Begrijpend leesprestaties onderzocht - Een analyse op basis van Cito dataretour

SAMENVATTING

In dit artikel wordt onderzocht hoe de begrijpend leesprestaties van basisschoolleerlingen zich hebben ontwikkeld in de periode 2006/2007 tot en met 2011/2012. In het onderzoek is gebruikgemaakt van de toetsresultaten van ruim 1.3 miljoen leerlingen op de LVS-toetsen Begrijpend lezen. Analyses laten zien dat de prestaties van leerlingen binnen een afnamemoment de laatste jaren niet zijn veranderd. De begrijpend leesprestaties nemen in de verschillende cohorten wel consequent en in even sterke mate toe vanaf het eerste afnamemoment aan het einde van groep 3 tot en met het laatste afnamemoment medio groep 8. Tevens komt naar voren dat de leerlingen die aan het normeringsonderzoek deelnamen soms iets zwakker of beter presteerden dan de leerlingen die de toetsen in schooljaar 2011/2012 maakten. Als gevolg van dit verschil worden de relatieve prestaties van leerlingen soms onder- of overschat. Door gebruik te maken van een *embedded field* toetsdesign en door de normen regelmatig bij te stellen, kan dit probleem gedeeltelijk ondervangen worden.

we proberen op basis van begrijpend lezen uit te leggen hoe toetsresultaten kunnen worden weergegeven en geïnterpreteerd, en met welke onzekerheden bij de interpretatie rekening moet worden gehouden. Daarnaast geven we aan waarom een regelmatige check en actualisering van normgegevens wenselijk is. Aangezien dit laatste volledig nieuw is, zullen veel van de lezers hiermee te maken krijgen

1 Introductie

In een recent wetsvoorstel dat is ingediend door de Minister van Onderwijs, Cultuur en Wetenschap (OCW) worden scholen in het primair onderwijs verplicht om gebruik te maken van een leerlingvolgsysteem. Van scholen wordt verwacht dat zij in hun onderwijs uitgaan van leerstandaarden en lesdoelen, op basis van een leerlingvolgsysteem informatie verzamelen over het leerproces, deze informatie vastleggen voor nadere analyse en interpretatie, en op basis hiervan beslissingen nemen over het vervolg van het onderwijs. Met de LVS-toetsen uit het Cito Volgsysteem primair en speciaal onderwijs kunnen de schoolvorderingen van leerlingen systematisch worden gevolgd. Leerling-, groeps- en schooloverzichten geven de leerkracht, de intern begeleider en de school houvast bij het plannen en evalueren van het onderwijsaanbod. Hoewel de LVS-toetsen van Cito in tegenstelling tot methodegebonden toetsen een onafhankelijk en objectief beeld proberen te geven van de prestaties op leerling-, groeps- en schoolniveau, roepen de uitkomsten ervan met enige regelmaat vragen op. Soms blijken leerkrachten en scholen het lastig te vinden om de uit-

komsten om te zetten naar beslissingen en bijbehorende verbetermaatregelen (Ledoux, Blok & Bogaard, 2009; Reezigt, Houtveen & Grift, 2002). In andere gevallen stemmen de uitkomsten tot teleurstelling omdat leerkrachten en intern begeleiders hogere verwachtingen hadden. Dergelijke vragen kunnen via een gerichte training in het gebruik van leerlingvolgsystemen beantwoord worden. De validiteit van het leerlingvolgsysteem van Cito wordt af en toe ook betwist, soms door scholen maar vaker door onderwijsadviesdiensten en/of wetenschappelijk onderzoekers.

Recent hebben Keijsers, Van der Horst en Wolgram (2012) melding gemaakt van een dip in de begrijpend leesprestaties van leerlingen in groep 6 van de basisschool. Waar leerlingen in groep 4 en 5 nog goed presteerden, lijken de prestaties op het afnamemoment *medio 6* opeens systematisch terug te lopen. De verklaring voor deze zogenaamde 'leesdip' wordt in eerste instantie gezocht in het onderwijs dat scholen aan leerlingen aanbieden. Na een nadere analyse achten Keijsers et al. (2012) een dergelijke verklaring echter niet erg aannemelijk, omdat de teruggang in prestaties bij vrijwel alle scholen in hun onderzoeksgroep zichtbaar was. Bovendien is er geen reden om aan te nemen dat scholen in groep 4 en 5 wel kwalitatief hoogwaardig onderwijs aanbieden en in groep 6 niet meer. Vermoedelijk hebben Keijsers et al. (2012) gelijk in hun afweging. Het is de vraag waar de sterke teruggang in prestaties die Keijsers et al. (2012) signaleren dan wel door veroorzaakt wordt. Wordt het opmerkelijke onderzoeksresultaat bijvoorbeeld veroorzaakt doordat de onderzoeksgroep van Keijsers et al. (2012) niet representatief was met betrekking tot relevante achtergrondkenmerken als *leeftijd*, *etniciteit*, *sekse* en *regio*? (zie bijvoorbeeld, Evers, Lucassen, Meijer & Sijtsma, 2010). Of zijn er, zoals Keijsers et al. (2012) stellen, vergissingen gemaakt bij de ontwikkeling en normering van de LVS-toetsen Begrijpend lezen? Het is belangrijk hier duidelijkheid over te geven, omdat leerkrachten en intern begeleiders de resultaten gebruiken in hun onderwijs, en de onderwijsinspectie de begrijpend leesresultaten meeneemt in de schoolevaluatie (zie Inspectie van het onderwijs, 2012).

Het afgelopen jaar heeft Cito onderzoek gedaan naar de prestaties van leerlingen in de groepen 3 tot en met 8 op het gebied van begrijpend lezen. In dat onderzoek is gebruikgemaakt van Cito dataretour. Via Cito dataretour sturen basisscholen jaarlijks op vrijwillige basis hun toetsresultaten naar Cito voor onderzoeksdoeleinden. Het opsturen van de resultaten is volledig geautomatiseerd via het Computerprogramma LOVS. Verreweg de meeste basisscholen geven dan ook gehoor aan de oproep die Cito jaarlijks doet. Via Cito dataretour zijn inmiddels toetsgegevens beschikbaar van honderdduizenden leerlingen op verschillende toetsen. Via statistische weging of speciale *sampling*-technieken kan ervoor gezorgd worden dat het databestand dat samengesteld wordt uit Cito dataretour representatief is voor de Nederlandse leerlingpopulatie. In de volgende paragraaf laten we eerst zien hoe toetsresultaten op het gebied van begrijpend lezen beoordeeld en geïnterpreteerd kunnen worden. Dat doen we in eerste instantie aan de hand van de werkwijze die in het leerlingvolgsysteem van het Cito gehanteerd wordt. We laten ook enkele alternatieve werkwijzen zien. Daarna gaan we in op de begrijpend leesprestaties van de leerlingen. We laten zien hoe de leerlingen tijdens het normeringsonderzoek presteerden op de LVS-toetsen Begrijpend lezen en brengen in kaart in hoeverre de begrijpend leesprestaties in de periode 2006/2007 tot en met 2011/2012 veranderd zijn. De uitkomsten op basis van Cito dataretour worden gerelateerd aan de bevindingen van Keijsers et al. (2012).

Kortom, we proberen op basis van begrijpend lezen uit te leggen hoe toetsresultaten kunnen worden weergegeven en geïnterpreteerd, en met welke onzekerheden bij de interpretatie rekening moet worden gehouden. Daarnaast geven we aan waarom een regelmatige check en actualisering van normgegevens wenselijk is. Aangezien dit laatste volledig nieuw is, zullen veel van de lezers hiermee te maken krijgen.

2 Beoordelen en interpreteren van begrijpend leesscores

De afname van een LVS-toets Begrijpend lezen levert in feite niets meer op dan een aantal scorepunten op een toets. Als een leerling bijvoorbeeld 31 van de 50 opgaven correct weet te beantwoorden en elke opgave telt even zwaar mee, dan is de begrijpend leesscore voor de leerling gelijk aan 31. We weten dan nog niet of een begrijpend leesscore van 31 een goede of een slechte prestatie is. De score krijgt pas betekenis door deze te vergelijken met een norm. De score kunnen we beoordelen door:

1 de score te vergelijken met de scores van andere leerlingen uit de doelgroep
2 de score te vergelijken met scores die de leerling bij eerdere afnamemomenten behaald heeft, of
3 de score 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 en in het derde geval van absoluut normeren. In het leerlingvolgsysteem van Cito 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 aselecte steekproef van ten minste 400 leerlingen die binnen de doelgroep vallen (Evers et al., 2010). In de derde en laatste stap wordt de normschaal geconstrueerd. In de literatuur worden verschillende normschalen voorgesteld. We bespreken hier de percentielschaal, de genormaliseerde schaal en de ontwikkelingsschaal.

De percentielschaal is vermoedelijk de eenvoudigste normschaal. De cumulatieve scoreverdeling vormt het uitgangspunt bij de constructie van deze normschaal. Een percentielschaal verdeelt de scoreverdeling in 100 delen van gelijke grootte. In het eerste percentiel zitten de 1 procent laagste scores en in het honderdste percentiel de 1 procent hoogste scores (Sanders & Verstralen, 2010). In de praktijk kan een begrijpend leesscore van 31 bijvoorbeeld in het 65ste percentiel vallen. We weten dan dat 65 procent van de leerlingen in Nederland hetzelfde of lager scoort. Hoewel een percentielscore eenvoudig te interpreteren is, kleven er ook nadelen aan. Bij scores rond het gemiddelde worden kleine verschillen in prestatie in de percentielscore namelijk uit elkaar getrokken. Daarom wordt de percentielschaal vaak ingedikt tot een deciel-, kwintiel- of kwartielschaal. Dat is bij de LVS-toetsen Begrijpend lezen ook gedaan. Niveaucategorieën I tot en met V gaan bijvoorbeeld uit van een kwintielschaal met vijf gelijke groepen: I = ver boven het gemiddelde (20 procent), II = boven het gemiddelde (20 procent), III = de gemiddelde groep leerlingen (20 procent), IV = onder het gemiddelde (20 procent), en V = ver onder het gemiddelde (20 procent). Normschalen met relatief weinig categorieën zijn over het algemeen gemakkelijker te gebruiken dan normschalen met relatief veel categorieën. Bovendien zijn ze ook toepasbaar bij kortere toetsen en speelt schijnnauwkeurigheid een minder prominente rol.

De genormaliseerde schaal kan gebruikt worden als alternatief voor de percentielschaal. Bij dit type schaal worden scores op basis van het gemiddelde en de standaarddeviatie via een lineaire transformatie omgezet naar een normscore. De transformatie wordt zo gekozen dat de prestatie gemakkelijk vergeleken kan worden met de prestaties van andere leerlingen in Nederland. Een veel voorkomende genormaliseerde schaal is de Z-schaal met een gemiddelde van 0 en een standaarddeviatie van 1. Een positieve Z-score wijst op een prestatie boven het gemiddelde. Een negatieve Z-score wijst op een prestatie beneden het gemiddelde. Hoe hoger (positief) of lager (negatief) de Z-score, hoe meer de prestatie van het gemiddelde afwijkt. Naast de Z-schaal wordt ook gebruikgemaakt van de IQ-schaal met een gemiddelde van 100 en een standaarddeviatie van 15, en de T-schaal met een gemiddelde van 50 en standaarddeviatie van 10. Keijsers et al. (2012) gaan in hun beschrijving van de begrijpend leesprestaties van leerlingen uit van de laatstgenoemde schaal. Hoewel de genormaliseerde schaal gemakkelijk is toe te passen en een aantal voordelen met zich meebrengt in vergelijking met de percentielschaal, zien we dit type schaal bijna nooit terug in leerlingvolgsystemen. Daar is een eenvoudige verklaring voor. In een leerlingvolgsysteem worden leerlingen herhaald gemeten. Bij elk afnamemoment is er sprake van een ander gemiddelde en een andere standaarddeviatie. Als we de lineaire transformatie per afnamemoment kiezen op basis van het geobserveerde gemiddelde en de bijbehorende standaarddeviatie, kunnen we geen groei meer waarnemen. Op elk afnamemoment verwachten we immers exact hetzelfde gemiddelde (bijvoorbeeld 50) en exact dezelfde standaarddeviatie (bijvoorbeeld 10). Een grafiek met de afnamemomenten op de x-as en de toetsresultaten op de y-as krijgt bij een transformatie die specifiek is voor een afnamemoment dus een bijzondere interpretatie: het is een afwijkingsgrafiek en geen groeigrafiek.

Ten slotte kunnen scores van leerlingen op het gebied van begrijpend lezen beoordeeld worden via een zogeheten ontwikkelingsschaal. Bij toepassing van een ontwikkelingsschaal vergelijken we een leerling niet met andere leerlingen in Nederland, maar met zichzelf. Ontwikkelingsschalen gaan ervan uit dat leerlingen herhaald gemeten worden. Bij een meting wordt bij voorkeur gebruikgemaakt van een toets die qua moeilijkheidsgraad aansluit bij het vaardigheidsniveau van de leerling (Wilson, 2005). Dit betekent dat bij de verschillende afnamemomenten in de tijd verschillende toetsen worden gebruikt. De scores die leerlingen behalen op deze toetsen zijn niet direct met elkaar te vergelijken. Want is een leerling die op afnamemoment *medio 4* 25 opgaven goed maakt op de ene toets en op afnamemoment *einde 4* 27 opgaven goed maakt op een andere toets nu vooruitgegaan? In leerlingvolgsystemen is het dan ook niet betekenisvol om de (absolute) groei van leerlingen te volgen met een set losse toetsen. Het is beter om gebruik te maken van meetschalen waarmee de groei van leerlingen gevolgd kan worden onafhankelijk van de toetsopgaven die bij een toetsafname aan de leerlingen wordt voorgelegd. Dergelijke meetschalen kunnen onder andere geconstrueerd worden met behulp van meetmodellen uit de item respons theorie (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991). Als het meetmodel geldt voor een verzameling opgaven kunnen de scores van de leerlingen via het onderliggende meetmodel gecorrigeerd worden voor de moeilijkheidsgraad van de toets. De gecorrigeerde scores worden in de praktijk meestal vaardigheidsscores genoemd. Bij de ontwikkeling van de LVS-toetsen Begrijpend lezen is deze werkwijze toegepast. De begrijpend leesprestaties worden met de LVS-toetsen van Cito dus gemeten op basis van een ontwikkelingsschaal en niet via een losse set toetsen. Het voordeel is dat de groei van individuele leerlingen ook daadwerkelijk tot uitdrukking komt in de toetsresultaten. Er is sprake van een vooruitgang als de score *hoger* is dan de sco-

re op het voorgaande afnamemoment. Er is sprake van een achteruitgang als de score *lager* is dan de score op het voorgaande afnamemoment. Bij toepassing van de percentielschaal of de genormaliseerde schaal wijst een hogere of lagere score niet altijd op een daadwerkelijke voor- of achteruitgang in prestatie. Een hogere of lagere score wijst alleen op een verschuiving in de relatieve positie van een leerling ten opzichte van de normgroep. Vanzelfsprekend kan de percentielschaal (of eventueel de genormaliseerde schaal) naast de ontwikkelingsschaal gebruikt worden. De percentielschaal geeft dan informatie over de prestaties van de leerling in vergelijking met andere leerlingen in Nederland. De ontwikkelingsschaal laat zien hoeveel een individuele leerling groeit in de tijd en kan gebruikt worden om leerdoelen te formuleren en evalueren.

3 Prestaties van leerlingen op het gebied van begrijpend lezen

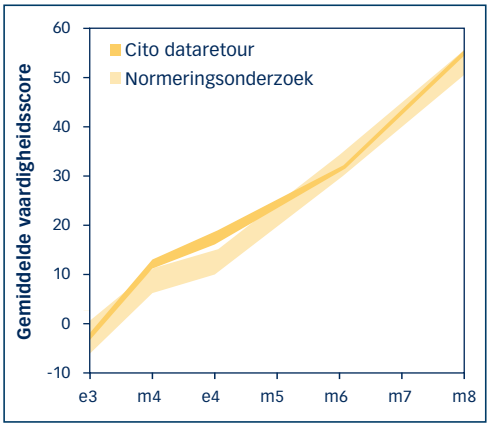
De begrijpend leesprestaties van leerlingen zijn geanalyseerd met behulp van Cito dataretour. Er is gebruikgemaakt van gegevens uit de periode van 2006/2007 tot en met 2011/2012 voor afnamemomenten *einde 3* tot en met *medio 8*. De hoeveelheid beschikbare gegevens per afnamemoment varieerde van $N = 12.679$ voor afnamemoment *medio 4* in schooljaar 2006/2007 tot $N = 80.507$ voor afnamemoment *einde 4* in schooljaar 2008/2009. Gemiddeld gezien waren er per afnamemoment in een schooljaar resultaten beschikbaar van 49.600 leerlingen. De data die verzameld zijn tijdens het normeringsonderzoek zijn ook meegenomen in de analyses. In totaal hebben 5.090 leerlingen deelgenomen aan dit onderzoek. Het deelnemersaantal was met $N = 443$ het kleinst voor afnamemoment *einde 3* en met $N = 988$ het grootst voor afnamemoment *medio 7*. In de eerste stap is via een cross-cohort analyse nagegaan in hoeverre de begrijpend leesprestaties van leerlingen in de afgelopen jaren veranderd zijn. In de analyse zijn álle leerlingen meegenomen die tot de doelgroep van een bepaald afnamemoment behoorden. Bij de analyse voor bijvoorbeeld afnamemoment *medio 7* zijn dus ook de resultaten van de leerlingen gebruikt die de toets bedoeld voor een hoger (*medio 8*) of lager (*medio 6*) leerjaar hebben gemaakt. Daarnaast is er in de analyse voor gekozen om niet te rekenen met de geschatte vaardigheidsscores van de leerlingen, omdat deze scores mogelijk leiden tot *bias* in de resultaten (Wu, 2005). In plaats daarvan is voor elke leerling op basis van de toetsscore en het item respons model een *plausible value* gegenereerd (Béguin & Glas, 2001; Maris & Bechger, 2005; Marsman, Maris, Bechger & Glas, 2011; Mislevy, 1991). *Plausible values* geven 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 is onwenselijk om geschatte vaardigheidsscores te analyseren.

Tabel 1 laat zien hoe cohorten leerlingen de laatste jaren gepresteerd hebben op de LVS-toetsen Begrijpend lezen. We zien een zeer constant prestatieniveau in de tijd voor de verschillende afnamemomenten. Bij afnamemoment *medio 4* zien we bijvoorbeeld dat de leerlingen in schooljaar 2006/2007 een gemiddelde vaardigheidsscore van 13.13 behaalden. In de volgende schooljaren zien we nauwelijks verschuivingen. In schooljaar 2011/2012 was de gemiddelde vaardigheidsscore bijvoorbeeld gelijk aan 12.92. Tabel 1 laat tevens zien dat de leerlingen tijdens het normeringsonderzoek (NO) soms zwakker presteerden. Vooral op afnamemomenten *medio 4*, *einde 4* en *medio 5* scoorden de leerlingen opmerkelijk laag in vergelijking met de leerlingen die aan de échte afnamen hebben deelgenomen. Het prestatieverschil bedraagt ongeveer .25 standaarddeviatie.

Tabel 1 **Begrijpend leesprestaties in de periode 2006/2007 – 2011/2012**

			Schooljaar					
Tijd	**Stat.**	**NO**	**06/07**	**07/08**	**08/09**	**09/10**	**10/11**	**11/12**
e3	*M*	-2.40	-2.36	-2.34	-2.05	-2.19	-1.95	-1.67
	SD	15.63	13.85	13.73	13.75	13.67	13.65	13.74
m4	*M*	8.91	13.13	11.98	12.46	12.45	12.58	12.92
	SD	14.29	12.03	12.46	12.31	12.29	12.22	12.15
e4	*M*	13.18	17.44	17.08	17.75	17.96	18.29	18.42
	SD	15.28	13.13	13.18	13.02	12.95	12.86	12.97
m5	*M*	22.32	----	25.47	25.35	25.46	25.51	25.66
	SD	13.91	----	12.69	12.62	12.44	12.42	12.37
m6	*M*	33.13	----	----	32.21	31.98	32.11	32.44
	SD	13.24	----	----	11.54	11.50	11.38	11.38
m7	*M*	44.54	----	----	----	44.47	44.66	44.82
	SD	14.44	----	----	----	12.72	12.68	12.67
m8	*M*	53.80	----	----	----	----	55.35	56.02
	SD	15.21	----	----	----	----	15.58	15.58

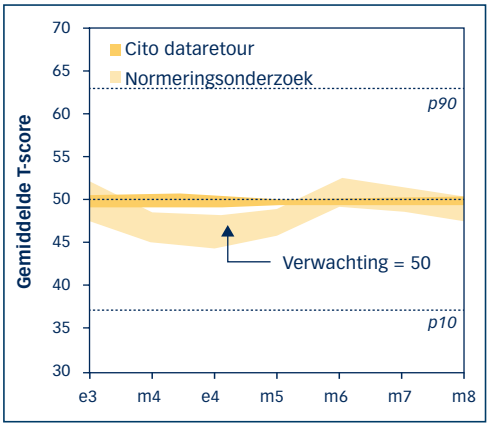
In een vervolgstap is de groei van leerlingen op het gebied van begrijpend lezen in kaart gebracht vanaf het einde van groep 3 tot en met het midden van groep 8. Figuur 1 laat per afnamemoment het gemiddelde resultaat zien. In Figuur 1 is rekening gehouden met de onzekerheid die bij het gemiddelde hoort. Dat is belangrijk, omdat elke steekproef onzekerheid met zich meebrengt. Die onzekerheid is in steekproeven waarin leerlingen genest zijn binnen scholen over het algemeen groter dan in *random* steekproeven uit de totale leerlingpopulatie (Cochran, 1977). Bij de berekening van de onzekerheid is dan ook rekening gehouden met deze geneste structuur van de data. We zien dat de begrijpend leesprestaties gestaag toenemen gedurende de basisschoolperiode. Leerlingen maken relatief veel groei door aan het begin van de basisschoolperiode, dan neemt de groei iets af, en tegen het einde van de basisschoolperiode is de groei weer iets groter. De resultaten van de leerlingen die hebben deelgenomen aan het normeringsonderzoek verschillen statistisch gezien vaak niet van de resultaten van de leerlingen die de toets hebben gemaakt als een échte afname, na uitgave van de LVS-toetsen Begrijpend lezen. De betrouwbaarheidsintervallen voor Cito dataretour en het normeringsonderzoek overlappen immers bij veel afnamemomenten. Bij afnamemomenten *medio 4*, *einde 4* en *medio 5* is er wel sprake van een relevant prestatieverschil. Dat is ook zichtbaar in Figuur 1 doordat daar geen overlap is tussen de weergegeven betrouwbaarheidsintervallen.



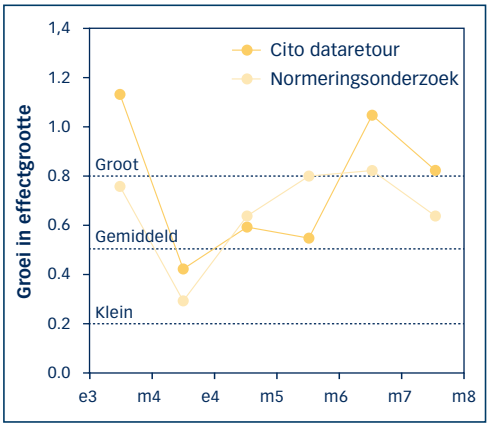
Figuur 1 **Vaardigheidsgroei gedurende de basisschoolperiode bij begrijpend lezen**

Figuur 1 geeft inzicht in de begrijpend leesprestaties van leerlingen op een otwikkelingsschaal. Zoals al eerder is aangegeven kunnen toetsresultaten ook anders weergegeven worden. Figuur 2 laat dezelfde resultaten zien als Figuur 1, maar dan op de T-schaal met Cito dataretour als referentiegroep. Wederom is ook de onzekerheidsmarge weergegeven. Op het eerste gezicht lijkt er sprake te zijn van een dip op

de afnamemomenten *medio 4*, *einde 4* en *medio 5*. Een dergelijke conclusie volgde eerder niet uit Figuur 1. Deze conclusie is dan ook onjuist. Figuur 2 geeft de begrijpend leesprestaties namelijk niet weer in een groeigrafiek, maar in een afwijkingsgrafiek. Door de toetsresultaten per afnamemoment via een lineaire transformatie over te zetten naar de T-schaal verwachten we op elk afnamemoment een gemiddelde van 50 en een standaarddeviatie van 10. Een T-score groter dan 50 betekent dat een leerling beter presteert dan de gemiddelde leerling in de referentiegroep. Een T-score kleiner dan 50 betekent dat een leerling zwakker presteert dan de gemiddelde leerling in de referentiegroep. Met dit in het achterhoofd zien we in Figuur 2 net als in Figuur 1 dat leerlingen tijdens het normeringsonderzoek op sommige afnamemomenten gemiddeld zwakker presteerden dan leerlingen in Cito dataretour. De conclusie die volgt uit Figuur 2 is dan ook exact dezelfde als de conclusie die volgt uit Figuur 1. Het is duidelijk dat op basis van Figuur 2 geen uitspraak gedaan kan worden over de mate van groei of achteruitgang in prestaties gedurende de basisschoolperiode.



Figuur 2 **Begrijpend leesprestaties in de proeftoetsing vergeleken met Cito dataretour**



Figuur 3 **Mate van groei bij begrijpend lezen uitgedrukt in effectgroottes**

Een nadeel van Figuren 1 en 2 is dat ze geen informatie geven over de sterkte van de groei die leerlingen doormaken. In Figuur 1 kunnen we niet zien of een groei van 10 punten groot of klein is. In Figuur 2 wordt de sterkte van de groei uitsluitend relatief gekwantificeerd. Effectgroottes geven wel weer hoe sterk leerlingen in absolute zin groeien. In Figuur 3 worden de prestatieverschillen tussen opeenvolgende afnamemomenten uitgedrukt in een effectgrootte. Een van .20 kan beschouwd worden als een klein effect, een van .50 als een gemiddeld effect, en een van .80 als een groot effect (Cohen, 1988). Op het eerste gezicht lijkt er sprake te zijn van een sterke teruggang in prestatie tussen afnamemomenten *medio 4* en *einde 4*. Daarna lijken de prestaties van leerlingen zich langzaam te herstellen. Een dergelijke conclusie is echter onjuist. Figuur 3 geeft namelijk weer hoe sterk de groei is tussen opeenvolgende afnamemomenten. We zien dat de groei die leerlingen doormaken in het midden van de basisschoolperiode wat kleiner is (rond .50) dan aan het begin en het einde van de basisschoolperiode (rond 1.00). Van een teruggang in prestatie is nergens sprake. De effectgrootte zou namelijk negatief zijn als de gemiddelde prestatie bij een afnamemoment lager zou zijn dan de gemiddelde prestatie bij het voorgaande afnamemoment.

Hoewel Figuren 1, 2 en 3 sterk van elkaar verschillen leiden ze bij een correcte interpretatie tot dezelfde conclusies. De prestaties van leerlingen op het gebied van begrijpend lezen nemen consequent toe van afnamemoment naar afnamemoment (duidelijk zichtbaar in Figuur 1). De groei is groot aan het

begin en aan het einde van de basisschoolperiode en in de tussenliggende periode is de groei wat kleiner (zichtbaar in Figuur 1 en 3). We zien ook dat de prestaties van leerlingen tijdens het normeringsonderzoek verschillen van de prestaties van leerlingen in Cito dataretour (zichtbaar in alle drie figuren). Dit laatste heeft geen gevolgen voor toetsresultaten die worden weergegeven op een ontwikkelingsschaal. Als toetsresultaten afgebeeld worden op de percentielschaal of een genormaliseerde schaal kunnen er wel problemen ontstaan bij de interpretatie. Het relatieve niveau van leerlingen wordt namelijk onderschat of overschat. Zo worden de scores uit de scoreverdeling van het normeringsonderzoek gebruikt om de grenzen van de niveaucategorieën I tot en met V te bepalen. Als de scoreverdeling van de normeringsgroep enigszins afwijkt van de scoreverdeling van de leerlingen in de échte afnamen, kunnen er verschillen ontstaan. Tabel 2 laat zien wat het gevolg is in termen van niveaucategorieën I tot en met V. We zien dat het geobserveerde percentage leerlingen in een categorie soms tamelijk sterk afwijkt van het verwachte percentage in een categorie. Waar we op afnamemoment *medio 4* bijvoorbeeld verwachten dat 20 procent van de leerlingen niveau V behaalt, zien we in de praktijk dat slechts 9.4 procent van de leerlingen in deze categorie scoort. Om zicht te krijgen op de relevantie van de verschillen is de effectgrootte uitgerekend (Cohen, 1988). Een van .10 kan beschouwd worden als een klein effect, een van .30 als een gemiddeld effect, en een van .50 als een groot effect. Bij de LVS-toetsen Begrijpend lezen zijn de verschillen te interpreteren als klein tot gemiddeld.

Tabel 2 **Percentage leerlingen in niveaus I tot en met V uitgesplitst naar afnamemoment**

	Afnamemoment						
Niveau	**e3**	**m4**	**e4**	**m5**	**m6**	**m7**	**m8**
V	15.9	9.4	8.3	11.9	16.0	15.6	16.7
IV	21.8	16.3	16.8	16.1	25.4	22.0	18.8
III	22.3	23.6	22.4	21.5	23.1	22.5	18.3
II	22.4	26.9	25.3	25.1	20.2	22.0	20.7
I	17.6	23.8	27.2	25.4	15.3	17.9	25.5
N	263447	293495	293807	250385	184401	119486	42146
φ	.136	.317	.342	.263	.197	.138	.151

Als leerlingen bij een normeringsonderzoek anders presteren dan bij échte afnamen die meetellen, mogen we verwachten dat de geobserveerde percentages in niveaucategorieën I tot en met V afwijken van de verwachte percentages. Tabel 2 laat echter zien dat er ook sprake is van afwijkingen als de prestaties van beide groepen leerlingen wel met elkaar in de pas lopen. Zo vonden we voor afnamemoment *medio 6* in Figuur 1 nauwelijks verschillen tussen het normeringsonderzoek en de échte afnames. Toch scoort slechts 15.3 procent van de leerlingen in categorie I en is de effectgrootte voor afnamemoment *medio 6* gelijk aan .197, wat duidt op een klein effect. Om meer grip te krijgen op de relevantie van de gevonden verschillen in Tabel 2 is een simulatie uitgevoerd. In de eerste stap zijn geneste data gegenereerd voor 20 scholen met 20 leerlingen. Een aantal van 20 x 20 = 400 leerlingen is volgens de richtlijnen van de COTAN voldoende om een toets te normeren. De intraklassecorrelatie, een maat die aangeeft in welke mate leerlingen binnen een school op elkaar lijken, is gelijkgesteld aan .15. In onderwijskundig onderzoek vinden we vaak intraklassecorrelaties die liggen tussen .05 en .25 (Schochet, 2008; Snijders & Bosker, 1993). Op basis van de gesimuleerde steekproef met 400 leerlingen zijn de

grenzen voor niveaucategorieën I tot en met V opnieuw bepaald. Vervolgens zijn de normen toegepast op een tweede gesimuleerde steekproef met N = 100.000 leerlingen. De scores voor de leerlingen in de eerste en de tweede steekproef zijn uit exact dezelfde verdeling getrokken. Tabel 3 laat zien in hoeverre het geobserveerde percentage in elke categorie verschilt van het verwachte percentage als gegarandeerd is dat zowel het ware gemiddelde als de ware standaarddeviatie in beide steekproeven gelijk is. De simulatie is zeven keer herhaald. Ook in de simulatie vinden we soms behoorlijk grote verschillen. Dit betekent dat we op basis van een geneste steekproef met 400 leerlingen niet mogen verwachten dat het geobserveerde percentage in niveaucategorieën I tot en met V altijd dicht bij 20 procent ligt. Blijkbaar brengt een steekproef met 400 leerlingen vrij veel onzekerheid met zich mee en zijn de afwijkingen in Tabel 2 minder opmerkelijk dan aanvankelijk gedacht.

Tabel 3 **Percentage leerlingen in niveaus I tot en met V bij een gesimuleerde steekproef met N = 400**

	Simulatie						
Niveau	**1**	**2**	**3**	**4**	**5**	**6**	**7**
V	12.2	21.4	23.0	18.0	19.7	22.2	18.2
IV	19.8	17.3	20.5	18.1	20.0	17.7	23.9
III	21.7	17.9	23.8	21.4	23.5	22.0	19.1
II	26.0	18.9	20.4	23.8	18.1	22.5	21.5
I	20.2	24.5	12.3	18.7	18.7	15.6	17.2
N	100000	100000	100000	100000	100000	100000	100000
φ	.224	.133	.205	.113	.094	.140	.121

Als de onderzoeksresultaten aanleiding geven om de normen bij te stellen, dan wordt dit gedaan. Ook de normen voor de tweede generatie LVS-toetsen Begrijpend lezen zijn recent bijgesteld. Op deze manier hopen we beter dan in het verleden in te kunnen spelen op onderwijskundige en maatschappelijke veranderingen. De nieuwe aanpak vraagt wel om een andere denkwijze bij scholen. Waar normen voorheen 10 tot 15 jaar geldig waren, vindt er nu met grotere regelmaat een bijstelling plaats

4 Conclusies en discussie

De resultaten van dit onderzoek laten zien dat de prestaties van leerlingen op het gebied van begrijpend lezen de laatste jaren niet veranderd zijn wanneer we per afnamemoment kijken. Wel nemen de prestaties in alle cohorten consequent en in even sterke mate toe vanaf het eerste afnamemoment aan het einde van groep 3 tot en met het laatste afnamemoment medio groep 8. Aan het begin en aan het einde van de basisschoolperiode maken leerlingen een sterkere ontwikkeling door dan in de tussenliggende periode. Dit resultaat komt tot uitdrukking in de effectgrootte die aanvankelijk dicht bij 1.20 ligt, dan enige tijd rond .50 schommelt en dan weer toeneemt tot .81. Deze verschuiving in ontwikkeling aan het eind van de basisschoolperiode is deels te verklaren door de grotere tijdsintervallen. Maar ook als we hiervoor corrigeren, zien we dat leerlingen in de periode van *medio 5* naar *medio 6* net wat minder groei doormaken (= .55) dan in de andere perioden (1.04 - .81).

Het onderzoek laat tevens zien dat de prestaties van de normgroep soms afwijken van de prestaties die leerlingen bij échte afnamen behalen. We hebben gezien dat dit prestatieverschil problematisch kan zijn als toetsscores van leerlingen op basis van de prestaties van de normgroep afgebeeld worden op een percentielschaal of een genormaliseerde schaal. Als de prestaties van de normgroep onwerkelijk hoog of laag liggen, is een onder- of overschatting van het relatieve niveau van leerlingen het gevolg. In het geval dat de prestaties van de normgroep bij álle afnamemomenten hoger of lager liggen, dan wordt het niveau systematisch over- of onderschat en lijkt er op het eerste gezicht niets aan de hand te zijn. In het geval dat de prestaties van de normgroep soms wel en soms niet hoger of lager liggen, zal het relatieve prestatieniveau van leerlingen onrealistisch sterk fluctueren in de tijd. De interpretatie van toetsresultaten wordt daardoor ernstig bemoeilijkt.

Het effect van de afwijkende prestaties van de normgroep bij sommige afnamemomenten is in kaart gebracht door een vergelijking te maken tussen het geobserveerde en het verwachte percentage leerlingen in niveaucategorieën I tot en met V. We vonden behoorlijk grote verschillen. Bij afnamemomenten *medio 4*, *einde 4* en *medio 5* scoorden opmerkelijk veel leerlingen in niveau I. Bij afnamemomenten *medio 6* en *medio 7* liep het percentage leerlingen in niveau I terug. Het percentage leerlingen in niveau V was bij alle afnamemomenten kleiner dan verwacht. De resultaten zijn goed te verklaren. De relatieve prestaties van leerlingen worden bij afnamemomenten *medio 4*, *einde 4* en *medio 5* overschat door de zwakke prestaties van de normgroep. Er scoren relatief veel leerlingen in niveaucategorieën II, III en IV, omdat de standaarddeviatie in het normeringsonderzoek bij alle afnamemomenten groter was dan bij de échte afnamen. Simulaties lieten bovendien zien dat we van een steekproef met 400 leerlingen niet mogen verwachten dat het geobserveerde percentage in niveaucategorieën I tot en met V altijd dicht bij 20 procent ligt. Ook in de simulaties week de geobserveerde verdeling over niveaucategorieën I tot en met V soms behoorlijk af van de ware verdeling. Niettemin hebben Keijsers et al. (2012) gelijk in hun constatering dat de relatieve prestaties van leerlingen in eerste instantie iets overschat worden (*medio 4*, *einde 4* en *medio 5*) en later mogelijk iets onderschat (*medio 6* en *medio 7*). Het onderhavige onderzoek laat duidelijk zien wat de oorzaak is van de over- en onderschatting van de relatieve prestaties van leerlingen. Het zijn de zwakke begrijpend leesprestaties van de normgroep die zorgen voor de fluctuatie in het relatieve niveau. Van een zogeheten (absolute) dip in de begrijpend leesprestaties van leerlingen is geen sprake, hoewel sommige figuren die suggestie in eerste instantie mogelijk wel wekken. In Figuren 2 en 3 lijkt er bijvoorbeeld sprake te zijn van een prestatiedip in groep 4. Dat deze dip uitsluitend het gevolg is van de wijze waarop de resultaten worden weergegeven, wordt duidelijk in Figuur 1. Om onjuiste conclusies te voorkomen, is het belangrijk dat resultaten duidelijk gepresenteerd en toegelicht worden.

Het is opmerkelijk dat de toetsresultaten van de normgroep soms sterk afwijken van de toetsresultaten die leerlingen tijdens échte afnamen behalen. Het prestatieverschil kan deels verklaard worden vanuit de onzekerheid die elke steekproef met zich meebrengt (zie ook Keuning & Béguin, 2012). Daarnaast zijn ten minste twee andere verklaringen mogelijk.
Ten eerste kan het prestatieverschil veroorzaakt worden door een discrepantie tussen de afnamesituatie tijdens het normeringsonderzoek en de feitelijke afnamesituatie na uitgave. Ondanks dat de toetsinhoud en afname-instructies gelijk blijven,

is niet altijd te vermijden dat de afnamesituatie verandert. Het is goed mogelijk dat leerlingen tijdens het *low stake* normeringsonderzoek minder goed hun best doen dan tijdens de meer *high stake* afnamen die na uitgave plaatsvinden (cf. Hemker, 2012; Wise & DeMars, 2005). In dat geval kunnen de *low stake* condities tijdens het normeringsonderzoek dus leiden tot *bias* in de normen. Ook kan de manier waarop toetsen gebruikt worden over de jaren heen zichtbaar en onzichtbaar verschuiven. De laatste jaren zijn de LVS-toetsen Begrijpend lezen bijvoorbeeld een steeds grotere rol gaan spelen in de interne en externe verantwoording die scholen afleggen. De toetsen waren voorheen echte *low stake* toetsen. Tegenwoordig ervaren zowel leerlingen, leerkrachten als scholen de toetsen meer en meer als *high stake* toetsen. Ten tweede kan het prestatieverschil veroorzaakt worden door een verandering in didactiek. Leerlingen zijn de laatste jaren steeds vluchtiger gaan lezen, terwijl de LVS-toetsen Begrijpend lezen vragen om een behoorlijk lange concentratie. Dit zijn leerlingen niet meer zo gewend. Daarnaast hebben leesmethoden geprobeerd om het begrijpend lezen aantrekkelijker te maken door kortere en actuelere teksten aan te bieden. In de LVS-toetsen Begrijpend lezen worden nog betrekkelijk lange teksten aangeboden. Leerkrachten zijn zich ook bewuster geworden van hiaten op het gebied van begrijpend lezen bij leerlingen en hebben hun onderwijskundig-didactische aanpak hierop aangepast. Al deze ontwikkelingen kunnen leiden tot veranderingen in prestatie, niet alleen bij begrijpend lezen, maar ook bij andere vakgebieden.

Hoewel het gesignaleerde prestatieverschil bij de LVS-toetsen Begrijpend lezen vrij gemakkelijk te verklaren is, ontstaan er mogelijk wel problemen in het onderwijs. Leerkrachten nemen immers fluctuaties in het relatieve niveau van leerlingen waar die niet zijn toe te schrijven aan een ontwikkelingsachterstand of -voorsprong, maar het gevolg zijn van *bias* in de normen. Om dit probleem in de toekomst zoveel mogelijk te beperken, voert Cito twee veranderingen door in het leerlingvolgsysteem. Ten eerste wordt bij de ontwikkeling van de derde generatie LVS-toetsen niet langer gebruikgemaakt van *standalone* toetsdesigns. In plaats daarvan worden zogeheten *embedded field* onderzoeken georganiseerd waarin nieuw ontwikkeld materiaal meedraait in de bestaande toetscyclus (zie Schmeiser & Welch, 2006). De leerlingen weten op voorhand niet welke opgaven nieuw zijn en welke opgaven meetellen voor hun resultaat. Tevens zijn zij niet op de hoogte van het feit dat de toets óók voor onderzoeksdoeleinden wordt ingezet. Op deze manier wordt de discrepantie tussen de afnamesituatie tijdens de proeftoetsing en de feitelijke afnamesituatie na uitgave tot een minimum gereduceerd. Motivatie-effecten zijn dan zo goed als uit te sluiten. Dat deze aanpak lijkt te werken, zien we reeds terug bij de LVS-toetsen Begrijpend lezen. In de hogere leerjaren is de normering bij wijze van experiment namelijk mede gebaseerd op toetsresultaten die bij échte afname verzameld zijn. Waar het prestatieverschil tussen de normgroep en Cito dataretour in de lagere leerjaren vrij groot is, lopen de prestaties van beide groepen in de hogere leerjaren in statistisch opzicht mooi in de pas. Ten tweede wordt de actualiteit van de normen met ingang van schooljaar 2013/2014 jaarlijks gecheckt op basis van Cito dataretour. Als de onderzoeksresultaten aanleiding geven om de normen bij te stellen, dan wordt dit gedaan. Ook de normen voor de tweede generatie LVS-toetsen Begrijpend lezen zijn recent bijgesteld. Op deze manier hopen we beter dan in het verleden in te kunnen spelen op onderwijskundige en maatschappelijke veranderingen. De nieuwe aanpak vraagt wel om een andere denkwijze bij scholen. Waar normen voorheen 10 tot 15 jaar geldig waren, vindt er nu met grotere regelmaat een bijstelling plaats.

GERAADPLEEGDE LITERATUUR EN WEBLINKS

Zie hiervoor http://tvodigitaal.nl – januari – 'Artikelen, Columns, Mededelingen'.

OVER DE AUTEUR

Jos Keuning is onderwijskundige en in 2008 gepromoveerd in de sociale wetenschappen op een onderzoek dat zich richtte op de lees- en spellingontwikkeling van kinderen gedurende de basisschoolperiode. Na zijn promotie is hij als methodoloog gaan werken bij het Psychometrisch Onderzoekcentrum van Cito. In die functie is hij betrokken geweest bij projecten die gericht waren op de ontwikkeling van tests voor specifieke doelgroepen. Daarnaast doet hij onderzoek naar leerrendementsverwachting en schooleffectiviteit.

Maartje Hilte is psychologe en in 2009 gepromoveerd in de sociale wetenschappen op een onderzoek dat zich richtte op effecten van spellingoefeningen op de prestaties van leerlingen in de basisschoolperiode. Na haar promotie is zij als toetsdeskundige gaan werken bij Cito. Zij is in deze functie betrokken bij de ontwikkeling van diverse taaltests voor leerlingen in het primair onderwijs. Zo is zij betrokken bij de ontwikkeling van de toetsen Begrijpend lezen, Begrijpend luisteren en Woordenschat uit het Cito Volgsysteem primair en speciaal onderwijs.
E-mail: Maartje.Hilte@cito.nl

Anke Weekers is orthopedagoog en onderwijskundige en is twee jaar werkzaam geweest als orthopedagoog (specialisatie leerstoornissen). In 2009 is ze gepromoveerd in de sociale wetenschappen op een onderzoek dat zich richtte op het modelleren van data verzameld met attitudeschalen en persoonlijkheidsvragenlijsten. Na haar promotie is ze als methodoloog gaan werken bij het Psychometrisch Onderzoekscentrum van Cito. In die functie is ze betrokken bij projecten die gericht zijn op de ontwikkeling van toetsen begrijpend lezen, schrijven en taalverzorging in basisonderwijs, speciaal (basis)onderwijs en voortgezet onderwijs.