Een Leesdip in groep 6

Bestaat hij of bestaat hij niet?

Kenniscentrum Begrijpend Lezen
Peter Wolfgram
Willem van der Horst
Ernst Keijsers

November 2012

INHOUD

1.	Vraagstelling	3
2.	Werkwijze	4
2.1	Steekproef	4
2.2	Gegevensverzameling	4
2.3	Bestandsbewerkingen	4
2.4	Analyse	5
3.	Resultaten	6
3.1	Steekproef	6
3.2	Toetsscores op leerlingniveau	7
3.3	Indeling in CITO-categorieën	8
3.4	Toetsscores op schoolniveau	9
4.	Conclusies	11

CED-Groep
Kenniscentrum Begrijpend lezen
Peter Wolfgram
Willem van der Horst
Ernst Keijsers

1 VRAAGSTELLING

Veel scholen en onderwijsadviseurs hebben opgemerkt dat de scores in groep 6 op de toets begrijpend lezen (LBL) uit het Leerling en onderwijsvolgsysteem van het Cito (LOVS) in vergelijking met de normen lager zijn dan in de leerjaren ervoor. Omdat dit ook na enkele jaren gebruik van de LBL blijft voortduren, heeft het Kenniscentrum Begrijpend Lezen besloten de scores van een grote groep scholen op te vragen om te onderzoeken of de feitelijke gegevens overeenkomen met die waarnemingen. De vraagstelling voor dit onderzoek luidt:

> *Zijn de scores die met de LBL worden verkregen in groep 6 vergeleken met de normen belangrijk lager dan in de overige leerjaren?*

De verschillende toetsmodules van groep 3 tot en met groep 8 van de LBL kunnen samen beschouwd worden als één toets voor de vaardigheid begrijpend lezen. Het aantal goede antwoorden wordt omgezet in een score op één vaardigheidsschaal. Vanaf groep 3 neemt normaal gesproken de vaardigheidsscore ieder jaar toe.
Met een landelijk representatieve steekproef heeft het CITO voor iedere toetsmodule een kwartielverdeling bepaald, waarbij binnen het laatste kwartiel ook nog het laagste deciel wordt onderscheiden: de zogenaamde ABCDE-scores. Dit is de relatieve normering. Leerkrachten kunnen zien hoe de toetsprestaties van hun leerlingen zich verhouden tot de landelijke steekproef.

Bij iedere afname van de toets hopen leerkrachten dat de toetsprestaties ten opzichte van de landelijke normen beter zijn dan bij de vorige. Het onderwijs is dan vruchtbaar geweest. Het probleem dat nu gemeld wordt is dat in groep 6 de resultaten meestal juist lager zijn dan in groep 5. Als dit werkelijk over de hele linie het geval zou zijn, dan kan dit betekenen dat er vraagtekens bij de normen gezet kunnen worden. Moeten de scholen zich wel zorgen maken over lagere toetsprestaties in groep 6?

In dit onderzoek worden de toetsscores van een groot aantal leerlingen geanalyseerd zodat kan worden vastgesteld of de waarnemingen, namelijk dat de toetsprestaties in groep 6 inderdaad beduidend lager zijn dan in andere leerjaren, kunnen worden bevestigd.

2 WERKWIJZE

2.1 Steekproef

In het kader van dit onderzoek zijn data op een efficiënte manier verzameld. Dat wil vooral zeggen dat het, in het bestek van dit onderzoek, onmogelijk was om een landelijk representatieve steekproef samen te stellen en daarvandaan de benodigde gegevens te verzamelen. Tijd en geld waren daarvoor eenvoudigweg niet voldoende beschikbaar. In plaats daarvan zijn scholen benaderd die voor ons *bereikbaar* waren. Het gaat dan om scholen waarmee de CED-Groep een adviesrelatie onderhoudt. 60 van die scholen zijn benaderd door de taaladviseurs van de CED-Groep met het verzoek om deel te nemen aan het onderzoek. 45 scholen hebben hun medewerking toegezegd en data voor het onderzoek aangeleverd, daarmee aangevend hoe belangrijk het onderzoek in het veld wordt gevonden. De scholen die niet deelnamen verwezen in de regel naar overwegingen inzake privacy en naar drukke andere werkzaamheden.

2.2 Gegevensverzameling

De scholen hebben hun leerling- en toetsscorebestand voor dit onderzoek opgestuurd naar de Toetsservice van de CED-Groep. Dit resulteert in een gegevensbestand met:

- een nieuw identiteitsnummer voor iedere leerling;
- een schoolcode (brin);
- de vaardigheidsscores van iedere leerling op de LBL-modules en/of de TBL;
- de schooljaren waarin de vaardigheidsscores met de verschillende modules zijn vastgesteld;
- de kalenderdata van afname van de toetsmodules.

2.3 Bestandsbewerkingen

De vaardigheidsscores zijn omgezet naar een gestandaardiseerde score aan de hand van de normering. Hiervoor is de t-score gebruikt. Dit is een scoreverdeling waarbij het landelijk gemiddelde 50 is en de landelijke standaarddeviatie 10.

Het oorspronkelijke bestand bestaat uit iets meer dan 48000 records, voor elke leerling evenveel records als hij toetsscores heeft. Zie tabel 2.1 voor een overzicht.

Tabel 2.1: Verdeling van records over leerjaren

leerjaar	aantal	%	Cumulatief %
0	11	,0	,0
1	54	,1	,1
2	123	,3	,4
3	4717	9,8	10,2
4	12838	26,7	36,9
5	8590	17,8	54,7
6	7934	16,5	71,2
7	8198	17,0	88,2
8	5617	11,7	99,9
9	7	,0	99,9
99	1	,0	99,9
NULL	58	,1	100,0
Totaal	48148	100,0	

Om diverse redenen bleek een aantal records onbruikbaar:

- De records met vreemde leerjaren zijn verwijderd. In de tabel gaat het om de leerjaren 0, 9, 99 en NULL. Ook de records met de leerjaren 1 en 2 zijn verwijderd.
- Aansluitend bleek dat de gemaakte toetsmodules niet altijd overeenstemden met het leerjaar van leerlingen. Dat kunnen invoerfouten zijn, maar die records kunnen ook leerlingen betreffen die voor- of achterlopen op hun groepsgenoten. Noch invoerfouten noch scores van exceptionele leerlingen kunnen in dit onderzoek worden gebruikt, dus ook die records zijn verwijderd.
- Een betrekkelijk groot aantal records, meer dan 15000, is verwijderd omdat die de Toets Begrijpend Lezen (TBL) betroffen. De TBL is de voorganger van de LBL en wordt op een aantal van de deelnemende scholen nog gebruikt in met name de hogere groepen. De scores op beide toetsen zijn niet zonder meer te vergelijken en het gebruik van beide sets records zou het trekken van conclusies uit het onderzoek bemoeilijken, cq. onmogelijk maken.
- Een beperkt aantal records werd ons aangeleverd als behorend bij de LBL6E en LBL7E. Dit zijn geen LBL-toetsen maar waarschijnlijk entree-toetsen. Ook die zijn verwijderd.
- Tenslotte werden meer dan 1300 records verwijderd omdat die, met uitzondering van de toetsscores, identiek waren aan een ander record. Het gaat hier om leerlingen die binnen één jaar twee keer zijn getoetst met dezelfde toets. De records met de hoogste scores zijn bewaard

en in ons onderzoek betrokken.

Uiteindelijk resteert er een bestand van 28606[1] records die in het onderzoek konden worden betrokken. Dit bestand is omgezet naar een leerlingbestand met voor iedere leerling één record. Onderstaande tabel toont het aantal records per toetsafname.

Tabel 2.3: Overzicht van toetsafnames per leerjaar

		leerjaar						
		3	4	5	6	7	8	Totaal
toets	LBL3E	4369	0	0	0	0	0	4369
	LBL4M	0	5776	0	0	0	0	5776
	LBL4E	0	5436	0	0	0	0	5436
	LBL5M	0	0	5146	0	0	0	5146
	LBL6M	0	0	0	4121	0	0	4121
	LBL7M	0	0	0	0	3150	0	3150
	LBL8B	0	0	0	0	0	608	608
Totaal		4369	11212	5146	4121	3150	608	28606

In de tabel is zichtbaar dat het aantal records afneemt naarmate de leerjaren hoger worden. De verklaring is dat scholen de LBL, als opvolger van de TBL, gefaseerd invoeren, beginnend in laagste groepen. Het resultaat is onder meer dat het aantal records van de LBL8B te klein is om verder in de analyses te betrekken.

2.4 Analyse

Het doel is te onderzoeken of de scores op de LBL in groep 6 onevenredig laag zijn. Daartoe worden de t-scores in groep 6 vergeleken met de scores in groep 5 en groep 7 met de t-toets voor afhankelijke steekproeven. Het is niet goed mogelijk om de scores van alle leerjaren met elkaar te vergelijken, omdat de gegevens betrekking hebben op de periode waarin scholen overschakelden van de TBL naar de LBL. Daardoor wisselen de aantallen leerlingen per leerjaar. Door de scores in aanpalende leerjaren met elkaar te vergelijken betreffen de t-toetsen zoveel mogelijk leerlingen. De t-toets wordt zowel op leerling- als op schoolniveau uitgevoerd.

[1] In het Didactief-artikel over dit onderzoek (november 2012) staat abusievelijk 30558 vermeld.

3 RESULTATEN

3.1 Steekproef

Er zijn van 45 scholen gegevens ontvangen. Daarvan liggen er 19 in Rotterdam en 28 in Zuid-Holland. De drie noordelijke provincies en Noord-Holland, Zeeland en Limburg ontbreken in de steekproef. Naast Rotterdam kunnen Utrecht, Almere, Tilburg, Apeldoorn, Dordrecht en Zwolle als grote gemeenten worden beschouwd. Daarnaast zijn er een aantal kleine plaatsen in de steekproef vertegenwoordigd. Zie tabel 3.1.1 Gezien de wijze waarop de steekproef tot stand is gekomen kan niet van een landelijk representatieve steekproef worden gesproken.

Tabel 3.1.1 Steekproef, aantal scholen per plaats en provincie

		Aantal scholen	Gemiddelde schoolgrootte	Totaal aantal leerlingen
Flevoland	Almere	1	733	733
	Totaal	1	733	733
Gelderland	Apeldoorn	1	160	160
	Asperen	2	189	378
	Didam	1	158	158
	Totaal	4	174	696
Noord-Brabant	Tilburg	2	190	379
	Totaal	2	190	379
Overijssel	Genemuiden	2	331	661
	Punthorst	1	178	178
	Staphorst	4	293	1171
	Zwolle	1	218	218
	Totaal	8	279	2228
Utrecht	Utrecht	1	103	103
	Zeist	1	538	538
	Totaal	2	321	641
Zuid-Holland	Barendrecht	1	351	351
	Bergschenhoek	1	354	354
	Dordrecht	2	205	409
	Hardinxveld-Giessendam	1	224	224
	Maassluis	1	597	597
	Pernis Rotterdam	1	419	419
	Ridderkerk	1	268	268
	Rotterdam	19	357	6790
	Vlaardingen	1	110	110
	Totaal	28	340	9522
Totaal		45	316	14199

De gemiddelde schoolgrootte is 316, de kleinste school heeft 82 leerlingen en de grootste 906.

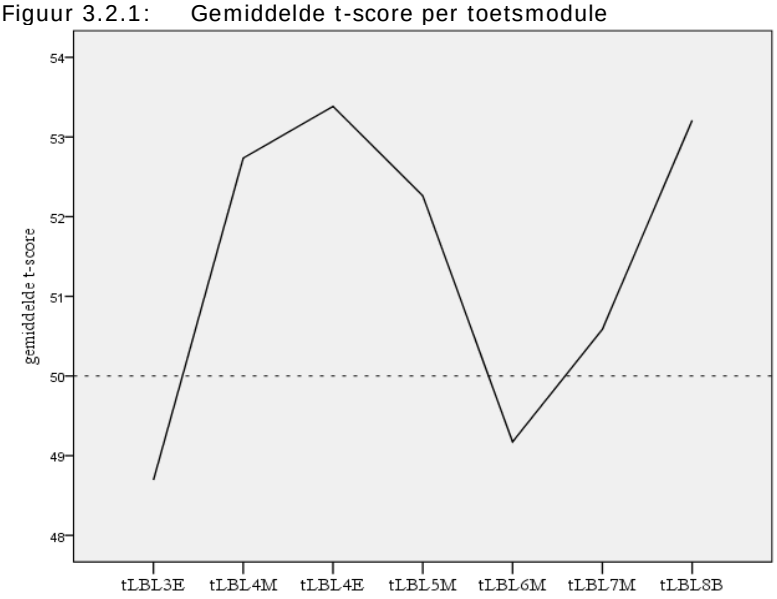
3.2 Toetsscores op leerlingniveau

In Tabel 3.2.1 staan enkele beschrijvende statistieken. De gemiddelde t-scores variëren tussen 48,7 en 53,4. De standaarddeviaties komen dicht in de buurt van de verwachte waarde 10.
Figuur 3.2.1 toont de gemiddelden per module.

Tabel 3.2.1: t-scores, aantal waarnemingen, gemiddelden, standaarddeviatie en range per toetsmodule

toets	N	M	Sd	Min	Max
LBL3E	4369	48,7	9,6	19	88
LBL4M	5776	52,7	8,7	25	94
LBL4E	5436	53,4	10,0	24	107
LBL5M	5146	52,3	9,2	24	103
LBL6M	4121	49,2	9,6	17	97
LBL7M	3150	50,6	10,8	11	108
LBL8B	608	53,2	12,4	21	97

Figuur 3.2.1: Gemiddelde t-score per toetsmodule

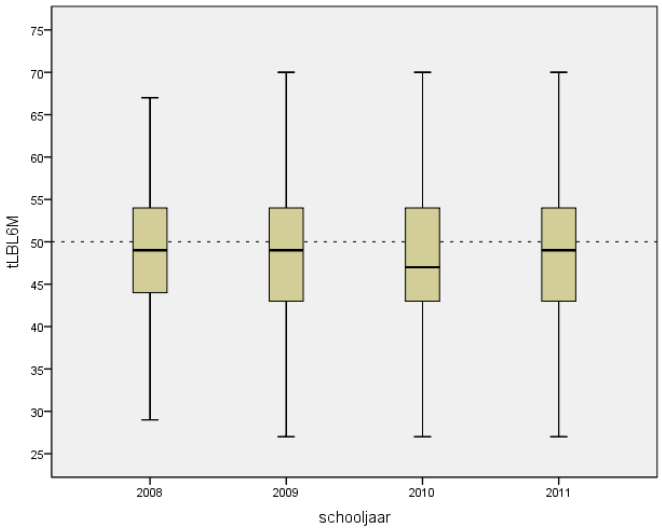


De sterke daling van relatieve scores in groep 6 lijkt inderdaad aanwezig te zijn. In groep 4 en 5 is de gemiddelde score hoger dan 52 in groep 6 lager dan 50. Voordat we die verwachting toetsen, onderzoeken we eerst of de scores in de verschillende afnamejaren op hetzelfde niveau liggen. Als de scores tussen afnamejaren verschillen hebben verschillen in scores tussen toetsmodules immers weinig betekenis.
In tabel 3.2.2 en figuur 3.2.2 laten we de verschillen tussen afnamejaren van de LBL6M zien. De gemiddelden liggen alle lager dan 50, ze variëren tussen 48,8 en 49,7.

Tabel 3.2.2: t-scores LBL6M per afnamejaar

	Aantal	Gemiddelde score	Sd
2008	246	49,22	8,633
2009	1028	48,97	9,585
2010	1398	48,82	9,649
2011	1448	49,65	9,694
Totaal	4120	49,17	9,595

Figuur 3.2.2: t-scores LBL6M per afnamejaar



Met één-weg-variantieanalyse is onderzocht of er verschillen tussen de schooljaren zijn. Dat is niet het geval (F = 1,99, p = 0,114), de variantie binnen groepen is veel groter dan de variantie tussen groepen (tabel 3.2.3). De toetsscores liggen dus in elk van de schooljaren op hetzelfde niveau.

Tabel 3.2.3 Resultaten één-weg-varantieanalyse tussen schooljaren

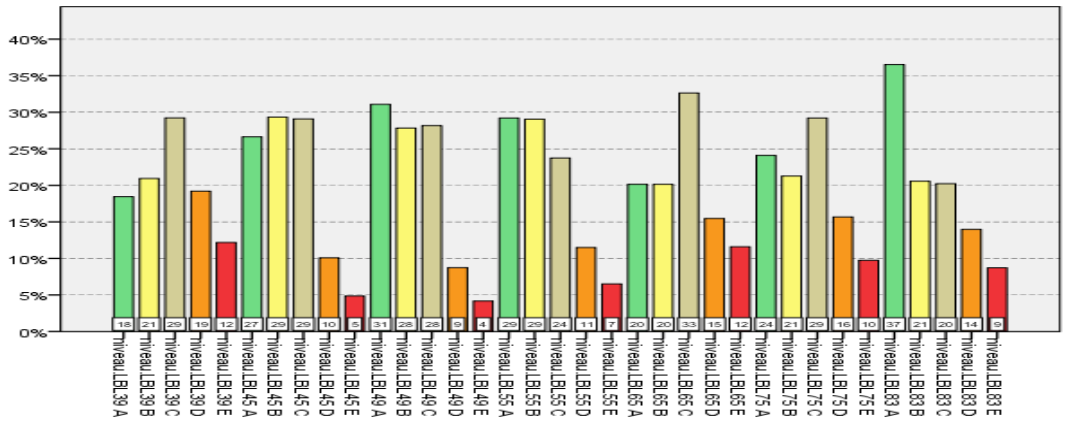
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	548,751	3	182,917	1,988	,114
Within Groups	378660,859	4116	91,997		
Total	379209,609	4119			

Nu kan onderzocht worden of de t-scores in groep 6 verschillen van die in de groepen 5 en 7. Dat is inderdaad het geval. Beide verschillen zijn significant (t = 26,3, respectievelijk -10,2, in beide gevallen is p < 0,001). Het verschil tussen de gemiddelde score aan het einde van groep 5 en die van midden groep 6 is 3,3. Het verschil tussen de gemiddelde score van midden groep 6 en die van midden groep 7 is -1,6. Hoe groot zijn deze verschillen? Volgens Cohen is de effectgrootte van het eerste verschil 0,33, wat als een klein effect moet worden beschouwd en die van het tweede verschil 0,02, een verwaarloosbaar effect. Met andere woorden het verschil tussen 5 en 6 is aanwezig maar klein en het verschil tussen 6 en 7 is nauwelijks aanwezig.

3.3 De CITO-categorieën

De resultaten op leerlingniveau kunnen ook worden weergegeven aan de hand van indeling van de scores in CITO-categorieën. Die indeling staat hieronder in Figuur 3.3.1.

Figuur 3.3.1 Indeling in CITO-categorieën



Niet verwonderlijk laat figuur 3.3.1 eenzelfde beeld zien. De indeling is relevant omdat leerkrachten immers vooral zo het niveau van hun klas, en veranderingen daarin, bekijken. Zichtbaar is dat in groep 6 het percentage A- en B-leerlingen (de groene en gele staaf samen) dramatisch is gedaald. Behoort in groep 5 nog bijna 60% van de leerlingen tot één van deze categorieën, in groep 6 is dat nog maar 40%. De percentages van zowel de C-, D- als E-leerlingen zijn toegenomen.

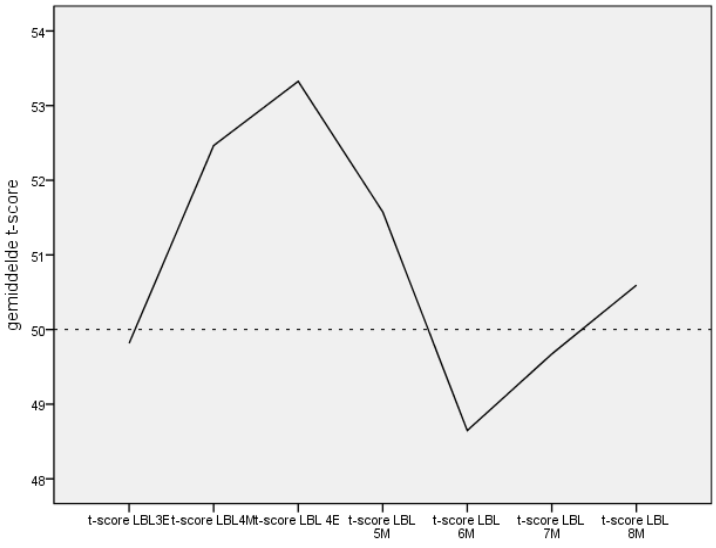
3.4 Toetsscores op schoolniveau

Grote scholen bepalen het niveau van de individuele leerlingscores in dit onderzoek meer dan kleinere scholen. Als we op het niveau van scholen kijken wordt dat effect teniet gedaan, alle scholen dragen in gelijke mate bij. In tabel 3.3.1 worden de gemiddelde scores vermeld. Meer dan de helft van de scholen neemt de LBL8B (nog) niet af. De gemiddelde scores verschillen enigszins van die op leerlingenniveau, maar de algemene trend is dezelfde: de scores in de groepen 3, 6 en 7 zijn de laagste. De standaarddeviaties zijn belangrijk kleiner omdat het hier om verschillen tussen scholen gaat. Die zijn nu eenmaal kleiner dan verschillen tussen leerlingen.

Tabel 3.4.1: LBL-scores op schoolniveau, aantal scholen, gemiddelden en standaarddeviaties

	Aantal scholen		Gemiddelde	
	Valid	Missing	score	s.d.
t-score LBL3E	40	5	49,8	5,20
t-score LBL4M	44	1	52,5	3,61
t-score LBL 4E	44	1	53,3	4,34
t-score LBL 5M	45	0	51,6	3,91
t-score LBL 6M	45	0	48,6	3,84
t-score LBL 7M	42	3	49,7	4,41
t-score LBL 8M	17	28	50,6	7,31

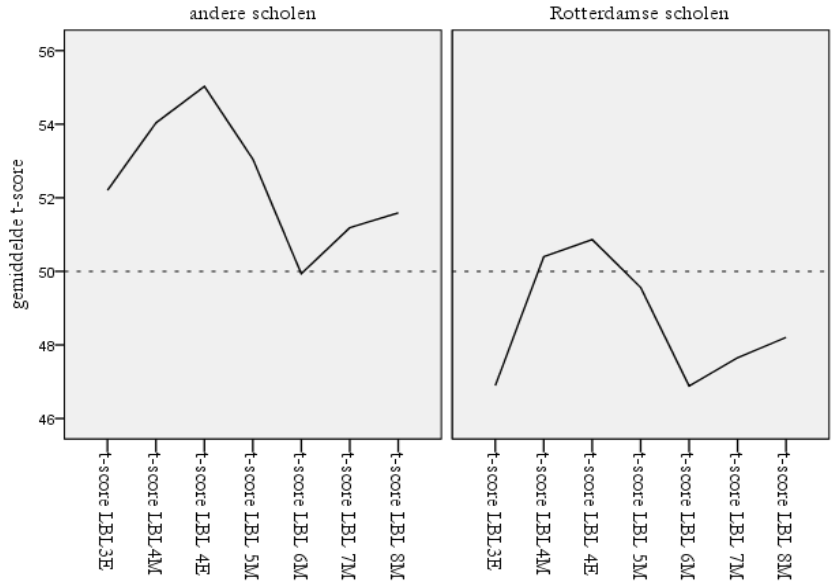
Figuur 3.4.1 Gemiddelde LBL-scores op schoolniveau



De daling van 5E naar 6M is goed zichtbaar. We toetsen de verschillen tussen 5E en 6M, en 6M en 7M met de t-toets voor afhankelijke steekproeven. Beide toetsen laten een significant verschil zien ($t = 9,5$ ($p < 0,001$), respectievelijk -3,3 ($p = 0,002$). Hoe groot zijn de verschillen? Het verschil tussen 5E en 6M kan als een groot effect beschouwd worden ($d = 0,76$), de effectgrootte van het verschil tussen 6M en 7M is bijna verwaarloosbaar ($d = 2,8$). De waarnemingen van scholen dat de scores in groep 6 aanzienlijk lager zijn dan die in groep 5 wordt dus ook hier bevestigd.

Een groot deel van de steekproef bestaat uit Rotterdamse scholen. De scores van die scholen hebben dus veel invloed op de resultaten. Om te onderzoeken in hoeverre dit onze conclusies beïnvloedt, vergelijken we hier de Rotterdamse scores met die van de overige scholen. In figuur 3.3.2 wordt het verloop van de gemiddelde scores van Rotterdamse scholen naast die van de andere scholen gezet.

Figuur 3.4.2: Gemiddelde schoolscores van Rotterdamse en andere scholen



Het scoreverloop van beide groepen scholen is bijna gelijk (de scores op de 8M betreft maar weinig scholen), zij het dat de scores van de Rotterdamse scholen op een lager niveau liggen. Het belangrijkste is echter te constateren dat in beide groepen er een daling is van gemiddelde scores van groep 5 naar groep 6. De achteruitgang in scores is dus een algemeen verschijnsel.

4. Conclusies

Verklaringen voor de grilligheid kunnen worden gezocht in het onderwijs, in ons onderzoek en in de toets. Dat de *kwaliteit van het onderwijs* de oorzaak zou zijn vinden we niet erg aannemelijk. Waarom zou er in groep 4 en 5 heel goed onderwijs worden gegeven en in de andere groepen veel minder goed? Dit is des te onwaarschijnlijker omdat bijna alle onderzoeksscholen achteruit gaan van groep 5 naar 6. Slechts 4 van de 46 boeken vooruitgang. Ook de vooruitgang van 3 naar 4 was algemeen, namelijk op 43 scholen. De grilligheid is structureel.

Op de *kwaliteit van ons onderzoek* is wel kritiek mogelijk. We vinden een grondig en verantwoord vervolgonderzoek nodig omdat onze steekproef niet representatief is. Maar naar onze vaste overtuiging geeft ons onderzoek wél een stevige indicatie dat er echt iets aan de hand is. Zo blijken de tendensen in de verschillende gemeten kalenderjaren hetzelfde te zijn. En wanneer we de 19 Rotterdamse scholen afzonderen van de niet-Rotterdamse, dan blijkt in beide groepen scholen de dip zichtbaar te zijn, zij het op de Rotterdamse scholen op een lager niveau. Wederom structurele grilligheid.

Resteert de *kwaliteit van toets LBL*. Sommige leerkrachten veronderstellen dat de toegenomen complexiteit en de lengte van de teksten in de groep 6-toets verantwoordelijk zijn voor de dip. Leerlingen zouden ervan schrikken. Maar als wij de twee toetsen analyseren, lijken de groep 6-tekstjes qua tekstlengte eerder wat makkelijker geworden. Andere leerkrachten denken dat de leerlingen onvoldoende met informatieve teksten hebben geoefend om de groep 6-toetsen goed aan te kunnen. Ook dat is niet erg plausibel, omdat de opbouw van beide toetsen evenwichtig is.

Eerder geloven wij dat, wanneer vervolgonderzoek hetzelfde resultaat oplevert, de onderzoeksgroep waarop Cito de normering van de LBL heeft gebaseerd aan een grondige inspectie moet worden onderworpen. Dat is urgent omdat Cito op dit moment een derde generatie lovs-toetsen aan het ontwikkelen is én omdat leerkrachten betrouwbare toetsuitslagen nodig hebben voor hun onderwijspraktijk.