wordt geen weging toegepast. Als bepaalde groepen leerlingen onder- of overgerepresenteerd zijn, wordt gekeken of de prestaties van de onderscheiden groepen leerlingen van elkaar verschillen. Als dit het geval is, wordt de hernormeringssteekproef voor de betreffende achtergrondvariabele teruggewogen. Anders wordt alsnog besloten om geen statistische weging toe te passen. Bij de analyses die in schooljaar 2012/2013 hebben plaatsgevonden is gebleken dat statische weging in een enkel geval niet nodig is. In andere gevallen volstond een weging naar *schooltype*. Er is nog geen reden geweest om een hernormeringssteekproef op meer dan één achtergrondvariabelen terug te wegen.

Wij denken dat de representativiteit van de hernormeringssteekproef voldoende gewaarborgd is als de beschreven werkwijze gevolgd wordt. Niettemin blijft er in zekere zin sprake van een *sample of convenience*, omdat er alleen data zijn van scholen die het Computerprogramma LOVS van Cito gebruiken. Volgens ons is dit echter geen reden om af te zien van het checken, en mogelijk herzien, van normen op basis van de hernormeringssteekproeven die geconstrueerd kunnen worden op basis van Cito dataretour. De bereidheid van scholen om deel te nemen aan normeringsonderzoeken blijkt in de praktijk bijvoorbeeld zeer beperkt te zijn. Soms is de deelnamebereidheid van scholen minder dan 20 procent. In dat geval is er sprake van een enorme zelfselectie van scholen. Het is zeer wel denkbaar dat deze zelfselectie een groter effect heeft op de normen dan het gegeven dat de normen gebaseerd worden op gegevens van scholen die het Computerprogramma LOVS gebruiken. Daarnaast bepaalt de omvang van een normeringssteekproef hoe nauwkeurig normen bepaald kunnen worden. Geneste steekproeven met ten minste 400 leerlingen blijken vrij veel onzekerheid met zich mee te brengen (Keuning, Hilte, Weekers, 2014).

Via Cito dataretour is het mogelijk om normen te baseren op de gegevens van duizenden scholen met honderdduizenden leerlingen. Het kan haast niet zo zijn dat normen die gebaseerd zijn op échte toetsafnamen bij honderdduizenden leerlingen van mindere kwaliteit zijn dan normen die gebaseerd zijn op een normeringsonderzoek bij enkele honderden leerlingen. Om meer zekerheid te krijgen over deze aanname hebben we analyses uitgevoerd met betrekking tot het niveau van leervorderingen in groep 8 van de

scholen in Cito dataretour. Dit deden we door de prestaties van de dataretourscholen op de Cito Eindtoets Basisonderwijs te vergelijken met die van scholen die niet in Cito dataretour zitten. We keken hierbij zowel naar de algemene prestaties op deze toets als naar de prestaties op het specifieke leerstofdomein (i.e. rekenen-wiskunde, begrijpend lezen en spelling). De afwijkingen bleken zeer klein en verwaarloosbaar, namelijk 1 tot 3 procent van de (gepoolde) standaarddeviatie voor de specifieke leerstofdomeinen en 3% voor de totaalscore op de eindtoets. De kans dat we door de inzet van Cito dataretour te maken krijgen met een systematisch afwijkende normeringssteekproef is daarmee bijzonder klein.

4.4 Praktijkvoorbeeld

In schooljaar 2012/2013 zijn de normen gecheckt van LVS-II *Rekenen-Wiskunde*, *Begrijpend lezen* en *Spelling*. In alle gevallen was er reden om de normen te herzien. De normen zijn vanaf 1 september 2013 beschikbaar in het Computerprogramma LOVS. Ter illustratie bespreken we in deze paragraaf gedetailleerd de uitkomsten voor de eerste twee afnamemomenten van LVS-II *Rekenen-wiskunde*. Tabellen 4.1 tot en met 4.4 bevatten de resultaten van de representativiteitsanalyse die heeft plaatsgevonden.

Tabel 4.1 Representativiteit naar regio voor LVS-II *Rekenen-Wiskunde* medio 3 en einde 3

Groep	Categorie	Medio 3			Einde 3			
		Jaar 09/10	Jaar 10/11	Jaar 11/12	Jaar 09/10	Jaar 10/11	Jaar 11/12	
Populatie	Noord	10.33	10.29	10.20	10.33	10.29	10.20	
	Oost	22.76	22.74	22.70	22.76	22.74	22.70	
	West	46.55	46.78	47.06	46.55	46.78	47.06	
	Zuid	20.36	20.19	20.04	20.36	20.19	20.04	
Dataretour	Noord	12.37	13.57	13.64	12.83	13.74	14.27	
	Oost	26.78	24.64	24.72	26.76	24.33	24.25	
	West	36.48	36.09	33.51	35.32	35.73	31.48	
	Zuid	24.37	25.69	28.13	25.09	26.20	30.00	
	χ^2	2926.93	2927.11	3788.50	χ^2	3285.23	2971.76	4354.02
	df	3	3	3	df	3	3	3
	N	71695	56900	44542	N	64228	52463	36775
	p	.00	.00	.00	p	.00	.00	.00
	ϕ	.20	.23	.29	ϕ	.23	.24	.34

Tabel 4.2 Representativiteit naar urbanisatiegraad voor LVS-II Rekenen-Wiskunde medio 3 en einde 3

Groep	Categorie	Medio 3			Einde 3		
		Jaar 09/10	Jaar 10/11	Jaar 11/12	Jaar 09/10	Jaar 10/11	Jaar 11/12
Populatie	Platteland	57.00	56.70	56.26	57.00	56.70	56.26
	Stad	43.00	43.30	43.74	43.00	43.30	43.74
Dataretour	Platteland	62.91	62.37	62.70	63.81	63.06	63.62
	Stad	37.09	37.63	37.30	36.19	36.94	36.38
χ^2		1022.50	745.77	751.30	χ^2	1215.90	863.24
df		1	1	1	df	1	1
N		71695	56900	44542	N	64228	52463
p		.00	.00	.00	p	.00	.00
ϕ		.12	.11	.13	ϕ	.14	.15

We zien dat de hernormeringssteekproef voor zowel medio 3 als einde 3 representatief is met betrekking tot de variabele *seks* (zie tabel 4.4). De effectgrootte ϕ varieert van .00 tot .01. De twee hernormeringssteekproeven zijn niet voldoende representatief als het om de variabelen *regio* (zie tabel 4.1) en *urbanisatiegraad* (zie tabel 4.2) gaat ($\phi > .10$). De noordelijke, zuidelijke en oostelijke provincies zijn enigszins oververtegenwoordigd in beide normeringssteekproeven en de westelijke provincies zijn ondervertegenwoordigd. Dit heeft vermoedelijk zijn weerslag op de verdeling naar stad en platteland; de verstedelijkte gebieden zijn ondervertegenwoordigd. De conclusie bij de variabele *schooltype* is niet eenduidig. Op het eerste gezicht vertoont de verdeling van leerlingen over de vier schooltypen in beide normeringssteekproeven grote gelijkenis met de verdeling van leerlingen in de populatie. De effectgrootte ϕ is in de meeste gevallen dan ook kleiner dan .10. Schooljaar 2010/2011 vormt bij de normeringssteekproef voor eind groep 3 een uitzondering. De scholen met weinig achterstandsleerlingen zijn in dit schooljaar iets oververtegenwoordigd.

Tabel 4.3 Representativiteit naar schooltype voor LVS-II Rekenen-Wiskunde medio 3 en einde 3

Groep	Categorie	Medio 3			Einde 3		
		Jaar 09/10	Jaar 10/11	Jaar 11/12	Jaar 09/10	Jaar 10/11	Jaar 11/12
Populatie	[0, .10)	56.69	58.88	60.58	56.69	58.88	60.58
	[.10, .25)	28.68	27.29	26.41	28.68	27.29	26.41
	[.25, .40)	7.29	6.96	6.59	7.29	6.96	6.59
	[.40, 1]	7.35	6.87	6.42	7.35	6.87	6.42
Dataretour	[0, .10)	60.47	62.23	63.50	60.61	63.24	63.03
	[.10, .25)	27.30	25.87	25.88	27.29	25.18	26.51
	[.25, .40)	6.60	6.80	5.72	6.62	6.53	5.65
	[.40, 1]	5.63	5.10	4.90	5.49	5.05	4.81
	χ^2	563.26	411.73	279.63	χ^2	560.75	523.55
	df	3	3	3	df	3	3
	N	71695	56900	44542	N	64228	52463
	p	.00	.00	.00	p	.00	.00
	ϕ	.09	.09	.08	ϕ	.09	.10

Tabel 4.4 Representativiteit naar sekse voor LVS-II Rekenen-Wiskunde medio 3 en einde 3

Groep	Categorie	Medio 3			Einde 3		
		Jaar 09/10	Jaar 10/11	Jaar 11/12	Jaar 09/10	Jaar 10/11	Jaar 11/12
Populatie	Jongen	50.49	50.44	50.42	50.49	50.44	50.42
	Meisje	49.51	49.56	49.58	49.51	49.56	49.58
Dataretour	Jongen	50.46	50.31	50.62	50.51	50.21	50.40
	Meisje	49.54	49.69	49.38	49.49	49.79	49.60
	χ^2	.03	.38	.70	χ^2	.01	1.13
	df	1	1	1	df	1	1
	N	71364	56690	44288	N	63922	52231
	p	.87	.54	.40	p	.94	.29
	ϕ	.00	.00	.00	ϕ	.00	.01

Tabel 4.5 laat via een regressieanalyse zien in hoeverre prestatieverschillen tussen leerlingen verklaard kunnen worden vanuit de variabelen *schooljaar*, *schooltype*, *regio*, *urbanisatiegraad* en *sekse*. De effecten zijn in één analyse geschat. Dit betekent dat we in Tabel 4.5 het unieke effect van een bepaalde variabele kunnen zien, gecorrigeerd voor de effecten van de overige variabelen. De meeste variabelen hebben geen betekenisvol effect op de prestaties van leerlingen. We zien weliswaar bij zowel medio groep 3 als eind groep 3 dat meisjes iets zwakker presteren dan jongens, en dat leerlingen op het platteland iets beter presteren dan leerlingen in de stad, maar de effecten zijn niet betekenisvol ($d < .20$). Ook *regio* en *schooljaar* blijken nauwelijks de prestaties van leerlingen te beïnvloeden. Het effect van *schooltype* is wel

relevant. Leerlingen op scholen met veel achterstandsl leerlingen presteren duidelijk zwakker dan leerlingen op scholen met minder achterstandsl leerlingen.

Tabel 4.5 Uitkomsten regressieanalyse voor LVS-II Rekenen-Wiskunde medio 3 en einde 3

Variabele	Categorie	Medio 3				Einde 3			
		N	effect	z	d	N	effect	z	d
Schooljaar	09/10	71695	----	----	----	64228	----	----	----
	10/11	56900	.56	7.26	.04	52463	.56	7.38	.04
	11/12	44542	1.38	16.65	.10	36775	1.20	14.14	.09
Schooltype	[0, .10)	107050	----	----	----	95287	----	----	----
	[.10, .25)	45820	-2.01	-26.15	-.15	40484	-1.77	-22.90	-.14
	[.25, .40)	11144	-5.05	-36.46	-.37	9756	-4.20	-30.04	-.33
	[.40, 1]	9123	-8.40	-55.16	-.62	7939	-7.00	-45.48	-.54
Regio	Noord	22669	----	----	----	20695	----	----	----
	Oost	44238	-.42	-3.67	-.03	38870	-.77	-6.83	-.06
	West	61617	.01	.07	.00	53009	-.62	-5.49	-.05
	Zuid	44613	.49	4.34	.04	40892	-.25	-2.29	-.02
Urbanisatie	Platteland	108524	----	----	----	97463	----	----	----
	Stad	64613	-.79	-10.54	-.06	56003	-.67	-8.83	-.05
Sekse	Jongen	86948	----	----	----	76950	----	----	----
	Meisje	85394	-2.02	-30.68	-.15	75791	-1.71	-25.85	-.13

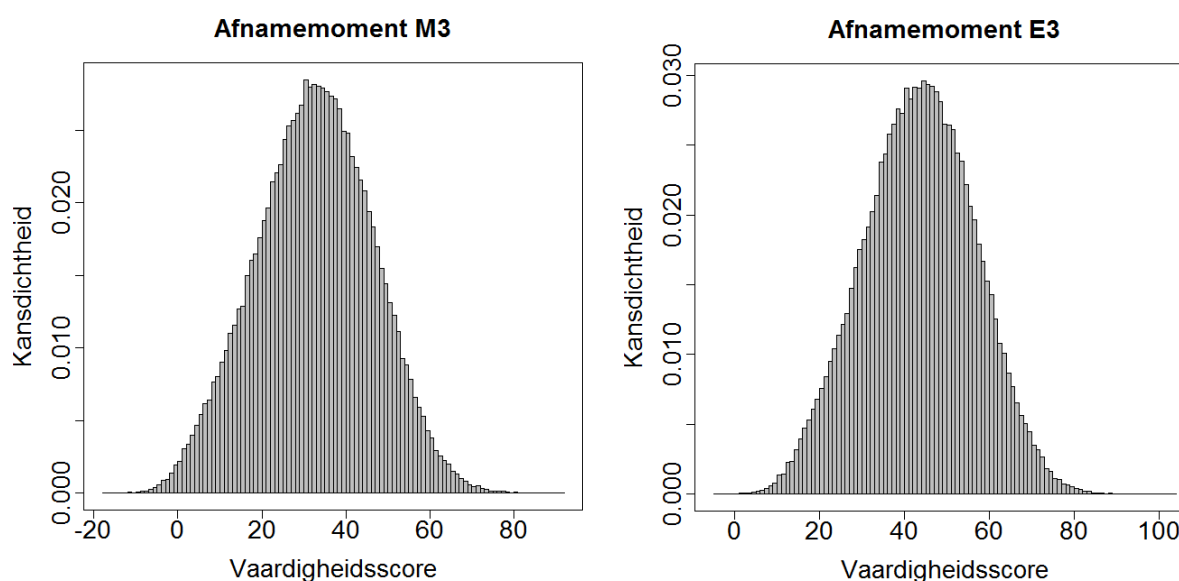
Op basis van Tabellen 4.1 tot en met 4.5 kan geconcludeerd worden dat de hernormeringssteekproef voor medio groep 3 niet statistisch gewogen hoeft te worden. De hernormeringssteekproef is weliswaar onvoldoende representatief met betrekking tot de variabelen *regio* (Tabel 4.1) en *urbanisatiegraad* (Tabel 4.2), maar beide variabelen blijken geen betekenisvol effect te hebben op de prestaties van leerlingen (Tabel 4.5). Voor de hernormeringssteekproef voor eind groep 3 geldt grotendeels dezelfde redenering. De variabele *schooltype* bleek bij dit afnamemoment echter in één schooljaar niet representatief vertegenwoordigd te zijn. Tevens bleek de variabele effect te hebben op de prestaties van leerlingen. Daarom is besloten om de hernormeringssteekproef voor eind groep 3 in zijn geheel te wege voor *schooltype*. Tabel 4.6 geeft de gewichten die zijn toegekend aan de verschillende groepen leerlingen. Daarbij wordt er standaard voor gezorgd dat de drie schooljaren een gelijk aandeel hebben in de normering. Daarom kunnen de gewichten binnen een schooljaar allemaal kleiner of groter zijn dan 1 en vinden we in de tabel ook gewichten bij de hernormeringssteekproef voor medio groep 3 (ook al is daar niet gewogen voor andere achtergrondvariabelen dan schooljaar).

Tabel 4.6 Gehanteerde gewichten bij de hernormering van LVS-II Rekenen-Wiskunde

Steekproef	Variabele	Categorie	Jaar 09/10	Jaar 10/11	Jaar 11/12
Medio 3	----	----	.805	1.014	1.296
Einde 3	Schooltype	[0, .10)	.745	.908	1.337
		[.10, .25)	.837	1.057	1.386
		[.25, .40)	.877	1.038	1.625
		[.40, 1]	1.067	1.328	1.857

Als de gewichten bepaald zijn, wordt op basis van de *plausible values* die zijn toegekend aan de leerlingen in de hernormeringssteekproef een gewogen histogram gemaakt. Figuur 4.1 laat de verdeling van *plausible values* zien voor de eerste twee afnamemomenten van LVS-II Rekenen-Wiskunde. We zien dat de *plausible values* in dit voorbeeld op zowel medio groep 3 als eind groep 3 een normale verdeling volgen. Bij het berekenen van de percentielen die horen bij de vaardigheidsindelingen A tot en met E enerzijds en I tot en met V anderzijds, gaan we echter uit van de inverse van de **empirische** cumulatieve verdelingsfunctie. Dit betekent dat we uitgaan van wat we hebben waargenomen en niet van een veronderstelde normaalverdeling. Als de vaardigheid van leerlingen niet normaal verdeeld is, zien we dat dus terug in de percentielen die we bepalen.

Figuur 4.1 Scoreverdelingen op basis van plausible values voor LVS-II Rekenen-Wiskunde



Tabel 4.7 geeft de hernormeringsgegevens voor LVS-II Rekenen-Wiskunde. Zoals verwacht mag worden, scoren de leerlingen bij afnamemoment eind groep 3 hoger dan bij afnamemoment medio groep 3. De percentielen liggen bovendien behoorlijk ver uit elkaar. Dit betekent dat het zonder meer mogelijk is om leerlingen op basis van de ontwikkelde toetsen voor medio groep 3 en eind groep 3 in te delen in de onderscheiden niveaugroepen. De prestaties van leerlingen blijken een behoorlijke verschuiving doorgeemaakt te hebben ten opzichte van het normeringsonderzoek. Op afnamemoment medio groep 3 scoren leerlingen gemiddeld 6.21 punten hoger en op afnamemoment eind groep 3 ligt de score gemiddeld 8.40 punten hoger (zie voor de oorspronkelijke normen, Janssen, Verhelst, Engelen & Scheltens, 2010). In termen van effectgrootte betekent dit een gemiddelde tot grote verschuiving: (medio 3) $d = .48$, (einde 3)

$d = .61$. Voor de andere afnamemomenten is de verschuiving minder groot. Dit komt vermoedelijk doordat de toetsen voor medio groep 3 en eind groep 3 al langer op de markt zijn dan de toetsen voor de overige afnamemomenten.

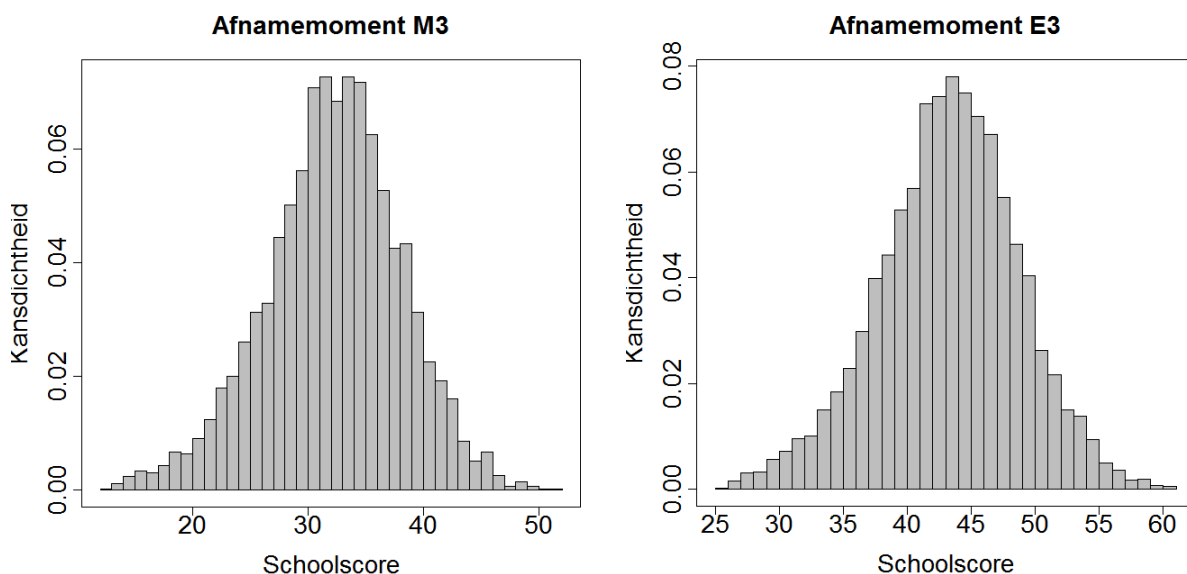
Tabel 4.7 Hernormeringsgegevens op leerlingniveau voor LVS-II Rekenen-Wiskunde medio 3 en einde 3

Tijd	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
medio 3	32.5	13.9	-.258	-.028	14.5	20.5	22.5	29.5	32.5	36.5	42.5	44.5
einde 3	43.4	13.1	-.258	-.031	26.5	32.5	34.5	40.5	43.5	47.5	52.5	54.5

Noot: K = Kurtosis, S = Skewness

Bij het hernormeren op schoolniveau gaan we te werk op de manier zoals beschreven in paragraaf 3.6. Dit betekent dat we eerst kijken naar de verdeling van schoolgemiddelden en dan een *intercept-only* multilevel model schatten in een bootstrap procedure. Bij het trekken van scholen en leerlingen houden we rekening met de gewichten die eerder zijn toegekend (zie Tabel 4.6). Figuur 4.2 laat de gewogen verdeling van schoolgemiddelden zien voor de eerste twee afnamemomenten van LVS-II *Rekenen-Wiskunde*. De verdelingen zijn gebaseerd op de gemiddelden van 6625 (medio 3) en 5914 (einde 3) scholen. Bij het schatten van het multilevel model en het berekenen van de percentielen veronderstellen we dat de schoolgemiddelden normaal verdeeld zijn. Deze aanname lijkt voor zowel medio groep 3 als eind groep 3 verdedigbaar.

Figuur 4.2 Verdeling van de schoolgemiddelden voor LVS-III Rekenen-Wiskunde



Tabellen 4.8 en 4.9 laten zien tot welke resultaten de bootstrap procedure heeft geleid bij de eerste twee afnamemomenten van LVS-II *Rekenen-Wiskunde*. We zien dat de uitkomsten erg stabiel zijn. Met de vermelde school- en leerlingaantallen had een kleiner aantal bootstrap replicaties in feite ook volstaan. Uit de intraklassecorrelatie en het design effect blijkt dat een multilevel analyse op beide afnamemomenten wenselijk is. De intraklassecorrelatie ligt zowel op medio groep 3 als eind groep 3 boven de .04 en het design effect is bij beide afnamemomenten duidelijk groter dan 2.

Tabel 4.8 Uitkomsten multilevel analyse voor LVS-II Rekenen-Wiskunde medio 3

Replicatie	N_y	N_p	γ_{00}	τ	σ	ρ	Design eff.
1	6625	172346	32.03	5.50	13.30	.15	4.65
2	6625	172971	32.11	5.59	13.31	.15	4.77
3	6625	170787	32.18	5.50	13.32	.15	4.62
4	6625	172585	32.14	5.46	13.34	.14	4.61
5	6625	172501	32.17	5.59	13.25	.15	4.78
...	...	...	...	...	...	...	...
15	6625	170244	32.17	5.53	13.30	.15	4.63
16	6625	170664	32.09	5.50	13.28	.15	4.62
17	6625	172815	32.09	5.47	13.32	.14	4.61
18	6625	171998	32.06	5.50	13.33	.15	4.62
19	6625	171447	31.97	5.36	13.32	.14	4.46
20	6625	169760	32.11	5.53	13.25	.15	4.64
Eindresultaat			32.14	5.51	13.30	.15	4.64

Tabel 4.9 Uitkomsten multilevel analyse voor LVS-II Rekenen-Wiskunde einde 3

Replicatie	N_y	N_p	γ_{00}	τ	σ	ρ	Design eff.
1	5914	150347	43.01	5.13	12.83	.14	4.37
2	5914	150334	43.27	5.09	12.79	.14	4.35
3	5914	149906	43.20	5.13	12.83	.14	4.36
4	5914	151282	43.20	5.02	12.86	.13	4.24
5	5914	151168	43.18	5.15	12.80	.14	4.41
...	...	...	...	...	...	...	...
15	5914	149659	43.25	5.09	12.80	.14	4.33
16	5914	151286	43.15	5.23	12.84	.14	4.49
17	5914	149568	43.09	5.09	12.84	.14	4.30
18	5914	151906	43.10	5.15	12.87	.14	4.41
19	5914	151197	43.06	5.05	12.81	.14	4.32
20	5914	150496	43.12	5.12	12.82	.14	4.35
Eindresultaat			43.15	5.10	12.83	.14	4.33

Op basis van het algemene gemiddelde γ_{00} en de niveau 2 standaarddeviatie τ die in Tabel 4.8 en 4.9 in de rij *Eindresultaat* staan, berekenen we de percentielen die horen bij de vaardigheidsindelingen A tot en met E, en I tot en met V. Tabel 4.10 geeft de hernormingsgegevens op schoolniveau voor LVS-II *Rekenen-Wiskunde*. Zoals we mochten verwachten is de spreiding op schoolniveau kleiner dan op leerlingniveau. De percentielen komen dus dichter bij elkaar te liggen. De afstanden zijn echter nog wel groot genoeg om scholen zinvol te classificeren in niveaus A tot en met E of I tot en met V. De nauwkeurigheid waarmee we scholen kunnen classificeren zal echter kleiner zijn dan de nauwkeurigheid waarmee we leerlingen kunnen classificeren. Ten opzichte van het normeringsonderzoek hebben ook de scholen, zoals we mochten verwachten op basis van de prestaties van de leerlingen, een behoorlijke sprong gemaakt in prestatie. De effectgrootte d is voor beide afnamemomenten vanwege de kleinere variantie op schoolniveau echter aanzienlijk groter dan eerder bij de leerlingen. Voor medio groep 3 is de effectgrootte gelijk aan 1.11 en

voor eind groep 3 gelijk aan 1.57. Op zowel leerling- als schoolniveau was bijstelling van de normen dus dringend gewenst.

Tabel 4.10 *Hernormeringsgegevens op schoolniveau voor LVS-II Rekenen-Wiskunde medio 3 en einde 3*

Tijd	M	SD	K	S	P10	P20	P25	P40	P50	P60	P75	P80
medio 3	32.1	5.5	----	----	25.1	27.5	28.4	30.7	32.1	33.5	35.9	36.8
einde 3	43.2	5.1	----	----	36.6	38.9	39.7	41.9	43.2	44.4	46.6	47.4

Noot: K = Kurtosis, S = Skewness

4.5 Perspectief

De analyses voor LVS-II *Rekenen-Wiskunde*, *Begrijpend lezen* en *Spelling* laten zien dat leerlingen de toetsen vandaag de dag beter maken dan tijdens het normeringsonderzoek. Daarom zijn de normen voor deze toetsen op 1 september 2013 herzien. Dat is eerder dan de herziening die al gepland was in het kader van de inhoudelijke update van de LVS-toetsen (LVS-III). Deze herziening van de normen heeft gevolgen voor het onderwijs. De prestaties van scholen en leerlingen worden met de invoering van nieuwe normen immers anders geduïd. Een leerling heeft een hogere score nodig om tot de beste 20 procent van zijn leeftijdgenoten te behoren dan voorheen. Een overschatting van het niveau van een leerling of school lijkt prettig, maar naast het feit dat de toetsing geen objectief en eerlijk beeld geeft, kan deze op langere termijn tot verrassingen leiden. De scores op de centrale eindtoets of in het voortgezet onderwijs kunnen bijvoorbeeld tegenvallen. Om deze reden gaat Cito de normen van alle toetsen elk jaar checken. De check impliceert **niet** dat de normen van de verschillende LVS-toetsen vanaf nu ook jaarlijks aangepast gaan worden. Dat hangt onder meer af van de prestatieverschuivingen die de komende jaren bij LVS-II en LVS-III te zien zijn. Daarnaast is nog onvoldoende duidelijk in hoeverre het vanuit praktisch en maatschappelijk oogpunt wenselijk en uitvoerbaar is om normeringen geregeld aan te passen. Cito doet daar momenteel onderzoek naar.

Vooruitlopend op het onderzoek beschrijven we in deze paragraaf ter illustratie drie scenario's die we kunnen volgen bij het checken en bijstellen van normeringen. In de beschrijving van de scenario's gaan we uit van de leerlingcohorten zoals die gedefinieerd zijn in Tabel 4.11.

Tabel 4.11 Definitie leerlingcohorten

Cohort	Groep					
	3	4	5	6	7	8
1						05/06
2					05/06	06/07
3				05/06	06/07	07/08
4			05/06	06/07	07/08	08/09
5		05/06	06/07	07/08	08/09	09/10
6	05/06	06/07	07/08	08/09	09/10	10/11
7	06/07	07/08	08/09	09/10	10/11	11/12
8	07/08	08/09	09/10	10/11	11/12	12/13
9	08/09	09/10	10/11	11/12	12/13	13/14
10	09/10	10/11	11/12	12/13	13/14	14/15
11	10/11	11/12	12/13	13/14	14/15	15/16
12	11/12	12/13	13/14	14/15	15/16	16/17
13	12/13	13/14	14/15	15/16	16/17	17/18

We zien in Tabel 4.11 dat leerlingen van cohort 8 in schooljaar 2007/2008 instromen in groep 3. Cohort 8 verlaat de basisschool in schooljaar 2012/2013. Voor de andere cohorten geldt hetzelfde principe. Een cohort heeft dus betrekking op een groep leerlingen die in een bepaald schooljaar instromen in het basisonderwijs in groep 3. Het is mogelijk om zo'n cohort exacter te omschrijven in termen van bijvoorbeeld geboortedatums. Een dergelijke verfijning laten we hier achterwege. Een verdere verfijning heeft in het licht van de opzet van Tabel 4.11 ook weinig zin, omdat een cohort gedurende zijn gang door het basisonderwijs verandert; sommige leerlingen doubleren en andere slaan een klas over. Niettemin mogen we aannemen dat de cohorten ondanks de verschuivingen over de jaren heen qua samenstelling globaal vergelijkbaar blijven.

Scenario 1: incidentele bijstelling

Dit scenario gaat ervan uit dat elk jaar, op basis van een nader te definiëren beslisregel, bepaald wordt of er reden is om de normen bij te stellen. In dat geval zal bij kleine prestatieverschuivingen besloten worden om de normen niet bij te stellen. Als op een gegeven moment blijkt dat de kleine prestatieverschuivingen in elk schooljaar samen (dat wil zeggen, cumulatief) tot één grote prestatieverschuiving leiden, zullen de normen wel bijgesteld worden. Een nadeel van deze aanpak is dat de prestatieverschuivingen eerst enige tijd genegeerd worden en dan in één klap zorgen voor een relatief grote verandering in de normen. Het praktijkvoorbeeld in paragraaf 4.4 heeft laten zien hoe groot de verandering kan zijn. Grote veranderingen in normeringen zijn vervelend en moeilijk uit te leggen aan het onderwijsveld. Bovendien heeft een incidentele verandering in de normen tot gevolg dat leerlingen in een bepaalde groep niet altijd met dezelfde leerlingpopulatie (of hetzelfde cohort) vergeleken worden.

Tabellen 4.12 en 4.13 laten aan de hand van een voorbeeld zien welke effect een driejaarlijkse hernormering zou hebben op het referentiecohort. In het voorbeeld veronderstellen we dat er in schooljaar 2006/2007 een toets op de markt komt (blauw gearceerd in Tabel 4.11) die in schooljaar 2005/2006 genormeerd is voor de groepen 3 tot en met 8. Als de normering voor een langere periode hetzelfde blijft, betekent dit dat cohorten 1 tot en met 6 de referentiegroepen vormen. In Tabel 4.12 zien we dat leerlingen in groep 3 vergeleken worden met leerlingen uit cohort 6, in groep 4 met leerlingen uit cohort 5, en in groep 8 met leerlingen uit cohort 1. Als de normering in schooljaar 2009/2010 en 2012/2013 wordt bijgesteld (grijs gearceerd in Tabel 4.11), verandert de situatie. Afhankelijk van het tijdstip waarop de leerlingen in een bepaalde groep zitten, is er een ander referentiecohort. In Tabel 4.13 zien we dat leerlingen die in

schooljaar 2006/2007 starten in groep 3 (cohort 7) gedurende de basisschoolperiode afwisselend vergeleken worden met leerlingen uit cohorten 4, 5 en 6. Leerlingen die in schooljaar 2008/2009 starten in groep 3 (cohort 9) worden daarentegen afwisselend vergeleken met leerlingen uit cohorten 6, 7 en 8. Het is de vraag in hoeverre het een probleem is dat het referentiecohort wisselt. De relatieve prestaties van groepen leerlingen zullen in ieder geval niet gemakkelijk over de tijd te vergelijken zijn. Daarnaast zullen scholen en leerkrachten het vermoedelijk lastig vinden om te accepteren dat de interpretatie van een vaardigheidsscore die is vastgesteld niet definitief is.

Tabel 4.12 Referentiecohort bij een vaste normering over de tijd

Cohort	Groep					
	3	4	5	6	7	8
7	6	5	4	3	2	1
8	6	5	4	3	2	1
9	6	5	4	3	2	1
10	6	5	4	3	2	1
11	6	5	4	3	2	1
12	6	5	4	3	2	1

Tabel 4.13 Referentiecohort bij een driejaarlijkse bijstelling

Cohort	Groep					
	3	4	5	6	7	8
7	6	5	4	6	5	4
8	6	5	7	6	5	7
9	6	8	7	6	8	7
10	9	8	7	9	8	7
11	9	8	10	9	8	10
12	9	11	10	9	11	10

Scenario 2: jaarlijkse bijstelling (cohortnormering)

Dit scenario gaat ervan uit dat de normen van een toets elk jaar opnieuw vastgesteld worden. Op deze manier krijgt elk leerlingcohort een eigen normering. Tabel 4.14 laat zien hoe dit scenario in de praktijk zou uitpakken. De scores van leerlingen die in schooljaar 2006/2007 (cohort 7) starten in groep 3 worden vergeleken met de scores van de leerlingen die een jaar eerder in groep 3 zaten (cohort 6). Voor het volgende schooljaar geldt exact hetzelfde. De scores van de leerlingen worden opnieuw vergeleken met de scores van het vorige cohort. Waar leerlingen bij een vaste normering over de tijd, en zeker bij een normering die incidenteel wordt bijgesteld, steeds met een ander leerlingcohort worden vergeleken, is er bij een zogeheten cohortnormering dus sprake van een vast referentiekader. De leerlingen die op een bepaald moment in de tijd met een eigen (maatschappelijk en historisch bepaalde) stand van zaken (context) instromen in het onderwijs worden vergeleken met leerlingen die een jaar eerder zijn ingestroomd.

Tabel 4.14 Referentiecohort bij een jaarlijkse bijstelling

Cohort	Groep					
	3	4	5	6	7	8
7	6	6	6	6	6	6
8	7	7	7	7	7	7
9	8	8	8	8	8	8
10	9	9	9	9	9	9
11	10	10	10	10	10	10
12	11	11	11	11	11	11

Een belangrijk voordeel van een cohortnormering is dat er niet elk jaar op basis van een arbitraire beslisregel bepaald hoeft te worden of de normen bijstelling behoeven. Ook is naar scholen en leerkrachten gemakkelijker uit te leggen waarom normen bijgesteld worden. Van relatief grote (plotselinge) veranderingen in de normen zal bovendien geen sprake zijn. Naast een aantal voordelen brengt een cohortnormering echter ook een aantal problemen met zich mee die voornamelijk praktisch van aard zijn, maar daarom niet minder ingrijpend. Bijvoorbeeld: hoe kunnen scholen lange termijn doelen stellen voor prestatieverbetering en hoe kunnen ze prestaties van opeenvolgende cohorten op hun eigen school met elkaar vergelijken? Voor rapportagesystemen vraagt cohortnormering een andere verwerkingssystematiek en mogelijk ook nieuwe rapportagevormen.

Scenario 3: geen bijstelling

In dit scenario houden we vast aan de oorspronkelijke initiële normering, ongeacht de omvang van de verschuivingen die zich voordoen, per jaar, dan wel cumulatief over een reeks van jaren. Dit scenario komt overeen met het oorspronkelijke uitgangspunt dat vernieuwing van de normen pas aan de orde is bij invoering van een nieuwe generatie toetsen. In dit scenario worden leerlingen steeds vergeleken met één referentiegroep die in de tijd gezien steeds verder in het verleden komt te liggen. Ook hier blijft het referentiekader dus ongewijzigd, maar raakt het in steeds hogere mate verouderd. Een belangrijk voordeel van dit scenario is dat vaardigheidsscores een eenduidige interpretatie hebben en we blijven aansluiten bij de wijze waarop het onderwijs momenteel is ingericht. Streefdoelen kunnen zonder problemen beschreven worden in termen van relatieve normscores en ook in beoordelingskaders kan men uit blijven gaan van relatieve grenzen. Een groot nadeel is echter dat we de relatieve prestaties van leerlingen in toenemende mate overschatten en we mogelijk onjuiste conclusies trekken over het (relatieve) prestatieniveau van leerlingen en scholen.

Kiezen uit drie scenario's

Tot nu toe was een regelmatige (eens per 6-8 jaar) vernieuwing van de inhoud van de toetsen met een bijbehorende nieuwe normering gebruikelijk. Daarmee kwam Cito ruimschoots tegemoet aan de criteria die de COTAN hanteert met betrekking tot de veroudering van normen. Bovendien was het op deze manier mogelijk om ook de toetsinhoud bij de tijd te houden (vergelijk de verschillende generaties in het LVS). Nu we weten uit analyses van de beschikbare gegevens dat een zekere slijtage van de normen aanwezig is, die zich wellicht sneller voltrekt dan vermoed, is het wellicht nodig om een meer frequente herziening van de normen aan te houden. Welk scenario men ook kiest, er is steeds sprake van voor- en nadelen, waarbij een voordeel van het ene scenario in de regel een nadeel van een ander scenario impliceert. Bij elk scenario zal bekeken moeten worden hoe men omgaat met allerlei uitzonderingssituaties (hoe ga je om met leerlingen die doubleren, niet in een specifieke jaargroep zitten, een andere toets maken of dezelfde toets twee keer maken) en wat de consequenties zijn voor bijvoorbeeld het Computerprogramma LOVS en de wijze van rapporteren. Cito is bezig om alle praktische en meer fundamentele consequenties van verschillende scenario's in kaart te brengen en wil hierbij niet over één nacht ijs gaan.

Vanuit theoretisch oogpunt lijkt het tweede scenario het gemakkelijkst te verdedigen. In dat geval worden de normen jaarlijks bijgesteld, ongeacht de omvang van de prestatieverschuivingen die zich voordoen. In afwachting van de empirische gegevens onderzoekt Cito momenteel de haalbaarheid van dit scenario naast die van andere scenario's. Mogelijk brengt een cohortnormering of welke andere tussentijdse aanpassing van de normen dan ook meer nadelen dan voordelen met zich mee. De haalbaarheid is onder meer afhankelijk van de flexibiliteit van de rapportages die gewenst zijn in een leerlingvolgsysteem en de vraag of een regelmatige bijstelling van normen kan rekenen op draagvlak onder leerkrachten en scholen. Omdat de gevolgen van de invoering van een cohortnormering of een andere vorm van regelmatige hernormering op dit moment nog niet goed in hun volle omvang kunnen worden overzien, kiest Cito er op dit moment voor om de normen voor LVS-II *Rekenen-Wiskunde*, *Begrijpend lezen* en *Spelling* eenmalig te herzien en de normen te blijven monitoren. Een definitieve keuze is dus nog niet gemaakt.

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.
- Bloom, H.S., Bos, J.M., & Lee, S., (1999) Using cluster random assignment to measure program impacts. *Evaluation Review*, 23, 445-469.
- Bloom, H.S., Richburg-Hayes, L., & Black, A.R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, 29, 30-59.
- Brophy, J., & Ames, C. (2005). *NAEP testing for twelfth graders: Motivational issues*. A paper prepared for the National Assessment Governing Board, August.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Erlbaum.
- 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.
- Eggen, T.J.H.M., (2004). *Contributions to the theory and practice of computerized adaptive testing*. Arnhem: Cito.
- Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.
- Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam, NIP/COTAN.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hedges, L.V., & Hedberg, E.C. (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis*, 29, 60-87.
- Hemker, B.T. (2012). *What's in it for me?* Paper presented at the 8th Conference of the International Test Commission (ITC). Amsterdam, July 3-5 2012.
- 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.
- Hox, J.J., & Maas, C.J.M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157-174.
- Hox, J.J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.

- 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., Wouda, J. & Verbruggen, I. (in voorbereiding). *Wetenschappelijke verantwoording van LVS-III Rekenen-Wiskunde*. Arnhem: Cito.
- Keuning, J. & Hilte, M., Weekers, A. (2014). Begrijpend leesprestaties onderzocht: Een analyse op basis van Cito dataretour. *Tijdschrift voor orthopedagogiek*, 1, 2-13.
- Keuning, J. (2011). *Normeren op schoolniveau met Cito dataretour*. Arnhem: Cito.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.
- Koretz, D. & Béguin, A. (2010). Self-Monitoring Assessments for Educational Accountability Systems. *Measurement*, 8, 92-109.
- Maris, G. & Bechger, T.M. (2005). An introduction to the DA-T gibbs sampler for the two-parameter logistic (2pl) model and beyond. *Psicologica*, 26, 327-352.
- Marsman, M., Maris, G., Bechger, T.M., & Glas, C.A.W. (2011). *A Conditional Composition Algorithm for Latent Regression*. Arnhem: Cito.
- Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.
- Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P.V. Marsden (Ed.), *Sociological Methodology* (pp. 267-316). Washington, DC: American Sociological Association.
- Onderwijsinspectie (2011):
<http://www.onderwijsinspectie.nl/actueel/nieuwsbrieven/details/Vragen+en+antwoorden+over+de+wijzigingen+in+de+opbrengstbeoordeling.html>
- Popham, W.J. (1999). Where large scale assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice*, 18, 13-17.
- Sanders, P. & Verstralen, H. (2010). Het beoordelen van toetsscores. In P. Sanders (ed.), *Toetsen op school* (pp. 143-155). Arnhem: Cito.
- Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R.L. Brennan (Ed.) *Educational Measurement* (pp. 307-353). Washington, DC: American Council on Education.
- Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics*, 33, 62-87
- Shepard, L. A. (1990). *Inflated Test Score Gains: is it Old Norms or Teaching to the Test? CSE Technical Report 307*. University of Colorado, UCLA Center for Research on Evaluation, Standards, and Student Testing.
- Snijders, T.A.B. & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 237-260.
- 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.

Van Berkel, S., Hilde, M., Engelen, R. Kamphuis, F., Kleintjes, F. & Krom, R. (2010). *Wetenschappelijke verantwoording van de LOVS-toetsen Woordenschat groep 3 t/m 5*. Arnhem: Cito.

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

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

Verhelst, N.D., & Glas, C.A.W. (1995). 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. & Verstralen, H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL): Theory and a computer program*. Arnhem: Cito.

Wise, S.L. & C.E. DeMars (2005), Low Examinee effort in Low Stakes Assessments: Problems and Potential Solutions. *Educational Assessment*, 10, p. 1-17.

Wu. M. (2005). The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation*, 31, 114-128.

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