

Zicht op Leerwinst

Landenstudies

In aanvulling bij het boek:

Zicht op leerwinst. Scenario's voor gestandaardiseerd toetsen

Maarten Penninckx, Jan Vanhoof, Amy Quintelier, Sven De Maeyer & Peter Van Petegem

2017 - Acco: Leuven.



INHOUDSTAFEL

Resultaten Internationale studie	4
1. Case Australië	4
1.1 Inleiding tot het onderwijssysteem in Australië	4
1.2 Gestandaardiseerde toetsen op federaal niveau	8
1.3 Gestandaardiseerde toetsen op state/territory-niveau	22
1.4 Toetsen vanuit toetsinstituten	34
1.5 Leerwinst en toegevoegde waarde	41
2. Case Engeland - Centre of Evaluation and Monitoring	43
2.1. Inleiding tot het onderwijssysteem in Engeland	43
2.2 Verplichte gestandaardiseerde toetsen in Engeland	45
2.3 Toetsen binnen Centre for Evaluation and Monitoring	52
2.4 Impact van het leerwinstmonitoringssysteem van CEM	61
2.5 Onwenselijke neveneffecten	65
2.6 Evoluties en toekomstperspectieven binnen CEM	66
2.7 Leerwinst en toegevoegde waarden binnen toetsen van CEM	68
3. Case Nederland (pilotproject Leerwinst en Toegevoegde waarde in het primair onderwijs)	71
3.1 Inleiding tot het onderwijssysteem in Nederland	71
3.2. Gestandaardiseerde toetsen in Nederland	76
3.3. Leerlingvolgsysteem CITO en pilotproject 'Leerwinstmonitoring'	89
3.4. Leerwinst en toegevoegde waarde	110
4. Case Noorwegen (inclusief focus op Oslo)	111
4.1 Inleiding tot het onderwijssysteem in Noorwegen	111
4.2. Gestandaardiseerde toetsen in Noorwegen (federaal)	115
4.3. Gestandaardiseerde toetsen in de stad Oslo	136
4.4. Leerwinst en toegevoegde waarde	149
5. Conclusie: Aandachtspunten voor Vlaanderen vanuit de internationale studie	151
5.1 Beslissing over implementatie	151
5.2 Eigenschappen van het leer(winst)monitoringssysteem	156

5.3 Leerwinst	161
5.4 Toegevoegde waarde.....	163
5.5 Rapportering.....	164
5.6 Ondersteuning van scholen.....	165
5.7 Financiële inschatting is bijzonder moeilijk	168
Literatuurlijst	169

RESULTATEN INTERNATIONALE STUDIE

In dit deel geven we in de eerste vier secties een horizontale analyse waarin elk van de vier bezochte leer(winst)monitoringssystemen uitvoerig besproken wordt. Achtereenvolgens ligt de focus op de case studies Australië, Engeland (met focus op het leer(winst)monitoringssysteem van het Centre of Evaluation and Monitoring), Nederland (met focus op het pilootproject 'Leerwinst en Toegevoegde Waarde in het primair onderwijs') en Noorwegen (inclusief focus op het schoolbestuur van Oslo). We geven daarbij steeds een korte inleiding in de context van het onderwijssysteem en beschrijven daarna het gebruik van gestandaardiseerde toetsen, inclusief een inschatting van de impact ervan. Elk van deze case studies is opgebouwd op basis van een gerichte literatuurstudie, gesprekken tijdens het bezoek ter plaatse en tijdens het bezoek vergaarde grijze literatuur. Aan elk van de respondenten wordt een code toegekend, die wordt gebruikt bij de rapportering in dit deel. Bijlage 4 geeft de lezer een overzicht van de verschillende respondenten, waarbij evenwel de anonimiteit van de respondenten wordt bewaakt.

Uit elk van deze case studies konden uiteenlopende aandachtspunten worden gehaald voor de eventuele implementatie en het ontwerp van een leer(winst)monitoringssysteem in Vlaanderen. Een overzicht van deze aandachtspunten wordt besproken in de afsluitende vijfde sectie van dit deel.

1. CASE AUSTRALIË

Australië heeft een lange traditie inzake gestandaardiseerde toetsen. Er wordt door overheden en scholen veelvuldig van gestandaardiseerde toetsen gebruik gemaakt. Het was dan ook nodig om een selectie te maken en een weloverwogen focus te bepalen. Daarbij werd ernaar gestreefd om een zo breed mogelijk gamma aan toetsen te bestuderen met het oog op representativiteit voor het volledige toetsenaanbod. We hebben aandacht voor toetsen vanuit de nationale overheid, van de overheden van de verschillende staten/territories, en van privé-initiatieven. Zowel verplichte gestandaardiseerde toetsen als toetsen die door scholen vrijwillig worden afgenomen, komen aan bod.

1.1 Inleiding tot het onderwijssysteem in Australië¹

Australië is een gefederaliseerd land: elk van de zes staten en twee territories hebben een autonoom onderwijssysteem². De financiering of subsidiëring van scholen gebeurt evenwel gedeeltelijk op het nationale niveau. Ook wordt sinds enkele jaren een nationaal curriculum gehanteerd.

¹ Gebaseerd op: ACARA (2014); Australian Government: Department of Education Employment and Workplace Relations (2010); Ministerial Council on Education Employment Training and Youth Affairs (2008); Santiago, Donaldson, Herman, en Shewbridge (2011) + informatie tijdens gesprekken in case studie.

² Het zeer kleine 'Jervis Bay Territory' heeft geen autonoom functionerend onderwijssysteem, maar wordt (net als voor justitie en welzijn) ondergebracht bij 'Australian Capital Territory'.

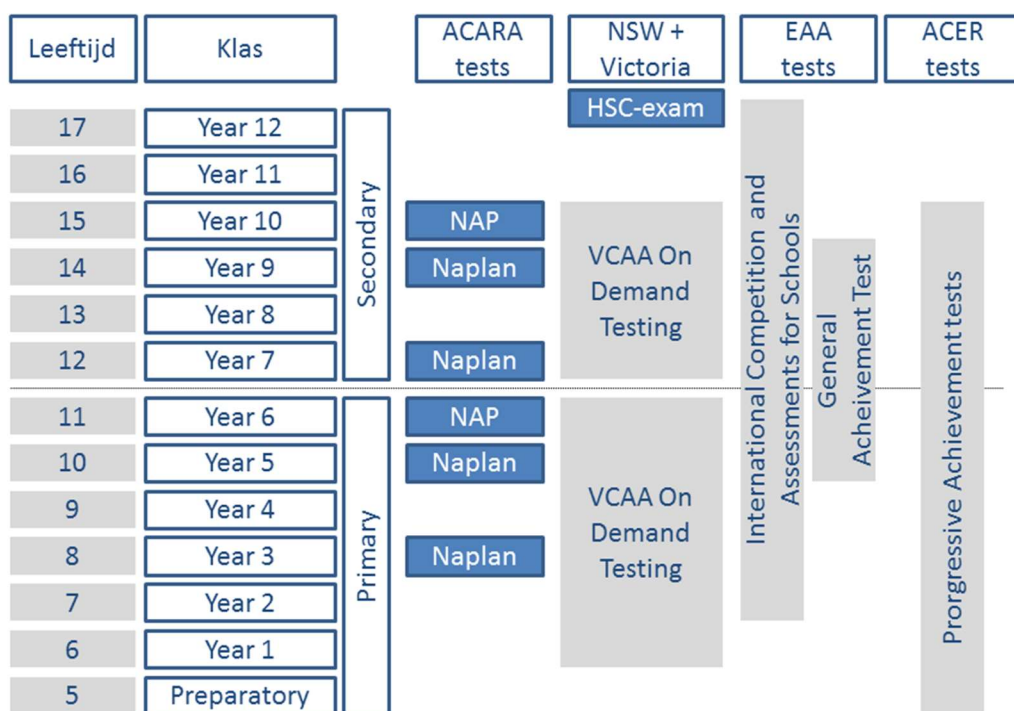
De visie op onderwijs wordt bepaald door de in 2008 door de federale regering goedgekeurde *Melbourne Declaration on Educational Goals for Young Australians*. In deze verklaring worden twee doelstellingen vooropgesteld voor het Australische onderwijs, met name: het Australische onderwijs bevordert gelijkheid en uitmuntendheid; en alle jonge Australiërs worden succesvolle 'learners', zelfverzekerde en creatieve individuen, en actieve en geïnformeerde burgers.

1.1.1 Onderwijsstructuur

Er zijn verschillen tussen staten/ *territories* inzake de onderwijsstructuur van het leerplichtonderwijs:

- In drie staten en twee territories omvat het lager onderwijs zeven jaren (één jaar kleuteronderwijs³ en leerjaren 1 tot en met 6). Het secundair onderwijs (in '*high schools*') omvat de leerjaren 7 tot en met 12, met uitzondering van Tasmania, waar het secundair onderwijs maar tot het tiende leerjaar duurt, gevolgd door de vervolgjaren (buiten de leerplicht) 11 en 12 (in '*colleges*').
- In de andere drie staten omvat het lager onderwijs het jaar kleuteronderwijs en de leerjaren 1 tot en met 7. Het secundair onderwijs begint er in het achtste leerjaar en duurt tot en met het twaalfde leerjaar.

Figuur 10 geeft een overzicht van de onderwijsstructuur in de hierboven als eerst vermelde staten en territories, inclusief de gestandaardiseerde toetsen die in deze case studie zullen besproken worden.



Figuur 10. Onderwijsstructuur in Australië, inclusief gestandaardiseerde toetsen

³ De benaming van dit jaar 'kleuteronderwijs' varieert tussen staten/territories. Gebruikte benamingen zijn 'pre-primary', 'transition', 'preparatory', 'reception' en 'kindergarten'.

Het leerplichtonderwijs beslaat het lager onderwijs en (delen van) het secundair onderwijs. Er is enige variatie in de leeftijd van kinderen waarvoor de leerplicht geldt. De start van de leerplicht is bepaald van vier jaar en vijf maanden; tot vijf jaar. De leerplicht eindigt in sommige staten/territories op 16 jaar, in andere op 17 jaar.

Het secundair onderwijs is erg comprehensief van aard: leerlingen worden niet ingedeeld in verschillende studierichtingen. In het elfde en twaalfde leerjaar worden veel keuzevakken aangeboden, die eerder academisch of eerder beroepsgericht van aard zijn. De enige centrale verplichting in de meeste staten/territories is dat de leerlingen twee vakken ('units') Engels volgen.

1.1.2 Leerlingen en scholen

In elk van de staten/territories zijn er zowel openbare scholen als private scholen. De openbare scholen vallen onder de verantwoordelijkheid van de Minister van Onderwijs van de betreffende staat/territory. De private scholen kunnen opgedeeld worden in 'katholieke scholen' en 'onafhankelijke scholen'. In elke staat/territory zijn de openbare scholen ruim in de meerderheid: tussen 66% en 80% van de scholen (zie Tabel 8)⁴ zijn openbare scholen; en tussen 58% en 72% van de leerlingen gaan naar deze openbare scholen (zie Tabel 9).

Tabel 8

Aantal scholen in Australië in 2014 (Australian Bureau of Statistics, 2015)

	Openbare scholen		Katholieke scholen		Onafhankelijke scholen		Totaal	
New South Wales	2165	70%	587	19%	331	11%	3083	100%
Victoria	1524	69%	489	22%	204	9%	2217	100%
Queensland	1232	72%	297	17%	186	11%	1715	100%
South Australia	525	73%	103	14%	93	13%	721	100%
Western Australia	775	72%	162	15%	139	13%	1076	100%
Tasmania	192	75%	37	14%	28	11%	257	100%
Northern Territory	152	80%	17	9%	21	11%	190	100%
Australian Capital Territory	86	66%	30	23%	14	11%	130	100%
Totaal Australia	6651	71%	1722	18%	1016	11%	9389	100%

⁴ Alle cijfermatige data zijn gebaseerd op Australian Bureau of Statistics (2015). Het betreft cijfers uit 2014.

Tabel 9

Aantal leerlingen in Australië in 2014 (Australian Bureau of Statistics, 2015)

	Openbare scholen		Katholieke scholen		Onafhankelijke scholen		Totaal	
New South Wales	762244	65%	253941	22%	150556	13%	1166741	100%
Victoria	563549	63%	204750	23%	128736	14%	897035	100%
Queensland	515501	67%	142913	18%	114895	15%	773309	100%
South Australia	165144	64%	48442	19%	46297	18%	259883	100%
Western Australia	258627	66%	69350	18%	63833	16%	391810	100%
Tasmania	55864	70%	14825	19%	9209	12%	79898	100%
Northern Territory	29395	72%	5053	12%	6199	15%	40647	100%
Australian Capital Territory	37205	58%	17992	28%	9245	14%	64442	100%
Totaal Australia	2387529	65%	757266	21%	528970	14%	3673765	100%

1.1.3 Curriculum

Tot 2011 werkten de scholen volgens een curriculum ontwikkeld in hun staat/territory. Sindsdien is er (zoals aangegeven in de *Melbourne Declaration on Educational Goals for Young Australians*) geleidelijk een nationaal curriculum ontwikkeld van het kleuteronderwijs tot en met het tiende (of twaalfde) leerjaar voor Engels, wiskunde, wetenschappen, kunstvakken, 'gezondheid en lichamelijke opvoeding', economie, vreemde talen, geschiedenis, technologie en '*work studies*'. In bijna alle vakken is dit nationale curriculum momenteel van kracht⁵.

Het nationale curriculum wordt ontwikkeld door het in 2008 opgerichte onafhankelijke agentschap ACARA (*Australian Curriculum, Assessment and Reporting Authority*). Zoals blijkt uit de naam van het agentschap, is ACARA ook verantwoordelijk voor een nationaal evaluatieprogramma en voor de rapportering van 'vergelijkbare data' over alle scholen (zie sectie 1.2).

Daarnaast zijn er curricula op het niveau van de verschillende staten/territories, en zijn er grote verschillen in hoe deze worden ingevuld. Terwijl er in bijvoorbeeld New South Wales een zeer gedetailleerd curriculum wordt uitgewerkt dat veel bijkomende specificatie biedt binnen het kader van het nationale curriculum, beperkt men zich in andere staten/territories tot het erkennen van het nationale curriculum, waarbij scholen zelf verdere invulling dienen te geven.

⁵ De enige uitzondering betreft het curriculum voor vijf (optionele) vreemde talen. In september 2016 moet ook het curriculum deze vreemde talen (waaronder klassieke talen en gebarentaal) operationeel zijn. De curricula voor de elf andere vreemde talen, is wel reeds van kracht.

1.1.4 Externe kwaliteitszorg

Australië kent geen centrale inspectie. Elke staat/territory heeft wel een systeem van externe kwaliteitszorg waarbij reviewers de openbare scholen bezoeken en controleren. Er zijn evenwel grote verschillen tussen staten/territories inzake de frequentie en de manier waarop deze controles worden uitgevoerd, alsook inzake wie de functie van 'reviewer' vervult: dit kan variëren van ambtenaren tot 'succesvolle' professionals of mensen met een academische achtergrond. Daarnaast heeft ook de koepelorganisatie voor katholieke scholen (*National Catholic Education Commission*) soortgelijke reviewers.

Ook de nationale gestandaardiseerde toetsen (NAPLAN, zie sectie 1.2.2) passen binnen de (externe) kwaliteitszorg van scholen.

1.2 Gestandaardiseerde toetsen op federaal niveau⁶

In het kader van de *Melbourne Declaration on Educational Goals for Young Australians*, werd onder de verantwoordelijkheid van ACARA het *National Assessment Program* (NAP) ontwikkeld. NAP bestaat uit twee luiken: de 'NAP Sample Assessments' en de 'National Assessment Program – Literacy and Numeracy' (NAPLAN). Daarnaast valt ook de deelname aan internationale studies (bijvoorbeeld PISA en TIMSS) onder het beleid van het NAP, maar deze liggen buiten het bestek van dit rapport.

1.2.1 Federaal niveau: NAP Sample Assessments

De 'sample assessments' van het *National Assessment Program* (NAP) worden sinds 2003 jaarlijks afgenomen in een steekproef van scholen in het zesde en tiende leerjaar. De toetsen meten afwisselend 'vaardigheid in informatie- en communicatietechnologie', 'wetenschappelijke vaardigheid' (enkel in het zesde leerjaar) en 'burgerschapzin', zodat elk van deze drie domeinen éénmaal om de drie jaar wordt getoetst. Deze toetsen zijn voornamelijk bedoeld om een nationaal overzicht te krijgen van de stand van zaken en vooruitgang inzake deze domeinen. Daarnaast kunnen scholen die niet tot de steekproef behoren, ook zelf kiezen om de toetsen af te nemen als bron van informatie binnen hun eigen kwaliteitszorg.

Sinds 2013 (waarin 'burgerschapzin' aan de beurt was) gebeurt de toetsafname online (met een back-up via USB-sticks). De steekproef bevatte in de toetsafname van 2015 ongeveer 12.000 leerlingen van ongeveer 700 scholen⁷ (nooit meer dan 20 leerlingen per school). Gewoonlijk worden niet alle items aan elke deelnemende leerling voorgelegd, zodat een ruimere scope van onderwerpen kan worden getoetst.

⁶ Gebaseerd op: ACARA (2012, 2013, 2014, 2015, s.d.); Australian Government: Department of Education Employment and Workplace Relations (2010); Cook (2015); Fraillon, Schulz, Nixon, en Gebhardt (2014); Klenowski (2014); Masters, Rowley, Ainley, en Khoo (2008); Ministerial Council on Education Employment Training and Youth Affairs (2008); Redden en Low (2012); Senate Standing Committee on Education and Employment (2014); Wu (2010) + informatie tijdens gesprekken in case studie.

⁷ De meest recente exacte cijfers die ter beschikking zijn, betreffen de toets van 2013. Toen namen 11.255 leerlingen uit 671 scholen deel.

De NAP Sample Assessments leiden tot een jaarlijks rapport over de leerprestaties van de Australische leerlingen en kunnen zo een input geven aan het onderwijsbeleid in de verschillende staten/territories. De rapportering geeft (voor zowel het zesde als tiende leerjaar) een overzicht van algemene resultaten van leerlingen op verschillende subdomeinen en een vergelijking met eerdere toetsafnames. De resultaten van verschillende staten/territories worden tevens naast elkaar geplaatst. Daarnaast worden de resultaten besproken voor jongens/meisjes, voor '*Indigenous*' leerlingen⁸, voor leerlingen met een andere moedertaal dan Engels en voor leerlingen die buiten Australië zijn geboren, en voor scholen in verschillende geografische locaties (stedelijk, provinciaal, of landelijk).

De toetsen worden elk jaar vernieuwd. Alle items worden (via een Rasch-analyse) op één schaal ingeschaald. Dit geldt zowel voor items in de toets van het zesde als voor de toets van het tiende leerjaar, waardoor de verschillen tussen het zesde en tiende leerjaar doorheen de tijd kan worden opgevolgd. Ongeveer 20% van de items wordt behouden van de voorgaande toetsafname, zodat er trends doorheen de tijd geïdentificeerd kunnen worden. De selectie van deze 'anker-items' is gebaseerd op een data-analyse van de items in de voorgaande toets. In een eerste fase van de toetsontwikkeling worden deze anker-items bepaald en worden nieuwe items toegevoegd. In een tweede fase worden de items in een online systeem gebracht en vindt een eerste kwaliteitscontrole plaats. De derde fase omvat een '*field trial*' met gemiddeld 2000 leerlingen in ongeveer 50 scholen, verdeeld over drie staten/territories. Op basis van de resultaten van deze '*field trial*' worden in een vierde fase alle items geëvalueerd. Deze eerste vier fasen nemen 18 maanden in beslag. De vijfde fase omvat dan de voorbereiding voor de afname, de toetsafname zelf, en de correctie van de toetsen. In een zesde en laatste fase worden de data geanalyseerd en rapporten geschreven. Deze twee laatste fasen nemen samen ongeveer 12 maanden in beslag.

1.2.2 Federaal niveau: NAPLAN

De NAPLAN-toetsen worden van alle leerlingen van het derde, vijfde, zevende en negende leerjaar afgenomen tijdens de vierde maand van het schooljaar. Een uitzondering kan gemaakt worden voor leerlingen met een '*significant or complex disability*' en voor niet-Engelstalige leerlingen die minder dan een jaar in Australië verblijven. Bovendien kunnen ouders een niet-deelname wegens 'religieuze overwegingen of filosofische tegenkanting tegen toetsen' invoeren.

Jaarlijks nemen meer dan één miljoen leerlingen deel aan de NAPLAN-toetsen. De toetsen meten in elk van de vier leerjaren de vaardigheid van alle leerlingen in vier leerdomeinen: schrijven, lezen, '*language conventions*' (grammatica, spelling) en rekenen. De implementatie van NAPLAN in 2008 had tot gevolg dat verschillende toetsprogramma's op staat/territory-niveau werden afgeschaft (bijvoorbeeld de *Achievement Improvement Monitor* van Victoria).

⁸ De term '*Indigenous*' kan worden vertaald als 'autochtoon'. Met de autochtone bevolking bedoelt men in Australië de Aboriginal-bevolking.

Verantwoordingsgerichte en ontwikkelingsgerichte doelstellingen

Met NAPLAN worden drie doelen nagestreefd, met name om:

1. overheden en schoolbesturen in staat te stellen om goed en slecht functionerende scholen te identificeren;
2. transparantie te bieden aan ouders over de performantie van verschillende scholen, en zo de betrokkenheid van ouders bij de school van hun kinderen te versterken;
3. scholen (en andere relevante stakeholders) informatie te bieden die hen helpt binnen hun interne kwaliteitszorg. Aan scholen wil men een antwoord bieden op drie vragen:
 - Hoe goed doen we het?
 - Hoe goed groeien we?
 - Hoe goed doen we het, en hoe goed groeien we in vergelijking met andere scholen?

De eerste twee doelen passen binnen het verantwoordingsgerichte perspectief op het mesoniveau (de school) en beogen als gevolg van deze verantwoording een indirect ontwikkelingsgericht doel: overheden en ouders kunnen op die manier een geïnformeerde actieve partner worden binnen de ontwikkeling van de school. De nadruk op verantwoording betekent evenwel niet dat scholen rechtstreeks kunnen afgerekend worden op de resultaten: er zijn immers geen automatische consequenties verbonden aan de toetsresultaten. Praktijken die gebezigd worden in andere landen (bijvoorbeeld intensievere inspectie, sluiting van scholen, ...) worden expliciet vermeden om neveneffecten als een onevenwichtige voorbereiding, teaching to the test en vervalsingen te vermijden. Daarnaast hebben de toetsresultaten in theorie ook geen gevolgen voor de individuele leerlingen: het is immers niet de bedoeling dat de individuele resultaten een gewicht krijgen in het kader van beslissingen inzake de verdere schoolloopbaan van leerlingen⁹. Hiermee hoopt men om ook het neveneffect van 'verhoogde stress voor leerlingen' te vermijden. ACARA benadrukt de vrijblijvendheid van de toetsresultaten sterk in haar communicatie met scholen¹⁰.

De derde doelstelling is ontwikkelingsgericht van aard, en is zowel gericht op het microniveau (bron van informatie voor leraren over de prestaties van leerlingen) als op het mesoniveau (bron van informatie voor de interne kwaliteitszorg van scholen) en het macroniveau (bron van informatie voor overheden om het onderwijs optimaal te ondersteunen).

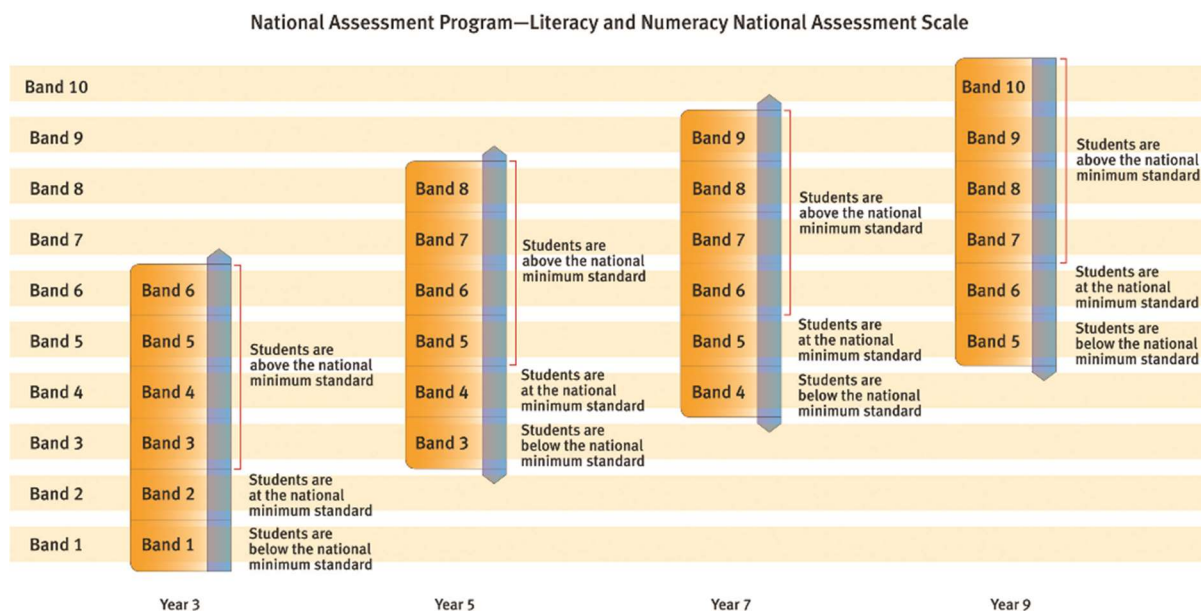
Toetsontwikkeling

ACARA coördineert de ontwikkeling van NAPLAN in overleg en samenwerking met deskundigen op het gebied van lezen, schrijven en rekenen. De eerste toetsen in 2008 waren (bij gebrek aan een nationaal curriculum op dat ogenblik) niet gebonden waren aan een overkoepelend curriculum. Tegenwoordig is dat wel het geval. Men meent dat de koppeling aan het nationale curriculum een essentiële voorwaarde is om er zeker van te zijn dat men toetst wat door leraren aan leerlingen wordt aangeleerd.

⁹ Een uitzondering hierop vormen de scholen in Western Australia, zoals we verder in deze sectie (zie 'neveneffecten') zullen verduidelijken.

¹⁰ Een goed voorbeeld hiervan is de brochure die ter beschikking is op http://www.nap.edu.au/verve/_resources/Acara_NAPLAN_Infographic%28V4-2%29.pdf

De toetsen worden elk jaar volledig vernieuwd. Via RASCH-analyse wordt de moeilijkheidsgraad van elk item bepaald. Op die manier kan de moeilijkheidsgraad van de verschillende toetsen worden vergeleken, alsook van de toetsen van verschillende leerjaren (binnen hetzelfde leerdomein). Op basis van de moeilijkheidsgraad van de items kan elke leerling toegewezen worden aan een bepaald vaardigheidsniveau (van 1 tot 10). Dit wordt visueel voorgesteld in Figuur 11.



Figuur 11. Vaardigheidsniveaus van de verschillende NAPLAN-toetsen

De toetsitems zijn allemaal meerkeuzevragen of vragen met een kort antwoord (met uitzondering van de toets voor schrijven). De open vragen (in de toets voor schrijven) worden beoordeeld door daartoe specifiek opgeleide ‘markers’. De toetsen worden totnogtoe op papier afgenomen.

Schrijf- en leesvaardigheid worden getoetst, niet enkel omdat deze makkelijk(er) te toetsen zijn, maar ook omdat deze vaardigheden als fundamentele basisvaardigheden worden beschouwd (zoals vastgelegd in de *Melbourne Declaration on Educational Goals for Young Australians*). Bovendien meent ACARA dat er een sterke relatie is tussen de prestaties van leerlingen in deze vaardigheden en de prestaties in andere leerdomeinen.

“If you’re gonna do a program that is focused on more literacy and numeracy, you have to acknowledge that schools do do a lot more, but from the government’s perspective, these are basic skills, basic rights.” (interview, respondent AU2)

De inhoud van de toetsen is relatief gemakkelijk, en meer dan 80% van de leerlingen slaagt erin om een vooropgestelde standaard te behalen. Toch is er de aanbeveling van een Onderzoekscommissie in de Australische Senaat om de inhoud van NAPLAN meer diepgravend te maken, waarbij meer complexe vaardigheden worden getoetst.

De ontwikkeling en review van toetsen duurt ongeveer 18 tot 24 maanden. Het proces bestaat uit verschillende stappen en is zo ontwikkeld dat een hoge standaard wordt gewaarborgd:

- in het licht van kwaliteitszorg worden in een eerste fase richtlijnen voor de ontwikkeling van de toetsen opgesteld en beoordeeld;
- toets-ontwikkelaars ontwerpen items die voldoen aan de toets-specificaties. ACARA besteedt dit proces uit aan organisaties met ervaring op het gebied van toetsontwikkeling;
- review van toetsitems, waarbij diverse partijen de geformuleerde items beoordelen in het licht van de overeenkomst met het curriculum en andere relevante bepalingen;
- representatieve steekproeven van studenten participeren in een pilot van de toets. De resultaten van de pilot worden geanalyseerd en vragen die voldoen aan de specificaties worden behouden;
- advisering van experts: *ACARA's Expert Advisory Panel* beoordeelt de toets, analyseert de resultaten van de pilot en adviseert voor het finaliseren van de toets;
- resultaten van NAPLAN toetsen uit verschillende jaren worden met elkaar vergeleken en beoordeeld op dezelfde schaal. Een tweede nationaal onderzoek met een representatieve steekproef van studenten wordt opgezet ten behoeve van het 'vergelijkingsprogramma';
- afname van de toetsen;
- experts op het gebied van onderwijsmeting analyseren de resultaten van de toetsen op verschillende manieren;
- resultaten van de NAPLAN toetsen worden gepubliceerd.

Toetsafname

De NAPLAN-toetsen worden ontwikkeld door ACARA, maar de afname en het toekennen van scores gebeurt onder de verantwoordelijkheid van de '*test administration authorities*' van de verschillende staten/territories op basis van de door ACARA ontworpen richtlijnen.

De NAPLAN-toetsen worden in alle scholen op dezelfde dagen afgenomen. De afname van de verschillende toetsen is gespreid over drie dagen, gevolgd door een inhaaldag waarop leerlingen die eerder afwezig waren, alsnog de toetsen kunnen afleggen. In het licht van efficiënte en gelijke afname van de toetsen, voorziet NAPLAN strikte voorschriften in het '*National Protocol for Test Administration*' over hoe en wanneer de toetsafname dient te gebeuren. Er zijn voorschriften inzake de 'veiligheid van het toetsmateriaal', de uniformiteit van de toetscondities, de informatie die wordt gegeven aan leerlingen, en de aanpassingen die kunnen gebeuren voor leerlingen met een beperking.

Publieke rapportering op schoolniveau: My School

In de *Melbourne Declaration on Educational Goals for Young Australians* werd bepaald dat er publiek moest worden gerapporteerd over scholen in de verschillende sectoren en in alle staten/territories. De toetsresultaten van NAPLAN worden sinds 2010 publiek bekend gemaakt via de website *My School*¹¹. Met als doel om transparantie aan te bieden inzake zoveel mogelijk aspecten van het onderwijs, biedt *My School* over elke school in Australië contextuele informatie (bijvoorbeeld socio-economische status

¹¹ <http://myschool.edu.au/>

leerlinginstroom), informatie over de financiële capaciteit (bijvoorbeeld de subsidiëring en financiën) alsook informatie over de 'academische prestaties' (bijvoorbeeld de leeruitkomsten, de graad van aanwezigheid, en de uitstroom van de leerlingen). De leerprestaties en de leerwinst van leerlingen (tussen het derde en vijfde jaar, het vijfde en zevende jaar, en het zevende en negende jaar) kunnen worden opgevolgd en vergeleken met eerdere cohorten van leerlingen in dezelfde school, of met andere scholen. De vergelijking met andere 'gelijkaardige' scholen gebeurt op basis van de *Index of Community Socio-Educational Advantage* (ICSEA).

"You can build in fairness into the system by how you compare schools. You don't list schools A to Z but by student populations or by other facts." (interview, respondent AU2)

De ICSEA houdt rekening met socio-economische factoren van de leerlingen (beroep en onderwijsniveau van ouders), de geografische ligging van de school en het aandeel van 'Indigenous' leerlingen. Daarbuiten wordt er evenwel op geen andere manier 'toegevoegde waarde' berekend van scholen, omdat dit te complex wordt geacht voor zowel scholen als voor het bredere publiek. ACARA maakt ook geen rankings op.

De aangeboden transparantie van *My School* heeft expliciet zowel een verantwoordingsgericht als een ontwikkelingsgericht perspectief. Men verwacht dat scholen aan de hand van deze informatie hun eigen kwaliteit kunnen en willen verbeteren. Tegelijkertijd vindt men het belangrijk om via deze website transparantie aan te bieden aan het publiek en tegelijkertijd scholen verantwoording te laten afleggen voor de leerprestaties en de leerwinst van hun leerlingen. Deze transparantie wordt onder meer verantwoord door een enquête die werd uitgevoerd door Salesforce (in opdracht van het *Department of Education, Employment and Workplace Relations*) waarbij 83,2% van de ondervraagde ouders aangaf dat deze informatie over scholen publiek beschikbaar zou moeten zijn. Zelfs 96,9% meende dat ze over deze informatie zouden moeten beschikken over de school waar hun eigen kinderen schoollopen¹².

Men probeert evenwel de kans op rankings te minimaliseren door het vergelijken van scholen te beperken. Zo kan een gebruiker in *My School* op een zelfde grafiek slechts vijf scholen selecteren¹³. Wanneer individuen websites opzetten of publicaties uitbrengen vanuit 'ongewenste doelstellingen' wordt er druk uitgeoefend op de initiatiefnemer, en indien mogelijk worden er gerechtelijke stappen ondernomen om de ongewenste publicaties terug te roepen. Dit geldt evenwel niet voor bijvoorbeeld de pers. Ongenuanceerde rankings van scholen komen in de pers echter slechts in beperkte mate voor.

¹² Salesforce (2008) in Australian Government: Department of Education Employment and Workplace Relations (2010).

¹³ Om rankings te vermijden, heeft een recent rapport van een Onderzoekscommissie in de Senaat geadviseerd om deze mogelijkheid om verschillende scholen met elkaar te vergelijken, in een volgende fase te verwijderen (Senate Standing Committee on Education and Employment, 2014).

Australië heeft in het algemeen een sterke traditie van openbaarheid van gegevens in alle publieke domeinen. Daardoor lijkt het voor verschillende stakeholders ook meer vanzelfsprekend dat toetsresultaten publiek beschikbaar zijn, zoals blijkt uit een recent onderzoeksrapport (in opdracht van ACARA) naar de aanvaarding van My School:

"Parents had generally not given this issue [het publieke karakter van de data beschikbaar op My School] prior thought and, when questioned, found it difficult to see a reason why the information on the My School website would not be freely available in the public domain." (Colmar Brunton, 2014, p. 5)

"Parents, principals, partners and other stakeholders all accept that the community has a right to know the outcomes of NAPLAN testing and the other school-related information available on the website" (Colmar Brunton, 2014, p. 6).

My School werd van meteen na de implementatie uitvoerig gebruikt: in de eerste 12 maanden na de lancering van de website, werden meer dan 17 miljoen *page views* geregistreerd. Uit het eerder vermelde onderzoek in opdracht van ACARA blijkt dat het beschikbaar stellen van deze schooldata via *My School* overigens niet echt een breuk was met de eerdere traditie: ook voor 2010 waren deze gegevens reeds beschikbaar voor de geïnteresseerde ouder, vaak via de communicatiekanalen van de school zelf. De implementatie van *My School* maakt deze gegevens enkel nog meer toegankelijk. *My School* wordt dan ook sterk gewaardeerd door de ouders, al zijn er mogelijkheden om de data op een nog meer gebruiksvriendelijke manier aan te bieden, en raadt het onderzoeksrapport aan om nog meer duiding te voorzien bij de statistische gegevens.

Rechtstreekse rapportering aan ouders

De toetsresultaten worden drie maanden na de afname van de toetsen ook aan de ouders van leerlingen bezorgd. Naast een algemene score van de leerling, krijgen ouders informatie voor elk van de items in de toets. Door de gemeenschappelijke schaal die in de verschillende toetsen wordt gebruikt, kunnen ouders de voortgang van hun kinderen opvolgen.

Bovendien toont dit rapport niet enkel de resultaten van de betreffende leerling, maar ook het nationale gemiddelde en enkele ijkpunten (het twintigste en tachtigste percentiel) van alle leerlingen, en het gemiddelde van de school van de betreffende leerling.

Impact

Ondanks de uiteenlopende doelstellingen, wordt NAPLAN door de verschillende actoren in het onderwijsveld voornamelijk als een middel tot verantwoording van scholen gepercipieerd. Dit heeft sterk te maken met de publieke rapportering via *MySchool*.

Microniveau

Wat de ontwikkelingsgerichte doelstelling betreft, zijn verschillende respondenten het erover eens dat de toetsresultaten op microniveau (als bron van informatie voor leraren over de vaardigheid van leerlingen) onbruikbaar zijn, in de eerste plaats door de lange periode tussen de afname van de toets en het ontvangen van de resultaten. Bovendien is het bekend dat steeds vaker de minst sterke

leerlingen worden 'vrijgesteld' van toetsafname (zie sectie over neveneffecten), waardoor de diagnostische waarde van de toets voor deze belangrijke doelgroep helemaal wegvalt.

In secundaire scholen zijn de NAPLAN-scores soms wel van invloed op het microniveau. Zowel de data van het vijfde leerjaar, als de data van het zevende leerjaar, kunnen als zinvolle informatie beschouwd worden over wat een instromende leerling al dan niet kan. Hiervoor worden evenwel vaker ook andere gestandaardiseerde toetsen gebruikt (zie secties 1.3.2 en 1.4).

Macroniveau

Een analyse van de resultaten van 2008 tot 2015 leert bovendien dat de introductie van NAPLAN niet heeft gezorgd voor een globale verbetering van de leerlingenprestaties, en evenmin de kloof in de prestaties tussen Indigenous en andere studenten heeft verkleind. Dit duidt er op dat NAPLAN ook op macroniveau weinig effectief is. Het gebrek aan effect op leeruitkomsten werd door de media bij de publicatie van de resultaten in 2015 sterk in de verf gezet.¹⁴ Dat NAPLAN daarom geen meerwaarde heeft, wordt evenwel betwist door ACARA, omdat eerder een conceptueel effect bij scholen en bij het ruimere publiek wordt beoogd, eerder dan een rechtstreekse impact op leerlingresultaten:

"The headlines were: 'after all these years, Naplan hasn't improved education'... Well: it wasn't meant to!" (interview, respondent AU1)

Bovendien leidt NAPLAN ook niet tot een verantwoording op het niveau van de staten/territories, omdat onder meer door *My School* de focus sterk komt te liggen op de verantwoording van individuele scholen (data op *My School* worden immers niet geaggregeerd op een hoger niveau dan de school). Toch gebeurt het dat de regering van een staat/territory de resultaten aangrijpt om het eigen succes te beklemtonen¹⁵.

Mesoniveau

Hoewel empirisch onderzoek ontbreekt, blijkt uit de interviews dat NAPLAN zorgt voor een conceptueel effect op mesoniveau: directies en leraren worden door de data aangezet tot reflectie en een dialoog over hun onderwijskwaliteit. Zelfs wanneer de resultaten weinig verrassend zijn, wordt de bevestiging van het eigen idee als waardevol beschouwd - dat heeft niet in het minst te maken met de overzichtelijke beschikbaarheid van longitudinale data:

"It's a good message for us [a primary school] that we're staying on track, that we keep progressing; that's why we use the data." (interview, respondent AU13)

Ondanks het gebrek aan effect op macroniveau, meent men bij ACARA dat NAPLAN op sommige scholen een impact op de onderwijskwaliteit heeft: *"some schools go up, some schools go down"* (interview, respondent AU2). Door NAPLAN is, volgens de organisatie, in de eerste plaats de discussie over onderwijskwaliteit in scholen en omtrent scholen versterkt door het lokale raamwerk te

¹⁴ Bijvoorbeeld: B. Brennan (2015); Dodd (2015); A. Smith (2015)

¹⁵ Een voorbeeld hiervan is het persbericht van de regering van Australian Capital Territory Government (2014) beschikbaar op http://www.cmd.act.gov.au/open_government/inform/act_government_media_releases/burch/2014/act-students-top-the-nation-in-naplan

verbreden met gegevens op nationaal niveau. Bovendien werden veel scholen voor het eerst geconfronteerd met leerwinstgegevens. Daarom is men ervan overtuigd dat men een sterk conceptueel effect bereikt op mesoniveau in de scholen die sterke resultaten behalen, maar waarbij de leerwinst relatief beperkt is (de '*coasting schools*'). In tegenstelling tot bij vorige programma's - die zelden of nooit leerwinst in kaart brachten - worden deze scholen er nu immers mee geconfronteerd dat ze hun goede resultaten vooral dienen toe te schrijven aan hun gunstige leerlinginstroom, en slechts in beperkte mate aan hun eigen inbreng. Empirische evidentie voor deze stelling van ACARA dat de discussie in en tussen scholen is versterkt, ontbreekt evenwel.

Ten slotte meent men bij ACARA dat de betrokkenheid van ouders bij de ontwikkeling van scholen versterkt werd, omdat deze ouders nu beschikken over een geïnformeerde basis. Deze stelling wordt evenwel door andere respondenten en door het eerder aangehaalde onderzoek van Colmar Brunton (2010) gecontesteerd.

Ten slotte wordt wel een 'strategisch effect' gerapporteerd: het is gebruikelijk voor scholen dat goede NAPLAN-resultaten publiek in de verf worden gezet.

Kritiek en neveneffecten

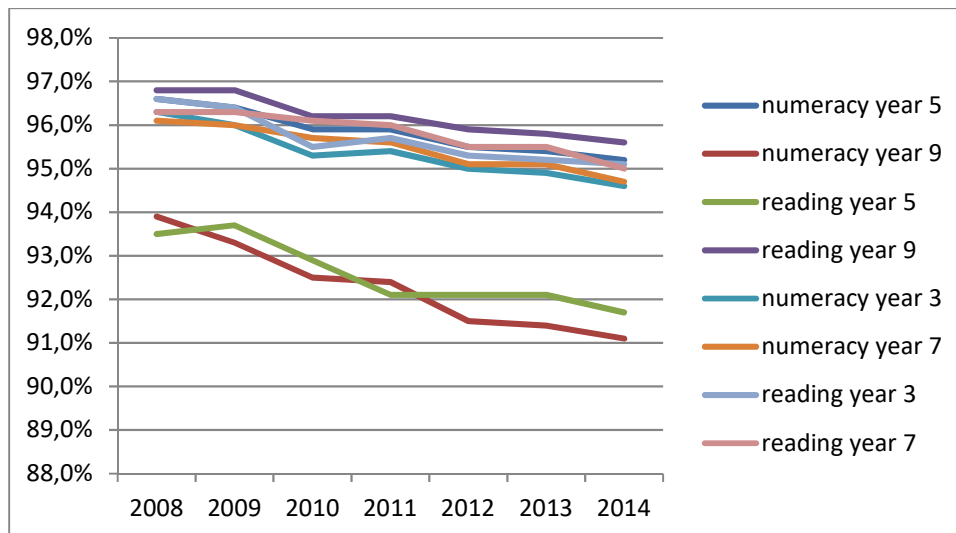
De mate waarin er zich onwenselijke neveneffecten voordoen, hangt sterk samen met de perceptie van de inzet ('*stakes*') van NAPLAN-toetsen voor scholen, leraren en leerlingen.

Lage of hoge inzet?

Zoals hierboven gesteld, tracht ACARA de '*stakes*' voor scholen en voor leerlingen zo laag mogelijk te houden door zo weinig mogelijk consequenties aan de resultaten te verbinden. Zo werden in 2014 slechts 80 incidenten uit 73 scholen gerapporteerd waarbij er een mogelijke overtreding was van de *National Protocols for Test Administration* (waarvan slechts 21 incidenten in verband met (mogelijke) fraude), al is men er zich van bewust dat mogelijk niet alle incidenten worden gemeld.

Toch wordt uit verschillende andere bronnen vernomen dat de NAPLAN-toetsen helemaal niet zo vrijblijvend zijn voor de school. Onder meer door de publieke rapportering ervaren vele scholen de inzet als relatief hoog. Zeker in de sterk competitieve markt van scholen in de grote steden, is het belangrijk om een goed publiek rapport te krijgen. Het jaar na jaar groter wordend aandeel van leerlingen die worden 'vrijgesteld' van deelname aan NAPLAN (zie Figuur 12), is volgens critici een duidelijk signaal hiervan; net als de groeiende markt van externe consultants die door scholen worden ingehuurd om NAPLAN voor te bereiden.

Momenteel woedt bovendien in verschillende staten/territories een discussie of er op schoolniveau geen consequenties aan de resultaten moet worden verbonden. Onder meer in Victoria werd recent door het kabinet van Minister van Onderwijs James Merlino aangekondigd dat werd onderzocht hoe men de financiering van scholen afhankelijk kan maken van de NAPLAN-resultaten, waarbij 'goede scholen' kunnen beloond worden.



Figuur 12. Graad van deelname aan NAPLAN tests lees- en rekenvaardigheid in de periode 2008-2014¹⁶

Bovendien wordt uit verschillende bronnen buiten ACARA vernomen dat een omvangrijk aantal scholen de toetsen op eigen initiatief wel hanteert als een bron van informatie bij de evaluatie van leraren.

Daarnaast hebben de resultaten van de NAPLAN-toetsen soms ook een impact op de studieloopbaan van leerlingen. Zo worden de resultaten van het negende leerjaar in de staat Western Australia sinds 2015 ook op een andere manier gebruikt. Leerlingen die op de NAPLAN-toetsen een bepaalde score behalen, zijn immers vrijgesteld van een high-stakes toets die door andere leerlingen in het tiende leerjaar dient te worden afgenomen om te bepalen of ze voor een diploma in aanmerking komen. Dit leidt tot een aanzienlijk hogere beleving van stress bij de leerlingen in het negende leerjaar in deze staat. Opmerkelijk is evenwel ook dat de resultaten van de leerlingen in deze staat opvallend hoger lagen dan in andere staten - een bevinding die in voorgaande jaren niet kon worden vastgesteld. Wellicht is deze plotse groei in leerlingenprestaties toe te schrijven aan de extrinsieke motivatie van leerlingen om een goed resultaat te behalen op de NAPLAN-toetsen in het negende jaar in Western Australia: de hogere inzet voor leerlingen leidde tot een sterkere motivatie en bijgevolg betere resultaten.

Het verantwoordingsgericht gebruik op microniveau is evenwel niet beperkt tot Western Australia: ook bepaalde selectieve scholen gebruiken de resultaten op NAPLAN in hun toelatingsbeleid.

“There is a growing market in terms of families taking their children to coaching for the purpose of succeeding in NAPLAN. You will find that a lot of non-government [secondary] schools, particularly the ‘lead ones’ are looking to NAPLAN scores as one of the indicators for entering or getting a space in their school.” (interview, respondent AU5)

¹⁶ Op basis van data van ACARA op <http://reports.acara.edu.au/Home/Participation>

Betrouwbaarheid en validiteit

Desondanks zijn er grote bezorgdheden bij ouders, scholen en andere stakeholders omtrent de betrouwbaarheid en de validiteit van de resultaten die NAPLAN oplevert. Men twijfelt er aan of NAPLAN een goede weergave kan zijn van een school, aangezien sommige scholen meer dan andere een grote nadruk leggen op het 'coachen' van leerlingen voor de NAPLAN-toetsen (*teaching to the test*, of extensieve toetsvoorbereiding) of de minst sterke leerlingen niet laten deelnemen aan de toetsen (zie ook bij 'neveneffecten').

Daarnaast is er vanuit academische hoek kritiek op de betrouwbaarheid (de meetfouten zouden te groot zijn om betrouwbare uitspraken te doen) en de validiteit van de toets bij de meting van leerprestaties van Indigenous leerlingen¹⁷. Er wordt betoogd dat deze leerlingen - waarvan Engels vaak niet de moedertaal is - benadeeld worden door het testformat. Een ander probleem is dat er relatief veel kleine scholen zijn in afgelegen gebieden van Australië. Voor deze kleine scholen is het moeilijk om betrouwbare data op schoolniveau te genereren. Daarnaast blijkt uit de bevraging van Colmar Brunton (2010) dat stakeholders vragen hebben bij de factoren die in rekening worden genomen bij het bepalen van de ICSEA (de maat om vergelijkbare scholen te selecteren)¹⁸.

Gevaar voor verkeerde interpretaties

Directies en andere stakeholders beklemtonen in het eerder vermelde onderzoek door Colmar Brunton (2010) het gevaar dat ouders de schoolkeuze voor hun kinderen laten beïnvloeden door informatie die inaccuraat kan zijn (cf. de hierboven vermelde bezorgdheid in verband met de betrouwbaarheid en validiteit van de resultaten van NAPLAN), of makkelijk tot verkeerde interpretaties kan leiden. Het is onduidelijk of ouders voldoende betekenis kunnen geven aan de gepresenteerde data - een bezorgdheid die na een bevraging van ouders door de onderzoekers van Colmar Brunton wordt bestempeld als een (gedeeltelijk) terechte bezorgdheid. Mogelijke aanleidingen tot verkeerde interpretaties zijn dat de ICSEA als maat voor de leerling- en contextkenmerken van de school een erg moeilijk te interpreteren begrip blijkt voor ouders en dat de resultaten van leerlingen in het zevende leerjaar als maat kunnen worden geïnterpreteerd voor de kwaliteit van een secundaire school, terwijl leerlingen op dat ogenblik nog maar vier maanden op deze school aanwezig zijn. Voor verschillende stakeholders weegt dit risico op verkeerde interpretaties door ouders zelfs zwaarder dan de gepercipieerde voordelen van NAPLAN. Toch blijkt dat *My School* slechts één van de verschillende informatiebronnen is die door de meeste ouders worden gebruikt bij de bepaling van hun schoolkeuze.

Teaching to the test

Wanneer met *teaching to the test* wordt bedoeld dat men zich meer focust op kwaliteitsvol onderwijs gericht op de basisvaardigheden van lees- en schrijfvaardigheid (zonder de andere delen van het curriculum te negeren), zijn de verschillende respondenten bij de interviews het erover eens dat dit geen probleem is: deze vaardigheden zijn, zeker in het lager onderwijs, aanvaard als de essentiële bouwstenen van het onderwijsproces. We horen bij geen enkele respondent een pleidooi voor bredere toetsing, onder meer om de toetsdruk op leerlingen en scholen niet te verhogen.

¹⁷ Bijvoorbeeld Klenowski (2014); Wu (2010)

¹⁸ Factoren die volgens directies en andere stakeholders mee in de ICSEA zouden moeten worden verrekend, zijn de geografische locatie en de schoolgrootte.

"The biggest emphasis obviously in Australian school is: literacy and numeracy. Other subjects like science get a lot of promotion, but there's not a lot of money around for them, so schools have to try as hard as they can to be able to develop curriculums around those courses. But then again, if you have other subjects that you're actually gonna be testing, then you're testing just for the sake of testing all the time - which I don't think is a good thing. NAPLAN is in three days or four days and it's finished. If you would add in other subjects, you would need two weeks." (interview, respondent AU13).

Toch ziet men, in beperkte mate, een andere invulling van teaching to the test, met name: dat een aantal leraren in de maand voor de afname van de toetsen zich sterk verliezen in specifieke toetsvoorbereiding; en dat het curriculum wordt vernauwd tot enkel die zaken die getoetst zullen worden (bijvoorbeeld het onderwijs in schrijfvaardigheid wordt verengd tot 'persuasive writing' omdat men weet dat dit aan bod zal komen tijdens NAPLAN).

Volgens ACARA moeten bovenstaande voorbeelden van teaching to the test echter toegeschreven worden aan het slechte functioneren van deze leraren:

"But if you do that, you have a bigger problem than just the test scores." (interview, respondent AU2)

Deze uitspraak wordt gebaseerd op het (door ACARA gepercipieerde) low-stakes karakter van de toetsen, en het feit dat de standaard vrij laag is: omdat de toets in de vierde maand van een lopend schooljaar wordt afgenomen, wordt immers voornamelijk leerstof van het voorbije leerjaar getoetst. De visie van ACARA wordt slechts gedeeltelijk gedragen door een recente onderzoekscommissie in de Australische Senaat, die er op wijst dat deze neveneffecten dan wel mogelijk te wijten zijn aan een verkeerde interpretatie van de bedoelde inzet van NAPLAN, maar dat ze daarom niet minder belangrijk zijn:

"To suggest that changes in the classroom are not as a result of NAPLAN is not taking full responsibility for the profound impact that standardized testing can have. This in itself is not a reason not to test, but it is something that educational authorities need to be cognizant of in providing support to schools as part of the NAPLAN process." (Senate Standing Committee on Education and Employment, 2014, p. 14)

Om een vernauwing van het curriculum te vermijden wordt bij ACARA sterk het belang benadrukt om als overheid nadrukkelijk te erkennen dat een school meer doet dan enkel lees- en rekenvaardigheden aanleren, maar tegelijk te beklemtonen dat lees- en rekenvaardigheid in die mate fundamentele vaardigheden zijn dat ze als 'basisrecht' voor elke leerlingen moeten beschouwd worden en daarom extra aandacht verdienen.

Streaming, ongelijkheid en exclusie

Daarnaast wordt door onderzoek een sterkere 'streaming' en een grotere homogeniteit in groepssamenstelling vastgesteld als gevolg van de druk uitgeoefend door NAPLAN om goede

resultaten te behalen; wat door de onderzoekster in kwestie wordt genoemd als *“another perverse effect of the current approach to accountability”* (Klenowski, 2014, p. 451). Een andere kritiek op NAPLAN luidt dat het de ongelijkheid tussen *Indigenous* leerlingen en andere leerlingen nog vergroot, enerzijds omdat het testformat weinig is afgestemd om de leercultuur van de eerste groep, anderzijds omdat ouders van de midden- en hogere klassen gemakkelijker de informatie van *My School* zullen gebruiken om hun school aan te zetten tot een kwaliteitsverbetering.

Stress

Dat NAPLAN bij een zeker aandeel van de leerlingen een zekere stress oplevert, vindt men op zich niet zo erg bij ACARA. Niet alleen meent men dat door het (gepercipieerde) low-stakes karakter de verhoging van stress niet zo groot is, maar ook omdat men dit beschouwt als een deel van en voorbereiding op het leven. Vaak zijn het volgens ACARA de ouders of de leraren die de toetsafname meer stresserend maken, eerder dan de NAPLAN toets zelf. Ook hier deelt de Senaatscommissie niet geheel de mening van ACARA:

“The committee accepts that anxiety will be a factor for some students in any testing environment. What makes NAPLAN of particular significance is the perception of NAPLAN as a high-stakes test.” (Senate Standing Committee on Education and Employment, 2014, pp. 16-17)

Financiën

ACARA ontvangt een jaarlijkse subsidie van 12,175 miljoen Australische Dollar (AUD) (ongeveer 7.750 miljoen €) voor 50% van haar werking van de federale overheid. De andere 50% komt van de overheden van staten/territories¹⁹. Bij ACARA maakt men zich sterk dat minder dan 0,5% van het onderwijsbudget gespendeerd wordt aan NAPLAN en de NAP sample assessments.

Daarnaast valt de disseminatie en de correctie van de NAPLAN-toetsen onder de verantwoordelijkheid van de staten/territories: deze kost is dus niet inbegrepen in bovenstaande cijfers van ACARA. Een vaak in de media gebruikt cijfer is dat de volledige ontwikkeling en afname van NAPLAN-toetsen jaarlijks 100 miljoen AUD (ongeveer 67,3 miljoen €) kost.

Voor de ontwikkeling en het onderhoud van de My School website werd daarnaast een budget van 1,507,907 AUD (ongeveer 960,000€) besteed aan externen in het eerste boekjaar (2010-2011), en daarna tussen 707.000 en 747.000 AUD (tussen 450,000 en 475,000€) voor elk van de volgende jaren.

Toekomstplannen: online en adaptief toetsing

ACARA is zich ervan bewust dat de impact van NAPLAN onder enkele voorwaarden groter zou kunnen zijn voor de kwaliteitszorg van scholen en voor het leren van leerlingen. Men plant daarom om het toetsaanbod aan verschillende wijzigingen te onderwerpen:

¹⁹ Op basis van budgetinschattingen beschikbaar gemaakt door Senate Standing Committee on Education Employment and Workplace Relations (2013)

"[Respondent AU2] said teachers wouldn't like it... he was right. We're gonna raise the discussion about the use of data. We are also planning to make NAPLAN more useful for teachers, so our full cohort tests, NAPLAN, a million students across the country, what we are doing, we're gonna move it online, we gonna use a form of adaptive testing, we gonna get the results back to them more quickly and we're going to get, you know, stronger links back to the curriculum." (interview, respondent AU1)

De doelstelling van ACARA is om vanaf 2017 NAPLAN online te kunnen aanbieden. Deze doelstelling werd bevestigd in de aanbevelingen van een recente onderzoekscommissie in de Australische Senaat (Senate Standing Committee on Education and Employment, 2014). Individuele staten/territories zullen evenwel autonoom kunnen beslissen vanaf welk moment ze de online toetsing zullen implementeren. In het licht van deze doelstelling werd reeds in drie opeenvolgende schooljaren (2012, 2013 en 2014) pilootonderzoek uitgevoerd, waarbij onder meer werd nagegaan in welke mate items anders werden beantwoord door leerlingen wanneer ze worden gepresenteerd in verschillende modi (papier versus online), de mate waarin de automatische beoordeling effectief is, en de mate waarin verschillende antwoordmogelijkheden en toepassingen (drop-down menu's, items bewegen over het scherm, tekst aanduiden en een audio- of videobestand afspelen) effectief zijn. Door online toetsing beoogt men betere meetresultaten te verkrijgen, en sneller te kunnen rapporteren over de resultaten. Uit een deelonderzoek van de pilootstudies waarbij de houding van leerlingen werd nagegaan, wordt geconcludeerd dat leerlingen geen voorbehoud hebben ten opzichte van de online toets. Integendeel, leerlingen voelen zich meer zelfverzekerd en meer betrokken bij de online toets in vergelijking met een toets op papier.

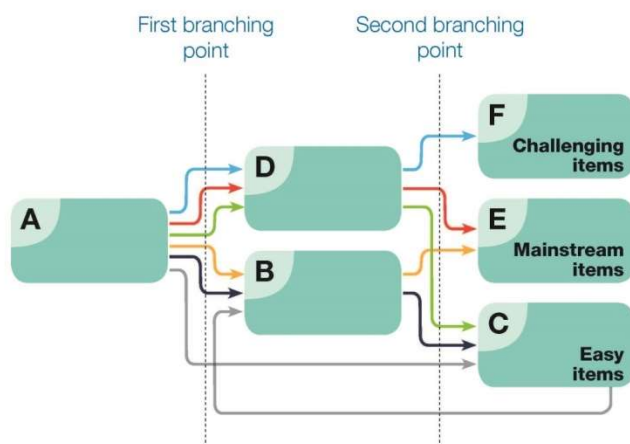
Daarnaast zouden bij online toetsing de items in de rapportering meteen kunnen gelinkt worden aan delen van het curriculum. Dit laatste is meteen een aanzet tot een ander voornemen, namelijk om aan leraren bronnen aan te bieden om met bepaalde moeilijkheden, die blijken uit de NAPLAN-resultaten, om te gaan. Met andere woorden wil ACARA aan de geautomatiseerde data ook concreet ondersteuningsmateriaal koppelen. Dit beoogt men voornamelijk op school- en op klasniveau (en minder op het niveau van individuele leerlingen). Er zijn evenwel nog geen concrete stappen in die richting gezet.

Ten slotte meent men dat bij online toetsing de score voor schrijfvaardigheid ook kan bepaald worden door de geschreven antwoorden van leerlingen in de andere toetsen. Of andersom, zou de aparte spelling- en grammatica-toets kunnen wegvallen door ze te integreren in de toets voor schrijfvaardigheid.

Bovendien wil men samen met de overgang naar online toetsen, ook de overgang maken naar adaptief toetsen. Dit werd reeds uitgetoetst in de tweede pilootstudie (2013) waarbij 23.000 toetsen in 250 verschillende scholen werden afgenomen om na te gaan in welke mate de online toets betere informatie oplevert over *'underperforming students'*.

De plannen betreffen een getrapte adaptieve toets (*'branching'*), waarbij de leerling drie van zes itemsets krijgt aangeboden (zie Figuur 13). Alle leerlingen beginnen bij itemset A. Op basis van hun

antwoorden krijgen ze nadien de makkelijkere itemset B of de moeilijker itemset D aangeboden. Nadat leerlingen itemset B of D volledig hebben afgewerkt, wordt hen op basis van hun voorlopige prestaties de itemset C, E of F aangeboden met respectievelijk een lage, medium en hoge moeilijkheidsgraad. Op die manier kan een grotere precisie mogelijk worden gemaakt bij de diagnosticering van sterke en zwakke leerlingen.



Figuur 13. Testdesign getrappt adaptieve toetsing²⁰

1.3 Gestandaardiseerde toetsen op state/territory-niveau ²¹

Daarnaast worden door elke staat/territory verschillende toetsen (al dan niet verplicht) aangeboden. We onderzoeken deze toetsen aan de hand van twee case studies: de summatieve toetsen voor leerlingen aan het eind van het secundair onderwijs in New South Wales, en de formatieve ‘On Demand’ toetsen die worden aangeboden in de staat Victoria.

1.3.1 *Summatieve toetsen op niveau van de individuele staten - case New South Wales*

BOSTES (*Board of Studies, Teaching and Educational Standards NSW*) is verantwoordelijk voor het curriculum, de evaluatie, het lesgeven en het bepalen van de standaarden die in de scholen in de staat New South Wales worden nagestreefd. Tot de operationele taken behoort onder meer het ontwerpen van het curriculum, de ontwikkeling en afname van de eindexamens, en de afname van de NAPLAN tests. Wat betreft dat laatste, is het eigenlijke werk van BOSTES beperkt: net als de meeste andere staten/territories, wordt hiervoor een externe firma ingeschakeld (met name Pearson Research and Assessment). Deze firma zorgt voor het printen en verdelen van de examens, voor de correctie en de rapportering van NAPLAN. Aangezien deze taak door verschillende staten/territories toegekend wordt aan dezelfde firma, wordt de gelijkheid in het toekennen van scores aan leerlingen over de verschillen staten/territories gegarandeerd.

²⁰ Bron: <http://www.nap.edu.au/online-assessment/research-and-development/tailored-tests.html>

²¹ Gebaseerd op: BOSTES (2015a, 2015b); O'Mara (s.d.); Tertiary Institutions Service Centre (s.d.); Universities Admissions Centre (2015); Victorian Curriculum and Assessment Authority (2013) + informatie tijdens gesprekken in case studie.

HSC-examen

In verband met de taak 'ontwikkeling en afname van de eindexamens', ontwikkelt BOSTES jaarlijks voor elk van de 129 mogelijk gekozen vakken een apart examen. Elk domein wordt getoetst omdat elk vak in principe gelijk gewaardeerd wordt. Dit zijn de HSC-examens²². Deze HSC-examens dienen door alle leerlingen van het twaalfde leerjaar te worden afgelegd (al kunnen leerlingen van het elfde leerjaar zich ook reeds voor bepaalde examens inschrijven). Ook in elk van de andere staten/territories worden eindexamens afgenomen van alle leerlingen in het twaalfde leerjaar, weliswaar met erg uiteenlopende benamingen (bijvoorbeeld 'Queensland Core Skills Test', 'General Achievement Test' (in Victoria) en 'Western Australia Certificate of Education Exams'). De examens zijn schriftelijk, met uitzondering van een aantal vakken waarin observatie noodzakelijk is (bijvoorbeeld drama, muziek, spreekvaardigheid in een vreemde taal,...).

"It can include a written examination, it can also include a performance or a project, or an artwork if you do these things. Then students send it in to the markers, sometimes, or we go to the school and mark it in the school. So we have teams of markers traveling around the state, going to listen to students' play or watch them dance... And there are quality mechanisms in place to make the marking as consistent as possible."
(interview, respondent AU6)

Studenten hebben een vrij grote keuze inzake de vakken waarin ze examen afleggen; Er zijn examens in maar liefst 129 vakken, gespreid over de leerdomeinen Engels (het enige leerdomein waarvan leerlingen verplicht een examen moeten afleggen), Wiskunde, Wetenschappen, Technologie, Creatieve Kunsten, Persoonlijke ontwikkeling, Gezondheid en Lichamelijke Ontwikkeling, Samenleving en Omgeving, Talen, en Beroepsgericht onderwijs.

Elke leerling wordt op basis van zijn score in elk van de afgelegde examens toegewezen tot een bepaald vaardigheidsniveau ('band'). Dat vaardigheidsniveau varieert van 1 tot 6, waarbij 6 het hoogste niveau is. Leerlingen in vaardigheidsniveau 1 zijn niet geslaagd voor het examen.

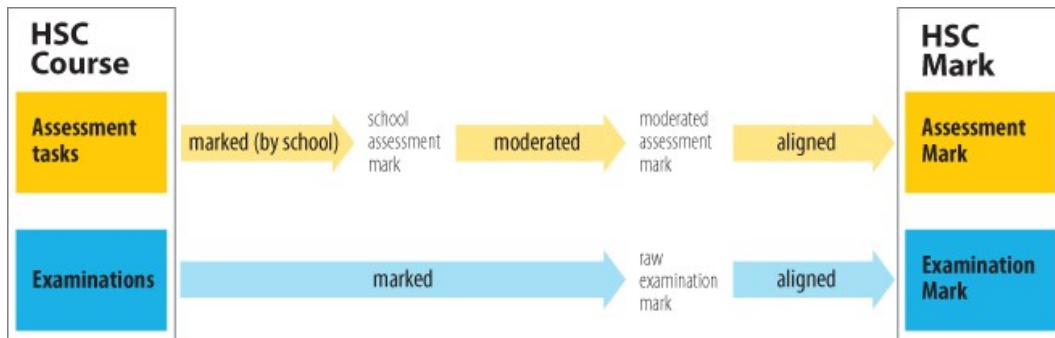
HSC-score

De resultaten voor het HSC-examen bepalen voor 50% de individuele HSC-score die aan elke leerling per leerdomein aan het eind van het secundair onderwijs wordt toegekend (zie Figuur 14). De andere 50% wordt ingevuld door een cijfer dat door de school op basis van haar eigen evaluaties wordt doorgegeven ('Assessment Tasks'). Het is niet zo dat er een wantrouwen bestaat bij BOSTES over het oordeel van de school, maar men meent dat een centraal examen noodzakelijk is om een beoordeling te maken die betrouwbaar is overheen verschillende scholen:

"Schools are very good at knowing how well their students have done compared to each other, but not necessarily good at knowing how they compare to students in other schools. And also it will provide a bit of public confidence that everybody is on an equal footing." (interview, respondent AU3)

²² HSC is een acroniem voor 'Higher Secondary College'

Wat betreft het cijfer van de school, wordt door BOSTES evenwel een correctie toegepast om de gelijkheid tussen scholen te garanderen, en om legitimiteit toe te kennen aan de evaluaties door scholen²³. Ten slotte gebeurt op elk van de scores nog een correctie ('*alignment*') om de vergelijkbaarheid overheen cohorten van studenten te garanderen. Uit de feedback uit secundaire scholen blijkt evenwel dat deze berekening door veel leraren en leerlingen als weinig transparant wordt omschreven.



Figuur 14. Bepalen van de HSC-beoordeling in verschillende stappen²⁴

De score toebedeeld door scholen kan bepaald worden door uiteenlopende evaluatiemethoden. In de praktijk ziet men echter dat meer en meer scholen niet enkel voorbereidingstoetsen bestellen (ontwikkeld door externe organisaties, of door de koepelorganisatie van katholieke scholen) om leerlingen voor te bereiden op de HSC-toetsen, maar ook dat de resultaten van deze voorbereidende toets mee worden verrekend in de 'school-eigen' assessment score die bijdraagt tot de finale HSC-score.

Verantwoordingsgerichte evaluatie op niveau van leerlingen: slagen voor secundair onderwijs en toegang tot universiteiten

De HSC-examens zijn expliciet gericht op het summatief beoordelen van leerlingen: samen met de interne toetsen van de scholen, hangt het slagen van een leerling voor het secundair onderwijs af van zijn/haar score op deze examens. BOSTES geeft een individueel HSC-certificaat aan alle leerlingen die slagen voor het secundair onderwijs. Daarnaast worden ook de namen van de beste leerlingen in elk leerdomein bekendgemaakt op de zogenaamde '*merit list*²⁵'. Leerlingen die het secundair onderwijs verlaten voor de HSC-examens krijgen een overzicht van de behaalde resultaten doorheen het secundair onderwijs (RoSA, *Record of School Achievement*).

²³ Bij deze 'moderation' wordt in drie stappen gewerkt:

1. Het gemiddelde van de schoolgroep op het 'schoolexamen' wordt gelijkgesteld met het gemiddelde van de schoolgroep op het HSC-examen
2. De beste score op het schoolexamen wordt gelijkgesteld met de beste score van deze school op het HSC-examen
3. De scores van de andere leerlingen worden ingevuld, waarbij de 'ranking' en de relatieve verschillen tussen leerlingen gelijk worden gehouden. Waar mogelijk, heeft de laagst scorende leerling op het schoolexamen de score van de laagst scorende leerlingen van deze school op het HSC-examen, maar bij verschillende spreidingen is dit niet steeds mogelijk.

²⁴ Bron: BOSTES (2015a)

²⁵ Op de website <http://www.boardofstudies.nsw.edu.au/hsc-results/hsc-merit-lists.html>

De score op de HSC-examens bepaalt ook de ATAR-score²⁶ van de leerlingen die na het secundair naar een universiteit willen doorstromen. De ATAR is een maat voor de prestaties van leerlingen die universiteiten toelaat om een ranking te maken van de kandidaat-instromende studenten. Op basis hiervan (en rekening houdend met het aantal beschikbare plaatsen) wordt beslist welke leerlingen tot universitaire programma's worden toegelaten: de ATAR-scores worden door universiteiten gebruikt bij de toelating van studenten in hun programma's²⁷, al gebruiken recent sommige universiteiten ook andere criteria. De ATAR wordt dan ook enkel bekend gemaakt aan de leerlingen zelf, en aan de universiteiten waar ze zich als kandidaat-student hebben aangemeld. De secundaire scholen beschikken niet over de ATAR van hun oud-leerlingen.

De ATAR-score is gebaseerd op een relatieve ranking: het is met name een percentielscore, die aangeeft hoe de examenprestaties van leerlingen in het twaalfde jaar te vergelijken zijn met andere leerlingen van hetzelfde cohort (die in hetzelfde jaar aan het zevende jaar zijn begonnen). De score gaat van 'minder dan 30' tot 99,95. Deze hoogste score kan gelezen worden als: op 10.000 studenten, scoren slechts 5 leerlingen beter dan deze leerling. Bij de berekening van de ATAR-score wordt rekening gehouden met het gewicht van bepaalde leerdomeinen.

Uit cijfers van New South Wales geldt dat jaarlijks ongeveer 80% van de leerlingen die hun diploma secundair onderwijs hebben behaald, zich kandidaat stellen voor een plaats aan een universiteit. Gemiddeld krijgt 45% een plaats aan een universiteit toegewezen, de andere 35% gaat naar het technisch volwassenenonderwijs (Technical and Further Training, TAFE) of naar de arbeidsmarkt.

Ontwikkeling van HSC-examens

De ontwikkeling van de HSC-examen gebeurt door ervaren leraren en academici. De toetsen worden voorgelegd aan ervaren evaluatoren, aan copy editors, en aan experts in het curriculum. Belangrijke aandachtspunten zijn een aangepast taalgebruik (geschikt voor verschillende culturen) en een gepast aanbod voor leerlingen met een beperking. De toetsontwikkeling duurt tussen 12 en 18 maanden. Een '*consultative committee*' is verantwoordelijk voor de kwaliteit van het gehele proces.

De cesuur-scores tussen de verschillende vaardigheidsniveaus worden bepaald door (oud-)leraren met ervaring in het lesgeven en evalueren van het betreffende vak, en met ervaring in het maken van beoordelingen ten aanzien van vastgestelde standaarden.

Correctie

De examens worden handmatig verbeterd door ervaren evaluatoren. Er wordt veel zorg besteed aan de betrouwbaarheid van het toekennen van punten aan leerlingen.

Elke 'kandidaat-evaluator' doorloopt eerst een proeffase, waarin zijn/haar beoordelingen worden opgevolgd door een 'senior evaluator', daarbij ondersteund door statistische rapporten die patronen

²⁶ ATAR is een acroniem voor Australian Tertiary Admission Rank

²⁷ Queensland is een uitzondering: in deze staat wordt geen ATAR-score toegekend, maar wordt op een vergelijkbare manier een 'Overall Position' aan elke leerling toegekend.

van beoordelingen van elk van de 'kandidaat-evaluatoren' weergeven. Aan het eind van de proeffase beslist deze senior evaluator of de kandidaat-evaluator effectief examens mag verbeteren.

Vragen met een kort antwoord worden door één evaluator beoordeeld op basis van een overzicht van mogelijke correcte antwoorden. Bij vragen met een langer antwoord, worden door twee evaluatoren onafhankelijke beoordelingen gemaakt. Indien er een te grote discrepantie is tussen beide beoordelingen, wordt een derde evaluator gevraagd een bijkomende beoordeling te maken. Volgens de bepalingen van vastgelegde '*discrepancy resolution*' procedures, wordt onder leiding van de senior evaluator bepaald welke twee beoordelingen uiteindelijk zullen worden gebruikt.

Andere maatregelen om de betrouwbaarheid van de beoordelingen te garanderen, zijn de volgende: van elke evaluator wordt een aantal beoordelingen onafhankelijk ook door de senior evaluator beoordeeld; statistische rapporten tonen de beoordelingspatronen van elke evaluator in elke sessie aan; en '*control scripts*' worden gebruikt waarbij elke evaluator in een bepaald team een identiek lang antwoord verbetert en waarbij verschillen tussen de beoordelingen worden onderzocht.

Bij niet-schriftelijke toetsen moeten strikte procedures en criteria ervoor zorgen dat evaluatoren een zo consistent mogelijk oordeel maken, al geeft men bij BOSTES toe dat dit niet makkelijk is.

Daarnaast waakt men er ook over dat de moeilijkheidsgraad en de inschaling van leerlingen in de zes vaardigheidsniveaus overheen de jaren consistent gebeurt.

"The Committee always uses the same syllabus, and last year's exam is used as an exemplar, they're trying to be consistent in terms of difficulty, but sometimes, one year might be unexpectedly harder than previous years. But the standard setting process - the idea is that if the exam for unknown reasons is a little bit more difficult, the standard to get into 'band 6' might be a couple of marks different to still maintain the same standard." (interview, respondent AU4)

Rapportering voor scholen

De resultaten worden voor scholen bekend gemaakt via een online monitoringssysteem, met name SMART (School Measurement, Assessment, Reporting Toolkit). Door middel van SMART worden alle scholen opgevolgd. Het is een interactief systeem, dat scholen toelaat om bijvoorbeeld specifieke groepen binnen de school te vergelijken met andere groepen. Naast de scores op het HSC-examen, zijn ook van elke leerling de resultaten beschikbaar op de NAPLAN-toetsen en op de ESSA²⁸. Bovendien worden de prestaties van alle leerlingen ook vergeleken met een 'verwachte score', die berekend wordt op basis van hun eerdere prestaties. Aangezien het monitoringssysteem op niveau van de staat wordt georganiseerd, is het ook mogelijk om de informatie van neveninstromende leerlingen van de vorige school gemakkelijk op te vragen.

²⁸ ESSA is een acroniem voor Essential Secondary Science Assessment, een verplichte toets voor alle leerlingen in New South Wales in jaar 8.

Publieke rapportering van resultaten

Van elke school wordt bekend gemaakt welk percentage van de leerlingen in de sterkste categorie terechtkomen (vaardigheidsniveau 6). Daarnaast worden de namen van de beste leerlingen in elk leerdomein gepubliceerd op de eerder genoemde 'merit list'.

Over de ATAR worden geen gegevens bekend gemaakt - dit gebeurde wel in het verleden, maar leidde tot ongewenste neveneffecten als rankings gepubliceerd in de media. Momenteel geven universiteiten wel nog een overzicht van welke leerlingen in welke studierichting zijn ingeschreven.

Impact

Respondenten zijn het erover eens dat scholen en leraren zich sterk concentreren op de resultaten van het HSC-examen (of gelijkaardig examen in andere staten/territories) om hun onderwijs te verbeteren voor de volgende cohorten van leerlingen.

"During 'staff development days', they'd be dominated in high schools by analysis of last year's results and what it means for the planning." (interview, respondent AU3)

Sommige scholen huren consultants in om de toetsresultaten te helpen interpreteren en om te zetten in initiatieven voor kwaliteitsverbetering. Verder krijgt men regelmatig de feedback van leraren die zijn geëngageerd bij de ontwikkeling en bij de correctie van de HSC-examens, dat dit een ideale vorm van professionele ontwikkeling is.

Organisaties die programma's van professionele ontwikkeling voor leraren aanbieden, geven ook aan dat hun aanbod dat ingaat op HSC-examens (bijvoorbeeld hoe leerlingen best voorbereiden op het examen, hoe de resultaten gebruiken voor kwaliteitsverbetering) zeer populair zijn.

Neveneffecten

Verschillende respondenten bevestigen dat het HSC-examen leidt tot zeer grote stress bij leerlingen, door het uitgesproken high-stakes karakter van de toets. Gevraagd of dit het geval is, nuanceert men dat bij BOSTES zelf evenwel:

"Well it is stressful for pupils. Some students are stressed that they're gonna get caught cheating." (interview, respondent AU3)

Naast het belang van de HSC-examens voor de toelating tot universiteiten en de 'merit list', wordt de inzet nog verhoogd door verschillende andere praktijken, zoals:

- Leerlingen die voor vijf vakken in 'vaardigheidsniveau 6' terechtkomen, worden 'all-rounders' genoemd en krijgen daarvoor een getuigschrift van de Eerste Minister.
- Wie een beroep in het onderwijs wil, moet minimaal drie 'vaardigheidsniveau 5' scores behalen.

Beide bovenstaande praktijken hebben tot gevolg dat bepaalde leerlingen zich richten op vijf, respectievelijk drie vakken om het beoogde vaardigheidsniveau te behalen, terwijl de scores in andere vakken voor hen niet belangrijk zijn. Naast de verhoogde stress zijn de gedocumenteerde

neveneffecten zijn op leerlingniveau dat op voorhand uitgewerkte antwoorden onder leerlingen worden verspreid, en deze antwoorden worden uit het hoofd geleerd;

“There is in some subjects a little bit of an industry in terms of, they believe they require pre-prepared responses, and students memorize them and use them, no matter what the question is, that’s than the answer that they give.” (interview, respondent AU6)

Een andere onwenselijk neveneffect is dat leerlingen hun keuze voor vakken in het elfde en twaalfde jaar laten afhangen van gepercipieerde opportuniteiten, eerder dan door hun interesse of capaciteiten. De gepercipieerde opportuniteiten worden vaak beïnvloed door onterechte assumpties over welke vakken zwaarder doorwegen bij de berekening van de ATAR, wat genoemd wordt als ‘*trying to game the system*’:

“Students would see that some courses were scaled high, because they had a very able candidature; very high achieving students would choose physics or high-level maths. And students would think that if they’d choose that course, they will be scaled up - which was not right. Now some students are thinking the opposite and think they’re gonna get a better result if they go down and be compared with a shallower pool - and that doesn’t work particularly well either.” (interview, respondent AU6)

Ten slotte wordt ook vrijwillig zittenblijven gerapporteerd: sommige leerlingen verkiezen om jaar 12 opnieuw te doen om bij een tweede poging betere resultaten te halen op hun eindexamen.

Ook op schoolniveau zijn veel onwenselijke neveneffecten ruim gedocumenteerd. Zo maken kranten rankings op van scholen op basis van het aantal ‘vaardigheidsniveau 6’ leerlingen, met veel onwenselijke neveneffecten tot gevolg. Scholen gebruiken de examenresultaten ook strategisch. Ze maken er reclame mee om (goede) leerlingen aan te trekken, wat verdere ongelijkheid in de hand werkt. Verder is er sprake van een vernauwing van het curriculum door de sterke gerichtheid van het onderwijs in jaar 11 en 12 op deze examens: in het elfde en twaalfde jaar blijkt het erg moeilijk voor scholen om tijd te voorzien voor de niet-meetbare aspecten van het curriculum, zelfs wanneer deze door de school zelf belangrijk worden geacht. In erg selectieve scholen is zowel het onderwijs als de toetsing reeds van in het zevende jaar gericht op het behalen van goede eindexamens. Daarnaast meent men dat de toetsen van de school ondergeschikt worden oaan de toetsen van het eindexamen:

“What you do internally - your internal exams at your school, they don’t tend to matter too much. The most important thing is that exam at the end.” (interview, respondent AU14)

Er zijn ook private aanbieders van toetsen die leerlingen voor het HSC-examen moeten voorbereiden.

“Schools are meant to have for each course three to five assessment tasks. They may have done projects or experiments, not necessarily tests. But I think for many schools one of these is the trial HSC-exam, which students will have done a month before,

which is very much like the HSC, it is an exam just like ours [van BOSTES]. There are a number of companies that develop them and mark them, and they're very popular."
(interview, respondent AU4)

Ook de koepelorganisatie voor katholieke scholen ontwikkelt een soortgelijk *trial HSC-exam* voor de katholieke scholen in Victoria. Hoewel de bezochte secundaire school aangaf om zich niet te richten op de eindexamens, bleken teleurstellende resultaten in 2012 de aanzet om leerlingen meer 'ervaring' te geven met vergelijkbare toetssituaties.

Ten slotte rapporteren verschillende respondenten ook dat in selectieve scholen worden weggestuurd indien ze een te groot risico betekenen voor de gemiddelde schoolscore op de eindexamens.

"So some schools, they are very selective about what students they have in year eleven and twelve. (...) We have students coming in from other schools at year eleven and twelve, and nine and ten, where other schools have said 'you are not at the standards we need you to be at. "(interview, respondent AU14)

Officieel worden deze studenten uitgesloten van hun oorspronkelijke scholen omwille van gedragsproblematiek, maar aan die officiële reden wordt weinig geloof gehecht:

"And their behaviour may not be perfect, but ... if their behaviour wasn't great, but they were very very good academically, they would probably still be at those schools."
(interview, respondent AU14)

Financieel

De ontwikkeling, afname, verbetering en rapportering van de 129 toetsen jaarlijks is behoorlijk duur. Een precies bedrag kon hier niet worden genoemd. De hoge kostprijs is deels te verklaren door het hoge aantal toetsen, maar wordt ook beïnvloed door het ter plaatse gaan van evaluatoren bij de examens van vakken die niet schriftelijk worden afgenomen (en waarbij observatie vereist is).

Toekomstplannen

Er zijn geen concrete plannen om veranderingen aan te brengen aan de huidige praktijk inzake het HSC-examen. Enkele ideeën die momenteel bediscussieerd worden, zijn: een uitbreiding naar een aanbod voor online toetsen, waarbij evenwel rekening moet worden gehouden met problemen als verschillen tussen scholen inzake IT-infrastructuur of netwerkcapaciteit; een uitbreiding naar het lager onderwijs, waar toetsing een eerder formatief doel zou hebben; een aanbod waarbij leerlingen reeds vroeger het examen kunnen afleggen; en ten slotte de mogelijkheid om het HSC-examen ook in andere landen af te nemen (momenteel is dit logistiek niet mogelijk). Net als bij NAPLAN, worden ook door BOSTES worden adaptieve toetsen momenteel uitgetoetst.

In verband met SMART (het online rapporteringsinstrument voor scholen) heeft men een tweeledige ambitie, met name om scholen de mogelijkheid te geven ook de interne toetsen te laten toevoegen; en om ook leerlingen en ouders toegang te geven tot de data die in dit monitoringssysteem beschikbaar zijn.

1.3.2 Toetsen op niveau van de individuele staten - case Victoria: On Demand Testing

Vooraleer NAPLAN werd ontwikkeld, werden in verschillende staten/territories gelijkaardige gestandaardiseerde toetsen gebruikt om de leerprestaties van leerlingen overheen verschillende scholen na te gaan. In Victoria ging het om de *Achievement Improvement Monitor* (AIM). De toetsen van de AIM worden niet meer afgenomen sinds de implementatie van NAPLAN, maar de toetsen worden nog wel jaarlijks herwerkt door de *Victorian Curriculum and Assessment Authority* (VCAA)²⁹ en als 'On Demand Testing'-toetsen aan scholen aangeboden³⁰. In 2013 verzamelde de VCAA 1,4 miljoen toetsafnames, verdeeld over 250.000 leerlingen in 1200 scholen.

On Demand Testing is een online programma dat leraren flexibel, en op eender welk moment, kunnen inzetten. Het programma kan ingezet worden bij individuele leerlingen of groepen van leerlingen. On Demand Testing voorziet toetsen die typisch kunnen worden afgenomen van het eerste tot het tiende jaar, met daarnaast nog enkele 'advanced level 10 tests'. Geadviseerd wordt om de toets te combineren met andere assessments. Het aanbod van de 'On Demand Testing'-toetsen is gratis voor scholen. Sinds de introductie van het nationale curriculum, wordt elk van de toetsen hieraan gelinkt.

Ontwikkelingsgerichte evaluatie

De doelstelling van On Demand Testing is uitgesproken ontwikkelingsgericht op micro- en mesoniveau: de toetsen zijn bedoeld om de formatieve evaluatie van leerlingen te ondersteunen, en om leraren zicht te geven op de effectiviteit van hun didactische aanpak. De toetsen zijn daartoe bewust geënt op het curriculum in de staat Victoria, omdat informatie over de leerprestaties van leerlingen inzake het curriculum meer informatie opleveren, aldus de VCAA:

"It gives them [teachers] a tool which places the students on the curriculum that they are actually teaching them, so it is an fact a little bit extra as opposed to giving them a tool that's gonna give them something on a scale they may not mean anything to the teacher." (interview, respondent AU9)

Beschikbare toetsen

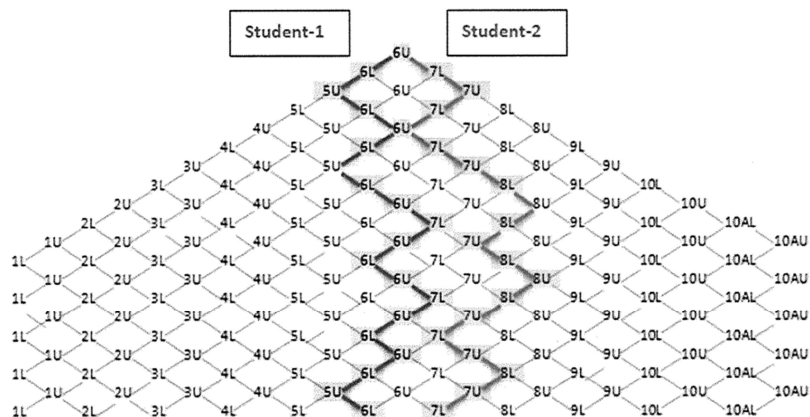
In totaal worden er 64 computer-adaptieve toetsen aangeboden in de leerdomeinen Engels en rekenen. In elke toets worden 30 of 60 items aan de leerlingen aangeboden, variërend van algemene tot specifieke toetsen. Het startitem is gebaseerd op het leerjaar van de leerling of de eindpositie van de leerling in de vorige toetsafname, maar de moeilijkheidsgraad kan snel stijgen of dalen. De toetsen met 30 items zijn 'Single Strand Adaptive Tests', waarbij afhankelijk van het antwoord dat een leerling geeft op een item (goed of fout) wordt de leerling doorverwezen naar een makkelijkere of moeilijkere vervolgvraag (zie Figuur 15). Toetsen met 60 items ('General Adaptive Tests') zijn ingedeeld in 20 sub-toetsen van telkens drie items, waarbij overgegaan wordt naar een moeilijkere subset indien

²⁹ VCAA heeft een functie in Victoria vergelijkbaar met BOSTES NSW in New South Wales, cfr sectie 1.3.1.

³⁰ Ook in verschillende andere staten is het een gangbare praktijk om eerder gebruikte toetsen (die door NAPLAN niet meer worden afgenomen) nu als een On Demand Test aan te bieden. In Victoria is de deelname evenwel groter dan in andere staten/territories waarvan de deelnamecijfers bekend zijn.

minimaal twee items correct werden beantwoord (en naar een makkelijkere subset indien de leerling nul of één item correct beantwoordde).

Dit adaptieve karakter van de toetsen vereist een zeer grote itembank die doorheen de lange geschiedenis van deze toetsen stelselmatig is opgebouwd: men beschikt momenteel over ongeveer 2000 items in de taaltoetsen, en 3000 items in de rekentoetsen. Alle toetsitems zijn ingeschaald op een Rasch schaal en worden gerandomiseerd en adaptief aangeboden.



Figuur 15. Computer adaptief toetsen: twee mogelijke paden die leerlingen kunnen volgen bij een Single Strand Adaptive Test³¹.

De meeste toetsitems zijn meerkeuzevragen. Er worden evenwel ook andere alternatieven gebruikt, namelijk korte open vragen, 'drag and drop'-vragen, vragen waarbij de leerling een on-screen calculator gebruikt en 'hot spot'-vragen. Omwille van het adaptieve karakter van de toets, is het een vereiste dat elk antwoord onmiddellijk en automatisch kan gecorrigeerd worden.

Kwaliteitszorg bij de toetsontwikkeling

Hoewel er momenteel geen nieuwe toetsen worden ontwikkeld, worden de items in de bestaande toetsen jaarlijks geëvalueerd. Daarbij wordt nagegaan of de items nog de juiste zaken meten (door externe inhoudsdeskundigen), of ze nog overeenkomen met eventuele wijzigingen in het curriculum, en of ze nog correct zijn gekalibreerd. Curriculumwijzigingen hebben ook tot gevolg dat de inschaling van gerelateerde items moet worden overgedaan. Bovendien worden nieuwe items uitgetoetst en gekalibreerd door ze gerandomiseerd tijdens de toetsafnames aan leerlingen aan te bieden.

Rapportering

On Demand Testing levert automatisch rapporten over resultaten van leerlingen en biedt directe feedback die leerlingen ondersteunt, aanmoedigt en motiveert in het leerproces. De rapportering van de toets wordt enkel aan de school bezorgd. De resultaten worden niet publiek bekend gemaakt, ook

³¹ Bron: O'Mara (s.d.)

ouders ontvangen geen individueel rapport. De rapportering voor de school voorziet een directe link van de items met het curriculumonderdeel dat getoetst wordt. De rapportering is interactief en houdt rekening met context- en inputfactoren van scholen. Bovendien zijn er ook longitudinale data beschikbaar, zowel op het individuele leerlingniveau als op het groepsniveau.

Impact

De praktijk leert dat scholen het hele jaar door een beroep doen op On Demand Testing. Vaak wordt een toets meerdere keren tijdens een schooljaar afgenomen om de voortgang van leerlingen na te gaan. Piekperiodes zijn evenwel het begin van het schooljaar; en net voor de afname van NAPLAN (als een oefening, maar ook om leerlingen gewend te laten raken aan het toets format).

Er bestaat geen empirisch onderzoek naar de impact van de On Demand toetsen. Zowel bij de VCAA als bij de bezochte scholen wordt evenwel benadrukt dat deze toetsen door hun adaptieve karakter en de directe feedback erg zinvolle informatie opleveren op het microniveau. Hierbij gaat het met name over het functioneren van de leerlingen in de klas, waarna indien nodig de instructie kan worden aangepast of individuele trajecten kunnen worden opgezet.

De resultaten zijn niet altijd verrassend - al komen er af en toe nieuwe inzichten over individuele leerlingen aan het licht - maar ook de bevestiging van de eigen inschatting van de leraren wordt als ondersteunend ervaren. Door het sterke differentiërende karakter van de adaptieve toetsen aan de uiteinden van de schaal, geven de toetsen vooral sterk diagnosticerende informatie over de zeer sterke en de zeer zwakke leerlingen.

"Teachers say they know where there students are, but they are focusing on the middle, they are missing students on either end of the scale, the upper and lower side. Specifically the teachers themselves would say that this kind of system shows them what's going on at the extremes in a specific way, and they didn't realize some students are struggling so much or that other students were so capable." (interview, respondent AU8)

Daarnaast leveren de toetsen een snel profiel op van leerlingen die nieuw instromen in de school.

"New intakes is another common use, so when a student first arrives, especially for secondary schools. You may want to see where they are at, you might want to place them." (interview, respondent AU9)

De directeur van de bezochte lagere school bevestigt dat de On Demand toetsing voor de meeste leerlingen geen nieuwe informatie oplevert wat betreft de leerprestaties, maar dat het leerwinstmonitoringssysteem helpt om de voortgang van leerlingen doorheen de tijd op te volgen:

"Are teachers surprised then by the results? I don't think so. But okay, it helps us as a school to continue to track how our students are going." (interview, respondent AU13)

Ten slotte worden de toetsen soms ook ingezet om op een meer objectieve manier te rapporteren aan ouders over de leerprestaties van de leerlingen.

Naast een impact op het microniveau wordt - weliswaar in mindere mate - een impact beschreven op het mesoniveau: de resultaten zijn een bron voor geïnformeerde dialoog tussen leraren en voor gezamenlijke afspraken over het curriculum.

"It also works as a collaborative tool for group discussion, so across teacher groups. The group of mathematics teachers would all discuss all the students, and strategies, and cooperate together looking at group data, so the aggregated data in the school at the larger scale whether it is problematic." (interview, respondent AU8)

Er wordt ook één nadeel van de adaptieve toetsen genoemd: de ervaring leert dat leraren het niet altijd makkelijk hebben om de resultaten van een adaptieve toets te interpreteren. Aanvankelijk merkte men bij VCAA dat leraren de rapportering verkeerd interpreteerden, omdat men naar de 'ruwe scores' van elke leerling keek. Het is evenwel mogelijk dat leerlingen door foutieve antwoorden aan het begin van de toets nadien relatief makkelijke items aangeboden kreeg. Dit euvel is na de nodige duiding evenwel verholpen, aldus de respondenten bij VCAA.

"In that sense, when a student has 20 out of 30 correct, teachers said you have 20 points. But this raw score says nothing about the ability of the student to the curriculum." (interview, respondent AU8)

Financieel

De toetsen zijn reeds enige tijd ontwikkeld. De jaarlijkse evaluatie van items, kalibratie, en de overstap naar een ander platform (zie toekomstplannen) vergt een inzet van 3,5 fulltime equivalent personeelsleden van VCAA. Dat is evenwel omdat de toetsen zelf volledig operationeel zijn (al worden regelmatig nieuwe items toegevoegd aan de bestaande toetsen). Om een volledig nieuwe toets te ontwikkelen rekent men op een ontwikkelperiode van 18 maanden.

Voor scholen zijn er geen kosten verbonden aan de afname van de toetsen.

Toekomstplannen

De grootste wijziging voor de nabije toekomst is van technische aard: het compatibel maken van de huidige - stilaan verouderde - software voor Ipads en Mac computers. Dit is niet alleen om de gebruiksvriendelijkheid te verhogen, maar ook omdat men meent dat elke leerling de toetsen op het meest vertrouwde devies moet kunnen afleggen om betrouwbare resultaten te krijgen:

"We're looking at if there is any difference using a keyboard, using no keyboard. As [respondent AU8] said, if a student is using a device day in day out, than that's the device they're the most comfortable using, than you would assume that there is impact on performance, because you're used to using that device on a day-to-day basis." (interview, respondent AU9)

Verder plant men voor 2016 een overstap naar een *central hosting* via een *cloud-based* platform. Daarnaast wil men op de iets langere termijn het aantal leerdomeinen dat door On Demand Testing wordt gedekt geleidelijk uitbreiden.

1.4 Toetsen vanuit toetsinstituten³²

Het is een algemeen aanvaarde praktijk voor Australische scholen om - naast de eerder beschreven gestandaardiseerde toetsen vanuit het federale niveau of state/territory-niveau - op eigen initiatief de leerlingen nog bijkomende gestandaardiseerde toetsen te laten afleggen. Soms betreft het '*full cohort tests*' waarbij alle leerlingen van een bepaald leerjaar worden getoetst, soms worden toetsen enkel van een selectieve groep leerlingen afgenomen. Dit laatste kan bijvoorbeeld om precieze diagnostische informatie in te winnen; of omdat de toets een schooloverstijgend competitief element heeft. In deze beschrijving richten we ons daarom tevens op de toetsen die door twee van de grootste toetsinstituten worden ontwikkeld, met name *Educational Assessment Australia*, en *Australian Centre for Educational Research*.

1.4.1 Toetsen van Educational Assessment Australia

Educational Assessment Australia (EAA) is een onafhankelijk onderdeel van de *University of New South Wales*. EAA wordt gecontracteerd voor de ontwikkeling, afname, verbetering en rapportering van allerlei gestandaardiseerde toetsen. Zo draagt EAA onder meer bij tot de *NAP Sample Assessments* van wetenschappelijke vaardigheid.

EAA heeft daarnaast twee 'eigen' producten die in het kader van dit rapport relevant zijn: de *International Competitions and Assessments for Schools* (ICAS) en de *General Achievement Test* (GAT).

International Competitions and Assessments for Schools (ICAS)

International Competitions and Assessments for Schools (ICAS) brengt vaardigheden van leerlingen in verschillende leerdomeinen in kaart. ICAS is beschikbaar voor de lagere en secundaire school, van jaar 2 tot jaar 13. De toetsen zijn ontworpen in zes leerdomeinen: digitale technologie, Engels, wiskunde, wetenschappen, spelling en schrijven. Niet alle toetsen zijn evenwel voor elk leerjaar beschikbaar.

Oorspronkelijk ontwikkeld als een toets om de sterkste leerlingen van een hele staat/territory te identificeren, wordt ICAS momenteel gebruikt als instrument om de huidige vaardigheid van leerlingen in verschillende leerdomeinen beter te begrijpen, maar ook om leerlingen te rangschikken op basis van hun prestaties. ICAS heeft bijgevolg een uitgesproken competitief element.

Toetsontwikkeling

Omwille van het competitie-element van ICAS, dienen de toetsen elk jaar volledig vernieuwd te worden. Elk jaar ontwikkelt een team van experts toetsen om het competentieniveau van de leerlingen te beoordelen, op basis van het nationale curriculum. Kwaliteit wordt nagestreefd door het gebruik van psychometrische metingen en IRT analyses. Items in de toets worden geschreven door medewerkers van EAA, ervaren leraren en experts in het betreffende leerdomein. Diverse bronnen

³² Gebaseerd op: Australian Council for Educational Research (2015); UNSW Global (2015) + informatie tijdens gesprekken in case studie.

worden geraadpleegd om ervoor te zorgen dat de items actueel, relevant en boeiend zijn voor de leerlingen.

In de ervaring van EAA lijken de meeste leerdomeinen zich relatief makkelijk te lenen tot IRT-inschaling van de items in de meeste leerdomeinen:

“Oddly, the model suits maths best. And then as you move away from maths as a subject, it’s getting more difficult. I think, once you go into year 10 or 11, you get more divergence, as you’re looking into more content-specific aspects that may play out differently. But really, in terms of that multidimensionality, in maths, is relatively small. It’s not a big problem really, the place we find the most difficulty is, I would say, any question that has significant spatial reasoning, visualizing 3D-objects, that’s when things start to wander off. The one that has its problem is spelling – that is English spelling. Both we and ACARA have problems with writing, equating fiction and non-fiction writing.” (interview, respondent AU7)

Omwille van het competitie-element in het monitoringssysteem van ICAS, is de gemiddelde moeilijkheidsgraad van de toetsen vrij hoog. Alle ICAS toetsen worden elk jaar gereviseerd door ervaren leraren. Zo zorgen zij ervoor dat de vaardigheden van leerlingen nauwkeurig gemeten worden en dat de toetsen relevant zijn in het licht van het Australische curriculum. Alle items worden tevens gereviseerd door een extern panel. Dit panel bestaat uit leraren en inhoudsdeskundigen, inclusief academici. Zij streven naar valide, betrouwbare en actuele items. Zodra de items zijn goedgekeurd, worden deze opgenomen in de toets. Er wordt een balans gecreëerd tussen de items op het gebied van moeilijkheidsgraad, onderwerp en getoetste vaardigheden.

Toetsafname

Elke toets wordt afgenomen op school met aanwezigheid van een leraar. De leraar ontvangt een handleiding met richtlijnen over de gewenste examencondities. Op die manier beoogt men gelijke omstandigheden in scholen te waarborgen.

Rapportering voor scholen

De school ontvangt zes weken na het afnemen van de toets een rapport met informatie over de prestatie, sterktes en zwaktes van elke leerling. De school heeft de mogelijkheid deze informatie door te sturen naar de ouders en/of te behouden in het portfolio van de leerling. Elk schoolrapport bevat de volgende onderdelen met zeer gedetailleerde informatie: een samenvatting van de resultaten van de gehele school; de resultaten op een gemeenschappelijke schaal; een samenvatting van elk jaar waarin leerlingen hebben deelgenomen; een analyse van vaardigheden per leerdomein (en per leerjaar); een analyse van moeilijkheidsgraden van de items (opgesplitst in de algemene moeilijkheidsgraad, en de moeilijkheidsgraad voor de leerlingen van de betreffende school); een analyse per leerling; resultaten van leerlingen op alfabetische volgorde; en resultaten van leerlingen op volgorde van prestatie.

Rapportering voor leerlingen

Elke leerling die participeert in ICAS ontvangt een individueel evaluatierapport over zijn/haar prestatie. In de rapportage komen de volgende punten aan bod: het antwoord van de leerling op een item, inclusief het juiste antwoord; de prestatie van de leerling op elk van de verschillende leerdomeinen; de groei van de leerling in elk jaar, over de afgelopen acht jaar; de prestatie van de leerling ten opzichte van de leerlingen in de regio. Een probleem daarbij is dat de referentiegroep waarmee leerlingen worden vergeleken, geen representatieve steekproef uit de volledige populatie van leerlingen in de hele regio is. Dit komt door het competitieve karakter van de toetsen, waardoor voornamelijk sterke leerlingen deelnemen:

“They get a comparison with everybody else who has done the test. And that’s an advantage and a disadvantage obviously, you get to see how you’ve done compared to anyone else who has done the test, but it’s not an average of everybody in Australia. It’s particularly the high-performing students, not so much in the younger years, but more so in the secondary schools. Overall, the pitch is higher, and some schools use it more for gifted and talented-type students.” (interview, respondent AU7)

Awards en erkenning

ICAS gebruikt een competitie-element waarbij elke leerling, naast een individueel rapport, tevens een getuigschrift ontvangt. Het percentage dat wordt toegeschreven per getuigschrift is afhankelijk van het aantal deelnemers. De certificering is onderverdeeld in:

- Getuigschrift met grote onderscheiding: Top 1% van alle deelnemers
- Getuigschrift met onderscheiding: Volgende 10% van alle deelnemers
- Krediet getuigschrift: volgende 25% van alle deelnemers
- Verdienste getuigschrift: volgende 10% van alle deelnemers
- Getuigschrift voor participatie: voor alle deelnemende participanten
- Principal’s Award: Het hoofd van de school kan deze award toekennen aan een leerling die een getuigschrift verdient maar niet binnen de eerste vier categorieën valt.

Een leerling met de hoogste score in een onderwerp in een bepaald jaar ontvangt een medaille en een bijbehorend getuigschrift. In 2015 zijn in totaal 514 leerlingen uit Australië en 100 leerlingen uit Nieuw-Zeeland en de Pacific erkend met een medaille voor hun uitzonderlijke prestatie. Leerlingen die een medaille hebben gewonnen in de laatste twee jaar van hun studie, zijn tevens gekwalificeerd voor een extra punt bij de berekening van de Australian Tertiary Admission Rank (ATAR) (zie sectie 1.3.1).

Impact

Er zijn geen empirische gegevens beschikbaar inzake het gebruik of de impact van ICAS. De vermoedelijke impact situeert zich voornamelijk op het microniveau, waarbij leerlingen een inschatting krijgen van hun eigen vaardigheid ten opzichte van de vaardigheid van andere leerlingen in andere scholen.

General Achievement Test

De *General Achievement Test (GAT)* is een twee uur durende toets ontwikkeld voor leerlingen in de leerjaren 5 tot en met 9. Hoewel de items van de toetsen elk jaar vernieuwd worden, kan het moment van de toetsafname door de school vrij bepaald worden. Eenzelfde toets kan zelfs meermaals van dezelfde leerling worden afgenomen om leerwinst na te gaan.

De toets kan worden ingezet voor het toetsen van nieuwe (potentiële) leerlingen, het bepalen van het beginniveau van een groep leerlingen binnen een bepaald leerdomein, of het op een betrouwbare wijze bepalen van de toekenning voor beurzen.

De GAT bestaat uit drie verschillende deelttoetsen: rekenvaardigheid (met 40 meerkeuzevragen), leesvaardigheid (40 meerkeuzevragen) en schrijfvaardigheid. In de laatste toets worden leerlingen gevraagd een schrijftaak uit te voeren. De schrijftaak wordt beoordeeld door ervaren beoordelaars aan de hand van zes vooropgestelde criteria.

Toetsontwikkeling

Het proces van toetsontwikkeling, en de kwaliteitszorg die hieraan is verbonden, is identiek als bij de ICAS (zie boven).

Toetsafname

De toetsen worden afgenomen door leraren van de scholen. Scholen kunnen de GAT op elk gewenst moment in het jaar gebruiken en de resultaten van de leerlingen opsturen naar de organisatie. Door de minder sterk uitgeschreven procedure voor de afname van de GAT, is de mate van betrouwbaarheid van de toetsresultaten evenwel ook kleiner, aldus de respondent bij EAA:

“The difference is that schools can do it when they want and in whatever way they want and they control how they use it and what they use it for. Which is another one of those ..., obviously, the more flexibility you give schools, the less you can tell them.”
(interview, respondent AU7)

Rapportering voor scholen/leerlingen

Tien dagen na ontvangst van de ingevulde toetsen krijgt de school per e-mail de rapportages doorgestuurd. Het rapport omvat een globale rapportering over de school, en een overzicht met scores van alle geteste leerlingen en de ranking in de school. Daarnaast zijn er ook individuele rapporten van leerlingen.

Impact

Er is geen globaal overzicht van de impact van GAT. Dat zeer veel scholen evenwel jaarlijks de toetsen bestellen, ziet EAA als een bevestiging dat de toetsen in scholen een positief effect hebben. Scholen gebruiken de GAT om na te gaan hoe een leerling scoort ten opzichte van de basisvaardigheid van deze leerling (gemeten via VCAA On Demand Testing), om zo hun onderwijs te verbeteren:

“we also use some of the statistics from our VCAA and the GAT - for every teacher, every class they teach they produce a graph and you can look at all of your students and see from the GAT there is a very high correlation between what study score they should get and what they actually do get - predicted versus actual - and we talk a lot with our teachers about ‘how many kids do you have above the line’ and to get more students above the line - that’s good teaching and learning.” (interview, respondent AU14)

Toekomstplannen

Er wordt onderzocht hoe ICAS en GAT online kunnen worden aangeboden? Er zijn echter nog geen concrete stappen in die richting genomen. Er is immers een groot voorbehoud wat betreft de IT-infrastructuur die aanwezig is in scholen (des te meer voor ICAS, omdat alle leerlingen de toets op hetzelfde ogenblik dienen af te leggen).

1.4.2 Toetsen van Australian Centre for Educational Research

Australian Centre for Educational Research (ACER) is een non-profit onderzoekscentrum in het brede veld van onderwijs. ACER werkt onafhankelijk van overheden, onderwijsinstellingen of andere onderzoeksinstellingen. In tegenstelling tot EAA, is ACER dus niet verbonden aan een universiteit.

ACER beschikt over een heel ruim arsenaal aan toetsen. Naast ‘school assessments’ - die hieronder allemaal kort worden beschreven - zijn er ook ‘higher education assessments’, ‘vocational, adult and workplace education assessments’, ‘organisational assessments’ en ‘clinical assessments’. Elk van die toetsen kan door scholen (of andere instellingen) aangekocht worden. Ook ontwikkelt ACER heel veel toetsen op maat van een opdrachtgever, waarbij regelmatig een getrappt adaptief design wordt gehanteerd.

We gaan hierna eerst in op de door scholen meest gebruikte toetsen, met name de Progressive Achievement Tests. Daarna geven we een korte beschrijving van elk van de andere ‘school assessments’ uit het aanbod van ACER.

Progressive Achievement Tests

De *ACER Progressive Achievement Tests (PAT)* zijn beschikbaar voor vier verschillende leerdomeinen, met name rekenen, wetenschappen, lezen, en ‘spelling, interpunctie en grammatica’. Elke toets is geschikt van het eerste, tweede of derde leerjaar (de subtest ‘leesvaardigheid’ zelfs al van het ‘*preparation year*’) tot in het tiende leerjaar. Alle toetsen zijn zowel online als op papier beschikbaar.

Doel van de toetsen

De PAT’s bieden snelle feedback aan leraren over de prestaties van leerlingen. De toetsen zijn sterk ontwikkelingsgericht en gericht op het microniveau. De toetsen informeren leraren over het leerproces van de leerlingen over een langere termijn, en helpen bij het identificeren van benodigde vervolgstappen of geschikte interventies. Toetsresultaten kunnen gebruikt worden op individueel of

groepsniveau in relatie tot een zelf- of groeps-criterium, met name voor het meten van verbeteringen in een bepaalde periode of het vergelijken van resultaten met de normgroep in Australië.

Het ontwikkelingsgericht werken wordt verder ondersteund door het ‘*PAT Resources Centre*’. Dit is een online platform speciaal ontwikkeld om in te spelen op behoeften van leraren die werken met PAT toetsen. Het platform biedt lesprogramma’s en middelen die gericht zijn op prestatieniveaus van leerlingen. De tools ondersteunen leraren in het informeren over en verbeteren van het leerproces.

Daarnaast biedt ACER ‘*PAT professional learning*’ aan voor leraren die zichzelf willen ontwikkelen in het gebruiken van PAT en het beter interpreteren van de resultaten. Via workshops, webinars en/of seminars krijgen leraren informatie over de ontwikkeling van toetsen tot aan leerstrategieën.

“If you want it to be an instrument of change, you need professional development, and you need materials, or both to help teachers to do that follow-up.” (interview, respondent 11)

Steeds meer wordt ook geïnvesteerd in zeer concrete materialen die door leraren meteen kunnen worden ingezet bij bepaalde profielen van leerlingen

“That’s when teachers start saying ‘what do you want me to do about it?’ Particularly if they’re not experienced teachers. So giving them support materials at that point is good.” (interview, respondent 11)

Toetsontwikkeling

PAT toetsen worden ontwikkeld door experts op het gebied van toetsontwikkeling in samenwerking met experts uit de psychometrie. Zij beogen een kwaliteitsvolle tool te ontwikkelen die onderzoek en diagnose naar het leerproces van leerlingen ondersteunt. De ontwikkeling van elke toets volgt een strenge procedure inzake inhoud en moeilijkheidsgraad. Er wordt zorgvuldig aandacht besteed aan de inhoud van de toetsen zodat relevante curriculum onderdelen terug komen in de toets. Elk toetsitem wordt geplaatst op een vaardigheidsschaal (ontworpen door middel van het RASCH-model).

Nieuwe edities van PAT toetsen worden permanent inhoudelijk verbeterd op basis van feedback van scholen. Nieuwe technologieën bieden kansen om de online toetsen verder te ontwikkelen.

Rapportering voor scholen/leerlingen

Scholen krijgen zowel individuele als groepsrapporten. Individuele rapporten geven rijke informatie over sterktes en zwaktes van leerlingen. Groepsrapporten bieden leraren inzicht in leerbehoeften van leerlingen en geven de mogelijkheid voor het identificeren van gemeenschappelijke moeilijkheden die aandacht vragen. Longitudinale rapporten bieden scholen de kans om leerlingprestaties over een bepaalde periode te bekijken en de effectiviteit van onderwijs in kaart te brengen.

PAT maakt gebruik van een normgroep waardoor scholen de kans krijgen om prestaties van leerlingen te vergelijken met een nationale steekproef.

Er zijn verschillende opties voor het verkrijgen van scores en rapportages: indien de toetsen online worden afgenomen, volgt een automatische correctie en wordt ook meteen een rapport gegenereerd. Bij afname van de PAT's op papier kunnen de toetsen door de leraren verbeterd worden aan de hand van een correctiesleutel en gedetailleerde richtlijnen bij het verbeteren. Een rapport wordt aangemaakt door het ingeven van de scores op een website.

Andere toetsen

ACER heeft nog uiteenlopende andere 'school assessments' in haar portfolio. De meeste van deze assessments hebben een diagnostisch doel, met name het identificeren van leerlingen in de school die boven of onder een bepaalde cesaar vallen en/of extra ondersteuning/uitdaging nodig hebben. Dit gebeurt via de *ACER Abstract Reasoning Tests* (vaardigheid logisch redeneren, gericht op sterk begaafde leerlingen), de *ELMs* (leesvaardigheid en rekenvaardigheid), de *eWrite* (schrijfvaardigheid), de *Higher Ability Selection Tests (HAST)* (leesvaardigheid, wiskundig redeneren, abstract redeneren en schrijfvaardigheid gericht op sterk begaafde leerlingen), de *Measures of Student Development* (schrijfvaardigheid en abstract, verbaal en kwantitatief redeneren), de *Neale Analysis of Reading Ability* (leesvaardigheid, met nadruk op nauwkeurigheid), de *Pattern and Structure Assessment (PASA)* (rekenbegrip bij hele jonge leerlingen tot jaar 2), de *PROBE Reading Assessment* (leesvaardigheid), de *Single Word Spelling Test* (spellingvaardigheid) en de *Test of Reading Comprehension* (leesvaardigheid). Twee toetsen hebben een uitgesproken internationaal karakter en zijn dus niet gelinkt aan het Australische curriculum, met name de *International Benchmark Test (IBT)* (Engels, rekenvaardigheid en wetenschappen) en de *International Schools' Assessment* (leesvaardigheid, schrijfvaardigheid, wetenschappelijke geletterdheid en wiskundige geletterdheid)

Twee toetsen hebben tot doel om leraren inzicht te geven niet-cognitieve aspecten, met name de *Social-Emotional Wellbeing (SEW) Survey* en de *School Life Questionnaire* (houding van leerlingen ten aanzien van de school, de leraren, het onderwijs en de medeleerlingen). De *ACER Scholarship Tests* heeft dan weer tot doel om scholen te ondersteunen bij het toekennen van een 'scholarship' (studiebeurs of verlaagd inschrijvingsgeld), en de *Compass Literacy and Numeracy Assessment* onderzoekt de leesvaardigheid en rekenvaardigheid van (jong) volwassenen die geen onderdeel uitmaken van formeel onderwijs.

Impact

De gestandaardiseerde toetsen blijken - uit de ervaring van ACER - zinvol te zijn voor leraren om een beter beeld te krijgen van hun leerlingen. PAT's zijn gelinkt aan het Australische curriculum, waardoor scholen niet enkel een idee krijgen van de prestaties van leerlingen, maar ook concrete informatie over op welke delen van het curriculum elk van de leerlingen nog vooruitgang kan of dient te boeken.

Hoewel lagere scholen de toetsen vaker gebruiken dan secundaire scholen, meent men dat de impact op secundaire scholen groter kan zijn.

"Good teachers may know in detail their students. But I think there are differences across the age ranges, because typically in primary schools you get teachers all day, and they know the student quite well, but by the time they start selecting classes and

teachers get to teach five different groups a day, that's harder to keep track of.”
(interview, respondent 11)

Scholen gebruiken PAT als aanvullende informatie op NAPLAN toetsen en de beoordeling van leerlingen door leraren. Vaak wordt de toets ook gebruikt als vervanging van de eigen (door leraren ontworpen) toetsen, om zo de kwaliteit van leerlingenevaluatie te verhogen.

“Teachers are not so good at designing assessments. If they make their own assessments, they fall into all of the test development traps of, you know, obvious correct answers, obvious distractors, they're using raw scores which don't take any account of difficulty of items,... So a carefully designed instrument can give a more detailed insight into pupils.” (interview, respondent 11)

PAT's worden ook vaak gebruikt door scholen in de aanloop naar NAPLAN, als voorbereiding (voor leerlingen op het toetsformat) en als diagnostisch instrument die aangeeft wat elke leerling nog extra nodig heeft in de laatste periode voor NAPLAN.

De strategie die vaak door gebruikers van toetsen van ACER wordt gehanteerd, is dat een meer algemene toets wordt afgenomen van alle leerlingen, en dat daarna een zeer specifieke diagnostische toets voor bepaalde leerlingen of bepaalde klasgroepen wordt afgenomen om een meer precies beeld te krijgen van een probleem dat door de eerste toets werd aan het licht werd gebracht.

Door de zeer concrete feedback die scholen krijgen van de automatische rapporten van de PAT, is het minder noodzakelijk dat scholen over een sterke datageletterdheid beschikken om de toetsresultaten om te zetten in initiatieven tot schoolverbetering. Op die manier kunnen de PAT's makkelijk als informatiebron ingezet worden voor differentiatie tussen leerlingen:

“We do them [PAT tests] before they come in at year seven and that is really to ascertain where the students are at, so they can be differentiated for.” (interview, respondent AU14).

Het geheel van toetsen (inclusief, maar niet exhaustief GAT, On Demand Toetsen en PAT's) is een informatiebron om voor elke leerling individuele voortgangsdoelen te bepalen en de selectie van vakken in het elfde en twaalfde jaar te informeren.

1.5 Leerwinst en toegevoegde waarde

1.5.1 Leerwinst

Verschillende toetsen die hierboven werden besproken, hebben een component waarbij longitudinale gegevens worden verzameld, en waarbij leerwinst in kaart wordt gebracht. In elk van deze gevallen gaat het om een *gain-score model* waarbij de prestaties van een leerling op moment X wordt vergeleken met de prestaties van een leerling op moment Y. De prestaties op verschillende

momenten, gemeten met verschillende toetsen, zijn vergelijkbaar omdat veelvuldig gebruik wordt gemaakt van de RASCH-methode om items van toetsen op een identieke schaal in te schalen.

Dit is het geval bij de volgende toetsen:

- NAPLAN: leerwinst door verschilscore over na twee jaren (jaar drie-vijf; vijf-zeven; zeven-negen);
- ook bij On Demand-Testing laat de gemeenschappelijke schaal toe om bij verschillende toetsafnames de leerwinst van leerlingen te analyseren;
- zowel rapportering over de toetsen van EAA als over de toetsen van ACER geeft standaard longitudinale gegevens weer wanneer leerlingen een toets meerdere keren (of de opvolgtoets voor opeenvolgende leerjaren) heeft afgenomen.

Er wordt geen leerwinst berekend bij de NAP Sample Assessments. Hier worden immers telkens andere cohorten van leerlingen getoetst. Ook bij de HSC-examens is er geen analyse van leerwinst: het gaat hier immers om toetsen van een summatieve aard, waarbij de nadruk net op de éénmalige prestatie ligt. Summatieve toetsen en leerwinstberekening hoeven elkaar evenwel niet uit te sluiten, zoals wordt aangetoond in het voorbeeld van ICAS en van de rapportering van de HSC-examens in SMART (waar de resultaten gekoppeld worden aan eerder afgenomen toetsen).

1.5.2 Toegevoegde waarde

Noch bij NAPLAN, noch bij de vele andere toetsen wordt de 'toegevoegde waarde' van scholen echt berekend. Meestal volstaat men met een vergelijking met de prestaties van leerlingen in andere scholen. Er zijn geen toetsen bekend waarbij de toegevoegde waarde van scholen op een meer geavanceerde manier wordt berekend. Hoewel alternatieven werden overwogen, werd er door ACARA voor gekozen om geen gebruik te maken van meer complexe maten van toegevoegde waarde omdat (1) het te onduidelijk is welke factoren hierin moeten worden meegenomen, en (2) het te moeilijk is om begrijpbaar te maken voor professionals zonder een statistische vooropleiding (waarmee zowel op leraren als op het bredere publiek wordt gedoeld).

"We're not doing the more complex valued added models, it's too black box for what we're trying to accomplish here." (interview, respondent AU2)

Ook beleidsdiscussies (op niveau van staten/territories) om toegevoegde waarde van scholen te meten zijn momenteel nog niet operationeel (of worden definitief niet geïmplementeerd) omwille van de complexiteit van deze berekening en de onduidelijkheid over welke factoren in deze berekening moeten worden meegenomen.

2. CASE ENGLAND - CENTRE OF EVALUATION AND MONITORING

Net als bij de andere cases, starten we deze sectie met een korte inleiding in het lokale onderwijssysteem. Bijkomend dienen we voor deze case een inzicht te geven in de verplichte gestandaardiseerde toetsen die in het Engelse onderwijs worden afgenomen. In de derde sectie bespreken we de verschillende gestandaardiseerde toetsen binnen het leerwinstmonitoringssysteem van het *Centre of Evaluation and Monitoring* (CEM). In de twee daaropvolgende secties belichten we de impact, en de evoluties en toekomstperspectieven met betrekking tot dit leerwinstmonitoringssysteem. We sluiten ook deze case af met een bespreking van leerwinst en toegevoegde waarde in deze case studie.

2.1. Inleiding tot het onderwijssysteem in Engeland³³

Engeland heeft traditioneel een sterk centraal gereguleerd onderwijssysteem. Het Engelse leerplichtonderwijs duurt van de leeftijd vanaf 5 jaar tot de leeftijd van 16 jaar. Daarnaast zijn 94% van de driejarige leerlingen en 99% van de vierjarige leerlingen reeds ingeschreven in een school. Volgens recente wetgeving wordt het einde van de leerplicht op 16 jaar behouden, maar moeten jongeren wel tot de leeftijd van 18 jaar aan een vorm van onderwijs of vorming deelnemen, hetzij formeel onderwijs, werkplekleren, of deeltijds onderwijs.

Leeftijd	Klas	Key stage Toetsen	CEM Toetsen	
17-18	Year 13	A-LEVEL	ALIS	
16-17	Year 12	Key Stage 5		
15-16	Year 11	GCSE	YELLIS	
14-15	Year 10	Key Stage 4		
13-14	Year 9	Key Stage 3	MIDYIS	INSIGHT
12-13	Year 8			
11-12	Year 7			
10-11	Year 6	KS2 test	PiPS KS2	InCAS
9-10	Year 5	Key Stage 2		
8-9	Year 4			
7-8	Year 3			
6-7	Year 2	KS1 test	PiPS KS1	
5-6	Year 1	Key Stage 1		
4-5	Reception	EYFS	PiPS baseline	
3-4	Nursery	Early Years Foundation St.	ASPECTS	

Figuur 16. Overzicht van het Engelse onderwijssysteem, inclusief gestandaardiseerde toetsen

³³ Gebaseerd op: Courtney (2013); DFES (2005); Earley (1997); English Government (2015a, 2015c, 2015d); Eurydice (2015); Isaacs (2010); Wilkins (2014) + informatie tijdens gesprekken in case studie.

2.1.1 Onderwijsstructuur

Het Engelse leerplichtonderwijs wordt onderverdeeld in vier sleutelfase of Key Stages; en wordt gevolgd door een vijfde sleutelfase (zie ook Figuur 16): Key Stage 1, voor leerlingen van 5 tot 7 jaar; Key Stage 2, voor leerlingen van 7 tot 11 jaar; Key Stage 3, voor leerlingen van 11 tot 14 jaar; Key Stage 4, voor leerlingen van 14 tot 16 jaar; en Key Stage 5, voor leerlingen van 16 tot 18 jaar.

Het leerplichtonderwijs wordt voorafgegaan door *Nursery education* (voor driejarige leerlingen), dat vaak wordt aangeboden door *Nursery Schools*. *Reception classes* (voor vierjarige leerlingen) worden meestal voorzien door lagere scholen. Het klassieke startjaar voor lager onderwijs blijft evenwel *year 1* (voor vijfjarige leerlingen). Sommige lagere scholen bieden nu ook al voorzieningen voor het jaar *Nursery education*.

2.1.2 Scholen in Engeland

Openbare scholen volgen het *National Curriculum*. Dit curriculum wordt bepaald door de overheid. Door de sterke graad van detail van het nationale curriculum vormt dit in hoge mate de basis van het onderwijssysteem. Sinds 2010 is er evenwel een groeiend aantal *academies* en door de overheid gesubsidieerde onafhankelijke scholen die kunnen afwijken van dit curriculum.

Openbare scholen of state schools (93% van de scholen) bieden, in tegenstelling tot private en openbare onafhankelijke scholen, gratis onderwijs aan. Er bestaan echter diverse vormen van openbare scholen:

- de meeste openbare scholen zijn *community schools* of *maintained schools*, die door de lokale overheid worden ingericht. Daarnaast zijn er ook *foundation schools*, ingericht door een onafhankelijk schoolbestuur. *Voluntary-aided schools* zijn vergelijkbaar met *foundation schools*. Hun schoolbestuur hangt een specifieke religieuze denominatie aan;
- *voluntary-controlled schools* zijn een tussenvorm tussen *community schools* en *voluntary-aided schools*;
- *academies*, tot slot, zijn publieke onafhankelijke scholen die van het nationale curriculum mogen afwijken. In het lager onderwijs zijn 11% van de scholen *academies*. Een bijzondere vorm hiervan zijn *free schools*: het betreft meestal vrij recent opgerichte scholen op basis van een vraag van ouders;
- in het secundair onderwijs zijn er zowel *comprehensive schools* die een breed curriculum aanbieden als meer selectieve *grammar schools* die een toelatingsbeleid hanteren. Dit laatste type van scholen komt slechts in beperkte mate voor (ongeveer 5%).

Daarnaast zijn er nog private scholen en openbare onafhankelijke scholen waar ouders wel voor het onderwijs van hun kinderen moeten betalen. Deze scholen mogen afwijken van het nationale curriculum. Voor deze scholen zijn er ook geen verplichtingen inzake gestandaardiseerde toetsen.

2.1.3 Verantwoordingsgerichte onderwijscontext

Sinds 1992 worden alle scholen regelmatig doorgelicht door het Inspectieorgaan OFSTED (*Office for Standards in Education*). Bij aanvang waren de inspecties, waarbij ook externe adviseurs werden betrokken, zeer high-stakes. Een kritische beoordeling kon immers aanleiding geven tot de sluiting van scholen of scholen kunnen worden geplaatst onder een '*special measures regime*' waarbij ze intensief werden opgevolgd (tot vijf inspecties in de twee jaar na de initiële inspectie). Hoewel de aanpak van de inspectie sinds 2000 wel enigszins is verzacht (onder meer door inclusie van zelfevaluaties in de externe evaluatie), kunnen inspecties nog steeds leiden tot ernstige consequenties voor scholen en hun personeelsleden. Zoals werd gedocumenteerd door uiteenlopende studies, is de emotionele impact van deze inspecties op leraren en directies dan ook bijzonder zwaar³⁴. Bij de openbare scholen hebben de resultaten van de leerlingen op de verplichte gestandaardiseerde toetsen een belangrijke invloed op deze inspecties en op de beoordeling van de inspectie over de kwaliteit van de school.

Door deze high-stakes inspecties, samen met de gestandaardiseerde toetsen die vaak vanuit een controlerend perspectief worden opgesteld, geldt de Engelse onderwijscontext als een onderwijssysteem dat sterk heeft ingezet op het verantwoordingsgerichte perspectief. De inspecties en de gestandaardiseerde toetsen worden door kritische auteurs vaak gezien als middelen waarmee de overheid de scholen een wantrouwige en beschuldigende houding etaleert ten aanzien van de scholen, scholen tracht te sturen en zelfs in een keurslijf tracht te dwingen³⁵.

2.2 Verplichte gestandaardiseerde toetsen in Engeland³⁶

We bespreken allereerst kort de rol van verplichte gestandaardiseerde toetsen in het Engelse onderwijssysteem, vervolgens komen enkele toekomstperspectieven en kritieken aan bod om vervolgens te besluiten met een korte sectie in verband met het ontwikkelingsproces van gestandaardiseerde toetsen.

2.2.1 Huidige situatie

Engeland heeft een 'assessmentcultuur' in het onderwijs: reeds lange tijd zijn leraren en scholen gewend aan het idee dat er externe, gestandaardiseerde toetsen worden afgenomen van leerlingen om de kwaliteit van het onderwijs op regelmatige tijdstippen na te gaan. De eerste echte nationaal afgenomen toetsen werden geïntroduceerd in 1988, bij de implementatie van de Education Reform Act. Leerlingen dienden aan het eind van elk van de onderwijsperioden (Key Stages) binnen het leerplichtonderwijs toetsen af te leggen. Algemeen gesteld worden in het leerplichtonderwijs verplichte toetsen afgenomen van alle leerlingen aan het eind van *year 2* (7 jaar, Key Stage 1), *year 6* (11 jaar, Key Stage 2), *year 11* (16 jaar, Key Stage 4) en buiten het leerplichtonderwijs in *year 13* (18 jaar, Key

³⁴ Bijvoorbeeld Brunsdon, Davies, en Shevlin (2006); Case, Case, en Catling (2000); Perryman (2007).

³⁵ Bijvoorbeeld Ball (2003); Courtney (2013); Perryman (2006); Troman, Jeffrey, en Raggi (2007).

³⁶ Gebaseerd op: English Government (2015a, 2015b); Eurydice (2009, 2015); Isaacs (2010); NFER (2015); Standaert (2014); Tymms en Albane (2002) + informatie tijdens gesprekken in case studie.

Stage 5). De toetsen aan het eind van *year 9* (14 jaar, Key Stage 3) worden sinds 2009 niet meer afgenomen, ervan uitgaande dat deze gestandaardiseerde toetsen vervangen zouden worden door evaluaties die door scholen zelf worden ontwikkeld.

Key Stage 1

De *National Curriculum Assessments* van Key Stage 1 zijn verplicht voor alle 'normaaltvorderende' zevenjarige leerlingen met betrekking tot lezen, schrijven, spreken en luisteren, rekenen en wetenschap. De school kan evenwel zelf beslissen welke toetsen worden afgenomen. De Key Stage 1 *tests* zijn bedoeld om aanvullende informatie te geven aan de leraar bij het vormen van een finaal '*assessment judgement*' van elke leerling aan het eind van Key Stage 1. Deze toetsen worden dan ook meestal aan het eind van Key Stage 1 afgenomen, maar de school kan er ook voor kiezen om dit vroeger te doen.

In *year 1* dient ook een *Phonics Screening Check* te worden afgenomen waarbij wordt nagegaan of leerlingen woorden fonetisch kunnen decoderen. Leerlingen die dit onvoldoende kunnen, dienen in *year 2* opnieuw een dergelijke toets af te leggen.

In 1998 werd aan de verplichte toetsen aanvullend een verplichte nulmeting toegevoegd, waarbij de kennis en vaardigheden van leerlingen aan het begin van hun schoolcarrière worden nagegaan. Het *EYFS (Early years Foundation Stage) Profile Handbook* stipuleert hoe leraren de ontwikkeling en de leerprestaties van elk kind in het *reception year* (4-jarige leerlingen) moeten beoordelen op basis van observaties en de kennis die het kind demonstreert. Op basis hiervan wordt een profiel opgemaakt, waarmee de leraar in *year 1* aan de slag kan om een responsief en gepast curriculum aan elke leerling aan te bieden.

Key Stage 2

Aan het eind van het basisonderwijs zijn *National Curriculum Assessments* verplicht voor de leerdomeinen Engels (moedertaal: lezen, grammatica, punctuatie en spelling) en wiskunde. De gestandaardiseerde toets 'wetenschappen' werd opgeheven in 2010. Ook hier worden de toetsresultaten beschouwd als een bron van aanvullende informatie voor het oordeel van de leraren. Deze toetsen hebben in principe geen invloed op de verdere loopbaan van leerlingen. Comprehensieve secundaire scholen hebben immers geen selectief instroombeleid (of althans niet op basis van leerprestaties), en in de resterende scholen (*grammar schools*) worden meestal toelatingstoetsen gebruikt die de scholen zelf hebben ontwikkeld of laten ontwikkelen.

Daarnaast zijn er facultatieve *national curriculum tests* voor acht-, negen- en tienjarige leerlingen waarmee de school kan nagaan hoe de voortgang van leerlingen zich verhoudt ten aanzien van het *national curriculum*, en als voorbereiding op de Key Stage 2 tests.

De *National Curriculum Assessments* voor het basisonderwijs worden ontwikkeld door de *Standards and Testing Agency* (STA). Dit agentschap staat ook in voor de correctie en het scoren van de toetsen.

Key Stage 4

De examens aan het eind van Key Stage 4 zijn meer bekend als de GCSE's³⁷. Deze examens zijn high-stakes, aangezien ze voor de leerlingen de toegang tot vervolgonderwijs bepalen. Hoewel ook de GCSE-toetsen sterk geënt zijn op het nationale curriculum, worden ze niet door de overheid zelf ontwikkeld. De ontwikkeling wordt uitgevoerd door vijf onafhankelijke examenagentschappen, met name: de *Assessment and Qualification Alliance (AQA)*; de *Council for the Curriculum, Examinations & Assessment (CCEA)*; *Edexcel, Oxford, Cambridge and RSA Examinations (OCR)* en de *Welsh Joint Education Committee (WJEC)*.

Scholen bepalen zelf van welk examenagentschap zij welke toetsen wensen af te nemen. In 2009, onder meer als gevolg van enkele wantoestanden bij de uitvoering van gestandaardiseerde toetsen, werd een instelling in het leven geroepen die toezicht houdt op de kwaliteit van kwalificaties en van de verplichten toetsen (OFQUAL, *Office of Qualifications and Examinations Regulation*). Ten opzichte van OFQUAL moeten de examenagentschappen zich verantwoorden voor de manier waarop zij eerlijke en vergelijkbare kwalificaties kunnen verzekeren, de manier waarop ze het bereiken van de vooropgestelde standaarden doorheen de toetsen kunnen nagaan en de manier waarop ze erover rapporteren. Ook dienen zij verantwoording af te leggen over de manier waarop examens verbeterd worden en de manier waarop aan de leerlingen wordt gerapporteerd. Een andere regulator is de *Joint Council for Qualifications (JCQ)*, van waaruit toezicht wordt uitgeoefend op administratieve vereisten inzake het verloop van examens en van accommodaties voor leerlingen met specifieke onderwijsbehoeften.

Er zijn GCSE-toetsen in uiteenlopende leerdomeinen beschikbaar. Leerlingen hoeven niet alle toetsen af te nemen: gewoonlijk legt een leerling examens in ongeveer tien leerdomeinen af. Dit maakt het mogelijk dat scholen differentiële curricula of leertrajecten aan leerlingen aanbieden, al is het verplicht om minimaal één GCSE in Engels, één GCSE in wiskunde en minimaal één GCSE in wetenschappen af te leggen. Door de meeste scholen wordt ook ICT als een verplicht GCSE vooropgesteld. De andere mogelijkheden zijn erg breed: er zijn verschillende GCSE's in talen (bijvoorbeeld Frans, Duits, Nederlands, Arabisch, ...), (humane) wetenschappen, economie, technologie en expressieve vakken.

Per leerdomein wordt vastgelegd hoe de evaluatie precies gebeurt, waarbij er telkens een externe component in de evaluatie is opgenomen. In verschillende leerdomeinen beslaat die externe component 100%, maar dat is niet noodzakelijk het geval: bijvoorbeeld bij 'art', 'music' en 'ICT' telt de externe proef voor 40% binnen het totaal van het GCSE; '*controlled assessments*' (ontwikkeld en uitgevoerd door de school) doorheen het laatste jaar, of twee laatste jaren, hebben een gewicht van 60%. In geen enkel leerdomein is de contributie van deze *controlled assessments* groter dan 60%.

De resultaten van het GCSE kunnen erg bepalend zijn voor de verdere studieloopbaan. In het hoger secundair onderwijs (leeftijd 17-18 jaar) wordt men vaak pas toegelaten wanneer men minimaal in vijf leerdomeinen een voldoende score heeft behaald, en mag men dan enkel in die leerdomeinen waarin men de voldoende score heeft behaald verder studeren (een voldoende score is minimaal een C op

³⁷ Acroniem voor General Certificates of Secondary Education

een scoresysteem van A* tot G). Ook sommige universiteiten houden bij de toelating rekening met de resultaten van potentiële studenten op het GCSE (waarbij vaak scores van A of A* worden verwacht), hoewel steeds meer universiteiten hun eigen ingangsproeven organiseren.

Key Stage 5

Traditioneel na het leerplichtonderwijs zijn er nog de zogenaamde *A-level tests* (*Advanced Level tests*) die de bovenbouw van het secundair onderwijs afsluiten. Deze *A-levels* werden traditioneel beschouwd als de ingangspoort naar universiteiten, maar, zoals eerder gesteld, laten universiteiten steeds vaker kandidaat-studenten hun eigen ingangsproeven afleggen.

Verantwoordingsgerichte doelen

De gestandaardiseerde nationale toetsen hebben traditioneel een hoog controlegericht gehalte voor scholen: scholen moeten zich verantwoorden voor de prestaties en voortgang van hun leerlingen. Sinds 1992 worden *league tables* of ranglijsten van scholen uitgegeven, gebaseerd op de resultaten die leerlingen in de verschillende scholen behalen. Het gaat dan voornamelijk om de resultaten van Key Stage 2 (lagere scholen) en Key Stage 4 (secundaire scholen). Het doel van deze vorm van transparantie is dat het ouders een beter (meer objectief) beeld zou geven over de prestaties van de school, om op die manier een beter geïnformeerde schoolkeuze toe te laten. Volgens het marktprincipe wordt dan verondersteld dat goed presterende scholen meer leerlingen zouden aantrekken, terwijl slecht functionerende scholen hierdoor een stimulans krijgen om een volgende keer betere resultaten voor te leggen.

In het begin van de jaren negentig van de vorige eeuw werd ook een inspectie-orgaan in het leven geroepen, dat tevens een zeer sterk verantwoordingsgericht perspectief innam bij de doorlichting van scholen (OFSTED, zie sectie 2.1.3). De openbaarheid van data en *league tables* leidden ertoe dat slecht presterende scholen openbaar worden gebrandmerkt, onder meer omdat de media hier veel aandacht aan besteden. Bovendien is het zo dat slecht functionerende scholen kunnen gesloten worden of, sinds het *Fresh Start* programma geïmplementeerd in 2000, overgenomen kunnen worden door andere scholen of door privé-organisaties met een sterke autonomie inzake personeelsbeleid (ontslagen en aanwervingen). Sinds 1997 kunnen leraren die slechte resultaten voorleggen ook makkelijker ontslagen worden.

De overheid gebruikt voor lagere scholen '*floor standards*': scholen wordt gebrandmerkt als *underperforming* wanneer hun leerlingen onder bepaalde minimale verwachtingen scoren op de Key Stage 2 tests. Deze minimale verwachtingen zijn;

- Minder dan 60% van de leerlingen behaalt minimaal niveau 4 in Engels en Nederlands
- Minder dan 92% van de leerlingen maakt de verwachte voortgang in Engels
- Minder dan 90% van de leerlingen maakt de verwachte voortgang in rekenen.

Voor secundaire scholen wordt worden de leerprestaties van leerlingen op zestienjarige leeftijd vergeleken met de prestaties op elfjarige leeftijd. Het gaat om een toegevoegde waarde-meting aangezien het doel was om op basis van de gemiddelde 'leerwinst' een uitspraak te doen over de

toegevoegde waarde van de school (in vergelijking met andere scholen). De leerwinst wordt berekend als een *gain score model* (zie Deel 3, sectie 3.2.2) en de toegevoegde waarde als het gemiddelde van de leerwinst van leerlingen volgens deze berekening.

Sinds 2006 werd er een zogenaamde '*Contextual Value Added*-meting gepubliceerd. In deze aangepaste rankings worden de leerprestaties van leerlingen op schoolniveau bijgesteld op basis van kenmerken van de leerlingen-instroom ('*eligibility for free meals*' als proxy voor armoede, etniciteit, sociaal milieu, geslacht en leerlingen met bijzondere onderwijsnoden). Een gevolg was evenwel dat de onderwijsinspectie evolueerde tot een erg technocratisch controleorgaan, met een bevestiging van de *contextual value added* tot gevolg. Recent worden de ranglijsten van scholen terug opgemaakt op basis van de 'ruwe score' van de gemiddelde leerlingen, zonder correctiefactoren. Variërende gewichten die aan de vijf correctiefactoren werden toegekend, zorgden immers voor spectaculaire wijzigingen aan de plaats van sommige scholen op de ranking.

Bovendien blijkt dat in vele scholen een context is ontstaan waarin ook leraren summatief worden beoordeeld aan de hand van de resultaten van leerlingen.

"Their performance management is results-driven, the performance management has to have in it a component on their performance in the classroom, and how much progress the classes are doing and what you are doing to ensure that you get maximum progress , so whether they get the next pay scale, depends on their performance. So their payment depends on the performance of pupils." (interview, respondent EN10)

Tendens naar meer ontwikkelingsgerichte doelen

Doorheen het meest recente decennium is er langzaam een verschuiving merkbaar van een verantwoordingsgericht gebruik naar een meer formatief (ontwikkelingsgericht) gebruik van de verplichte toetsen. Zo worden de resultaten op Key Stage 2 *tests* (einde basisonderwijs) vaak gebruikt door secundaire scholen om individuele leerdoelen voor elke leerling te bepalen. Secundaire scholen worden daar ook door de overheid toe aangezet, aangezien openbaar wordt gemaakt welke leervooruitgang men heeft kunnen boeken wanneer de prestaties van leerlingen op GCSE (Key Stage 4 aan het eind van secundair onderwijs) worden vergeleken met de prestaties op Key Stage 2. Op deze manier beoogt men ook de overgang van basis- naar secundair onderwijs te verzachten. Deze dynamiek past binnen de groeiende beweging naar *Assessment for Learning*, waarbij aan leerlingen niet enkel moet worden duidelijk gemaakt wat de doelen zijn van hun individuele leerproces en hoe ze zich verhouden ten opzichte van de gestelde doelen, maar waarbij ook de voortgang van leerlingen ten aanzien van deze leerdoelen nauwgezet wordt opgevolgd om waar nodig individuele ondersteuning te voorzien. Ook de afschaffing van de verplichte Key Stage 3 *tests* ten voordele van school-eigen toetsen kan gezien worden in het kader van een meer ontwikkelingsgerichte benadering van leerlingenevaluatie.

2.2.2 *Ontwikkeling van toetsen*

Voor de ontwikkeling van elk van de bovenstaande toetsen wordt ongeveer twee jaar ontwikkelingstijd gerekend. De ontwikkeling omvat onder meer de betrokkenheid van verschillende betrokken agentschappen, van inhoudsdeskundigen en van voormalige leraren. Panels van ervaren leraren hebben een rol in het bepalen van de moeilijkheidsgraad van de verschillende toetsitems en in het verlenen van advies inzake de verwachte minimumscores en andere cesuurbepalingen. Daarnaast worden ook scholen van bij de start van het ontwikkelingsproces betrokken door middel van trials van toetsvragen. Standaard wordt elke toets minimaal twee maal uitgeprobeerd alvorens deze op grote schaal wordt ingezet. Bovendien wordt naast de eigenlijke toets ook een reservetoets ontwikkeld in geval bijvoorbeeld een lek de betrouwbaarheid van de toets zou ondermijnen. Omwille van het gevaar dat items zouden uitlekken, worden de vragen voor GCSE en A-level evenwel niet op voorhand bij leerlingen uitgeprobeerd.

2.2.3 *Trends*

De nationale overheid heeft de intentie geuit om zowel de GCSE-toetsen als de A-levels meer 'uitdagend' te maken, zodat leerlingen beter voorbereid zijn op toekomstige studies of op het zoeken naar werk. Bovendien moet er zo beter gedifferentieerd kunnen worden tussen de resultaten van uitmuntende leerlingen.

Tegelijk is het voornemen geuit om het eindcijfer bij GCSE sterker afhankelijk te maken van de gestandaardiseerde toets, en dus een minder groot aandeel over te laten voor de '*controlled assessments*' die door de scholen worden opgesteld. Hierdoor moet de betrouwbaarheid van het cijfer worden versterkt.

Vanaf 2016 zijn ook '*baseline assessments*' verplicht aan de start van het *reception year*. Scholen moeten dan de leerwinst van leerlingen kunnen aantonen van bij aanvang van hun schoolcarrière tot aan het eind van het basisonderwijs (Key Stage 2). Scholen kunnen wel zelf kiezen welke toets ze hiervoor afnemen: naast de BASE toets van het *Centre of Evaluation and Monitoring* (zie 2.3), zijn er nog vijf andere toetsen geaccrediteerd (met name van het *Centre for Inspirational Learning*, *GL Assessment*, *Hodder Education*, *National Foundation for Educational Research* en *Speech Link Multimedia Ltd*).

2.2.4 *Impact*

Zoals ook blijkt uit de literatuurstudie, is er weinig evidentie over de echte impact van deze gestandaardiseerde toetsen op de onderwijskwaliteit. Wel werd aangetoond dat het slaagcijfer bij GCSE doorheen de jaren is gestegen. Critici beweren daarom dat het niveau van de toetsen is gedaald, maar uit internationaal vergelijkend onderzoek (bijvoorbeeld PISA) blijkt dat de standaarden opmerkelijk stabiel zijn gebleven, en dat leerlingen dus effectief beter presteren dan vroeger het geval was. Het is echter de vraag welke rol de hierboven beschreven toetsen spelen in dit verhaal.

Verder worden onder meer de gelijkgerichtheid inzake leerlingevaluatie overheen scholen, de informerende functie voor vervolgonderwijs, en de tijdswinst voor leraren (om niet meer zelf verantwoordelijk te zijn voor het opstellen en corrigeren van de summatieve evaluatie) als een belangrijke meerwaarde van de verplichte gestandaardiseerde toetsen gepercipieerd. Ook andere toetsen - formatieve toetsen waarvoor de school zelf verantwoordelijk zijn - worden regelmatig door (voornamelijk secundaire) scholen aangekocht vanuit toetsontwikkelaars, waarbij de toetsresultaten worden weergegeven in het licht van de verwachtingen op de GCSE of A-level examens, waarna het onderwijs weer kan worden bijgesteld. De examens geven dus duidelijke doelstellingen aan het onderwijs in scholen, en duidelijke data voor het gesprek van de school met ouders van leerlingen. Ook in de bezochte case school uit men zich als voorstander van het bovenstaand geschreven systeem van verplichte gestandaardiseerde toetsen:

"I think it's been for the best as the outcomes of the pupils have improved. Because schools are scrutinized more on them. You have a school where they are rigorous and they are pushing it at the child, and you may have another school where they are not so rigorous but they give us the pupils anyway. Where in this country, all of them sit the same test at the same age, and that will be used to get the feel and to know where people are, and because we're tested on that, schools really work hard for the best of the pupils." (interview, respondent EN10)

2.2.5 Kritische onderstroom

In het Engelse onderwijssysteem is een kritische onderstroom aanwezig, die beweert dat de druk van de verantwoordingsgerichte gestandaardiseerde toetsen scholen aanzet tot onwenselijk gedrag, zoals *teaching to the test* (met name: een focus van het onderwijs op de meetbare aspecten; door de brede invulling van de GCSE's op uiteenlopende leerdomeinen, is evenwel verzekerd dat alle leerdomeinen evenredig aan bod komen), alsook onwenselijk gedrag, zoals meer technocratisch en leraar-gericht onderwijs, minder enthousiasme voor experimentele vernieuwingen en creativiteit (ook '*ossification*' genoemd), een overmatige focus op korte-termijn doelen ten koste van lange-termijn ontwikkeling van vaardigheden en attitudes, en het uitvlakken van verschillen tussen scholen³⁸. Focusgroepen van leraren en directies lager onderwijs en leraren secundair onderwijs in een studie van Collins et al. (2010) uitten zich als voorstander van de afschaffing van Key Stage 2 toetsen om neveneffecten als *teaching to the test* te vermijden en meer aandacht te kunnen besteden aan minder getoetste vaardigheden. In 2016, vlak voor het verschijnen van dit rapport, riep de leraren-vakbond National Union of Teachers nog op tot een boycot van deze Key Stage 1-toetsen³⁹

Respondenten uit het Centre of Evaluation and Monitoring halen een eerder analyse aan waaruit bleek dat lagere scholen op onwenselijke manier de resultaten van de Key Stage 1 - *toetsen* maximaliseerden:

³⁸ Bijvoorbeeld Angus (2012); Brooks en Tough (2006); Hargreaves (1995); Nicholl en McLellan (2008); P. Smith (1993); Troman et al. (2007).

³⁹ Zie Coughlan (2016)

"We have seen it with the national tests. For example, we have some schools who have children go from 4 to 7, and others where they go to 4 to 11. There's a test at 7 [Key Stage 1 test]. So I compared the national tests at 7 for the schools that stopped at 7, with the schools that went on to 11, and they clearly got higher scores than they should have got compared to the PIPS test in those schools. They were definitely... cheating is not the right word, but 'excessibly' lenient." (interview, respondent EN3)

Bovendien beweren tegenstanders dat, ondanks de aandacht voor kwaliteitszorg, de toetsbetrouwbaarheid van de verplichte toetsen in sommige gevallen te wensen overlaat⁴⁰ en dat verschillende soorten vragen doorheen de tijd een longitudinale interpretatie onmogelijk maken⁴¹. Ten slotte wordt geargumenteed dat - eerder dan beleidsdiscussies te voeren met objectieve data - de resultaten van de toetsen bewust zo worden voorgesteld door politici of media om een eerder idee te staven⁴². Uit het schoolbezoek - in een school waar men zich algemeen voorstander noemt van de gestandaardiseerde toetsen - tijdens deze case studie, blijkt ook dat leraren niet langer het gevoel hebben dat ze door de overheid - en bijgevolg door het bredere publiek - vertrouwd worden:

"Trust? For British teachers, that is long gone... We know we do a good job. But we know that they don't trust us that we do a good job." (interview, respondent EN10).

Tegenstanders van het bovenstaand beschreven systeem, waarbij scholen zelfstandig een keuze kunnen maken tussen het aanbod van verschillende toetsontwikkelaars (zoals bij GCSE na KS4), claimen dat toetsontwikkeling hierdoor een erg lucratieve business is geworden. Deze claim wordt ondersteund door het feit dat de toetsontwikkende instanties ook eigen methodes en handboeken ter beschikking stellen die optimaal voorbereiden op de toets.

2.3 Toetsen binnen *Centre for Evaluation and Monitoring*³

In 1983 begon Carol Fitz-Gibbon met het *A-level Information System* (ALIS), een project vanuit het toenmalige Curriculum Evaluation and Management Centre (CEM, huidige benaming: Centre for Evaluation and Monitoring) van Durham University. ALIS was zelf al een opvolger van het COMBSE (Confidential Measurement Based Self Evaluation) project, dat nog een jaar eerder werd opgezet met het doel scholen te ondersteunen in hun interpretatie van de resultaten van hun leerlingen op de *A-levels*.

ALIS had tot doel om scholen inzicht te geven in de prestaties van hun leerlingen en om een vergelijkingsbasis met de prestaties van leerlingen van andere scholen aan te bieden. Binnen ALIS kregen scholen een eenvoudige leerwinstmeting van elke leerling, waarbij werd nagegaan in welke mate de scores op de A-levels (als achttienjarige leerlingen) verschilden van de score die voorspeld

⁴⁰ Aangetoond door een analyse van He en Opposs (2012) en Wiliam (2001)

⁴¹ Aangetoond voor Key Stage 2 toetsen inzake leesvaardigheid (Kispal, 2005)

⁴² Het betreft hier een kritiek geuit door Mansell (2013).

⁴³ Gebaseerd op: CEM (2014); Coe (2007); Tymms en Albone (2002); Tymms en Coe (2003) + website CEM (www.cem.org) + gesprekken tijdens case studie.

werd op basis van de scores op de GCSE's (als zestienjarigen). Daarnaast werden een aantal attitudes en percepties van de leerlingen via een door CEM ontwikkelde gestandaardiseerde toets bevestigd. Optioneel konden scholen sinds 1988 ook een *Test of Developed Abilities* (oorspronkelijke benaming: *International Test of Developed Abilities*) van elke leerling afnemen. Deze toets betreft een soort algemene vaardigheidsmeting en kon gebruikt worden als een tweede nulmeting (naast de GCSE).

Momenteel worden de toetsen jaarlijks door ongeveer 2500 scholen in het Verenigd Koninkrijk gebruikt (testafnames van ongeveer 650.000 leerlingen). Daarnaast worden de toetsen ook gebruikt in 70 landen over de hele wereld. De toetsen worden hiertoe aangepast aan de lokale omstandigheden van de opdrachtgever. De aanpassingen betreffen zowel de inhoud als de vormgeving, rekening houdend met culturele gebruiken, levensomstandigheden en lokale gevoeligheden. Bovendien dient ook de RASCH-inschaling van items in elke context opnieuw te gebeuren⁴⁴.

2.3.1 *Uitgangspunten van de gestandaardiseerde toetsen*

School improvement

Het uitgangspunt van elk van deze toetsen is dat kwaliteitsvolle informatie scholen en leraren kan helpen om bestaande of ontluikende problemen te identificeren middels het vroegtijdig diagnosticeren van *“pupils who are falling behind, coasting, or excelling”* (interview, respondent EN1).

De toetsen zijn dus duidelijk ontwikkeld vanuit een schoolontwikkelingsperspectief. Daartoe worden de resultaten van leerlingen vergeleken met de resultaten van andere leerlingen. Dit gebeurt ofwel door middel van een overzicht van waar de leerling zich situeert ten opzichte van leeftijdsgenoten, ofwel door voor elke leerling een inschatting te maken van de leeftijd waarop hij/zij presteert. De analyse op schoolniveau levert scholen een beeld op van hoe hun leerlingen zich verhouden ten opzichte van leerlingen van andere scholen.

Er wordt tot op zekere hoogte over gewaakt dat er geen verantwoordingsgericht gebruik van de toetsen wordt gemaakt: scholen mogen bijvoorbeeld de resultaten van hun leerlingen niet op hun website plaatsen. CEM gaat ook niet in op de occasionele vragen van ouders om toetsen te mogen inkijken of aankopen.

Grote vrijheid bij het inzetten van toetsen

Scholen kunnen de toetsen zelf naar eigen goeddunken inzetten: zowel beslissingen over de timing, als over de hoeveelheid van toetsen laat men bewust aan de ‘klanten’ over. Bovendien moeten niet alle toetsen van een geheel cohort van leerlingen worden afgenomen. In sommige scholen worden bepaalde toetsen van slechts een beperkt aantal leerlingen afgenomen, bijvoorbeeld in geval men meer diagnostische informatie wenst over de kennis of vaardigheid van een bepaalde leerlingen in een bepaald leerdomein.

⁴⁴ Dit gebeurt enkel indien de toets op een grote schaal wordt afgenomen.

Scholen kunnen individueel deelnemen, of per *local education authority* (of per andere groep van scholen). Wanneer een groep van scholen zich aandient, kan deze ervoor kiezen om de resultaten op schoolniveau te vergelijken met de gehele groep van scholen of met de groep van hen bekende scholen.

Scholen zijn uiteindelijk vrij om met de toetsresultaten te doen wat zij wenselijk achten: er is geen verplichting om deze of gene actie te ondernemen. In lijn met het eerdere uitgangspunt tracht men scholen wel te motiveren om de toetsresultaten enkel op een ontwikkelingsgerichte wijze in te zetten, maar de eindbeslissing ligt bij de scholen zelf:

"If I detect that it's gonna be used like that [for control/accountability], then I point that it's not what it's meant for, but once the door is closed, and the data's there, they can do with it whatever they want to do with it." (interview, respondent EN9)

Zo worden in de bezochte case school de resultaten van de afname van MidYis besproken met de ouders van de leerlingen, zodat er mogelijk voor leerlingen toch een zekere 'inzet' aan de toets verbonden is.

Begeleiding van scholen

Er wordt veel geïnvesteerd in opleiding en begeleiding van scholen. Scholen kunnen beroep doen op een helpdesk bij de voorbereiding of afname van de toetsen (voornamelijk technische ondersteuning). Daarnaast is er ook een ruim ondersteuningsaanbod voorzien na de toetsafname, gericht op de interpretatie van de resultaten. Een medewerker van CEM komt dan naar een school (of naar een studiedag waarop leden van verschillende scholen aanwezig zijn) en geeft algemene en individuele duiding over de resultaten. Deze ondersteuning wordt als een cruciale stap gezien om scholen in staat te stellen optimaal van de resultaten gebruik te maken.

Computer-based

De toetsen worden digitaal aangeboden. Oudere kinderen kunnen de toets zelfstandig afwerken, terwijl de leraar bij jongere kinderen de toets individueel per leerling afneemt en zelf aanklikt of een leerling het goede antwoord heeft aangeduid.

Op die manier zijn er ook bijvoorbeeld spreekoefeningen mogelijk, die wel voor alle leerlingen dezelfde zijn (gestandaardiseerd), maar waarbij het oordeel volledig aan de leraar wordt overgelaten. Omwille van de tijdsinvestering worden deze toetsen vaak door *teaching assistants* afgenomen. Er zijn bij jonge kinderen andere mogelijkheden dan 'correct' of 'wrong', zoals bijvoorbeeld 'refuse' (wanneer een leerling geen antwoord geeft) en 'speech' (waarbij wordt aangegeven dat de leerling wel een correct antwoord geeft, maar het antwoord verkeerd uitspreekt).

Buiten de gebruiksvriendelijkheid van computer-based toetsen, hebben deze toetsen ook het voordeel dat leraren zeer snel feedback krijgen over de leerprestaties van de leerlingen.

Ook tracht men zo dicht mogelijk aan te sluiten bij de lokale technische omstandigheden van de school. Het feit dat de toetsen van CEM op verschillende manieren kunnen worden afgenomen

(online, lokaal geïnstalleerd, hybride versie) wordt als een must omschreven om aan de noden van verschillende scholen tegemoet te komen. Een andere cruciale voorwaarde is de stabiliteit van het systeem. Daarom worden er ook in de online of hybride versie geen items opgeladen als de toets zelf reeds gestart is (met andere woorden: het gehele leerwinstmonitoringssysteem is volledig opgeladen vooraleer het eerste item wordt voorgesteld).

Interactief (computer-adaptive)

Aangezien de toetsen volledig computer-based zijn, bestaat ook de mogelijkheid om de aangeboden items sterk te richten op het vaardigheidsprofiel van elke individuele leerling,

Bij de meeste toetsen wordt een reeks items met progressieve moeilijkheidsgraad afgenomen waarbij een bepaalde reeks gestopt wordt nadat een leerling drie opeenvolgende foutieve antwoorden heeft gegeven. Op die manier wordt voorkomen dat leerlingen ontmoedigd worden door teveel foutieve antwoorden. Daarna wordt overgegaan naar een andere itemreeks of een ander leerdomein.

Bij andere toetsen (met name InCAS, zie sectie 2.3.2) wordt de moeilijkheidsgraad van de aangeboden items bepaald op basis van de eerdere antwoorden van leerlingen.

Een andere vorm van computer-adaptieve toetsen is dat items afhankelijk kunnen worden gemaakt van kenmerken van een leerling. Zo wordt bijvoorbeeld bij letterherkenning bij jonge kinderen de progressieve volgorde (van makkelijk naar moeilijk) niet enkel gebaseerd op onderzoek, maar ook op basis van de naam van de leerling. Zo is de eerste letter van de naam van het kind steeds de eerste letter die wordt aangeboden (en dus de letter die voor dit kind als de meest makkelijkste wordt beschouwd).

Inschatting op referentiegroep of op een leeftijd

Leerlingen worden ingeschat ten opzichte van referentiegroepen van leerlingen uit hetzelfde onderwijssysteem, volgens een normaalverdeling. De standaardisering van deze toetsen wordt zoveel mogelijk op één schaal gebracht (via Rasch-analyse). Voor INCAS bijvoorbeeld wordt elk item ingeschaald op een leeftijdsschaal: op welke leeftijd slaagt 50% van de leerlingen er in dit item correct op te lossen? (*'an age equivalent score'*). Een probleem met deze *age equivalent scores* is dat curriculumgebonden items soms plotse sprongen maken? en dat er dus een groot verschil is van welke items op een bepaalde leeftijd (per jaargroep) correct kunnen worden beantwoord. Daarom worden typische vragen die slechts kunnen worden opgelost wanneer een leerling een bepaalde techniek heeft aangeleerd (bijvoorbeeld probabiliteitsvraagstukken) slechts aangeboden vanaf een bepaalde leeftijd, zelfs wanneer een jongere leerling door goede antwoorden op eerdere vragen op een 'hoger' niveau terecht komt..

Wetenschappelijk onderbouwd

Een aparte onderzoeksafdeling binnen CEM zorgt ervoor dat de toetsen zijn afgestemd op de recente bevindingen binnen de ontwikkelingspsychologie. Toetsen zijn gebaseerd op onderzoek naar wat leerlingen kunnen en hoe toetsen kunnen afgenomen worden.

"I've just been speaking with one of my colleagues (...) talking about how to develop our assessment strategy to make sure that we take into account all the research and the data and educational policy and try to triangulate into one assessment." (interview, respondent EN7)

Kindvriendelijkheid

Voor jonge kinderen worden toetsen in een soort verhaal ingebed. Daarbij wordt er zorg voor gedragen dat, los van hoe ver de leerling geraakt in de opdrachten (cf. principe dat toets stopt binnen een bepaald thema na drie fouten), het verhaal toch op een zinvolle manier wordt afgesloten.

Toetsen voor de kleinste leerlingen duren maximaal 15 à 20 minuten. De leraar kan de toets ook op elk moment stopzetten wanneer de concentratie of interesse van de leerling verzwakt.

Gebruiksvriendelijkheid

Men streeft er naar om de werklast voor leraren en scholen tot een minimum te beperken. Daarnaast tracht men zo dicht mogelijk tegemoet te komen aan de lokale (technische) capaciteit van de school door verschillende manieren te voorzien om de toetsen af te nemen (met name lokaal geïnstalleerd, via website CEM, of een hybride versie). Ook wordt er sterk ingezet op begeleiding bij eventuele problemen (zie boven).

Op basis van de vele contacten van CEM-begeleiders met scholen wordt veel technische en inhoudelijke feedback verzameld; op basis van deze feedback kunnen aanpassingen aan de toetssoftware worden doorgevoerd om de gebruiksvriendelijkheid te verhogen.

Verbale instructie

Niet enkel omwille van bovenstaande beschreven principes van kind- en gebruiksvriendelijkheid, maar ook om de betrouwbaarheid en validiteit van de toetsen te verhogen, is een groot deel van de instructie in elk van de toetsen verbaal: de opdrachten worden door een stem voorgelezen.

De verbale instructie wordt uiteraard aangepast indien de toetsen in andere landen worden aangeboden (dit is niet altijd nodig, aangezien het soms om Engelse/internationale scholen in het buitenland gaat). Zelfs de toetsen voor Schotland of in andere Engelstalige landen zijn inhoudelijk dezelfde, maar worden telkens door een andere *voice artist* ingesproken, opdat het gebruikte accent aansluit bij het lokale gebruik.

Stimuleren van een onderzoekscultuur

Reeds van bij aanvang was het uitgangspunt de rationale van '*distributed research*', waarbij de ontvangers van de feedback (met name leraren in scholen) zelf actieve onderzoekers zijn en niet enkel beschouwd worden als passieve ontvangers: het onderzoek wordt gezien als een collaboratief proces, waarbij scholen verantwoordelijk zijn voor de interpretatie en de daaropvolgende actieplanning. Desalniettemin kan ook inzake deze laatste stappen een ondersteuning vanuit CEM worden ingezet. Het uiteindelijke doel echter blijft om een onderzoekscultuur binnen scholen te bewerkstelligen, waarbij leraren in scholen zelfstandig op zoek gaan naar data die hen helpen om hun leerlingen en hun onderwijseffectiviteit vanuit objectieve gegevens beter te begrijpen.

2.3.2 *Huidig toetsarsenaal*

Sinds 1991 heeft CEM een breed gamma aan gestandaardiseerde toetsen ontwikkeld dat wordt ingezet in diverse schoolse omgevingen. Zoals besproken in de voorgaande sectie, wordt elke toets digitaal aangeboden zodat directe feedback en computer-adaptieve toetsing mogelijk gemaakt worden. In wat volgt, wordt - per onderwijsniveau - een overzicht van de toetsen gegeven die voor alle scholen beschikbaar zijn.

Lagere scholen (inclusief *nursery education* en *reception year*)

Voor het pre-school onderwijs en onderwijs in de lagere school, zijn er verschillende toetssystemen beschikbaar, met name: ASPECTS, PIPS (PIPS Baseline, PIPS Key Stage 1, PIPS Key Stage 2) en InCAS.

ASPECTS

ASPECTS (Assessment Profile on Entry for Children and Toddlers) is een pre-school toetssysteem (een verzameling van gestandaardiseerde toetsen) voor leerlingen tot de leeftijd van vier jaar. ASPECTS wordt gewoonlijk afgenomen bij aanvang in het jaar *nursery education* als tool voor leraren om de basisvaardigheden van leerlingen die voor het eerst in het onderwijssysteem instromen, snel in kaart te kunnen brengen. ASPECTS geeft een overzicht van zowel kennis als vaardigheden van leerlingen in taal (bijvoorbeeld letterherkenning), rekenen (bijvoorbeeld concepten als groter/kleiner, getalherkenning, tellen, eenvoudige sommen) en 'fonologisch bewustzijn' (het correct kunnen nazeggen van bestaande of onbestaande woorden). Daarnaast biedt ASPECTS ook nog twee toetsen aan met betrekking tot respectievelijk 'persoonlijke, sociale en emotionele ontwikkeling' en 'motorische ontwikkeling'. Eerder dan toetsen, zijn deze twee laatste meer een observatielijst en een scorelijst voor leraren om het ontwikkelingsniveau van leerlingen in te schatten. Voor deze 'toets' wordt immers geen actieve inbreng van de leerlingen zelf verwacht, maar wordt het dagelijkse gedrag van de leerlingen in de klas op een systematische wijze in kaart gebracht.

PIPS Baseline

PIPS (*Performance Indicators in Primary School*) was één van de eerste producten van CEM, ter beschikking gesteld van scholen in 1994. 'PIPS Baseline' wordt afgenomen in het *reception year* en bestaat uit verschillende toetsen met betrekking tot taal, rekenen, fonologisch bewustzijn, kortetermijn-geheugen, 'persoonlijke, sociale en emotionele ontwikkeling', 'en 'gedragmatige aspecten'. Net als ASPECTS werkt ook PIPS Baseline bij de laatste twee 'toetsen' eerder als een registratiesysteem.

In tegenstelling tot de meeste andere toetsen (die standaard één maal per jaar worden afgenomen) krijgen de leerlingen twee maal de toetsen van ASPECTS en PIPS Baseline voorgeschoteld. Dit gebeurt respectievelijk bij de start van en aan het eind van respectievelijk de *nursery education* en het *reception year* (de tweede meting van ASPECTS kan ook bij aanvang van het *reception year* gebeuren). De 'eindtoets' begint waar de begintoeets eindigt: het beginniveau bij de eindtoets wordt dus voor elke leerlingen afzonderlijk bepaald op basis van de eerdere toetsresultaten. Voornamelijk van de

voorafname van de toets wordt een grote impact verwacht, omdat er aan het begin van de schoolloopbaan nog helemaal geen informatie beschikbaar is voor de leraren om te differentiëren:

“How do they know what the child knows when they come in? Maybe through working with the kids observations, but they do this within the first couple of weeks, it takes maybe six or eight weeks before they know, in the meantime the kids are ‘switched off from education’. Our idea is: you could take that time, or you could do a quick assessment and find out and then that means that you get faster information. Moreover, you get more information from a one-to-one assessment.” (interview, respondent EN1).

BASE zal vanaf 2015 PIPS Baseline vervangen. BASE is geaccrediteerd door het Engelse *Department for Education* als één van de toetsen die kan ingezet worden als verplichte nulmeting van de kennis en vaardigheden van elke leerling die in het leerplichtonderwijs instroomt (zie sectie 2.2.3).

PIPS Key Stage 1 en PIPS Key Stage 2

Na PIPS Baseline, zijn er nog ‘PIPS Key Stage 1’ en ‘PIPS Key Stage 2’. Beide toetssystemen bevatten verschillende gestandaardiseerde toetsen die ontwikkeld zijn om éénmaal per jaar te worden afgenomen. Naast toetsen die de prestaties van curriculum-gebonden rekenen en taal in kaart brengen, zijn er ook twee toetsen voorhanden voor respectievelijk ‘vocabularium’ en ‘non-verbale vaardigheid’ (voornamelijk patroonherkenning) als proxy’s voor ‘algemeen ontwikkelde vaardigheid’, en een toets inzake attitude ten aanzien van rekenen, lezen en de school in het algemeen.

Alle toetsen van PIPS zijn computer-adaptief opgevat. Indien een leerling binnen een bepaald thema drie maal op een rij een foutief antwoord geeft, wordt overgegaan naar een volgend thema.

InCAS

InCAS is een modulair opgezet geheel van verschillende toetsen, die gewoonlijk slechts bij een aantal leerlingen op een specifiek punt doorheen de schoolcarrière worden ingezet (bijvoorbeeld waar dyslexie of dyscalculie wordt vermoed), maar soms ook met meer regelmaat van elke leerling worden afgenomen. Elk item in InCAS is gerelateerd aan een bepaalde leeftijd waarop 50% van de leerlingen het betreffende item correct beantwoordt. Er bestaan niet-curriculum-gerelateerde toetsen in verband met lezen (woordherkenning, woordontleding, begrijpend lezen en spelling), rekenen (getallenleer, meetkunde, vraagstukken en hoofdrekenen met aparte toetsen voor optellingen, aftrekkingen, vermenigvuldigingen en delingen), algemeen ontwikkelde vaardigheid (uitgebreider arsenaal dan PIPS) en attitudes.

Behalve voor de toets in verband met attitudes, wordt per leerling en per toets een leeftijdsgerelateerde score berekend. Dit gebeurt door items van verschillende moeilijkheidsgraad aan te bieden. Net als PIPS is InCAS een computer-adaptief toetssysteem, maar op een meer complexe manier.

Zo krijgt bij InCAS de leerling als eerste opdracht een item voorgesteld overeenkomstig met zijn/haar leeftijd min twee jaar. Leerlingen die een item correct beantwoorden, krijgen vervolgens een item dat

moeilijker wordt geacht, bij een foutief antwoord is de moeilijkheidsgraad van het volgende item minder hoog. Hierdoor krijgen leerlingen niet alleen meer uitdagende taken. De sprong in moeilijkheidsgraad wordt verkleind naarmate de toetsafname vordert, wat resulteert in een progressieve verfijning van de inschatting van de actuele prestatie van de leerling. Door het gericht inzetten van de items op het vaardigheidsprofiel van de leerling, krijgt de toets ook een groter discriminerend vermogen.

Secundaire scholen

Voor secundaire scholen zijn er vier toetssystemen standaard beschikbaar, met name MidYis, YELLIS en INSIGHT voor het leerplichtonderwijs, en ALIS voor leerlingen in het hoger secundair onderwijs (Key Stage 5, na het leerplichtonderwijs).

MidYis

MidYis (*Middle years Information System*) wordt meestal als baseline voor het secundair onderwijs ingezet. MidYis vormt zo een extra informatiebron om geïndividualiseerde prestatiedoelen voor elke leerling op te stellen, naast de resultaten van de Key Stage 2 toetsen (aan het eind van het basisonderwijs). MidYis is een niet-curriculum gebonden toets in vocabularium, rekenvaardigheid, non-verbale vaardigheid, 'algemene vaardigheden' en attitudes ten aanzien van school en verschillende leerdomeinen. MidYis geeft een voorspelling voor elke leerling voor verschillende leerdomeinen aan het eind van Key Stage 3 en Key Stage 4 (GCSE): per leerling wordt per leerdomein getoond hoeveel kans hij/zij maakt op een eindscore A*, A, B, C, ...

YELLIS

YELLIS (*year Eleven Information System*) is ontwikkeld voor leerlingen in Key Stage 4. Naast een toets voor vocabularium, wiskunde, non-verbale vaardigheid en een algemene kennisvragenlijst, bevat YELLIS ook drie attitude-vragenlijsten, waarvan er twee voor de leerlingen en één voor hun ouders zijn. Ook YELLIS geeft scores van leerlingen in vergelijking met een ruime vergelijkingsbasis en een voorspelling voor GCSE. Nadat de GCSE toetsen zijn afgenomen, voorziet YELLIS feedback in verband met leerwinst en toegevoegde waarde.

INSIGHT

Een derde toets voor het secundair onderwijs is 'INSIGHT'. In tegenstelling tot MidYis en Yellis, is INSIGHT een curriculum-gebaseerde toets die wordt afgenomen aan het eind van Key Stage 3. INSIGHT bevat toetsen met betrekking tot lezen, wiskunde en wetenschap en aanvullend een 'algemeen ontwikkelde vaardigheid' (deze laatste is ook beschikbaar in MidYis). Ook op basis van INSIGHT wordt voor elk van de deelnemende leerlingen voorspellingen gemaakt voor GCSE.

ALIS

ALIS (*Advanced Level Information System*) is gericht op leerlingen in Key Stage 5. Op basis van de resultaten op GCSE enerzijds en van een eigen 'algemeen ontwikkelde vaardigheidstest' anderzijds, wordt een score voorspeld voor de resultaten voor de *A-Level* toetsen voor de verschillende leergebieden. Leerwinst en toegevoegde waarde worden algemeen en per leergebied berekend.

Daarnaast bevat ALIS ook nog een instrument waarmee leerlingen aan het eind van hun schoolcarrière de school en de verschillende vakken kunnen evalueren.

Andere toetsen binnen CEM

Naast de bovenvermelde toetsen die aan alle scholen worden aangeboden, kan CEM ook toetsen op vraag ontwikkelen. Dit gebeurt in opdracht van scholen, maar ook van andere instellingen of overheden. Het betreft in deze gevallen vaak toetsen met een specifiek doel, zoals toelatingsproeven, toetsen om een educatieve innovatie te evalueren, en bevestigingen in het kader van marktonderzoek.

2.3.3 Toetsontwikkeling

Kenmerkend in de ontwikkeling van elk van de toetsen is dat ze stap voor stap werden ontwikkeld en geïmplementeerd. PIPS begon bijvoorbeeld uitsluitend met een toets voor rekenen in het laatste jaar van het lager onderwijs. De uitbreiding was vaak geïnspireerd door actuele vragen van leraren in scholen (bijvoorbeeld leraren uit de *reception class* (vierjarige leerlingen) die niet enkel een *baseline assessment* wensten, maar ook een testafname aan het eind van het schooljaar om de vooruitgang van hun leerlingen in kaart te brengen).

De ontwikkeling van toetsen gebeurt volledig binnen CEM: er wordt geen beroep gedaan op externe professionals. Dit geldt zowel voor de inhoudelijke ontwikkeling, de technische ontwikkeling, als de vormgeving van de toetsen. Hiervoor werft CEM experts in verschillende leerdomen aan die onderzoek verrichten in het domein van leerpsychologie en vakdidactiek. De directe samenwerking tussen 'inhoudelijke' en 'technische' toetsontwikkelaars maakt een snellere iteratie en een synergie tussen beide processen mogelijk, die door alle betrokkenen sterk wordt gewaardeerd. Er wordt doorheen het ontwikkelingsproces wel feedback van leraren verzameld, wat de enige 'externe fase' in het ontwikkelingsproces is.

De computer-adaptieve toetsing vereist een zeer grote database van items (voornamelijk voor de conceptualisering van adaptieve toetsing zoals in InCAS gebeurt). Deze database groeit jaarlijks aan doordat er regelmatig nieuwe items doorheen de afgenomen toetsen worden gevalideerd. Regelmatig worden ook analyses gepland (frequentie onduidelijk) waarbij wordt nagegaan of de *age equivalent score* nog correct is en of afleiders bij meerkeuzetoetsen wel degelijk goede afleiders zijn.

2.3.4 Kwaliteitszorg

CEM heeft een beleid waarbij toetsen permanent kunnen worden geëvalueerd en bijgesteld. Nieuwe items kunnen op elk moment worden toegevoegd. Elk van de toetsen zou (in principe) om de zes jaar moeten herbekeken worden. In realiteit vonden we echter verschillende toetsen waarvan wordt aangegeven dat deze zes-jaar termijn reeds ruim overschreden is. Het blijkt dat er inderdaad een compromis wordt gezocht tussen enerzijds de nood aan regelmatige herstandaardisering, anderzijds de eraan verbonden werklast, des te meer omdat er geen grote 'item shift' wordt verwacht:

"We don't really expect that there will be rapid changes in the performance of pupils in the population all the time, there will be long-term trends but these will take long time"
(interview, respondent EN8)

Elke toets wordt eerst uitgetoetst bij een steekproef van scholen. Aan de hand van feedbacksheets voor leerlingen en/of leraren wordt bevestigd wat de beste en minst goede onderdelen van de toets zijn (zowel wat betreft inhoud, rapportering als design). Ook worden er regelmatig analyses gedaan om de waarde van de verschillende items na te gaan. Toetsontwikkeling wordt door de verschillende stappen een iteratief proces genoemd:

"Assessment development is an iterative process: we'll create assessments, we'll gather data, use the data to inform the assessment, to write new assessments, to improve our assessments, and modifying them all the time taking a huge databank that we have, which extends back many years [into account]." (interview, respondent EN7)

Van sommige toetsen wordt ook de toets-hertoetsbetrouwbaarheid of de predictieve validiteit nagegaan. De PIPS Baseline toets (vijfjarige leerlingen) blijkt een predictieve validiteit van 0.68 te hebben ten aanzien van de eindtoetsen van PIPS (elfjarige leerlingen).

Omwillen van mogelijke wijzigingen aan het curriculum, worden de items van de curriculumgebonden toetsen worden elk jaar opnieuw gestandaardiseerd (op de Rasch-schaal).

2.3.5 Kost

CEM is een financieel onafhankelijk centrum binnen de universiteit van Durham. Alle gemaakte personeels- en werkingskosten dienen dus via de verkoop van toetsen te worden 'terugverdiend'.

Voor scholen wordt een kost berekend per leerling: tussen 6 en 8,5 GBP (tussen 8 en 12 EUR) per leerling. Een basisschool met bijvoorbeeld 20 leerlingen per jaargroep die aan alle toetsen wil deelnemen (één toets voor *reception year*, PIPS KS1/2 + INCAS voor *year 1* tot en met *year 6*) komt op jaarbasis aan een bedrag van 2160 GBP (ongeveer 2940 EUR).

2.4 Impact van het leerwinstmonitoringssysteem van CEM

Er is momenteel geen beschreven impact-onderzoek van de toetsen van CEM beschikbaar. Er bestaat bijgevolg geen harde evidentie dat het leerwinstmonitoringssysteem effectief een positieve invloed heeft op het onderwijs. De impact die hieronder wordt beschreven, is volledig gebaseerd op wat mondeling tijdens de interviews werd vermeld (door zowel personeelsleden van CEM als van de gebruikers van de toetsen in een school). Wel wordt aangegeven dat de mate waarin deze positieve effecten zich manifesteren, sterk verschillen van school tot school, met een grote afhankelijkheid van het leiderschap en van de 'ethische omgang' van alle personeelsleden in de school met leerwinstmonitoringssysteem:

"I always say to schools: everyone has joined responsibility, whatever your job is, you have to add value each year. If you can do that, and everyone feels that they're not trying to make their kids look better so that the teachers' group next year looks worse, it's actually a combined effort - those are the schools that seem to be the most effective, if you generate that sort of ethos within the school." (interview, respondent EN2)

Een ander belangrijk aspect is dat scholen vrijwillig kiezen voor deelname aan het leerwinstmonitoringssysteem van CEM. Verschillende respondenten schrijven het succes van de toetsen van CEM dan ook toe aan de relatief kleinschalige opstart met een beperkt aantal scholen of zelf een beperkt aantal leraren in scholen, die nadien andere leraren en scholen hebben gemotiveerd om ook van het groeiende leerwinstmonitoringssysteem gebruik te maken.

"That brings us back to how you're going to introduce the system. I think of the example of Northern Ireland, where they gave the guarantee that they would not look at the data for schools at government level, and they've spent a lot of time training the teachers to do it small scales, trials, with enthusiastic people, who then spread their enthusiasm..." (interview, respondent EN4)

2.4.1 Positieve impact op microniveau

Nieuwe informatie

"Teachers don't know what they don't know" (interview, respondent EN3). Vaak is het maar door gebruik te maken van toetsen, dat men beseft dat er nood is aan of een mogelijkheid bestaat om via gestandaardiseerde data het onderwijs voor individuele leerlingen te versterken. Vaak kennen leraren de leerlingen al goed, en daarbij is de CEM-toets een goede bevestiging van wat ze al weten. In zulk geval betreft het een motiverende en objectiverende verificatie van de eigen inschatting van de leraar. Soms komen echter wel nuances naar boven (bijvoorbeeld waar leerlingen een disharmonisch profiel hebben, sterk in sommige leerdomeinen en minder sterk in andere). Regelmatig worden verrassingen gerapporteerd over individuele leerlingen.

Soms worden toetsen niet volledig ingezet, maar dienen ze enkel voor bepaalde groepen van leerlingen waarvoor men de toets het meest geschikt vindt (afhankelijk van het doel van de eigen school). De toetsen laten dus toe om specifiek 'in de diepte' diagnosticerende informatie te verzamelen over bepaalde leerlingen waarvan men een leerprobleem of een leerachterstand vermoedt.

Stellen van individuele doelen

De toetsen worden ook gebruikt als beginanalyse voor leerlingen bij aanvang van een Key Stage. Op basis van de resultaten van CEM worden in bepaalde scholen (onder meer de school die tijdens de case studie werd bezocht) sterk geïndividualiseerde doelstellingen vastgelegd op basis van de

resultaten van zowel de Key Stage 2 toetsen als MidYis⁴⁵. MidYis lijkt daarvoor een meer geschikte informatiebron. Door het verantwoordingsgerichte doel van de Key Stage 2 toetsen, is de ervaring van secundaire scholen immers vaak dat de lagere scholen op een oneigenlijke manier hebben toegewerkt naar de toetsen, waardoor deze geen correcte weergave bieden van wat leerlingen kennen en kunnen. Dit blijkt onder meer uit de praktijk in de door ons bezochte school:

"We use the CEM tests when they come in, along with the Key Stage 2 SAT's. What that does, it enables us to check the validity of the Key Stage testing because we honestly believe it's flawed. For various reasons. And what we find is a child would come in at a high level, and we started working with them; and the child just has not got a grasp. And that's down because they've had got a lot of support at Key Stage 2, but the wrong support, as they had not gotten to understand the concepts, but they've been trained to do the test . (...) So what the CEM test allows us to do, is to check the validity of that. And the first thing we do is see what the difference is between the Key Stage 2 and the CEM testing. In a lot of cases the discrepancy is they've got a high Key Stage 2 test but a low CEM test, so they could come in at band D [een lage prestatie]. Occasionally, you got a child who has got a low Key Stage 2 but his CEM test he's up at band A [een hoge prestatie]. Now at that point we start exploring why?" (interview, respondent EN10)

Uit de praktijk van deze case school blijkt ook hoe men de toetsresultaten op een gedetailleerd niveau analyseert om voor elke leerling een aangepast programma op te stellen, in de eerste plaats gericht op het wegwerken van eventuele probleempunten in de kennis en vaardigheid van elke leerling. Deze geïndividualiseerde interventies zijn vaak gebaseerd op de uitvoering van specifiek hiertoe aangekochte software. In deze school wordt de voortgang van leerlingen voortdurend opgevolgd door de leraren, en op schoolniveau besproken per trimester aan de hand van de gestelde individuele doelstellingen. Indien een leerling tijdens een trimester niet op zijn traject is op basis van beginniveau en actueel prestatieniveau, wordt uitvoerig besproken welke verschillende activiteiten verschillende leraren met de betreffende leerling hebben genomen, en eventueel bijkomende maatregelen voorgesteld. Indien bij een leerling twee trimesters op rij niet de verwachte voortgang kan worden vastgesteld, worden ook de ouders bij dit overleg met de verschillende leraren betrokken. Er is met andere woorden een zeer strikte opvolging van het leren van leerlingen, geïnformeerd door de toetsen van Key Stage 2 en CEM.

Evenwichtigere evaluatie

Gestandaardiseerde toetsen kunnen bias in evaluatie van leerlingen tegengaan. Zo wordt door verschillende respondenten gesteld dat bijvoorbeeld jongens, etnische minderheden en leerlingen met gedragsmoeilijkheden zonder gestandaardiseerde toetsen minder goede scores zouden krijgen.

⁴⁵ Het combineren van de resultaten van verplichte summatieve toetsen met de resultaten van een bijkomende algemene vaardigheidstoets, kan inderdaad de predictieve validiteit te opzichte van toetsen aan het eind van *Key Stage 3* en GCSE-examens sterk verhogen, zoals blijkt uit een onderzoek waarbij weliswaar een andere toets werd gebruikt (van toetsontwikkelaar GL Assessment) (Strand, 2006)

Bovendien leren leraren door de gestandaardiseerde toetsen op een meer objectieve wijze naar de leerprestaties van hun leerlingen kijken.

2.4.2 *Positieve impact op mesoniveau*

Nieuwe informatie

Nog meer dan op het microniveau, lijken de geaggregeerde data op schoolniveau vaak te leiden tot een verruimde blik op de sterktes en zwaktes van de school. De gewoonlijk gebruikte veronderstellingen blijken soms niet-valide assumpties. Een typisch voorbeeld is de school waarin de overtuiging leeft dat minder sterke leerprestaties van hun leerlingen te wijten zijn aan de kenmerken van de leerlinginput. Echter wanneer de resultaten van hun leerlingen in perspectief worden geplaatst ten opzicht van andere vergelijkbare scholen, worden sommige scholen gedwongen deze assumptie in vraag te stellen.

De school als leergemeenschap

Het feit dat deze toetsen beschikbaar zijn, maakt dat intercollegiale discussie wordt gestimuleerd. Er komt immers een concrete en objectieve case tot stand die mogelijks kan fungeren als basis voor peer consultatie mogelijk wordt. Een alternatief dat eerder werd uitgetest binnen CEM is een consultatiemodel waarbij verschillende schooldirecties geanonimiseerde rapporten bespreken.

2.4.3 *Positieve impact op macroniveau*

Onderzoeksdoeleinden

Door de grote schaal waarop de toetsen van CEM worden afgenomen, wordt zowel fundamenteel als zeer praktijk- en beleidsgericht onderzoek mogelijk gemaakt. Op basis van de CEM databank werd bijvoorbeeld fundamenteel onderzoek gedaan naar bepaalde kenmerken bij leerlingen met ADHD die een grote samenhang vertonen met onderpresteren (Tymms & Merrell, 2011), naar het belang van *educational authorities* voor leerlingenprestaties (Tymms et al., 2008), naar de impact van (een vroege instap in) het kleuteronderwijs op latere prestaties (Tymms, Merrell, & Henderson, 2000), naar de impact van huiswerk op nationale toetsen (Farrow, Tymms, & Henderson, 1999; Tymms & Fitz-Gibbon, 1992) en naar de invloed van mannelijke of vrouwelijke leraren voor jongens en meisjes (Carrington, Tymms, & Merrell, 2008)

Meer beleidsgericht kunnen trends in het onderwijs op longitudinale wijze worden nagegaan (bijvoorbeeld Merrell en Tymms, 2011) of wordt de mogelijkheid geschapen om de effecten van bepaalde onderwijskundige innovaties of experimenten makkelijker en objectiever na te gaan. Zo werd de effectiviteit nagegaan van een project met *peer tutoring* (bijvoorbeeld Tymms et al., 2011) en van een theater-drama project (Fleming, Merrell, & Tymms, 2004). Ook werd de effectiviteit van bepaalde ondersteunende maatregelen voor leerlingen met ADHD onderzocht (Tymms & Merrell, 2006).

Daarnaast zijn er - omdat de toetsen van CEM intussen in verschillende landen worden gebruikt - ook comparatieve studies uitgevoerd (bijvoorbeeld Merrell & Tymms, 2007).

Beleidsimpact

Mede door bovenstaande voorbeelden, is men en er binnen CEM ook sterk van overtuigd dat de informatie van eigen gestandaardiseerde toetsen ook op beleidsniveau meer informatie oplevert dan de informatie uit internationaal comparatieve toetsen (bijvoorbeeld PISA en TIMSS).

2.5 Onwenselijke neveneffecten

Hoewel er sterk over het ontwikkelingsgericht opzet van de toetsen wordt gewaakt, kan het niet uitgesloten worden dat er ook neveneffecten ontstaan door de vrijheid van scholen om de toetsen naar eigen goeddunken in te zetten. Ook zouden overheden toch de toetsresultaten kunnen opvragen en gebruiken binnen een verantwoordingsgericht perspectief, omdat dit nu eenmaal door de Engelse onderwijs traditie een gebruikelijke manier van het opvolgen van scholen is:

“In England, anything that’s been created as an assessment is potentially an accountability measure” (interview, respondent EN7).

Er is evenwel geen harde evidentie over de mate waarin de hieronder beschreven neveneffecten voorkomen, en het algemeen gedragen idee is dat deze op slechts erg beperkte schaal voorkomen.

2.5.1 Oneigenlijk gebruik

Er worden doorheen de gesprekken twee verschillende manieren van oneigenlijk gebruik gedocumenteerd. Ten eerste wordt vastgesteld dat scholen soms strategisch gebruik maken van de resultaten, ook al is dat duidelijk niet de bedoeling van de toetsen. Een type-voorbeeld hiervan is dat onafhankelijke scholen de (goede) resultaten gebruiken om leerlingen aan te trekken. Ten tweede ziet men bij CEM dat individuele leraren oneigenlijke analyses uitvoeren op basis van de data of dat verkeerde conclusies getrokken worden.

“I had a colleague who went into a school recently and the man their said I found a great way of doing things: I take the maths score and I add it to their reading score and I add it to their general ability score and I find the mean of all that and that is really helpful! It gives me a measure of every child’s need for priority. But what are you measuring? This guy couldn’t understand that by doing that you’re actually losing any information you have, but for him, he felt like he had a very good way of prioritizing kids. These scores are very different, it tells you nothing at all. But we’ve had to come to terms with the fact that people take our data and do all sorts of things with, and we don’t have control over that.” (interview, respondent EN2)

Teaching to the test

Een zekere mate van teaching to the test (met name een vernauwing van het curriculum door een sterkere focus op basisvaardigheden in rekenen en taal) wordt niet uitgesloten. Door de beperkte mate waarin dit voorkomt, wordt dit neveneffect zeker niet als een probleem beschouwd. Dat dit neveneffect slechts beperkt wordt gerapporteerd, zou niet alleen kunnen worden verklaard door het

ontwikkelingsgerichte perspectief, maar ook doordat de meeste toetsen niet gerelateerd zijn aan het curriculum.

2.5.2 *Teaching the test*

Het neveneffect van '*teaching the test*' (trainen van het soort vragen dat wordt gesteld of van de vragen zelf) hangt af van wat de leraar van het gebruik van de toets wil leren: indien dit informatief is (een beter beeld van wat leerlingen kunnen), zal dit neveneffect onbestaande zijn; wanneer men echter leerlingen - om welke reden dan ook - goed wil laten scoren, wordt erkend dat er een verleiding is *to teach the test*. Maar omdat de verschillende toetsen aan elkaar gekoppeld zijn, zal de verwachting voor de volgende toetsen alleen maar toenemen. Men vermoedt bijgevolg dat ook dit neveneffect zich slechts in beperkte mate voordoet.

2.6 Evoluties en toekomstperspectieven binnen CEM

In deze sectie lichten we eerst enkele evoluties toe inzake toetsen en toetsontwikkeling binnen CEM sinds de beginperiode. Daarna geven we een overzicht van enkele relevante huidige ontwikkelingen en toekomstperspectieven.

2.6.1 *Doorheen de voorbije decennia*

Sinds de ontwikkeling van het eerste project in de jaren '80 onderging het toetsenarsenaal verschillende ingrijpende wijzigingen en evoluties. De belangrijkste evoluties zijn:

- een expansie van toetsen voor uiteenlopende leeftijdscategorieën;
- computer-based toetsen, die toelaten om interactiever (computer-adaptief) te werken en om instant feedback te voorzien;
- een sterkere inzet op begeleiding van scholen;
- een graduele internationale ontwikkeling waarbij toetsen in meerdere landen worden aangeboden. Aanvankelijk was de expansie voornamelijk gericht op landen met een Angelsaksische onderwijs traditie (bijvoorbeeld Nieuw-Zeeland, Australië), maar nadien werden toetsen ook aangepast voor gebruik in heel verschillende onderwijscontexten (bijvoorbeeld Rusland). Hierdoor worden benchmarks tussen onderwijssystemen mogelijk. De PIPS Baseline werd reeds naar het Nederlands vertaald (OBIS – OnderBouwInformatieSysteem, is de vertaling van PIPS). Elke nieuwe versie dient bij de ontwikkeling geëvalueerd te worden inzake onderwijskundige en culturele aspecten (in zowel inhoud als design) en meestal wordt een nieuwe standaardisering opgemaakt;
- de ontwikkeling van toetsen op vraag (zie in sectie 2.3.2 'andere toetsen binnen CEM').

2.6.2 *Huidige ontwikkelingen en nabije toekomst*

Begeleiding

Gestandaardiseerde toetsen worden steeds vaker enkel gezien als het begin van een meer omvattend proces van training en begeleiding. Nog meer dan reeds het geval is, plant CEM om, op vraag van scholen, ook begeleidend advies te geven. Daarbij beperkt men zich niet enkel tot ondersteuning bij de interpretatie van de resultaten, maar worden concrete handelingsplannen gemaakt die bij specifieke problemen kunnen worden ingezet (individueel per school, of generiek bij een bepaald 'type van resultaten').

Toetsevaluatie

Recent werd de mogelijkheid geopend waarbij leraren/scholen(groepen) zelf hun toetsen kunnen insturen met het doel de betrouwbaarheid en validiteit ervan door CEM te laten nagaan. CEM geeft advies over hoe deze zelfontwikkelde toets kan verbeterd worden (afhankelijk van de precieze vraag van de 'klant').

Interactieve rapportering

Er wordt steeds meer werk gemaakt van (online) interactieve rapportering, waarbij de gebruiker

- zelf het niveau van data kan selecteren op leerling-, klas- of schoolniveau, of zelf groepen van leerlingen (binnen de klas of binnen de school) kan creëren;
- zelf de diepgang/algemeenheid van de data kan manipuleren;
- de benchmark (van leerlingen/scholen) mee kan manipuleren;
- verschillende voorstellingswijzen (figuren/tabellen) kan kiezen;
- andere tools kan aanpassen (bijvoorbeeld standaardisering voor gemiddelde score en normaalverdeling; meest frequent zijn gemiddelde = 100, SD=15; of gemiddelde = 50, SD=10).

Door deze interactiviteit neemt de verwachting toe dat scholen veel meer uit de eigen data kunnen halen.

Integratie van toetsen

Een project voor de nabije toekomst is om de verschillende toetsen van *reception year* tot aan het eind van de schoolcarrière te integreren en als één toetsenbatterij te beschouwen, en de verschillende items op eenzelfde schaal te integreren. Dit is vooral zinvol om de voortgang van leerlingen doorheen de schoolloopbaan in kaart te brengen (eerder dan op schoolniveau, omdat klasgroepen en schoolpopulaties veranderen).

CEM speelt momenteel in op de gepercipieerde vraag om nog meer toetsen, die gericht zijn op non-cognitieve leerprestaties, te ontwikkelen. Daar staat evenwel tegenover dat wordt aangegeven dat de huidige theoretische basis voor de attitudetoetsen relatief beperkt is. Bij de ontwikkeling van niet-cognitieve toetsen wordt steeds vaker binnen andere sectoren van het bedrijfsleven of binnen overheden gezocht naar goede praktijkvoorbeelden en/of een theoretische kennisbasis.

Daarnaast wordt er ook gewerkt aan gestandaardiseerde 'correctiesleutels' bij oefeningen die de motorische vaardigheden van (jonge) kinderen in kaart brengen. Daarbij wil men onder meer de evolutie inzake evenwicht en kracht in kaart brengen. De oefeningen zelf zijn gebaseerd op wetenschappelijke inzichten inzake de fysieke ontwikkeling van leerlingen. De standaardisering van de scoring van deze oefeningen is uiteraard moeilijker dan de cognitieve toetsen. Onder meer door videovoorbeelden streeft men naar gelijkgerichtheid in de beoordeling door verschillende leraren. Bij de huidige experimenten ligt de interbeoordelaarsbetrouwbaarheid meestal tussen 0,7 en 0,8. Deze ontwikkeling staat evenwel nog maar in de kinderschoenen. Het is de intentie om ook hiervan op termijn een referentiegroep samen te stellen op basis waarvan over elke leerling kan worden aangegeven of ze significant boven of onder het gemiddelde presteren.

Curriculum-gerelateerde toetsen

CEM speelt momenteel tevens in op de gepercipieerde vraag om meer curriculum-gerelateerde toetsen te ontwikkelen. Ook deze ontwikkelingen zijn - met uitzondering van bepaalde subtoetsen in PiPS en INSIGHT - voorlopig evenwel nog eerder beperkt.

Tablet

Naast de drie huidige versies (online, lokaal, hybride) wil men in de nabije toekomst ook een tablet-versie ontwikkelen.

2.7 Leerwinst en toegevoegde waarden binnen toetsen van CEM

De toetsen van CEM zijn sterk gericht op het microniveau, en dus eerder op leerwinst dan op toegevoegde waarde. Toch worden ook data op het mesoniveau gegeven aan scholen.

2.7.1 *Leerwinst bij de toetsen van CEM*

Verschiedende toetssystemen zijn sterk gericht op het in kaart brengen van de leerwinst van de leerlingen. Dit is één van de grootste sterktes van het leerwinstmonitoringssysteem, aldus verschillende respondenten.

"That we can track the child year on year, that is the power of these tests. It makes teachers and headteachers look in a completely different way to the kids." (interview, respondent EN2)

De verschillende afnames van de verschillende toetsen binnen PIPS (*Baseline*, Key Stage 1 en Key Stage 2) kunnen aan elkaar gekoppeld worden, zodat voor elke leerling in kaart wordt gebracht hoe de prestaties zich verhouden ten opzichte van eerdere meetmomenten. Hierbij wordt gekeken naar de algemene voortgang van leerlingen doorheen de tijd: het betreft een minder cijfermatige benadering, waarbij het algemene profiel (met meer datapunten) meer in rekening wordt gebracht. Zo wordt bijvoorbeeld bij INCAS de voortgang van leerlingen per leerdomein op een grafiek gepresenteerd. Afwijkingen als gevolg van meetfouten zijn daardoor makkelijker in perspectief te plaatsen (bijvoorbeeld een eenmalige sterke of zwakke prestatie ten opzichte van eerdere en volgende

prestaties). De leerwinst van leerlingen wordt hier gezien als het verschil tussen de behaalde score op een toets en de score op een voorgaande toets, zoals beschreven in de conceptualisering van 'leerwinst als groei'.⁴⁶

Daarnaast worden in bovenstaande toetssystemen ook de verwachte 'normale' groeilijnen binnen de populatie indicatief aangegeven, waarbij een maat van leerwinst wordt berekend op basis van regressielijnen en residuals, of als het verschil tussen een behaalde score op een toets en een voorspelde score op basis van een eerdere toets. Dit komt overeen met het *gain score model*.

In bepaalde toetssystemen van CEM wordt leerwinst berekend als het verschil tussen een voorspelde score op basis van een voorspelling een eerdere toets, en de score op een volgende toets, waarbij de twee toetsen niet dezelfde schaal hanteren. Op basis van elke toets in PIPS kan ook een voorspelling worden gemaakt voor de score van elke leerling op de verplichte eindtoets van Key Stage 1 en Key Stage 2. In het secundair onderwijs wordt leerwinst berekend op basis van het verschil van de resultaten op GCSE, INSIGHT of zelfs Key Stage 3 toetsen (door leraren ontwikkeld) ten opzichte van een voorspelde score op basis van de eerdere score op MidYis. Idem worden de resultaten van de *A-Levels* gebruikt om leerwinst te berekenen met YELLIS als beginmeting. Bovendien blijkt dat er ook scholen zijn die individuele streefdoelen voor leerlingen bepalen op basis van de toetsen van CEM (voornamelijk tijdens Key Stage 3); achteraf wordt met elke leerling de positieve of negatieve leerwinst geëvalueerd.

Het is opvallend dat deze voorspellingen zowel op curriculum-gerelateerde als niet-curriculum-gerelateerde toetsen kunnen gebaseerd worden. In dat laatste geval gaat het om een meting van een 'algemene vaardigheid'.

Bij de leerwinstberekeningen worden enkel eerdere leerprestaties in rekening gebracht, terwijl andere vaak gehanteerde factoren (geslacht, socio-economische status, thuistaal, ...) niet worden opgenomen. Deze keuze is deels pragmatisch van aard, maar deels ook ingegeven door het idee dat het discriminerend is om een verschillende leerwinst te verwachten van mensen met een verschillende achtergrond, wanneer het beginniveau reeds in rekening wordt genomen.

Aangezien in elk van deze voorbeelden de voorspelde scores worden bepaald aan de hand van eerdere toetsafnames (van eerdere cohorten van leerlingen) sluit deze benadering aan bij het *projection model* als conceptualisering van leerwinst. In het lager onderwijs - waar men vaker over data op verschillende meetmomenten beschikt - zou het in principe mogelijk zijn om de data van deze verschillende meetmomenten in één model te integreren, waarbij rekening kan gehouden worden met het eerder curvi-lineaire verloop van het leren van leerlingen. Momenteel is dat evenwel nog niet aan de orde, aangezien scholen hier totnogtoe ook geen vragende partij voor zijn.

⁴⁶ Zie Deel 3, sectie 3.2.2 voor een beschrijving van de conceptualisering van 'leerwinst als groei', van het *gain score model* en het *projection model*.

2.7.2 Toegevoegde waarde bij de toetsen van CEM

Bij de berekening van toegevoegde waarde wordt gebruik gemaakt van lineaire regressie.

“We use linear regression, so we’d have a standardized score as the predicted or taken score. For example, they are coming to school at year seven, they take a Midyis year seven test, and five years later they take the GCSE. We’ll match the two datasets together (...) and do a linear regression. (...) And then we look at the residuals for each child, so whether a child did better or worse than predicted by the regression line, and then we aggregate that to the school level. So it’s essentially a comparison of how did your school do in comparison to others.” (interview, respondent EN8)

De gemiddelden (en de spreiding) op schoolniveau worden vergeleken met dezelfde parameters van leerlingen in andere scholen, op basis van een voorafgaande nationale steekproef. Het betreft in de eerste plaats ‘statusmodellen’ (vergelijking van eenmalige leerprestaties), maar ook ‘gain score modellen’ waarbij de leerwinst van leerlingen in een school vergeleken wordt met de populatie (op basis van eerder verzamelde data)⁴⁷.

De referentiebasis waarmee scholen vergeleken worden, zijn de leerlingen van andere scholen die in het verleden de toets hebben afgenomen. Hier vindt evenwel nog een correctie op plaats, om de (niet-aselecte) steekproef meer representatief te maken voor alle scholen in Engeland: er nemen immers relatief veel ‘onafhankelijke scholen’ deel aan de CEM-toetsen (waar de leerlingen gemiddeld hogere scores behalen), zodat de resultaten enige bias kunnen vertonen. Er wordt geen gebruik gemaakt van corrigerende factoren (op basis van leerlinginstroom), maarde ervaring leert dat dit voor verschillende scholen ‘too black-box’ is.

In secundaire scholen is CEM beperkt tot toegevoegde waarde-metingen omdat men van veel scholen maar data op twee momenten ter beschikking heeft. Voor lagere scholen, waar men vaker over data van verschillende meetmomenten beschikt, zou men in principe ook meer complexe (multivariate) toegevoegde-waarde metingen kunnen maken, maar momenteel is dit nog niet aan de orde. Het is evenwel een moeilijk evenwicht, aangezien ook relatief eenvoudige berekeningen van toegevoegde waarde voor verschillende scholen al erg moeilijk te interpreteren zijn. Men sluit meer complexe toegevoegde-waarde-modellen evenwel niet uit als een mogelijke toekomstige evolutie. .

⁴⁷ Zie deel 2 voor conceptualisering van *statusmodel* en *gain score model*.

3. CASE NEDERLAND (PILOOTPROJECT LEERWINST EN TOEGEVOEGDE WAARDE IN HET PRIMAIR ONDERWIJS)

De case 'Nederland' verschilt enigszins van de drie andere internationale studies. Meer dan het in kaart brengen van de gebruiken inzake gestandaardiseerde toetsen en de toepassing ervan voor eventuele berekening van leerwinst en/of toegevoegde waarde, wordt de focus in deze case gelegd op het pilootproject 'Leerwinst en Toegevoegde Waarde in het Primair Onderwijs' dat in de periode 2012-2014 werd uitgevoerd door een projectgroep onder leiding van Prof. Janssens. De opbouw van de case is evenwel gelijkaardig als de andere cases: eerst komt een korte inleiding in het Nederlandse onderwijssysteem. Daarna geven we een overzicht van de gestandaardiseerde toetsen (met name de eindexamens) die in Nederland in gebruik zijn. Na deze context te hebben voorzien, zoomen we in op het bovengenoemde pilootproject.

3.1 Inleiding tot het onderwijssysteem in Nederland⁴⁸

Kinderen zijn in Nederland leerplichtig van 5 tot 16 jaar. De leerplicht is verspreid over het primair onderwijs en het voortgezet onderwijs, al maken geen van beide niveaus volledig deel uit van het systeem van leerplicht.

3.1.1 *Onderwijsstructuur*

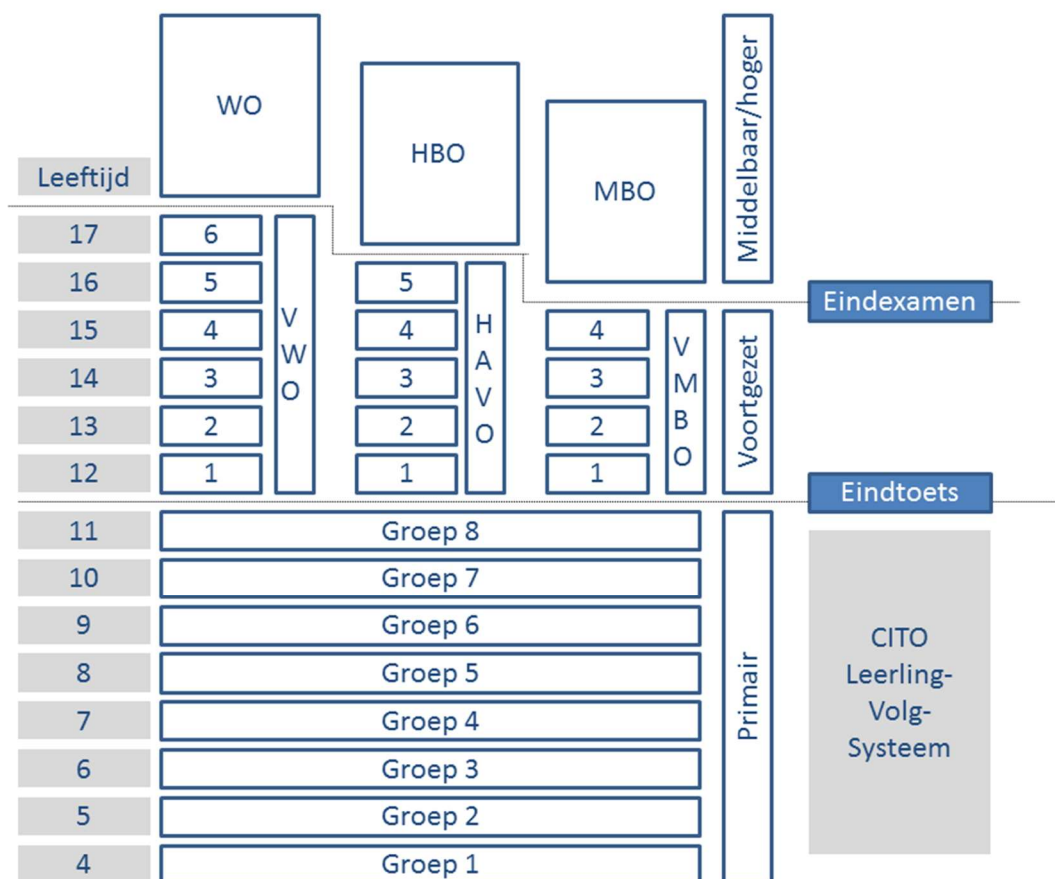
Het basisonderwijs is bestemd voor leerlingen in de leeftijd van 4 tot 12 jaar en is verplicht voor leerlingen vanaf 5 jaar. Het primair is normaliter ingedeeld in acht jaargroepen. Na groep 8 vervolgen leerlingen hun schoolloopbaan in het voortgezet onderwijs. Voortgezet onderwijs is bestemd voor leerlingen in de leeftijd van 12 tot 16, 17 of 18 jaar. Na afronding van het voortgezet onderwijs kunnen leerlingen doorstromen naar het middelbaar beroepsonderwijs of hoger onderwijs. Figuur 17 bevat een visueel overzicht van de onderwijsstructuur in Nederland.

Leerplichtonderwijs

Het primair onderwijs resulteert niet in een diploma, maar leerlingen krijgen wel een schooladvies voor het voortgezet onderwijs. De basisschool stelt in het laatste jaar per leerling een individueel advies voor het vervolgonderwijs op. Het schooladvies is de verwachting van de directeur van de basisschool over de geschiktheid van een leerling voor een bepaald type van voortgezet onderwijs. In principe hoeven ouders en scholen dit advies niet te volgen en kunnen zij hier van afwijken. In de praktijk blijkt echter dat de scholen voor voortgezet onderwijs het schooladvies doorgaans zeer nauwkeurig opvolgen.

De rol van de eindtoetsen primair onderwijs in het bepalen van dit advies is sinds 2015 grondig gewijzigd. Dit wordt uitvoerig besproken in sectie 3.2.1.

⁴⁸ Gebaseerd op: Doolaard (2013); EP-nuffic (2014); Inspectie van het Onderwijs (2015); Ministerie van Onderwijs Cultuur en Wetenschap (s.d.); Rijksoverheid (s.d.-a, s.d.-b, s.d.-c); Van Eck en Boogaerd (2007) + informatie tijdens gesprekken in case studie.



Figuur 17. Overzicht van het Nederlandse onderwijssysteem, inclusief gestandaardiseerde toetsen

Het voortgezet onderwijs wordt gekenmerkt door een sterke stratificatie van de leerlingenstroom en is onderverdeeld in de volgende types:

Voorbereidend Middelbaar Beroepsonderwijs (vmbo) duurt vier jaar en is bedoeld voor leerlingen in de leeftijd van 12 tot 16 jaar. Leerlingen met een vmbo diploma kunnen doorstromen naar het Middelbaar Beroepsonderwijs (mbo) of naar het vierde jaar van de havo. Binnen vmbo wordt gedifferentieerd tussen de volgende 'leerwegen' (in oplopende 'moeilijkheidsgraad'): basisberoepsgerichte leerweg, kaderberoepsgerichte leerweg, gemengde leerweg, en theoretische leerweg.

Hoger Algemeen Voortgezet Onderwijs (havo) duurt vijf jaar en is bedoeld voor leerlingen van 12 tot 17 jaar. Havo bestaat uit een eerste fase (drie jaar) en een tweede fase (twee jaar). Havo biedt algemene vorming en is een voorbereiding op het hoger beroepsonderwijs (hbo). Leerlingen kunnen na afronding van de havo ook doorstromen naar het vijfde jaar van het vwo.

Het *Voorbereidend Wetenschappelijk Onderwijs* (vwo) bereidt leerlingen voor op een vervolgstudie aan een universiteit. De opleiding duurt zes jaar en is bedoeld voor leerlingen van 12 tot 18 jaar. Het vwo bestaat uit twee fasen van telkens drie jaar.

Middelbaar of hoger (beroeps-)onderwijs

In het middel of hoger (beroeps-)onderwijs worden eveneens drie types onderscheiden.

Het *middelbaar beroepsonderwijs* (mbo) bereidt studenten voor op de beroepspraktijk of op een vervolgopleiding. De opleidingsduur is afhankelijk van de gekozen kwalificatie. Het mbo kent vier opleidingsniveaus, variërend van zes maanden tot vier jaar.

Het *hoger beroepsonderwijs* (hbo) is een op de praktijk gerichte vorm van hoger onderwijs met als voornaamste doel het overdragen van theoretische kennis en het ontwikkelen van vaardigheden in nauwe aansluiting op de beroepspraktijk. Een opleiding in het hbo duurt vier jaar en leidt tot de graad van bachelor in het desbetreffende vakgebied.

Een vwo of een propedeuse hbo (met eventueel aanvullende voorwaarden) bieden beiden toegang tot *wetenschappelijk onderwijs* (wo). Wo biedt onderwijsprogramma's aan met als voornaamste doel het zelfstandig beoefenen van de wetenschap of de beroepsmatige toepassing van wetenschappelijke kennis. Wo is ingedeeld in een bachelor opleiding van drie jaar gevolgd door een masteropleiding van één tot drie jaar.

Tabel 10

*Aantal leerlingen Nederlands voortgezet onderwijs per leerjaar en onderwijstype in 2014*⁴⁹

	Derde leerjaar		Vierde leerjaar		Vijfde leerjaar		Zesde leerjaar	
vmbo	79.600	45.9%	77.900	42.8%				
havo	44.700	25.7%	61.700	33.9%	53.400	56.7%		
vwo	43.600	25.1%	42.300	23.3%	40.700	43.3%	37.100	100%
ongedeeld ⁵⁰	5.700	3.3%						
Totaal aantal leerlingen	173.600	100%	181.900	100%	94.100	100%	37.100	100%

Uit Tabel 10 kan worden afgeleid dat in het derde leerjaar van het voortgezet onderwijs 45,9% van de leerlingen kiest voor een studierichting in het vmbo. Het havo en vwo krijgen elk ongeveer een kwart van de leerlingen. Cijfers zijn enkel beschikbaar voor het derde tot en met zesde leerjaar. Vaak zitten in het eerste en tweede leerjaar van het voortgezet onderwijs de verschillende onderwijstypes immers nog bij elkaar. Dit heeft te maken met het feit dat veel leerlingen dan nog in een brugklas zitten.

3.1.2 Leerlingen en scholen

In Nederland wordt een onderscheid gemaakt tussen openbaar onderwijs en (algemeen) bijzonder onderwijs. Openbare basisscholen zijn voor ieder kind toegankelijk en niet gebaseerd op een

⁴⁹ Gebaseerd op cijfers van www.stamos.nl

⁵⁰ Leerlingen van verschillende onderwijstypes door elkaar

godsdienst of levensovertuiging. Bijzondere scholen geven meestal les vanuit een godsdienstige of levensbeschouwelijke overtuiging. Algemeen bijzondere scholen geven les op grond van hun visie over onderwijs of opvoeding (bijvoorbeeld montessorischolen of daltonscholen).

Daarnaast is er speciaal onderwijs voor leerlingen met een lichamelijke of verstandelijke handicap, leerlingen met bepaalde stoornissen of gedragsproblemen, en voor slechthorende, slechtziende of langdurig zieke leerlingen. Het aantal leerlingen per schooltype is weergegeven in Tabel 11.

Tabel 11

*Aantal leerlingen per onderwijsniveau en schooltype in Nederland in 2014*⁵¹

	Primair onderwijs		Voortgezet onderwijs		Speciaal onderwijs	
Algemeen bijzonder	117.780	7.5%	129.896	13.1%	38.316	35.3%
Openbaar	462.617	29.6%	259.829	26.2%	21.732	20.0%
Bijzonder	983.528	62.8%	569.283	57.4%	48.384	44.6%
Samenwerking bijzonder/ algemeen	219	0.0%	26.488	2.7%	8.112	0.5%
Samenwerking bijzonder/openbaar	962	0.1%	3.342	0.3%	17.369	1.1%
Samenwerking openbaar/ algemeen			1.978	0.2%		
Totaal aantal leerlingen	1.565.106	100%	990.816	100%	133.913	100%

3.1.3 Curriculum

Het curriculum is gericht op het behalen van kerndoelen. Om de overgang tussen primair onderwijs en voortgezet onderwijs (en later het hoger en middelbaar onderwijs) te faciliteren, werden inzake Nederlands en rekenen zogenaamde 'referentieniveaus' uitgewerkt. Deze referentieniveaus bevatten zowel 'fundamentele niveaus' (minimumvereisten) als 'streefniveaus' (uitbreidingsdoelen).

Primair onderwijs

Elke leerling op de basisschool krijgt de wettelijk verplichte vakken Nederlands, Engels, rekenen en wiskunde, oriëntatie op jezelf en de wereld (bijvoorbeeld aardrijkskunde, geschiedenis), kunstzinnige oriëntatie (bijvoorbeeld muziek, tekenen en handvaardigheid), bewegingsonderwijs (bijvoorbeeld gymlessen), en seksualiteit en seksuele diversiteit. Enkel voor basisscholen in de provincie Friesland komt daar nog het verplichte vak Fries bij. Naast verplichte vakken bestaan er ook niet verplichte vakken die scholen kunnen aanbieden, zoals godsdienstles, Frans, Duits, wetenschap en techniek. De *kerndoelen voor het basisonderwijs* geven per vak aan wat een kind aan het eind van de basisschool

⁵¹ Gebaseerd op cijfers van www.stamos.nl

moet weten en kunnen. De overheid legt deze kerndoelen vast. Een school bepaalt zelf hoe de lessen in deze vakken worden gegeven en met welk lesmateriaal.

Voortgezet onderwijs

In de onderbouw (jaar 1 en 2 voor vmbo en jaar 1, 2 en 3 voor havo en vwo) van het voortgezet onderwijs zijn 58 globaal geformuleerde kerndoelen van kracht. Deze zijn geordend in zeven leerdomeinen: Nederlands, Engels, wiskunde, mens en natuur, mens en maatschappij, kunst en cultuur, en bewegen en sport. Scholen hebben daarmee de ruimte gekregen om zelf inhoudelijke keuzes te maken in het onderwijsprogramma voor leerlingen (op het gebied van vakken, projecten en leergebieden), rekening houdend met verplichte onderdelen. Voor vmbo, havo en vwo zijn overzichten gemaakt waarop per vak en per leergebied te zien is welke doorlopende lijn er is vanuit de kerndoelen in het primair onderwijs, via de kerndoelen van de onderbouw van het voortgezet onderwijs, naar ten slotte de examenprogramma's die de bovenbouw afronden.

Aan het einde van het tweede schooljaar kiest elke vmbo-leerling een sector (studierichting). Een sector bereidt vmbo studenten voor op hun vervolgopleiding en het werk dat ze na hun opleiding kunnen doen. Havo en vwo onderscheiden in de bovenbouw vier profielen: natuur en techniek, natuur en gezondheid, economie en maatschappij, en cultuur en maatschappij. Iedere havo- of vwo-leerling heeft een aantal verplichte vakken en twee of drie verplichte keuzevakken uit één van de vier profielen.

3.1.4 Interne kwaliteitszorg en 'opbrengstgericht werken'

Interne kwaliteitszorg is sinds het begin van de 21^e eeuw steeds meer een belangrijk thema geworden in het Nederlandse onderwijs. Uit onderzoek blijkt dat de mechanismen van interne kwaliteitszorg op vele scholen relatief goed zijn ontwikkeld⁵². De onderwijsinspectie geeft aan dat deze investering in interne kwaliteitszorg meer en meer vruchten begint af te werpen op het klasniveau. In haar onderwijsrapport spreekt de onderwijsinspectie van het ontstaan van een 'verbetercultuur' in scholen:

"Op scholen en opleidingen die een kwaliteitszorgsysteem gebruiken voor verbetering van de lessen en de eigen kwaliteit, ontstaan vaak een dynamische verbetercultuur en beter onderwijs." (Inspectie van het Onderwijs, 2015, p. 27)

Desalniettemin ondervindt men dat er nog grote verschillen zijn tussen scholen inzake de kracht van de mechanismen voor interne kwaliteitszorg. De scholen beschikken over relatief veel data (onder meer de leerresultaten van leerlingen volgens gestandaardiseerde toetsen), voornamelijk in het primair onderwijs. Deze data worden in veel scholen ingezet in een proces van schoolontwikkeling met het oog op het versterken van de leerresultaten van leerlingen. Omwille van die focus op het verhogen van de leerresultaten, wordt in het kader van interne kwaliteitszorg vaak gesproken over 'opbrengstgericht werken'.

⁵² Zie hiervoor onder meer de studie van Van Eck en Boogaerd (2007)

3.1.5 Externe kwaliteitszorg

De Inspectie van het Onderwijs bewaakt de kwaliteit van het onderwijs en stimuleert scholen en instellingen deze kwaliteit te verbeteren. Elke school in het primair of voortgezet onderwijs krijgt minimaal eens in de vier jaar bezoek van een inspecteur. Op basis van een jaarlijks risicoanalyse en eventueel nader onderzoek bepaalt de Onderwijsinspectie hoeveel toezicht een school werkelijk 'nodig' heeft, en of het toezicht moet worden geïntensiveerd. Bij de risicoanalyse onderzoekt de inspectie opbrengsten, zoals het resultaat van een centrale eindtoets of centrale examens; jaarstukken, zoals de schoolgids en financiële stukken; signalen, zoals meldingen van klachten en (negatieve) berichtgevingen in de media; en nalevingsaspecten, zoals onderwijstijd.

Als de analyse geen risico's laat zien, dan is er geen nader onderzoek of intensivering van het toezicht nodig. Bijgevolg krijgt de school een basistoezicht. Als de analyse wel risico's laat zien, dan vraagt de inspectie aanvullende informatie op. Zij onderzoeken de aard en achtergrond van de mogelijke risico's. Ook maakt de inspectie een uitgebreidere analyse van alle gegevens. De inspecteur bespreekt deze informatie met het schoolbestuur. Blijken er geen tekortkomingen te zijn, dan krijgt de school alsnog basistoezicht. Als er wel sprake is van mogelijke tekortkomingen, dan volgt een kwaliteitsonderzoek. Indien de opbrengsten drie achtereenvolgende jaren onder de norm liggen, volgt altijd een kwaliteitsonderzoek.

3.2. Gestandaardiseerde toetsen in Nederland⁵³

In dit deel geven we een overzicht van de eindexamens die aan het eind van het primair en aan het eind van het voortgezet onderwijs worden afgenomen. Deze eindexamens vormen immers een belangrijke contextfactor waarin het pilootproject 'Leerwinstmonitoring en Toegevoegde Waarde in het primair onderwijs' heeft plaatsgevonden.

3.2.1 Centrale Eindtoets in het primair onderwijs

Vanaf het schooljaar 2014 - 2015 is het voor alle leerlingen van groep 8 in het reguliere basisonderwijs verplicht om een eindtoets te maken. Voorheen was dit niet verplicht, al was het wel zeer gebruikelijk in een ruime meerderheid van de scholen. Scholen kunnen hiertoe kiezen om gebruik te maken van één van de eindtoetsen van verschillende aanbieders. Elk van de toetsen moet evenwel door het Ministerie van Onderwijs, Cultuur en Wetenschap worden goedgekeurd. In de praktijk gebeurt dit door een onafhankelijke overheidsorganisatie, het College voor Toetsen en Examens (CvTE). Het CvTE *"waarborgt namens de overheid de kwaliteit en het niveau van toetsen en examens en stelt scholen en instellingen in staat de afname van toetsen en examens vlekkeloos te laten verlopen"* (College voor Toetsen en Examens, s.d.-b).

⁵³ Gebaseerd op: A-Vision (2015); Alberts en Erens (2015); Bishop (1995); Bureau ICE (s.d.); Cito (2013a, z.j.-b); College voor Toetsen en Examens (2015a, 2015b, s.d.-a, s.d.-b); Inspectie van het Onderwijs (2014a, 2014b, 2015); Prokic-Breuer en Dronkers (2012) + informatie tijdens gesprekken in case studie.

Naast de beoordeling van de eindtoetsen van externe aanbieders, geeft het CvTE ook zelf de opdracht om een eindtoets te maken, met name de Centrale Eindtoets. De ontwikkeling daarvan gebeurt in handen van CITO. In 2015 werd in de scholen van 86% van de leerlingen gebruik gemaakt van de Centrale Eindtoets van CITO en het CvTE.

Elk van deze eindtoetsen primair onderwijs moet rapporteren over de mate waarin de leerlingen de referentieniveaus hebben bereikt. Een toelating door de Minister van Onderwijs voor het gebruik van een eindtoets, geldt meteen voor een periode van 4 jaar. Wel moet de inhoud van jaar tot jaar verschillend zijn. Daarom vindt er jaarlijks een controle door het CvTE plaats, waarbij gekeken wordt naar zowel betrouwbaarheid als validiteit van de eindtoetsen.

Er zijn twee alternatieve aanbieders van een eindtoets die na goedkeuring door het Ministerie van Onderwijs, Cultuur en Wetenschap door scholen mag worden gebruikt. De IEP Eindtoets van het onderwijsbureau ICE wordt gepromoot als een toets waarin (zuiverder dan in de andere beschikbare toetsen) de reken- en taalvaardigheid van leerlingen wordt gemeten, via een combinatie van meerkeuze- en open vragen. Route8 is ontwikkeld door het onderwijsbureau A-Vision en verschilt van de andere toetsen door haar computer-adaptieve karakter. Het is de enige eindtoets waarbij niet alle leerlingen identiek dezelfde items voorgeschoteld krijgen.

In de verdere beschrijving van de eindtoets primair onderwijs beperken we ons tot het meest gebruikte instrument, met name de eindtoets van CITO/CvTE.

Doel op microniveau (individuele leerlingen)

De Centrale eindtoets kan worden beschouwd als een toets die zich binnen een verantwoordingsgericht perspectief richt op het microniveau. Deze Centrale eindtoets past in het kader van het 'schooladvies', waarover we eerst kort toelichting geven. Basisscholen geven elk jaar vóór 1 maart aan alle leerlingen van groep 8 een (schriftelijk vastgelegd) schooladvies. Door middel van dit schooladvies wordt aangegeven welk type voortgezet onderwijs het beste bij een leerling past op basis van de aanleg en talenten van een leerling, de leerprestaties, de ontwikkeling tijdens de hele basisschoolperiode, de concentratie, motivatie en het doorzettingsvermogen van een leerling. Het schooladvies omvat niet meer dan twee leerwegen of schoolsoorten. Basisscholen mogen dus alleen nog een enkel (bijvoorbeeld havo) of dubbel (bijvoorbeeld havo-vwo) advies geven.

In tegenstelling tot de periode voor 2014, wordt het schooladvies van de basisschool onafhankelijk van de eindtoetsen primair onderwijs geformuleerd. Daartoe moet dit advies geformuleerd worden voor de afname van de eindtoetsen. De eindtoetsen kunnen het advies enkel positief bijstellen: indien bijvoorbeeld een leerling met het schooladvies 'havo' via de eindtoetsen bewijst toch het referentieniveau te bereiken van leerlingen die georiënteerd worden naar 'vwo', dient de school het schooladvies te heroverwegen, wat mogelijk kan leiden tot een positieve bijstelling. Andersom is dit niet mogelijk: een schooladvies 'vwo' kan niet op basis van de eindtoetsen worden bijgesteld naar 'havo'.

Het doel van de Centrale Eindtoets is bijgevolg het geven van een objectief ondersteunend advies (*'second opinion'*) aan de school over de keuze van het best passende type voortgezet onderwijs (vo). De Centrale Eindtoets meet wat een kind in acht jaar basisonderwijs geleerd heeft en zegt iets over de kansen op succes in de verschillende typen van het voortgezet onderwijs. Op zich is de toelating tot het voortgezet onderwijs evenwel niet afhankelijk van het resultaat op de eindtoets.

Doel op mesoniveau (scholen)

Ook op mesoniveau worden de resultaten van de eindtoetsen in een verantwoordingsgericht perspectief gebruikt. De Onderwijsinspectie maakt jaarlijks van elke basisschool een 'risicoanalyse'. Op basis van de resultaten van de uitkomsten van die risicoanalyse kan beslist worden om een school meer intensief te inspecteren. Er worden bij die risicoanalyse zeven 'opbrengstindicatoren' onderscheiden. Twee van deze zeven indicatoren zijn rechtstreeks gelinkt aan de studieresultaten van leerlingen:

- De resultaten van de leerlingen *aan het eind* van de basisschool liggen ten minste op het niveau dat op grond van de kenmerken van de leerlingenpopulatie mag worden verwacht; en
- de resultaten van de leerlingen voor Nederlandse taal en voor rekenen en wiskunde *tijdens de schoolperiode* liggen ten minste op het niveau dat op grond van de kenmerken van de leerlingenpopulatie mag worden verwacht. Deze tweede indicator wordt in kaart gebracht op basis van de resultaten van leerlingen op het leerlingvolgstelsel van CITO (of een andere leerlingvolgsysteem) en wordt bijgevolg besproken in sectie 3.3.1.

Wat betreft de eerste indicator, vergelijkt de inspectie de resultaten van de school met die van andere scholen met een overeenkomstig samengestelde leerlingenpopulatie. De eerste indicator wordt als volgt beoordeeld: een school krijgt de beoordeling 'goed' indien de leerresultaten van de leerlingen aan het eind van de basisschool drie schooljaren na elkaar ten minste op het niveau liggen dat op grond van de kenmerken van de leerlingenpopulatie mag worden verwacht; de school krijgt de beoordeling 'voldoende' indien de leerresultaten van de voorbije drie cohorten twee maal voldeden aan het bovenstaande criterium. Als de leerresultaten van de voorbije drie cohorten slechts één keer voldoende zijn, krijgt de school ook de beoordeling 'voldoende' indien het aantal in de toets betrokken leerlingen tien of meer bedraagt. Indien dat niet het geval is, dan betreft de inspectie de leerresultaten van groep 7 bij de beslissing om te bepalen of de school de beoordeling 'voldoende', dan wel 'onvoldoende' krijgt. De school krijgt sowieso de beoordeling 'onvoldoende' op deze indicator wanneer de resultaten aan het eind van de basisschool van geen enkele van de drie voorbije cohorten op het niveau liggen dat op grond van de kenmerken van de leerlingenpopulatie mag verwacht worden.

Aard van de toetsen

De inhoud van de Centrale Eindtoets van CITO wordt afgestemd op de referentieniveaus taal en rekenen en brengt in kaart welk referentieniveau de leerling in de verschillende leerdomeinen heeft behaald.

De Centrale Eindtoets van CITO duurt zes uur verdeeld over drie dagdelen. De eindtoets bestaat uit twee verplichte onderdelen: Nederlandse taal (lezen, woordenschat, schrijven, en taalverzorging) en rekenen (getallen, verhoudingen, meten en meetkunde, en verbanden). Daarnaast kunnen scholen kiezen voor het afnemen van het onderdeel wereldoriëntatie (ruimte, tijd, natuur en techniek, mens en samenleving). In 2015 legde 70% van alle leerlingen (in de scholen die kozen voor de Centrale Eindtoets van CITO en CvTE) een eindtoets in wereldoriëntatie af. Dit laatste onderdeel heeft evenwel geen invloed op het advies voor het voortgezet onderwijs.

Scholen kunnen (per leerling) kiezen voor een eindtoets op papier of voor de digitale versie. De onderdelen, de volgorde van de taken en het aantal opgaven zijn identiek in de papieren en de digitale eindtoets. De opgaven zelf zijn wel verschillend, omdat de digitale eindtoets verschillende, gelijkwaardige varianten kent, die gedurende een langere periode dan de papieren variant afgenomen kunnen worden. Voor het afnemen van een digitale eindtoets heeft een school voldoende computers nodig. Er is geen versie voor tablets beschikbaar.

De eindtoets voor taal en rekenen is ontwikkeld op twee verschillende niveaus. De twee versies heten 'eindtoets B (Basis)' en 'eindtoets N (Niveau)'. Beide versies bevatten dezelfde onderdelen en opgavenrubrieken en hetzelfde aantal opgaven. De eindtoets N is bestemd voor leerlingen van wie het schooladvies luidt dat ze het beste passen in vmbo met een basisberoepsgerichte leerweg of kaderberoepsgerichte leerweg. De eindtoets B is bestemd voor leerlingen voor wie het verwachte vervolgadvis type vmbo gemengde/theoretische leerweg of hoger zal zijn. Het onderdeel wereldoriëntatie is in principe geschikt voor alle leerlingen en wordt op één niveau afgenomen.

Vanaf 2018 zal het verschil tussen de toetsen B en N wegvallen. Men plant immers om vanaf dat moment adaptieve toetsen te kunnen aanbieden.

Ontwikkeling

Het CvTE is verantwoordelijk voor de uitvoering van de eindtoetsen. Deze worden in samenwerking met Stichting CITO gemaakt en bouwen voort op de Eindtoets Basisonderwijs van CITO, die voorheen facultatief aan scholen werd aangeboden. De toetsontwikkeling zelf is in handen van CITO. Voor elk toetsonderdeel werken een toetsdeskundige van CITO en gemiddeld drie leraren van groep 8 samen aan het maken van de opgaven. Voor de kwaliteit van een toets is het belangrijk dat deze aansluit bij de vaardigheden en beleavingswereld van leerlingen.

Om de betrouwbaarheid en de kwaliteit van de eindtoetsen te garanderen, gelden er voorwaarden voor bijvoorbeeld de inhoud en de afname van de eindtoets. Deze eisen gelden voor alle eindtoetsen. Zo gelden er eisen voor de wettelijk verplichte domeinen uit het Referentiekader Taal en Rekenen. Daarnaast werden ook kwaliteitseisen opgesteld voor de domeinen die toetsontwikkelaars optioneel in de eindtoets kunnen opnemen.

Alle opgaventypen in de centrale eindtoets zijn in de praktijk uitgetoetst, in een proefafname bij een representatieve groep leerlingen. Het betreft opgaventypes, en niet de exacte opgaven die uiteindelijk in de Centrale Eindtoets worden opgenomen. Uitgangspunt is dat alleen de opgaventypen in de toets

opgenomen worden waarvan in deze proeffase wordt uitgewezen dat ze meten wat ze behoren te meten. Uit de opgaven stelt het CvTE een evenwichtige toets samen. Dat gebeurt door zogenaamde vaststellingscommissies, waarin vakdeskundigen en leraren van groep 8 zetelen. Alle scholen kunnen zich kandidaat stellen om aan deze proefafname deel te nemen.

Afname

De eindtoets wordt jaarlijks tussen 15 april en 15 mei afgenomen op drie opeenvolgende schooldagen. Indien een school ook kiest voor het onderdeel wereldoriëntatie dan wordt die toets afgenomen aansluitend op de laatste toetstaak taal of rekenen van die dag. Leerlingen die afwezig zijn tijdens de toetsafname, kunnen aan een inhaalmoment deelnemen. Tijdens dit inhaalmoment zijn de toetsen evenwel enkel digitaal beschikbaar.

De directie van de school staat in voor de organisatie van het toezicht tijdens de toetsafname. De directie is er ook verantwoordelijk voor dat de toetsopgaven geheim worden gehouden tot het moment van de toetsafname. Voor de wijze van afname zijn uitgebreide procedurele voorschriften opgesteld.

Kwaliteitszorg

Om de toetsresultaten op waarde te schatten, analyseert CITO na de afname alle gegeven antwoorden. Op basis van de analyses adviseert CITO het CvTE over de normering. Het CvTE stelt vast hoeveel opgaven een leerling correct moet hebben beantwoord om een bepaalde standardscore te krijgen. Belangrijk hierbij is dat gelijke prestaties een gelijke waardering krijgen door de jaren heen. Ook als de moeilijkheidsgraad van de opgaven van jaar tot jaar een beetje varieert. Een score van 535 is dus het ene jaar evenveel waard als het andere jaar. Hiertoe is het op item response theorie (IRT) gebaseerde meetmodel essentieel. Dit meetmodel zorgt ervoor dat elke deelttoets is samengesteld uit opgaven die de operationalisatie vormen van eenzelfde unidimensionale, onderliggende deelvaardigheid (bijvoorbeeld spellingvaardigheid of begrijpend lezen). Opgaven die niet voldoen aan de betreffende eisen worden uit de opgavenbank verwijderd. Uiteindelijk wordt ook de toets als geheel op de mate van overeenkomst met het meetmodel onderzocht. Met behulp van hetzelfde meetmodel wordt er tevens voor gezorgd dat elke nieuwe Centrale Eindtoets niet alleen qua inhoud, maar ook in psychometrische zin volledig equivalent is met eerdere en latere versies.

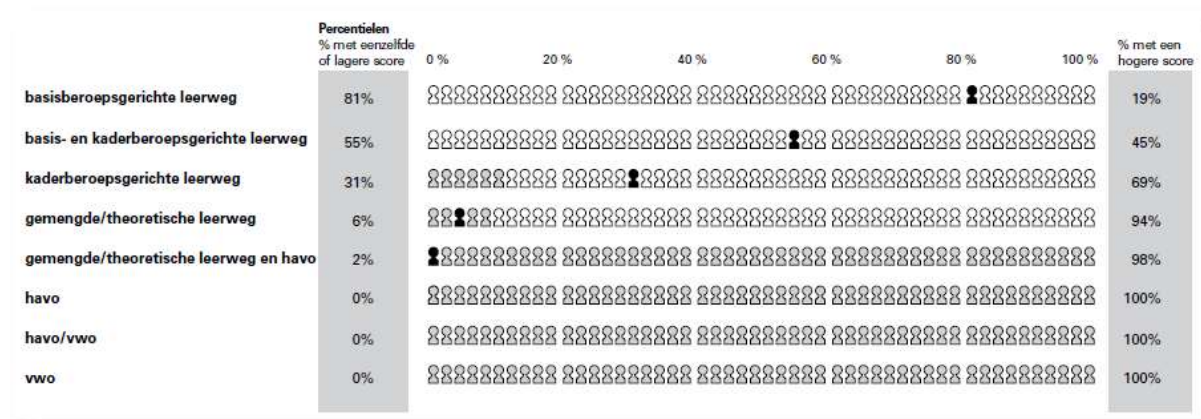
De inhoudelijke en psychometrische procedures dienen te resulteren in betrouwbare toetsen en deelttoetsen. Voor de totaalscore op de Centrale Eindtoets geldt voor de interne consistentie een waarde van .95 als norm, voor de deelscores op rekenen en taal een waarde van .90.

In de onderbouwing van de begrips- en criteriumvaliditeit wordt aandacht besteed aan modelpassing en kwaliteit van de opgaven, itembias, structuur (intercorrelaties), equivalentie, soortgenoot-, convergente en divergente validiteit, relatie met doorstroomadvies, verschillen tussen relevante subgroepen en samenhang met achtergrondkenmerken. Daarnaast wordt in samenwerking met het Centraal Bureau voor de Statistiek toelatings- en doorstroomonderzoek uitgevoerd en gerapporteerd ter onderbouwing van de criteriumvaliditeit of predictieve validiteit (de voorspellende waarde) van de Centrale Eindtoets.

Vanaf 2016 zal ook een ankeronderzoek worden uitgevoerd, waarbij toetsen van verschillende jaren op één schaal worden aangebracht. Zo wordt de vergelijkbaarheid van de toetsresultaten doorheen de jaren versterkt.

Rapportering

De school krijgt maximaal vier weken na de afname van de Centrale Eindtoets een overzicht van de resultaten. Per leerling worden zowel globale resultaten (inclusief eindresultaat tussen 500 en 550), als resultaten per onderdeel ter beschikking gesteld. Ook een overzicht van de beheersing van de referentieniveaus en een advies voor het beste passend type vervolgonderwijs wordt per leerling voorzien. Figuur 18 geeft hiervan een voorbeeld, waarbij van een bepaald leerdomein aangegeven wordt per onderwijstype hoeveel procent van de leerlingen die hierin terechtkomen lager of gelijk, dan wel hoger scoren dan de betreffende leerling: zo behaalt de leerling in dit voorbeeld een score die hoger (of gelijk is) dan de score van 81% van de leerlingen die kiezen voor vmbo basisberoepsgerichte leerweg. Het uiteindelijke advies is afhankelijk van het eindresultaat, waarbij onderscheid wordt gemaakt tussen vmbo basisberoepsgerichte leerweg (eindresultaat 500-523), vmbo kaderberoepsgerichte leerweg (524-528), vmbo gemengde of theoretische leerweg (529-536), havo (537-544) en vwo (545-550).



Figuur 18. Leerlingrapport van de Centrale Eindtoets (College voor Toetsen en Examens, 2015b)

Ten slotte worden ook geaggregeerde gegevens op schoolniveau ter beschikking gesteld, waarbij men de resultaten van de eigen school kan vergelijken met het landelijk gemiddelde. De wet op Openbaarheid van Bestuur bepaalt dat de resultaten openbaar beschikbaar moeten zijn.

3.2.2 Centraal examen voortgezet onderwijs

In het laatste jaar van het vmbo (jaar 4), havo (jaar 5) of vwo (jaar 6) leggen elk jaar ongeveer 200.000 leerlingen centrale examens af. Het eindcijfer van leerlingen aan het eind van het voortgezet onderwijs wordt samengesteld uit twee luiken. De helft van de score wordt toebedeeld onder

verantwoordelijkheid van de school - het zogeheten 'schoolexamen'. De andere helft wordt bepaald door de prestatie van de leerlingen op het centraal examen dat wordt vastgesteld door het College voor Toetsen en Examens (CvTE). Het eindcijfer voor een vak wordt bepaald door het gemiddelde cijfer voor het schoolexamen en het centraal examen.

De cijfers van het schoolexamen moeten door scholen worden doorgegeven voor de afname van het centraal examen, om fraude tegen te gaan (bijvoorbeeld indien het centraal examen een teleurstellend resultaat oplevert voor een bepaalde leerling). In tegenstelling tot in het primair onderwijs, zijn er geen verschillende aanbieders van het centraal examen voortgezet onderwijs.

Doel op microniveau (individuele leerlingen)

Het centraal examen is een formele toets aan het einde van het voortgezet onderwijs (vmbo, havo of vwo) in Nederland ten behoeve van individuele certificering. De toekomst van leerlingen wordt bepaald door het gemiddelde cijfer te berekenen van de schoolexamens en de centrale examens. Afhankelijk van dit cijfer wordt bepaald of een leerling een diploma van het voortgezet onderwijs krijgt. Hiertoe is een gemiddeld centraal examencijfer van 5,5/10 vereist. Bovendien moet ook de score in elk van de kernvakken op het centraal examen minimaal 5,5/10 zijn, waarbij per leerling één uitzondering wordt toegestaan - ook bij deze uitzondering moet men evenwel nog 5/10 halen.

Vroeger werd enkel gekeken naar het eindcijfer (gemiddelde van schoolexamen en centraal examen). Dit is evenwel bijgesteld, omdat leerlingen die hoge cijfers kregen op de schoolexamens zich aldus verschillende respondenten vaak niet meer gemotiveerd voelden om goed te presteren op het centraal examen. Om sterke leerlingen nog meer te belonen, krijgen leerlingen met een gemiddeld centraal examencijfer van 8/10 de vermelding 'cum laude' op hun diploma. Het primaire doel van het centraal examen is duidelijk verantwoordingsgericht op het microniveau.

Doel op mesoniveau (scholen)

Daarnaast is er evenwel ook op het mesoniveau een verantwoordingsgericht perspectief aan de resultaten op het centraal examen verbonden. De onderwijsinspectie zal immers ook in het voortgezet onderwijs de resultaten gebruiken om een *"de prestatie van de school vast te stellen"* (interview, respondent NL3) in een 'opbrengstenoordeel' in het kader van een risicoanalyse, ook al was dit niet het primaire doel van de centrale examens.

Dit opbrengstenoordeel wordt per school bepaald op het niveau van de onderwijstype (vmbo, met onderscheid tussen de vier verschillende 'leerwegen', havo en vwo). Daarbij wordt uitgegaan van vier indicatoren, met name: doorstroom van leerlingen in de onderbouw; doorstroom van leerlingen in de bovenbouw; gemiddeld cijfer op het eindexamen; en het driejaarsgemiddelde schoolexamen - centraal examen. Het gemiddeld cijfer op het eindexamen wordt berekend op basis van alle leerlingen van de laatste drie cohorten van alle vakken, met uitzondering van de beroepsgerichte vakken (op voorwaarde dat er minimaal 30 leerlingen in de school zijn met een examencijfer in het betreffende onderwijstype). Bij het bepalen van de norm die scholen moeten halen om een 'voldoende' gemiddeld cijfer op het eindexamen te behalen, wordt rekening gehouden met het percentage leerlingen met

lwoo-indicatie⁵⁴; het percentage leerlingen dat in een 'armoedeprobleem-cumulatie-gebied' woont; en het percentage leerlingen dat van een andere school komt en instroomt in leerjaar 3.

De vierde indicator houdt in dat per onderwijstype in de school jaarlijks de gemiddelde discrepantie wordt berekend tussen het cijfer van het schoolexamen en het cijfer van het centraal examen over de afgelopen drie jaren van alle leerlingen van alle vakken die met een centraal examen worden afgesloten. De scholen krijgen een beoordeling op een driepuntsschaal: 'zeer groot verschil', 'groot verschil' en 'gering verschil'. Deze indicator is in 2008 geïntroduceerd als antwoord op de bezorgdheid dat bepaalde scholen aan leerlingen hogere cijfers toekenden, waardoor er toch een onevenwicht bestond in de beoordeling van leerlingen in verschillende scholen.

"Wij constateerden dat die discrepantie steeds groter werd, maar aan één kant groter, dus die punten voor dat schoolexamen werden steeds hoger, of toch in verhouding met dat centraal examen, dat verschil werd op een gegeven moment vrij groot. Op een gegeven moment dreigt dat systeem natuurlijk onhoudbaar te worden, als dat verschil te groot wordt. En toen is er wel ingegrepen, dat was in 2008. Ten eerste zijn wij [de onderwijsinspectie] gaan controleren, dat verschil zijn wij ook gaan publiceren, op schoolniveau, en we kennen daar ook een oordeel aan toe. Als het schoolexamen te hoog wordt, dan zeggen we dat dat te groot is." (interview, respondent NL3)

Aard van de toetsen

Het Ministerie van Onderwijs, Cultuur en Wetenschappen bepaalt wat leerlingen moeten kunnen en kennen. In het examenprogramma is een omschrijving opgenomen van de examenstof voor ieder eindexamenvak, welk deel van de examenstof centraal zal worden geëxamineerd en over welke examenstof het schoolexamen zich uitstrekt. De examens zijn gericht op kennis, inzicht en vaardigheden.

Centrale examens bestaan zowel uit meerkeuzevragen als open vragen. Er zijn verschillende examens voor de verschillende onderwijstypes, en dat in verschillende leerdomeinen. Leerlingen kunnen in 7 of 8 vakken een centraal examen afleggen.

De meeste examens zijn klassieke pen-en-papier toetsen. Voor bepaalde algemene vakken in vmbo werden vanaf 2015 digitale examens uitgewerkt.

Ontwikkeling

Er zijn (in bijna alle gevallen) drie examens per vakgebied voor ieder niveau: een standaard examen, een uitgesteld examen (voor wie niet aanwezig kan zijn op het standaard examen) en een 'nood-examen'. In 2015 werden zo 442 pen-en-papier eindexamens ontwikkeld, en 22 digitale examens voor vmbo. In totaal worden daartoe jaarlijks ongeveer 14.000 examenopgaven ontwikkeld.

⁵⁴ Lwoo is een acroniem voor leerwegondersteunend onderwijs. Leerlingen met lwoo-indicatie volgen meestal regulier vmbo-onderwijs, maar hebben daarbij extra begeleiding nodig als gevolg van een leerachterstand, of op basis van IQ en/of een problematiek van socio-emotionele aard. Scholen krijgen hiervoor extra financiering.

Toetsen moeten omwille van het high-stakes karakter elk jaar volledig worden vernieuwd. Het CvTE besteedt de toetsontwikkeling uit aan de Stichting CITO. Elke toets wordt voorbereid door experts in toetsontwikkeling samen met drie of vier vakleraren. De toetsen worden daarna ter goedkeuring voorgelegd aan het CvTE, waar ze worden beoordeeld door een commissie. Ook in deze commissie zetelen vakleraren. Het proces om een toets te ontwikkelen beslaat twee à drie jaar.

Afname en correctie

Alle leerlingen in het laatste jaar van het voortgezet onderwijs nemen de eindexamens af. De Dienst Uitvoering Onderwijs (DUO) wordt door het CvTE gecontracteerd om de examens tijdig bij elk van de 1200 scholen af te leveren. De afname van de eindexamens gebeurt aan het eind van het schooljaar (meestal in mei). Daarnaast vinden er ook examens plaats die starten begin april en doorlopen tot half juni. Hierbij gaat het over praktijkexamens bij kunstvakken, de praktijkexamens voor beroepsgericht onderwijs en de digitale examens bij vmbo basisberoepsgerichte leerweg en vmbo kaderberoepsgerichte leerweg.

De schooldirectie is verantwoordelijk voor de afname. De onderwijsinspectie kan controleren of dit volgens de strikte procedures verloopt. Wanneer er onregelmatigheden plaatsvinden, beschikt de inspectie over het mandaat om - in het uiterste geval - een examen op een school ongeldig te verklaren.

De vakleraren corrigeren het examen van de eigen leerlingen. Daarna wordt elk van de examens ook door een leraar uit een andere school gecorrigeerd. Leraren die het oneens zijn met een vraag of antwoord kunnen contact opnemen met de Examenlijn van het CvTE. Alle meldingen worden door CvTE zorgvuldig op hun juistheid beoordeeld en indien nodig zal het CvTE de examinatoren nog nadere aanwijzingen voor het correctiewerk geven. De communicatie tussen scholen, CITO en externe leraar-beoordelaar van school- en examengegevens en inzake scores van leerlingen gebeurt via het specifiek daartoe ontwikkeld programma WOLF (Windows Optisch Leesbaar Formulier).

Nadat de schriftelijke examens zijn afgenomen, worden de opgaven en het correctievoorschrift op internet gepubliceerd. De opgaven van de digitale examens blijven geheim.

De normering van de toetsen gebeurt nadien door het CvTE. De normering zorgt ervoor dat de prestatie-eis van leerlingen elk jaar gelijk is, met name dat alle leerlingen die hetzelfde examen afleggen voor dezelfde score hetzelfde cijfer krijgen, en dat de prestatie om een voldoende te halen voor het centraal examen van jaar tot jaar gelijk is. De normering gebeurt door een referentiegroep die zowel het huidige examen als het examen van één van de voorgaande jaren maken.

Kwaliteitszorg

Bij de ontwikkeling wordt een beperkte groep betrokken, maar zijn er wel veel stadia verwerkt die de betrouwbaarheid en de validiteit van de toets moeten garanderen. De validiteit blijft evenwel beperkt, aangezien niet alle competenties worden getoetst: sommige competenties (die worden geacht moeilijker door een schriftelijke toets gemeten te worden) worden immers expliciet overgelaten aan het schoolexamen. De betrouwbaarheid van de toetsen blijkt wel erg hoog, al is ze voor het vak

Nederlands moeilijker te bereiken dan voor andere vakken. Het CvTE evalueert jaarlijks de examens, aan de hand van de melding op de examenlijn van het CvTE, en de eigen evaluatie van CITO. CITO evalueert aan de hand van een aantal vragen die leraren worden voorgelegd. De input wordt gebruikt om de examens te verbeteren. Ook maakt het CvTE gebruik van de evaluatie van de vakverenigingen. Op deze wijze ontstaat er een continu proces van uitwisseling met leraren en andere experts in het onderwijs.

Bij elk examen krijgen de betrokken leraren de kans om het examen te evalueren. Naast een algemene beoordeling van het examen, wordt gepeild naar het oordeel van leraren over de moeilijkheidsgraad, de lengte en de inhoudelijke aansluiting van het examen.

Rapportering

Alle leerlingen die hebben deelgenomen aan het centrale eindexamen ontvangen een cijferlijst met de eigen behaalde resultaten op zowel de centrale examens als op het schoolexamen. Scholen kunnen op de website van CITO geaggregeerde cijfers raadplegen voor het huidige schooljaar alsook van voorgaande schooljaren.

Schooltype/Leerweg	VWO	VWO	VWO	VWO	VWO
Soort examen	CSE	CSE	CSE	CSE	CSE
Vaknaam	Nederlands	Frans	Duits	Engels	Spaans
Totaal aantal kandidaten	38.496	15.049	20.707	38.625	1.701
Steekproefgrootte	30.957	13.102	15.917	32.751	1.539
Aantal vragen	36	44	43	40	44
P ¹ -waarde totale steekproef	0,56	0,65	0,64	0,72	0,73
P ¹ -waarde, alleen jongens	0,55	0,65	0,64	0,74	0,71
P ¹ -waarde, alleen meisjes	0,56	0,65	0,64	0,70	0,73
Maximumscore	64	50	49	48	51
Gemiddelde score	35,6	32,4	31,4	34,4	37,0
Normeringsterm	1,0	0,5	0,5	0,5	0,0
Gemiddeld cijfer	6,0	6,3	6,3	7,0	6,6
Percentage onvoldoendes (< 5,5)	24,4	24,9	18,8	11,2	16,6
Standaardafwijking	5,9	6,8	5,4	6,2	6,3
Betrouwbaarheid (asympt. GLB)	0,58	0,81	0,72	0,79	0,81
Standaardmeetfout	3,8	3,0	2,9	2,8	2,8

Figuur 19. Globale rapportering van taalvakken in het examenverslag

Ten slotte produceert de Stichting CITO ook jaarlijks het 'examenverslag'. Hierin geeft men op macroniveau inzage in het aantal afgelegde examens, de constructie van de examens, de verzameling van de afnamegegevens en de normering. Daarnaast is er een rapportering van het globale gemiddelde cijfer en aandeel van de leerlingen dat een onvoldoende haalt, per onderwijstype (vwo, havo, vmbo). Een voorbeeld hiervan is opgenomen in Figuur 19, waarbij de resultaten van vwo-studenten voor verschillende taalvakken zijn weergegeven. In deze figuur ziet men het aantal deelnemende leerlingen ('kandidaten') en het aantal gestelde vragen, p-waardes (een getal tussen 0 en 1 dat de gemiddelde moeilijkheidsgraad van de items weergeeft: hoe hoger de p-waarde, hoe makkelijker de items), maximaal te behalen score en de gemiddelde score van alle leerlingen, de normeringsterm (een ijkvariabele die na de afname van het examen wordt vastgesteld opdat het examen vergelijkbare eisen stelt als in voorgaande jaren), het gemiddeld examencijfer, het aandeel leerlingen dat 'zakt' voor het examen (leerlingen die niet de score 5,5/10 behalen), de

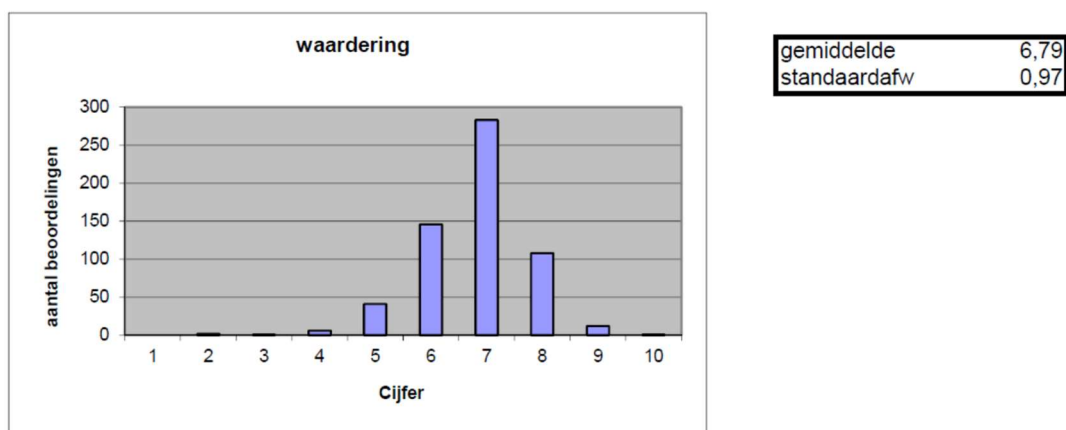
standaardafwijking, de betrouwbaarheid van de toets (door middel van de asymptotische GLB (*greatest lower bound*), met een score tussen 0 en 1, waarbij een hogere score duidt op een hogere betrouwbaarheid) en de standaardmeetfout die uit de betrouwbaarheid en de standaardafwijking kan worden afgeleid.

In een aparte applicatie van het examenverslag is weergegeven welke ruwe score overeenkomt met welk cijfer van 1 tot 10 (zie Figuur 20 voor het examen Duits in vwo). Hierin kan men tevens de verdeling van de leerlingen over alle scores bekijken.

Score	Cijfer	Percentiel-score	Z-score	Aantal per score
0	1,0	0,0	-5,8	0
1	1,1	0,0	-5,6	0
2	1,2	0,0	-5,4	0
3	1,3	0,0	-5,2	0
4	1,4	0,0	-5,0	0
5	1,5	0,0	-4,9	0
6	1,6	0,0	-4,7	0
7	1,8	0,0	-4,5	0
8	2,0	0,0	-4,3	0
9	2,2	0,0	-4,1	0
10	2,3	0,0	-3,9	0
11	2,5	0,0	-3,8	0
12	2,7	0,0	-3,6	1
13	2,9	0,0	-3,4	4
14	3,1	0,0	-3,2	6
15	3,3	0,0	-3,0	9
16	3,4	0,0	-2,8	19
17	3,6	0,0	-2,6	37
18	3,8	1,0	-2,5	64
19	4,0	1,0	-2,3	88
20	4,2	2,0	-2,1	135
21	4,4	3,0	-1,9	194
22	4,5	4,0	-1,7	249
23	4,7	6,0	-1,5	379
24	4,9	9,0	-1,4	479
25	5,1	12,0	-1,2	617
26	5,3	17,0	-1,0	706
27	5,5	21,0	-0,8	839
28	5,6	27,0	-0,6	990
29	5,8	33,0	-0,4	1029
30	6,0	40,0	-0,3	1137
31	6,2	48,0	-0,1	1177
32	6,4	55,0	0,1	1091
33	6,6	61,0	0,3	1077
34	6,7	68,0	0,5	1037
35	6,9	74,0	0,7	868
36	7,1	80,0	0,9	893
37	7,3	85,0	1,0	690
38	7,5	89,0	1,2	568
39	7,7	92,0	1,4	434
40	7,8	94,0	1,6	355
41	8,0	96,0	1,8	243
42	8,2	98,0	2,0	217
43	8,4	99,0	2,1	121
44	8,6	99,0	2,3	99
45	8,8	100,0	2,5	36
46	8,9	100,0	2,7	20
47	9,3	100,0	2,9	6
48	9,6	100,0	3,1	1
49	10,0	100,0	3,3	2

Figuur 20. Omzetting van scores naar cijfers voor examen Duits in vwo

Ten slotte rapporteert het examenverslag ook over de waardering van de kwaliteit van het centrale examen door de leraren van het betreffende vak. Een voorbeeld van deze algemene waardering (voor het examen Duits in vwo) is weergegeven in Figuur 21.



Figuur 21. Algemene waardering van het centrale examen Duits in vwo, door betrokken leraren

3.2.3 Impact

Op basis van de interviews lijkt het dat de eindtoetsen in zowel het primair als in het voortgezet onderwijs in Nederland breed aanvaard zijn als een maat om alle leerlingen op een gelijke manier te beoordelen (hoewel er nog kritiek is, zie sectie 3.2.4). Bovendien zijn verschillende stakeholders er tevens mee akkoord dat indien een inspectie gedifferentieerd wil werken op basis van een risicoanalyse, output-gegevens door middel van gestandaardiseerde toetsen hier zeker een bijdrage moeten hebben. Het verantwoordingsgericht gebruik van de resultaten op mesoniveau, lijkt op die manier ook breed aanvaard in het Nederlandse onderwijs.

In de bevroagde scholen laat men zich positief uit over de beleidswijziging, waardoor de eindtoetsen primair onderwijs later worden afgenomen (na bepaling van het schooladvies door de lagere school). Het schooladvies voor het onderwijstype voortgezet onderwijs voor elke leerling wordt nu immers sterker gedragen door een combinatie van eigen school-evaluaties en van evaluaties in het leerlingvolgsysteem doorheen groep 6, 7 en 8.

Of de toetsen rechtstreeks leiden tot betere leerprestaties, is evenwel onduidelijk. In een internationaal vergelijkend onderzoek⁵⁵ blijkt dat Nederlandse leerlingen in vergelijking met andere landen een hoge groepsdruk ('peer pressure') ervaren om goede prestaties te leveren op centrale examens. Leerlingen, vooral in havo en vwo, zetten elkaar onder druk om goede prestaties te leveren, zodat de kans op zittenblijven of heroriëntering kleiner wordt. Beide situaties worden immers door leerlingen en hun ouders als een publieke nederlaag gezien. Om deel te blijven uitmaken van een bepaalde vriendengroep op school, zouden de Nederlandse studenten harder werken dan studenten in de Verenigde Staten. In een andere comparatieve analyse⁵⁶ vonden onderzoekers evenwel dat de relatief sterke leerprestaties van leerlingen in Nederland niet statistisch konden worden verklaard door een effect van het eindexamen in het voortgezet onderwijs.

⁵⁵ Zie Bishop (1995)

⁵⁶ Zie Prokic-Breuer en Dronkers (2012)

3.2.4 Kritiek en neveneffecten

De openbaarheid van gegevens maakt dat kranten de resultaten publiceren en rankings van de prestaties van scholen opmaken. Dit leidt tot een hogere druk op scholen. De rankings maken bovendien oneigenlijk gebruik van de analyse van de onderwijsinspectie, omdat deze enkel gericht is op het identificeren van de scholen die mogelijk een risico lopen. De inschatting van de inspectie heeft niet het doel om een differentiërend oordeel uit te spreken over alle scholen die goed scoren (en dus geen risico lopen):

“Er komt veel meer nadruk te liggen op die toetsscores. Wat de inspectie doet, wij kijken alleen of een school boven of onder die ondergrens valt. De inspectie is niet geïnteresseerd of een school nu 540,1 of 539,6 scoort, maar je ziet partijen die ranglijstjes maken van die scores van scholen dus het moment dat je dat echt op hoog niveau openbaar maakt, dan krijg je ook weer weerstand van scholen tegen de toetsen, dat een school en een leerling ook meer is dan die toets. Maar voor ons, als een toetsscore over die norm valt, is dat niet meer belangrijk. Dan wordt er meer gekeken naar andere dingen.” (interview, respondent NL4)

Hoewel er vanuit het onderwijsveld veel kritiek is op de rankings van scholen, zijn ze in de huidige context onvermijdelijk, volgens verschillende respondenten:

“Ze kijken meer naar die Trouw en die Volkskrant die op onze data wat doen, dan naar ons hoor. Dat oordeel van ons staat op internet en daar ‘taalt’ men niet veel over. Maar als de pers bezig is, dan wordt daar veel meer stampij om gemaakt. Ik zie het wel als onvermijdelijk, die openbaarheid, we gaan niet meer terug daarin. In die zin... Het is natuurlijk als die gegevens eenmaal beschikbaar zijn en mensen werken daarmee, dan zij er altijd wel partijen die dat opeisen, je kunt dat ook niet voortdurend afschermen, je kan ranglijsten ook niet voorkomen, die kun je niet tegenhouden.” (interview, respondent NL3)

Het high-stakes karakter van de centrale examens maakt dat er zowel vanuit leerlingen als vanuit leraren regelmatig onwenselijk strategisch gedrag opduikt⁵⁷:

“Ieder jaar maken we wel kleine fraudes mee, en die komen soms ook wel bij ons terecht [bij de onderwijsinspectie], dat een docent vindt dat de leerlingen het niet goed genoeg hebben gedaan en dan zelf maar wat gaat verbeteren.” (interview, respondent NL3)

In het voortgezet onderwijs luidt daarnaast nog een kritiek dat scholen meer en meer geneigd zijn om het schoolexamen enkel te zien als een opstapje naar het centraal examen, waardoor beide evaluaties meer en meer op elkaar gelijken. Hierdoor gaat de eigenheid van het schoolexamen, dat een rijkere set aan competenties zou moeten kunnen beoordelen, verloren: zo worden in het vak Nederlands

⁵⁷ Het meest tot de verbeelding sprekende voorbeeld was de ‘examenroof’ in een Rotterdams college: <http://www.nu.nl/eindexamens/3499893/chronologie-examenfraude-ibn-ghaldoun-in-rotterdam.html>

luister-, spreek- en schrijfvaardigheid niet opgenomen in het centraal examen, en zou het schoolexamen deze lacune moeten opvangen.

Bovendien is er nog steeds een groep van leraren en ouders die tegen elke vorm van gestandaardiseerde toetsen is. Volgens de inschatting van de respondenten tijdens deze case studie is deze groep evenwel vrij klein.

3.3. Leerlingvolgsysteem CITO en pilootproject 'Leerwinstmonitoring'⁵⁸

Sinds schooljaar 2014-2015 zijn basisscholen verplicht te werken met een leerlingvolgsysteem (LVS). Een leerlingvolgsysteem bestaat uit toetsen die de hele basisschoolperiode lang de kennis en de vaardigheden van de leerlingen meten. Veel scholen kiezen daarbij voor het LVS van CITO dat al langer werd gebruikt⁵⁹.

3.3.1 Leerlingvolgsysteem CITO

Het Leerlingvolgsysteem van CITO omvat toetsen voor groep 3 tot en met groep 8 voor de leerdomeinen rekenen-wiskunde, spelling, technisch lezen, begrijpend lezen en woordenschat. Voor groep 6 tot en met groep 8 zijn er ook toetsen 'taalverzorging' met de onderdelen grammatica, interpunctie en spelling. Voor elk leerjaar zijn twee toetsen beschikbaar, die in principe in het midden en aan het einde van het schooljaar moeten worden afgenomen.

Doel: Ontwikkelingsgericht op micro- en mesoniveau

Het leerlingvolgsysteem ontwikkeld door CITO heeft een uitgesproken ontwikkelingsgericht doel: men wil informatie aanleveren die bijdraagt tot een optimale ontwikkeling van de leerlingen (microniveau) en de school (mesoniveau). De toetsen hebben onder andere als doel om een precieze indicatie te geven van het vaardigheidsniveau van de leerlingen op verschillende tijdstippen doorheen hun lagere schoolloopbaan, en om deze vaardigheidsniveaus te vergelijken met dat van leerlingen op hetzelfde moment in hun schoolloopbaan. De rationale van de toetsen van CITO-LVS en hoe ze geacht worden bij te dragen tot ontwikkeling, wordt uitgetekend in Figuur 22.

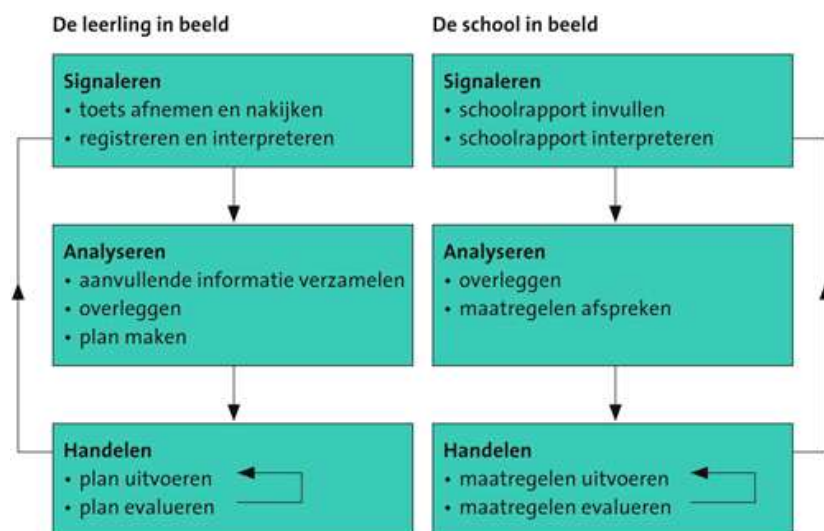
Verantwoordingsgericht doel op mesoniveau

Hoewel de toetsen van CITO-LVS daar in oorsprong niet voor zijn bedoeld, worden de resultaten op schoolniveau ook geplaatst in een verantwoordingsgericht perspectief. De Onderwijsinspectie maakt jaarlijks van elke basisschool een 'risicoanalyse'. Er worden zeven 'opbrengstindicatoren' onderscheiden. Eén van deze zeven indicatoren is rechtstreeks gelinkt aan de studieresultaten van leerlingen doorheen de schoolloopbaan, met name: de resultaten van de leerlingen voor Nederlandse taal en voor rekenen en wiskunde tijdens de schoolperiode liggen ten minste op het niveau dat op grond van de kenmerken van de leerlingenpopulatie mag worden verwacht. Deze tweede indicator

⁵⁸ Gebaseerd op: Cito (2013b, z.j.-a); Janssens et al. (2014) + informatie tijdens gesprekken in case studie.

⁵⁹ De eerste versie van deze CITO-LVS is ook in Vlaanderen beschikbaar. Er is evenwel geen aparte normering voor Vlaanderen beschikbaar.

heeft betrekking op de resultaten van leerlingen doorheen hun schoolloopbaan en wordt bijgevolg geïnformeerd door de resultaten uit het leerlingvolgsysteem van de school. Sinds 2014 moeten alle primaire scholen daarom van een leerlingvolgsysteem gebruik maken. Dit kan het leerlingvolgsysteem van CITO zijn, of een ander leerlingvolgsysteem dat voldoet aan de criteria van Commissie Test Aangelegenheden Nederland (COTAN) inzake validiteit, betrouwbaarheid en normering.



Figuur 22. Ontwikkelingsgericht werken met de CITO-LVS

Deze indicator wordt door de inspectie beoordeeld aan de hand van leerresultaten op de volgende toetsen: Technisch Lezen in groep 3 en groep 4; begrijpend lezen in groep 6; en rekenen en wiskunde in groep 4 en groep 6. Uitgangspunt is dat de beoordeling is gebaseerd op de leerresultaten van alle leerlingen in hetzelfde leerjaar. Daarnaast moeten alle te waarderen toetsen in hetzelfde schooljaar zijn afgenomen. Net als bij de vorige indicator wordt een uitspraak gedaan over de school gebaseerd op een vergelijking van de gemiddelde vaardigheidsscore in de school met de norm voor scholen met een gelijk percentage 'gewichtleerlingen'⁶⁰. Is de gemiddelde score lager dan de norm, dan is het oordeel op de betreffende toets onvoldoende. Ligt de gemiddelde score op of boven de norm, dan is het oordeel voldoende. Deze normen zijn uitgewerkt voor scholen die het CITO-LVS gebruiken en voor scholen die de Schoolvaardigheidstoetsen (SVT) toetsen van Boom Test Uitgevers hanteren.

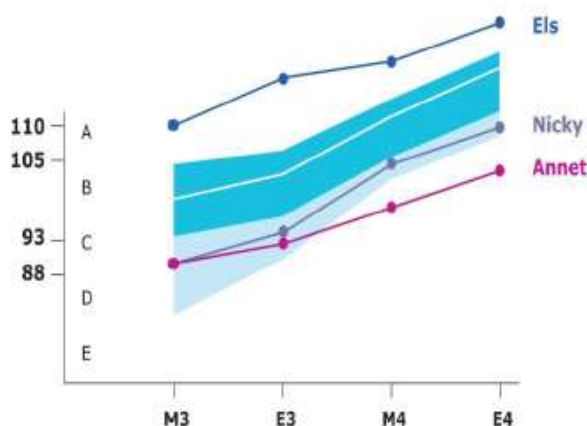
Afname en correctie

In principe nemen leerlingen van scholen die deelnamen aan de CITO-LVS, twee maal per jaar een toets af in elk van de voorziene leerdomeinen. Er zijn daartoe twee aparte toetsen voor elk leerjaar gemaakt, zoals bijvoorbeeld 'midden groep 3 (M3)' en 'eind groep 3 (E3)'. Individuele leerlingen kunnen ook een toets aangeboden krijgen op een hoger of lager niveau dan hun klasgenoten.

⁶⁰ Enkel bij technisch lezen wordt geen rekening gehouden met leerlingkenmerken, aangezien de invloed van deze kenmerken op het technisch lezen beperkt zou zijn (Inspectie van het Onderwijs, 2014a)

Aangezien alle toetsen binnen een leerdomein op dezelfde vaardigheidsschaal zijn ingeschaald (zie onder), bekomt men ook voor deze leerlingen een score die men kan vergelijken met de score van de andere leerlingen die de 'reguliere' toets hebben afgelegd.

Leraren moeten de toetsen zelf corrigeren en de toetsresultaten ingeven in een registratiesysteem. De toetsscore wordt daarna omgezet naar een vaardigheidsscore. Deze vaardigheidsscore duidt op een positie op een onderliggende vaardigheidsschaal. Deze vaardigheidsschaal maakt het mogelijk om - binnen een zelfde leerdomein - de resultaten van een leerling op verschillende toetsmomenten met elkaar te vergelijken en om de resultaten van leerlingen in dezelfde groep die verschillende toetsen hebben gemaakt te vergelijken. De leerwinst van leerlingen kan zo makkelijk in kaart worden gebracht. Een voorbeeld hiervan is Figuur 23, waarbij de toetsen midden groep 3 (M3) - eind groep 3 (E3) - midden groep 4 (M4) - eind groep 4 (E4) worden geprojecteerd, volgens de vaardigheidsscores van drie verschillende leerlingen.



Figuur 23. Leerwinst uit LVS CITO (Janssens et al., 2014)

Daarnaast zijn ook vaardigheidsniveaus opgenomen. De vaardigheidsniveaus zijn vooral in gebruik genomen om de data makkelijker betekenis te geven, omdat de vaardigheidsscores moeilijk te interpreteren waren:

“We hebben toetsen gemaakt voor elk half jaar, eentje voor M3, eentje voor E3, eentje voor M4 enzovoort, dan hebben we eigenlijk in kaart gebracht hoe de leerlingen daarop scoren, we geven ze een vaardigheidsscore. Alleen het vervelende aan die vaardigheidsscore is dat niemand weet hoe je dat moet interpreteren dat je kunt zeggen “ah, je hebt vaardigheidsscore 515 - ja, hartstikke leuk, maar hoe heb ik het dan gedaan?”(interview, respondent NL5)

Er zijn twee systemen van vaardigheidsniveaus in gebruik waarbij sommige scholen van het ene, en andere scholen van het andere systeem gebruik maken. Bij het ene systeem is niveau A de groep van

25% hoogst scorende leerlingen, niveau B percentiel 26 tot en met 50, en niveau C percentiel 51 tot en met 75. De 25% zwakst scorende leerlingen worden verdeeld in niveau D (15%) en niveau E (de laagste 10%). Het nadeel van dit systeem is dat er geen 'gemiddeld niveau' is: het midden ligt tussen niveau B en C. In een alternatief systeem (van I tot V) zijn er eveneens vijf groepen, maar daar zijn ze alle vijf even groot zijn: niveau I duidt daarbij op een vaardigheid behorende tot de 20% hoogst scorende leerlingen, en niveau V op een vaardigheid behorende tot de 20% laagst scorende leerlingen. Figuur 24 geeft een visuele weergave van de beide systemen van vaardigheidsniveaus.

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

Figuur 24. Verdeling van leerlingen over vaardigheidsniveaus (Cito, 2013b)

Een leraar kan aan de hand van de vaardigheidsniveaus makkelijk opvolgen hoe leerlingen evolueren (met name of ze binnen het eigen vaardigheidsniveau blijven, dan wel stijgen of dalen).

Rapportering

De rapportering laat de directies en leraren toe om op verschillende niveaus informatie op te vragen: op het niveau van individuele leerlingen, van klasgroepen, en van de volledige school. De rapportering is interactief van aard: zo kunnen leraren bijvoorbeeld groepen van leerlingen selecteren, bepaalde toetsen uitkiezen om te vergelijken, ...

Impact

De directies van beide bezochte scholen zijn van mening dat het CITO-LVS een enorm positieve impact heeft op de kwaliteit van het onderwijs in hun respectievelijke scholen. Het gebruik van CITO-LVS heeft ervoor gezorgd dat de scholen meer 'opbrengstgericht' (gericht op het behalen van optimale leerresultaten voor elke leerling) en meer 'handelingsgericht' (gericht op actie geïnformeerd door leerresultaten) zijn gaan werken.

In de ene bezochte school worden de resultaten van alle leerlingen na elke toetsafname besproken in een dialoog tussen schooldirectie en de klasleraar. In de andere bezochte school gebeurt hetzelfde,

maar neemt de interne begeleider de positie van de schooldirectie in. Op basis daarvan worden klassikale, of individuele acties vooropgezet.

“Dat kan betekenen: leertijduitbreiding; dus we gaan in plaats van één les rekenen per dag twee lessen geven, omdat we gewoon die eindnorm willen halen. En dat betekent ook dat sommige kinderen meer hulp krijgen.” (interview, respondent NL1)

In één van de bevraagde scholen worden de gegevens van het CITO-LVS van elke leerling beschouwd als de basis van een ‘overgangsgesprek’, waarin leraren van twee opeenvolgende leerjaren de voortgang van de leerlingen bespreken. In deze school worden de resultaten op CITO-LVS ook tijdens een overlegmoment met leraren besproken.

Verschillende scholen geven aan dat men door het gebruik van het leerlingvolgsysteem sterker gericht is op het aandeel van de eigen school in het leren van leerlingen:

“Wat het grootste winstpunt is, is dat we niet meer zeggen: ‘dat is een zwakke groep’, maar ‘wat kunnen we doen als leraren om het niveau omhoog te krijgen?’” (interview, respondent NL1)

Daarnaast worden leraren door de resultaten in het CITO-LVS ook aangemoedigd om niet enkel te focussen op leerlingen die het relatief moeilijk hebben, of op leerlingen die net wel of net niet een bepaalde grensnorm (bijvoorbeeld voor advies vmbo - havo) zouden bereiken. Doordat leraren halfjaarlijks ook geconfronteerd worden met de voortgang van het leren van de meest sterke leerlingen, wordt in de bevraagde scholen ook extra gefocust op het leerproces van deze leerlingen:

“Vroeger was het van: ‘nou, die goeie gaan wel goed’, maar nu van: ‘ja, die is gezakt van een A-plus naar een A, maar wat gaan wij er nu aan doen om te zorgen dat-ie weer terugkomt op de A-plus?’ (...) Dan zorgen wij dat-ie naast de methode ook nog iets extra’s krijgt.” (interview, respondent NL1)

Verder worden de resultaten van het CITO-LVS ook gebruikt in communicatie met ouders en met leerlingen. Uit de ervaring van de bevraagde schooldirecties blijkt dat ouders hierdoor sterker kunnen worden betrokken bij het leerproces van hun kinderen, en dat leerlingen zelf meer verantwoordelijkheid opnemen voor hun eigen voortgang (bijvoorbeeld door zelf doelen te stellen). Dit laatste heeft een positief effect op de werkhouding van voornamelijk de oudere leerlingen. Op deze manier komt ook het schooladvies dat aan het eind van het primair onderwijs wordt uitgesproken, slechts zelden als een verrassing voor leerlingen en ouders (wat vroeger vaker het geval was en soms tot conflict tussen scholen en ouders heeft geleid). Bovendien wint dit advies aan betrouwbaarheid door de beschikbaarheid van betrouwbare data doorheen de heel schoolloopbaan.

Dat steeds meer in termen van leerwinst, en daardoor minder exclusief in termen van leerprestaties wordt gedacht, wordt ervaren als een ‘meer positief verhaal’, dat motiverend is voor zowel leerlingen als leraren:

"Dit is positief, je kunt hier als leraar aantonen dat binnen het onvoldoende niveau wat jouw leerling heeft gehaald, dat hij nog steeds gegroeid is. (...) We kunnen aan leraren laten zien dat ze ook bij die leerlingen in 'vijf' toch nog het verschil kunnen maken!"
(interview, respondent NL2)

Een ander positief effect is dat de informele professionele dialoog omtrent leerlingen en leerprestaties door de regelmatige opvolging van leerlingen wordt gestimuleerd: *"Er wordt veel meer over onderwijs gepraat in de koffiekamer, dat vind ik mooi!"* (interview, respondent NL1).

Daarnaast zorgt het regelmatige opvolgen van de leerprestaties ook dat de (uitzonderlijke) leraren die hun job lichtzinnig opnemen, aldus de bevroegde directies ofwel zich harder moeten inspannen, ofwel een andere job (buiten het onderwijs) zoeken: *"Die mensen die denken: 'een makkelijke baan in het onderwijs, ik doe maar wat' - dat kan niet meer"* (interview, respondent NL1).

Beïnvloedende factoren

In beide bevroegde scholen worden twee factoren genoemd die maken dat de toetsen uit het CITO-LVS een grote positieve impact hebben op de onderwijskwaliteit: een eerste belangrijke factor is een directie die sterk op het gebruik van deze data inzet en er samen met leraren over in dialoog gaat. Daarnaast benoemt men in beide scholen een 'kritisch kantelmoment', waarbij een externe professional tijdens een studiedag werd ingehuurd om aan te tonen wat de mogelijkheden van deze data zijn, en hoe ze optimaal kunnen worden gebruikt. In beide scholen wordt die studiedag benoemd als de grens tussen 'vroeger' (toen men de resultaten enkel bekeek, maar deze geen impact hadden op het onderwijs) en 'nu' (met de verhoogde focus op opbrengstgericht en handelingsgericht werken). Deze externe professionals maken vaak ook 'overzichtstabellen' op van verwachte leerwinst op per leerlingcategorie (zie onder).

Om deze impact te bereiken, is verder het noodzakelijk dat de school betekenis kan geven aan de vaardigheidsscores. Toch is het voor scholen niet vanzelfsprekend om die betekenis zelf toe te kennen, omdat men op zoek is naar een soort vergelijkingsbasis die een beoordeling mogelijk maakt van een bepaalde groei tussen twee toetsen.

"Maar de vraag die nu vaak gesteld wordt: 'Ja, dit [de vaardigheidsscores] is allemaal net en aardig, maar hoe weet ik nu of ik voldoende gegroeid ben?' Ik kan dus bij deze leerling, hier zien we het aantal punten is omhoog gegaan, moet je daar tevreden mee zijn? Want hier zit ik een beetje onder in de curve, hier een beetje erboven, wat is hier nu 'voldoende' in?" (interview, respondent NL5)

Om die beoordeling mogelijk te maken, gebruiken scholen vaak overzichtstabellen die door private begeleiders of onderwijsconsultants zijn opgesteld om de leerwinst van leerlingen op te volgen. De overzichtstabellen geven een beeld van wat de verwachte leerwinst is voor een leerling die op een voorgaande toets binnen een bepaald vaardigheidsniveau heeft gescoord. Zo kan men bijvoorbeeld zien dat een leerling die acht punten is gestegen - waar de verwachte groei zes punten was - een hogere leerwinst heeft dan op basis van zijn voorgaande score mocht worden verwacht.

Het gebruik van deze overzichtstabellen lijkt wijd verspreid. Door de resultaten van het CITO-LVS, in combinatie met de overzichtstabellen, kan het leerproces van elke leerling beter worden opgevolgd. Volgens verschillende respondenten is dit niet alleen erg motiverend voor zowel zwakke als sterke leerlingen, ook voor leraren is dit een sterke basis om het maximale uit elke leerling naar boven te halen:

“Als die kinderen nu gewoon op de vijf [het laagste vaardigheidsniveau] blijven hangen, dan kunnen we nu toch zien hoe ze op hun eigen niveau zijn gestegen. En dat is echt wel heel stimulerend voor de leraren. Het heeft echt een impuls gegeven aan ons onderwijs.” (interview, respondent NL1)

Toch worden deze overzichtstabellen omwille van enkele beperkingen niet gesteund door CITO zelf.

In de praktijk zijn daar hele tabellen voor ontstaan die niet door CITO gemaakt zijn, waar wij ook onze twijfels bij hebben van ‘ja, is dat nou wel zo handig?’ (interview, respondent NL5)

Zo beargumenteert men bij CITO bijvoorbeeld dat de vaardigheidsniveaus te breed zijn om richtlijnen te geven inzake ‘te verwachten leerwinst’ voor alle leerlingen binnen een bepaald vaardigheidsniveau. De kritiek gaat dus voornamelijk op de relatief eenvoudige manier van berekening. Zo ziet men dat, afhankelijk van het systeem van indeling van vaardigheidsniveaus dat men hanteert (I tot V; of A tot E), leerlingen in het ene systeem wél voldoende leerwinst hebben geboekt, terwijl dezelfde leerwinst bij de indeling van de leerlingen in het andere systeem als onvoldoende wordt beoordeeld.

Aangezien deze overzichtsschema’s bijgevolg niet door CITO worden ondersteund, heeft men zich geëngageerd om mee te werken aan een meer ‘correcte’ opvolging van de leerwinst van leerlingen via het pilootproject ‘Leerwinst en toegevoegde waarde’. Dit project houdt in dat men de verwachte score niet zal bepalen voor ‘groepen’ van leerlingen van een gelijkaardig beginniveau, maar een verwachte score op basis van elk van de vaardigheidsscores zal bepalen (zie sectie 3.3.2). Het betreft dus een veel fijnmaziger maken van de informatie die uit de reeds bestaande tabellen af te leiden is.

Kritiek en neveneffecten

In diverse regio’s maken scholen voor primair en voortgezet onderwijs al jaren afspraken over de gegevens die bij het schooladvies worden betrokken. In sommige steden is afgesproken dat bij het schooladvies toetsen van het leerlingvolgsysteem een belangrijke rol krijgen, soms al vanaf groep 6. Verschillende stakeholders zijn evenwel kritisch ten opzichte van deze verantwoordingsgerichte inzet van de toetsen uit het leerlingvolgsysteem. Een eerste kritiek is dat als resultaten behaald in groep 6 al medebepalend zijn voor het advies (en dus de schoolloopbaan in het voortgezet onderwijs), de selectie van leerlingen al op een zeer vroeg moment begint. Voor bijvoorbeeld leerlingen die met een taalachterstand aan de school beginnen, vormt dit een evident nadeel. Een tweede risico is dat het schooladvies op basis van het leerlingvolgsysteem vaak berust op de vakken rekenen en begrijpend lezen, terwijl het schooladvies geïnformeerd zou moeten zijn door een breder scala van leerdomeinen. Een derde risico is dat zowel leerlingen als leraren en ouders de toetsen van het leerlingvolgsysteem

niet langer als een formatief gegeven maar eerder als summatieve instrumenten zullen beschouwen. Dit kan leiden tot neveneffecten die het ontwikkelingsgerichte doel van de toetsen in de weg staan.

“Een systeem als dit werkt natuurlijk ook strategisch gedrag in de hand, want wat scholen tevoren ook gaan doen, is dat ze gaan oefenen met de kinderen; dat schijnt wel veel te gebeuren, ja.” (interview, respondent NL4)

Vanuit CITO wordt gesteld dat bij de interpretatie van de vaardigheidsniveaus soms foutieve of oneigenlijke vormen van gebruik ontstaan:

“Je ziet ook ontstaan dat iedereen een 1 gaat nastreven. Of we zien doelen ontstaan: 1 is 20%, scholen die zeggen: 25% van onze leerlingen, of misschien wel 30% moet 1-scores gaan behalen. Dus die zien we ook terwijl dat natuurlijk niet reëel is.” (interview, respondent NL5)

Deze foutieve manier van gebruik tracht men tegen te gaan door het engagement in het pilootproject Leerwinst en Toegevoegde Waarde. Ook uit eerder onderzoek⁶¹ was al geconcludeerd dat de interpretatie van resultaten die uit het CITO-LVS kan worden afgeleid, niet vanzelfsprekend is voor de gebruikers ervan.

Daarnaast is het moeilijk voor scholen om de analyses om te zetten naar effectieve verbetertrajecten op micro- of mesoniveau.

“Maar mijn vraag is dan altijd: als je nu [volgens de leerwinstmeting] tien of elf punten moet groeien naar het volgende toetsmoment, weet je wat je dan moet doen in de praktijk? Nee!” (interview, respondent NL5)

Een andere kritiek komt van één van de bevroegde schooldirecties, die zich afvraagt of men door de combinatie van deze halfjaarlijkse gestandaardiseerde toetsen met de eindtoetsen van de lesmethodes, niet teveel in een toetscultuur is terechtgekomen:

“Ik vind dat we toch een beetje slaaf zijn geworden van die toetsen. Mensen zijn veel te veel aan het toetsen terwijl ze die tijd ook hadden gebruiken om instructie te geven aan het kind, of ook met het kind een werkje hadden kunnen maken.” (interview, respondent NL2)

Dat de onderwijsinspectie de toetsresultaten gebruikt binnen de risicoanalyse van scholen, wordt niet door iedereen gesteund. Net als in andere gevallen, wordt het samengaan van ontwikkelingsgerichte en verantwoordingsgerichte doelstellingen in vraag gesteld:

“Wat je nu wel hoort van scholen is “Ja, de buurschool, die oefent voor de toets hoor! en daardoor halen zij hogere scores, gaat de inspectie de normen van de LVS opschroeven. Moeten wij weer hoger scoren! Dus die buurschool oefent, en wij niet, dus komen wij bedrogen uit.” Dat is waar ze mee zitten. En het oorspronkelijke doel,

⁶¹ Zie Ehren & Visscher, 2011, in Janssens et al. (2014)

voor de leraren: hoe ontwikkelt de leerling zich door de jaren heen, dat gaat helemaal verloren. Dat kunnen ze zich niet meer permitteren!” (interview, respondent NL6)

De bevroagde directies staan vrij neutraal ten opzichte van het verantwoordingsgerichte gebruik van de resultaten van het CITO-LVS door de onderwijsinspectie. Men meent dat de neveneffecten in de eigen scholen relatief beperkt zijn.

3.3.2 *Pilootproject Leerwinst en Toegevoegde Waarde*

Door de (toenmalige) minister van Onderwijs, Cultuur en Wetenschap Marja Van Bijsterveldt werd besloten om in het schooljaar 2011-2012 voor het primair onderwijs te starten met een pilootproject inzake Leerwinst en Toegevoegde Waarde in het primair onderwijs. Naast het Ministerie waren de andere betrokken partners de Universiteit Twente, Rijksuniversiteit Groningen, de Inspectie van het Onderwijs, CITO, en de CED-groep. Deze laatste partner werd aangesproken voor praktische ondersteuning aan scholen tijdens het pilootproject en fungeerde daarom als eerste aanspreekpunt voor de deelnemende scholen.

Doelen

Het doel van de pilot was om na te gaan of een valide en betrouwbare maatstaf voor leerwinst of toegevoegde waarde zouden kunnen ontwikkeld worden, die ook praktisch bruikbaar zou zijn voor scholen. Daartoe werd op basis van de CITO-LVS-toetsen die door een steekproef van 45 scholen werden afgenomen, getracht de leerwinst van individuele leerlingen, en de toegevoegde waarde van scholen op een zo precies mogelijke manier in kaart te brengen. Het resultaat van zo een leerwinst- en toegevoegde waarde-berekening zou vanuit zowel een ontwikkelingsperspectief als vanuit een verantwoordingsperspectief op schoolniveau kunnen worden gebruikt.

Enerzijds zouden scholen de informatie kunnen gebruiken om hun globale onderwijs, alsook het onderwijs aan specifieke leerlingen, beter te analyseren dan wanneer enkel op leerprestaties beroep wordt gedaan. Door het beter in beeld brengen van leerwinst en toegevoegde waarde wordt scholen en schoolbesturen een versterkt inzicht geboden in de vorderingen van de leerlingen en de bijdrage die de school hieraan levert. Op die manier zou het opbrengstgericht werken van scholen kunnen worden versterkt. Anderzijds kunnen de maten voor leerwinst en toegevoegde waarde ook een belangrijke input vormen voor het werk van de onderwijsinspectie wanneer zij de kwaliteit van scholen extern beoordeelt.

Daarnaast blijkt uit de gesprekken met verschillende respondenten dat er ook een vorm van verantwoording op het macroniveau aan het pilootproject was verbonden, met name: de onderwijssector in Nederland die wou (of moest) aantonen wat de opbrengst van het huidige onderwijssysteem in termen van leerwinst is.

Resultaten leerwinst

Zoals beschreven in sectie 3.3.1, bekijken scholen leerwinst voornamelijk aan de hand van eigen interpretaties of aan de hand van overzichtstabellen die door externe begeleiders ter beschikking werden gesteld. Deze overzichtstabellen worden evenwel niet door CITO ondersteund.

De leerwinst werd in het pilootproject berekend aan de hand van drie verschillende modellen: het 'groeitempo-model', het 'seizoensgebonden leerwinstmodel' en het 'z-score model'.

Groeitempo-model

Het groeitempo-model is een voorbeeld van een *student growth percentile model*⁶², al zijn de verschillende toetsen in dit voorbeeld wel op een zelfde schaal ingeschaald en betreft het dus een *gain-based model*. Binnen dit model worden de vaardigheidsniveaus die door de CITO-LVS toets worden onderscheiden, gebruikt bij de nulmeting (de eerste toetsafname). Vervolgens wordt de leerwinst van de leerlingen (met name het verschil in vaardigheidsscore tussen eind- en nulmeting) vergeleken met de leerwinst van de leerlingen die in hetzelfde vaardigheidsniveau waren ingedeeld.

Door leerlingen enkel te vergelijken met leerlingen met hetzelfde beginniveau, wordt vermeden dat ongelijke vergelijkingen plaatsvinden: leerlingen die bij aanvang hoog presteren, kunnen bijvoorbeeld minder progressiemarge vertonen. Binnen elk van de begin-vaardigheidsniveaus worden leerlingen in dit groeitempo-model ingedeeld in een deelgroep met bovengemiddelde, gemiddelde, en ondergemiddelde leerwinst. Elk van deze drie deelgroepen bevat één derde van de leerlingen binnen de groep leerlingen met een gelijk begin-vaardigheidsniveau.

Wanneer deze analyse per school werd uitgevoerd, bleek dat er op de meeste scholen variaties waren inzake het aandeel van leerlingen dat zich in de drie categorieën (bovengemiddeld, gemiddeld, ondergemiddeld) bevond. Vaak ging het evenwel om kleine afwijkingen die te wijten zijn aan toevallige schommelingen, maar in sommige gevallen bleken de afwijkingen te groot om aan het toeval toe te schrijven. In beide richtingen werden zulke uitzonderingen vastgesteld: scholen waarin het aantal leerlingen in de hoogste categorie groter dan gemiddeld is, alsook scholen waar het omgekeerde het geval is. In de school waarover gerapporteerd wordt in Figuur 25, is er duidelijk sprake van een relatief gelijke verdeling over de drie categorieën.

Rekenen en Wiskunde Groep 3 - Groep 5		
	Aantal leerlingen	Percentage
bovengemiddeld	8	36,4%
gemiddeld	6	27,3%
ondergemiddeld	8	36,4%
	22	100,0%

Figuur 25. Rapportage aan scholen volgens het groeitempo-model (Janssens et al., 2014)

⁶² Zie Deel 3, sectie 3 voor meer informatie inzake deze term en volgende termen inzake modellen van leerwinst en toegevoegde waarde

Seizoensgebonden leerwinstmodel

Dit model is een uitbreiding van het groeitempo-model en kan dus ook gecatalogiseerd worden als een voorbeeld van een *student growth percentile model*. De extra gegevens die in dit model worden ingevoegd, komen van een herafname van de eindtoetsen (E3, E4, E5, ...) van het CITO-LVS (rekenen, spelling, technisch lezen). Deze herafname vond plaats meteen na de zomervakantie, zodat in kaart werd gebracht wat de evolutie was van de leerlingen tijdens de zomervakantie, een periode zonder input van de school. Het verschil in vaardigheidsscore tussen de tweede en eerste afname van de toets, wordt gezien als de (positieve of negatieve) leerwinst van de leerling in de zomervakantie. Ook hier wordt - net als in het groeitempo-model - in elk van de begin-vaardigheidsniveaus drie deelgroepen onderscheiden: met bovengemiddelde, gemiddelde, en ondergemiddelde leerwinst tijdens de zomervakantie.

Het verschil in vaardigheidsscore tijdens de zomervakantie werd daarna vergeleken met de leerwinst die werd geboekt tijdens het schooljaar. Op die manier levert dit seizoensgebonden leerwinstmodel extra informatie op om de toegevoegde waarde van een school in kaart te brengen: de leerwinst die immers tijdens de zomervakantie werd geboekt, kan niet worden toegeschreven aan de school, maar bijvoorbeeld wel aan natuurlijke groei of aan de inspanningen van ouders. Een gemiddelde negatieve leerwinst van leerlingen tijdens de zomermaanden (leerlingen die bijvoorbeeld aan het begin van groep 4 een lager niveau bereiken dan aan het eind van groep 3) betekent dan weer dat de school een extra inspanning moet doen om dezelfde leerwinst van leerlingen te kunnen bereiken in vergelijking met een school waarin de leerlingen een positieve leerwinst hebben tijdens de zomermaanden. Het seizoensgebonden model houdt hiermee rekening.

z-score model

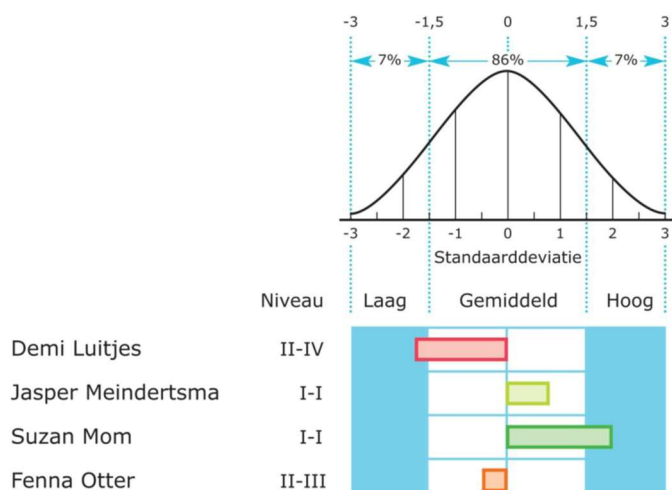
Het z-score model is een voorbeeld van een *gain-score model*. In dit model werd van elke leerling het verschil in vaardigheidsscore tussen het laatste afnamemoment en het eerste afnamemoment berekend.

Tijdens het pilootproject werden in dit model cijfers gebruikt van groep 3. Daarna werd de groei van elke leerling vergeleken met de groei van alle andere leerlingen met exact hetzelfde startniveau (niet in termen van vaardigheidsniveaus, maar de exacte vaardigheidsscore). Bij deze vergelijking werd gebruik gemaakt van een z-score, waarbij gekeken wordt naar zowel het gemiddelde eind-prestatieniveau, als naar de standaarddeviatie van alle leerlingen met een zelfde startniveau. Bij een z-score is de verwachte score altijd gelijk aan 0, en de standaardafwijking gelijk aan 1. Een leerling met een z-score van 0 heeft dus een leerwinst die volgens zijn beginniveau als 'te verwachten' mag worden beschouwd. Een leerling met een negatieve z-score heeft in principe een ondergemiddelde leerwinst, een leerling met een positieve z-score heeft in principe een bovengemiddelde leerwinst.

“Stel dat we zien van leerlingen die een ruwe score van 10 halen op een toets, daar hoort een vaardigheidsscore bij van [bijvoorbeeld] 30. Op toets 2 hebben die 9 goed gehaald, daar hoort een vaardigheidsscore van 40 bij; met een standaarddeviatie van 5. Als we dan de individuele score van een leerling hebben, kunnen we een z-score berekenen van elke leerling” (interview, respondent NL6)

Omdat een z-score van exact 0 slechts zelden zal voorkomen (enkel indien de leerling perfect de verwachte vaardigheidsscore heeft behaald), werd besloten om alle leerlingen waarvan de vaardigheidsscore aan het eind maximaal 1,5 standaarddeviatie afweek van de te verwachten score, te beschouwen als 'gemiddelde leerwinst'. Deze keuze betekent evenwel dat 86% van de leerlingen een 'gemiddelde leerwinst' hebben, terwijl telkens 7% van de leerlingen een bovengemiddelde of ondergemiddelde leerwinst hebben.

In de evaluatie door gebruikers, werd het z-score-model als het meest gebruiksvriendelijk beoordeeld. Aangezien het z-score-model rekening houdt met de precieze vaardigheidsscores, kan het leerwinst meer precies in kaart brengen dan het groeitempo-model (waar enkel met vaardigheidsniveaus werd rekening gehouden). Dit z-score-model heeft geleid tot een rapportage aan scholen over de leerwinst van elke leerling zoals in Figuur 26. In deze figuur wordt weergegeven op welk categorisch niveau een leerling (ten aanzien van de normgroep) presteert ('niveau'), en hoe de leerwinst moet gezien worden (ten aanzien van zelf-criterium). Dit laatste wordt via het rode of groene balkje aangegeven. De lengte van het balkje geeft aan in welke mate de leerwinst kleiner of groter is dan verwacht. In dit voorbeeld heeft één leerling een ondergemiddelde leerwinst, wat met een rode balk wordt aangeduid. Ook één leerling heeft een bovengemiddelde score, wat wordt aangeduid met een donkergroene balk.



Figuur 26. Leerwinst volgens het z-score-model (Janssens et al., 2014)

Resultaten toegevoegde waarde

Verder werden in het pilootproject uiteenlopende berekeningswijzen uitgetoetst om de toegevoegde waarde van scholen in kaart te brengen. In elk van deze modellen wordt de toegevoegde waarde berekend op basis van de leerwinst die door de leerlingen worden geboekt. Uiteindelijk werd over twee modellen uitvoerig gerapporteerd.

Vaardigheidsverschil-model

Bij het zogenaamde 'vaardigheidsverschil-model' is de berekening van de leerwinst van leerlingen gebaseerd op twee toetsen. Het verschil in de vaardigheidsniveaus tussen de twee toetsmomenten, is dan de leerwinst. De leerwinst wordt op schoolniveau gemodelleerd met een statistisch multilevel

regressiemodel, waarbij rekening wordt gehouden met de school en de jaargroepen waarin de leerlingen zijn gegroepeerd. Het eindrapport van het pilootproject geeft aan dat dit belangrijk is, met name om te vermijden dat eventuele verschillen tussen scholen inzake toegevoegde waarde te makkelijk (en ten onrechte) als statistisch significant worden beschouwd.

In het pilootproject werd besloten dat het vaardigheidsverschil-model uitstekend kan worden gebruikt in een verantwoordingsperspectief, om de bijdrage van de school tot de leerwinst van de leerlingen te beoordelen. Binnen een ontwikkelingsperspectief is de meerwaarde van het vaardigheidsverschil-model evenwel beperkt, omdat het maar een maat van toegevoegde waarde biedt op basis van een cohorte leerling die aan het eind van het leertraject in de school staat. In die zin biedt het volgens de projectgroep voor de school weinig houvast om eventueel nog bij te sturen op basis van de beschikbare informatie.

Scholen krijgen informatie over de gemiddelde leerwinst van hun leerlingen in een bepaalde cohorte, alsook van de leerlingen in de andere scholen die deelnamen aan het pilootproject. Door middel van de berekening van een 95% betrouwbaarheidsinterval werd nagegaan of de leerwinst in de school in de verschillende leerdomeinen hoger, gelijk, of lager lag dan in de andere deelnemende scholen (zie Figuur 27 voor een voorbeeldrapportering). In de rapportering wordt onderscheid gemaakt tussen een brutoscore en de netto-score - hierop gaan we in één van de volgende secties dieper in.

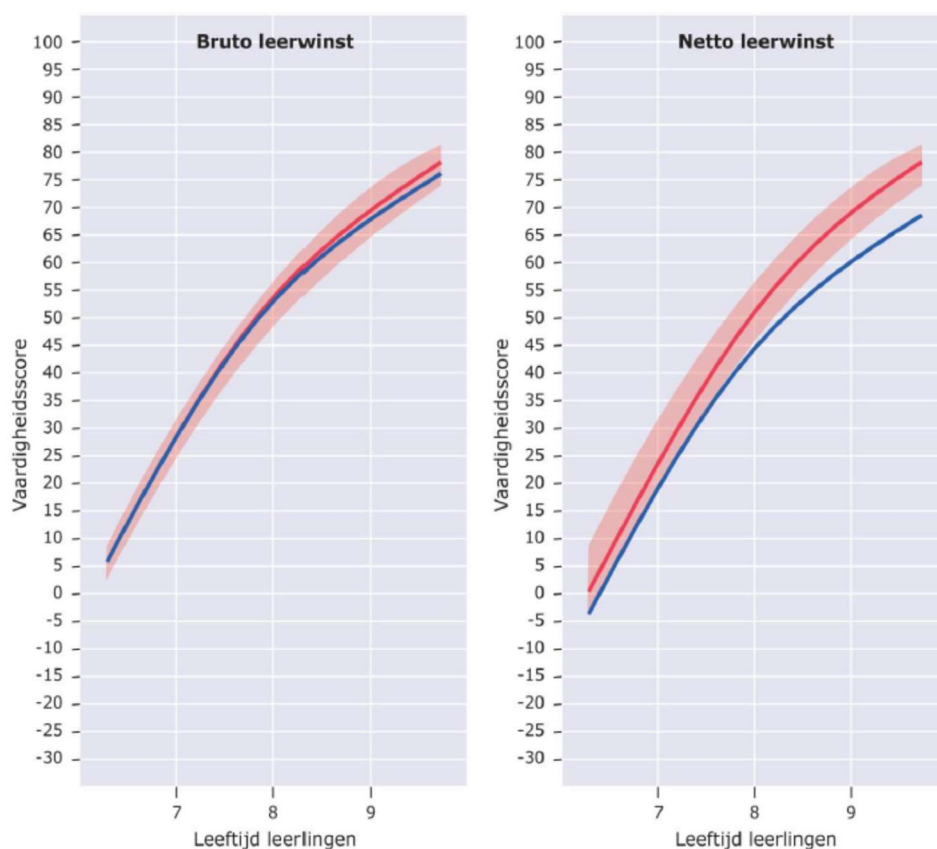
	Pilot-gemiddelde	Uw school (BI)	Afwijkend?
Rekenen-wiskunde: aantal beschikbare leerlingen: 42			
Bruto leerwinst	46,7	48,4 [44,6; 52,3]	gelijk
Netto leerwinst	45,0	47,3 [43,5; 51,2]	gelijk
Spelling: aantal beschikbare leerlingen: 40			
Bruto leerwinst	22,4	25,0 [22,8; 27,2]	hoger
Netto leerwinst	19,9	21,7 [19,5; 23,9]	gelijk
Technisch lezen (DMT): aantal beschikbare leerlingen: 42			
Bruto leerwinst	48,5	56,7 [51,7; 61,8]	hoger
Netto leerwinst	49,3	57,2 [51,8; 62,6]	hoger
Woordenschat: aantal beschikbare leerlingen: 42			
Bruto leerwinst	35,9	39,2 [33,2; 45,2]	gelijk
Netto leerwinst	28,2	33,5 [27,5; 39,5]	gelijk

Figuur 27. Toegevoegde waarde volgens het vaardigheidsverschil-model (Janssens et al., 2014)

Vaardigheidsgroei-model

Complexer is het zogenaamde ‘*vaardigheidsgroei*’-model waarbij met de prestaties van leerlingen op verschillende opeenvolgende toetsen binnen een zelfde leerdomein wordt rekening gehouden. Net als in het vaardigheidsverschil-model, wordt hier met een multilevel model gewerkt om de leerwinst op schoolniveau te berekenen. De toegevoegde waarde van de school over alle tussenliggende meetmomenten kan worden berekend, waardoor de school een inzicht krijgt op de momenten waarop

haar onderwijs het meest toevoegt aan de leerwinst van de leerlingen. Bovendien kunnen in het vaardigheidsgroei-model ook toetsresultaten gebruikt worden van in- en uitstromende leerlingen, in tegenstelling tot bij het vaardigheidsverschil-model. In principe kunnen de resultaten van elke leerling waarvan minimaal twee vaardigheidsscores zijn geregistreerd mee opgenomen worden in de berekening van de toegevoegde waarde van de school. Een voorbeeld van de rapportering is opgenomen in Figuur 28, waarbij de blauwe lijn de voortgang situeert van leerlingen in alle scholen van het pilootproject (met opnieuw onderscheid tussen netto en bruto leerwinst, zie 'Vaardigheidsverschil-model'), en de rode lijn het gemiddelde traject in de school waarover wordt gerapporteerd uittekent. Rond de rode lijn is het 95% betrouwbaarheidsinterval in licht rood weergegeven.



Figuur 28. Toegevoegde waarde volgens het vaardigheidsgroei-model (Janssens et al., 2014)

In tegenstelling tot het vaardigheidsverschil-model, krijgt de school bij deze meer complexe berekening wel reeds in de vroegere leerjaren een overzicht van hoe de toegevoegde waarde evolueert. Het pilootproject concludeert daarom dat het vaardigheidsgroei-model zowel in een ontwikkelings- als verantwoordingsperspectief kan worden ingezet.

Bruto versus netto leerwinst

De rapportering van het pilootproject naar scholen maakt vaak onderscheid tussen een netto en een bruto score van leerwinst. De bruto-score rapporteert over de ongecorrigeerde, ruwe data inzake leerwinst. Bij de netto-score is rekening gehouden met zogeheten *fairness* kenmerken. Na enkele proefperiodes werden volgende kenmerken behouden bij de berekening van de netto-score inzake

leerwinst: etniciteit, hoogste opleidingsniveau van de ouders, en aanwezigheid van ad(h)d, dyslexie, dyscalculie of autismespectrum-stoornissen. Bij de berekening van de netto leerwinst werd de score van leerlingen (in de hele populatie) met één van deze kenmerken afgezet ten opzichte van de leerwinst van alle andere leerlingen. De gemiddelde verschijscore werd bij elk van deze leerlingen bijgeteld.

In een aanvangsfase werden nog diverse andere kenmerken geselecteerd om de netto leerwinstscore te berekenen. Deze werden echter na enkele proefperioden gereduceerd tot de bovenstaande kenmerken. De redenen voor deze inperking waren divers: van sommige kenmerken bleek de input onbetrouwbaar, of was er een beter alternatief kenmerk voorhanden dat voldoende als proxy kon dienen; van andere kenmerken (met name 'geslacht' en 'overige (leer)stoornissen') kreeg men van de deelnemende scholen de feedback dat het niet wenselijk was deze op te nemen.

Overzicht van conclusies en aanbevelingen uit het pilootproject

Het is mogelijk leerwinst te berekenen

Aangezien elke school die deelnam aan het pilootproject ook systematisch halfjaarlijks de toetsen van CITO-LVS afnam, was voldaan aan twee belangrijke voorwaarden om leerwinst te kunnen berekenen: met name de aanwezigheid van een vaardigheidsschaal waarop de verschillende toetsen zijn gebaseerd, en de dekking van meerdere leerjaren in het toetssysteem. Door deze longitudinale gegevens over een langere periode is de leerwinstmonitoring en berekening van toegevoegde waarde gevoelig voor de kwaliteit van de geboden instructie. Gebruikers kunnen een relatie leggen tussen de hoogte van de scores en de kwaliteit van het gegeven onderwijs.

Binnen deze conclusie formuleert het eindrapport van het pilootproject twee kernachtige boodschappen, respectievelijk voor toetsontwikkelaars (aanbeveling 1) en voor scholen (aanbeveling 2)

Aanbeveling 1: Maak in toets- en schoolinformatiesystemen leerwinstberekeningen mogelijk die gebaseerd zijn op de vergelijking van de groei van leerlingen met een vergelijkbare uitgangssituatie. Gebruik hiervoor bij voorkeur toetsen waaraan vaardigheidsschalen ten grondslag liggen. Een bruikbaar alternatief is een leerwinstberekening op basis van de leercapaciteit of de intelligentie van leerlingen.

(...) ***Aanbeveling 2:*** Indien toets- en schoolinformatiesystemen de mogelijkheden bevatten om leerwinst te berekenen, ontwikkel een beleid voor de school om deze leerwinstberekeningen in te zetten voor schoolverbetering en in het bijzonder voor opbrengstgericht werken en voor de interne kwaliteitszorg. (Janssens et al., 2014, p. 93)

Leerwinstberekeningen kunnen door scholen zelf worden uitgevoerd

Aan de conclusie dat scholen zelf leerwinst kunnen berekenen, zijn evenwel verschillende voorwaarden verbonden, met name: het gebruik van toetsen gebaseerd op een vaardigheidsschaal; het gebruik van schoolinformatiesystemen die over de technische mogelijkheid beschikken om

leerwinst te berekenen; een verantwoordelijke omgang van de school met toetsafname en registratie van gegevens; en de mogelijkheid in het toets- en informatiesysteem om leerlingencohorten duidelijk af te bakenen. Aan deze conclusie worden zes aanbevelingen verbonden, waarvan twee ten aanzien van toetsontwikkelaars, met name op normtabellen op te stellen op basis van enerzijds leerwinst van leerlingen met een gelijk startniveau en anderzijds het gemiddelde van de leerwinst per school; en de aanbeveling om door middel van bijkomende toetsing de leerwinst tijdens de zomervakantie in kaart te brengen (aanbeveling 3 en 4). Drie aanbevelingen zijn gericht ten aanzien van aanbieders van toets- en schoolinformatiesystemen (aanbevelingen 5, 6 en 7), respectievelijk om in het leerwinstmonitoringssysteem de mogelijkheid in te bouwen voor scholen om zelf berekeningen uit te voeren op het niveau van de school, de klassen en de individuele leerlingen en daarbij vergelijkingen te maken; om leerlingencohorten op een flexibele manier door de school te laten definiëren, en om op elk van de drie vermelde niveaus leerwinst te laten rapporteren voor elke getoetste vaardigheid. Aanbeveling 8, ten slotte, wordt gericht ten aanzien van scholen en stelt dat de beschikbare toetsen op vaste momenten moeten worden afgenomen en de resultaten op systematische wijze moeten worden geregistreerd.

Het is mogelijk om per leerstofgebied de toegevoegde waarde van een school te schatten.

Door in de schatting van de toegevoegde waarde rekening te houden met de beginsituatie van leerlingen, met de leerwinst van de leerlingen in een school over een langere onderwijsperiode, alsook met verschillende fairness-kenmerken die door de projectgroep werden geïdentificeerd, is het pilootproject erin geslaagd om de toegevoegde waarde van scholen beter te benaderen dan wat totnogtoe mogelijk was binnen de reeds bestaande praktijk. Het meenemen van fairness-kenmerken is overigens niet zonder praktische problemen, zo blijkt uit de ervaring van het pilootproject:

“Het is privacy-matig niet altijd makkelijk dit soort informatie op te vragen, ook scholen willen dit niet altijd zo maar aanleveren” (interview, respondent NL6)

De berekening van de toegevoegde waarde is, aldus de projectgroep, het meest zinvol wanneer de periode waarover de toegevoegde waarde wordt berekend relatief beperkt is - bijvoorbeeld per twee leerjaren. De twee algemene aanbevelingen die horen bij deze conclusie, hebben betrekking op het te kiezen model en op de periode waarover een toegevoegde waarde-schatting opportuun is:

Aanbeveling 9: *Het vaardigheidsgroei-model biedt de meeste aanknopingspunten voor directe schoolverbetering en voor de beoordeling van de bijdrage van de school aan de leerwinst. Dit model biedt zowel de mogelijkheid om terug te kijken op de geleverde toegevoegde waarde, als vooruit te kijken naar de mogelijke ontwikkeling van de toegevoegde waarde. Het model kan dus in dienst staan van zowel het schoolverbeterings- als het accountabilityperspectief.*

Aanbeveling 10: *Neem als periodes voor de schatting van toegevoegde waarde betekenisvolle tijdseenheden, zoals een leerjaar of de onder- of bovenbouw van een basisschool. Hoe langer de periode waarover de toegevoegde waarde wordt geschat, hoe lastiger het is een verband*

te leggen met de eventuele bijdrage van de school aan de leerwinst en daaraan consequenties te verbinden (Janssens et al., 2014, p. 96).

Toegevoegde waarde-schattingen op basis van correctie van niet-schoolse factoren kunnen niet door de scholen zelf worden uitgevoerd.

De projectgroep acht het onmogelijk dat correcties voor de niet-schoolse factoren (zoals etniciteit, opleidingsniveau ouders,...) geautomatiseerd kunnen gebeuren door de toets- en schoolinformatiesystemen. Drie argumenten onderbouwen deze overtuiging: de gegevens van de niet-schoolse factoren van de hele populatie zouden dan ter beschikking moeten zijn in de schoolinformatiesystemen; men acht het onhaalbaar om de wetenschappelijke expertise die vereist is om de nodige multilevel analyses toe te passen, in deze schoolinformatiesystemen in te bouwen; en de modellen voor toegevoegde waarde zijn nog onvoldoende ontwikkeld, en de impact ervan op de resultaten te weinig onderzocht, om te spreken van een definitief model van toegevoegde waarde dat in de schoolinformatiesystemen kan worden geprogrammeerd.

Het is bij deze conclusie van belang op te merken dat men hier werkt met lokale geïnstalleerde systemen en niet met een online informatiesysteem, zoals gebruikelijk is bij sommige toetsen in andere onderwijssystemen. Mits ingebruikname van een online database zijn verschillende - zo niet alle - van de bovenstaande argumenten om te stellen dat toegevoegde waarde-schattingen niet rechtstreeks in het schoolinformatiesysteem beschikbaar kunnen zijn, niet meer van toepassing.

Verder onderzoek en ontwikkelingswerk is vereist naar modellen van toegevoegde waarde-schattingen

Hoewel de projectgroep de toegevoegde waarde van de verschillende scholen heeft kunnen berekenen op basis van de leerwinst van de leerlingen en informatie over de aanwezigheid van bepaalde niet-schoolse factoren bij leerlingen, zijn er nog problemen inzake zowel de complexiteit als de nauwkeurigheid van de berekeningswijzen die tijdens het pilootproject werden uitgewerkt. Aanvullend onderzoek acht men vooral aangewezen inzake methodisch-technische kwesties (bijvoorbeeld welke corrigerende factoren in rekening dienen te worden genomen bij de berekening van toegevoegde waarde) alsook inzake de impact op de betrouwbaarheid van de bepaling van de toegevoegde waarde van scholen als gevolg van de openbaarmaking ervan (en daaraan gerelateerde ongewenste effecten zoals strategisch gedrag van scholen).

De aanbevelingen 11 tot en met 13 zijn gerelateerd aan deze oproep voor aanvullend onderzoek, en pleiten voor bijkomend onderzoek naar relevante niet-schoolse factoren om de betrouwbaarheid en stabiliteit van de schattingen van toegevoegde waarde te versterken. Daarbij wordt aangegeven dat deze niet-schoolse factoren relatief eenvoudig beschikbaar moeten zijn. Bovendien moet volgens de laatste aanbeveling van de projectgroep onder meer duidelijk zijn hoe de toegevoegde waarde van de scholen zal gebruikt worden door de onderwijsinspectie.

Voorwaarden voor het in kaart brengen van leerwinst en toegevoegde waarde

In deze sectie sommen we ten slotte enkele ervaringen van de projectgroep op inzake voorwaarden voor het succesvol functioneren van monitoringssystemen van leerwinst en/of toegevoegde waarde:

kwaliteitsvolle toetsen, geschikte toets- en schoolinformatiesystemen, systematiek in afname en registratie, en de afbakening van de groep leerlingen waarmee rekening wordt gehouden.

De eerste voorwaarde luidt dat de school per leerstofgebied gebruikmaakt van kwalitatief goede toetsen, gebaseerd op een vaardigheidsschaal, die betrouwbare en valide gegevens opleveren. Dat impliceert onder meer dat de toetsafname gebeurt op de juiste momenten en volgens strikt opgevolgde voorschriften. Om te komen tot een betekenisvolle interpretatie van de leerwinst of toegevoegde waarde dienen deze maten gebaseerd te zijn op gegevens van toetsen uit één leerstofgebied. De betrouwbaarheid van de bepaling van de leerwinst en toegevoegde waarde worden vergroot naarmate deze is gebaseerd op meer dan twee metingen per leerling. Een betrouwbare interpretatie houdt onder meer in dat de resultaten van toetsen op verschillende schalen niet worden geaggregeerd. Door voor de leerstofgebieden afzonderlijk de leerwinst of toegevoegde waarde te bepalen wordt niet alleen recht gedaan aan de inhoudelijke verschillen tussen deze leerstofgebieden (en de differentiële effectiviteit over leerdomeinen), maar wordt de scholen ook meer gelegenheid geboden om op de leerwinst en toegevoegde waarde te sturen.

Het toetssysteem en informatiesysteem waarin de resultaten (leerwinst en/of toegevoegde waarde) wordt gerapporteerd, zijn idealiter overlappend. Daartoe moet het systeem in staat zijn om de nodige statistische analyses uit te voeren. Ook normgegevens om de leerwinst en toegevoegde waarde te beoordelen, dienen in dit informatiesysteem te zijn geïntegreerd.

Eerder werd reeds gewezen op het belang van betrouwbare afname van toetsen op vooraf vastgestelde momenten en volgens strikt opgevolgde voorschriften. Verder wordt tijdens het pilootproject ook het belang duidelijk van een systematische registratie van de resultaten: dit houdt in dat alle gegevens voor alle leerlingen op een correcte en coherente manier worden ingevoerd. Niet alleen bleken er grote verschillen inzake de 'invoerdiscipline' van scholen - de bereidwilligheid van scholen om gegevens over leerlingen (bijvoorbeeld over achtergrondvariabelen) te verzamelen en (correct) in te voeren - ook de afstemming van verschillende systemen die in verschillende scholen werden gebruikt zorgden voor een onvoorziene werklast.

Een andere belangrijke voorwaarde voor de berekening van leerwinst of toegevoegde waarde is dat de groep leerlingen waarvoor deze worden berekend, duidelijk is gedefinieerd. Deze afbakening van de leerlingengroep is niet steeds evident, met bijvoorbeeld in- en uitstromers in de school, met leerlingen die een jaar overdoen, en dergelijke.

3.3.3 Impact

In de rapportering van het pilootproject werd geconcludeerd dat:

"leerwinst en toegevoegde waarde een waardevolle bijdrage leveren aan de verdere ontwikkeling van het opbrengstgericht werken in scholen en aan de beoordeling van hun eigen leerprestaties. De groei van de prestaties van alle leerlingen is nauwkeuriger te bepalen en te beoordelen." (Janssens et al., 2014, p. 8).

Er blijkt uit de interviews met verschillende respondenten inderdaad dat scholen gemotiveerd worden om te denken in termen van leerwinst (eerder dan vanuit ruwe leerprestaties). Het blijft evenwel de vraag in welke mate het pilootproject hiertoe heeft bijgedragen. De scholen die werden bezocht, neigen er immers naar om de reeds eerder bestaande instrumenten te gebruiken om een maat voor leerwinst van elke leerling te bepalen - ook al zijn deze instrumenten minder betrouwbaar dan de rapportage ontwikkeld door het pilootproject. De hogere betrouwbaarheid weegt niet in alle gevallen op tegen de lagere gebruiksvriendelijkheid en de grotere complexiteit van de rapportage in het pilootproject.

De mate waarin scholen gebruik hebben gemaakt van de resultaten van het leerlingvolgsysteem en de rapporten van het pilootproject, hangt bovendien vaak erg af van één specifieke persoon in de school, en het belang dat deze persoon aan de opgeleverde informatie toekent. Hierbij wordt voornamelijk de rol van de interne begeleider van de school benadrukt.

We concluderen daarom dat de echte meerwaarde van het pilootproject voorlopig onduidelijk blijft. Het leerlingvolgsysteem heeft in de bevraagde scholen al sterk de focus gelegd op leerwinst, en de extra berekeningen en rapportering hebben daar - naar aanvoelen van de bevraagde directies - weinig aan toegevoegd. Toch hebben scholen door de vele instructies vanuit de projectgroep geleerd om de bestaande overzichtstabellen (die niet door CITO worden ondersteund) met een kritische blik te bekijken.

“Dat kijken naar leerwinst, dat deden we al. Voor mij als schoolleiding heeft het niet opgebracht wat ik had gehoopt. Wat ik nu wel weet, is dat dit [de overzichtstabellen van een private aanbieder] niet helemaal klopt: het is meer een indicatie, maar die klopt niet honderd procent, hebben ze maar daar uitgelegd. Maar het is een werkinstrument, als die niet honderd procent klopt... maakt niet zoveel uit.” (interview, respondent NL1).

Ook de andere directie geeft aan dat de meerwaarde van het pilootproject op de interne kwaliteitszorg in de school, of op de klasprocessen, niet groter was dan de meerwaarde van de afname van de CITO-LVS toetsen op zich.

Evaluatie rapportering in pilootproject

De inzichtelijkheid en bruikbaarheid van de rapportering waren twee uitgangspunten binnen het pilootproject. Ondanks verschillende evaluaties en bijstellingen, bleek dat de rapportering die aan het einde van het pilootproject was ontwikkeld, niet door elke school positief werd geëvalueerd. Dit wordt aangetoond door de eigen evaluatie van het pilootproject, waarbij de scholen werden bevraagd inzake hun beoordeling van de inzichtelijkheid en bruikbaarheid. De 33 ‘reguliere scholen’ in de steekproef die aan deze evaluatie deelnamen, beoordeelden de inzichtelijkheid van respectievelijk de leerwinst- en toegevoegde waarde-rapportering als een 6,9/10 en 6,5/10, met respectievelijk 9% en 15% van de scholen die een onvoldoende gaven (4/10 of lager). De beoordeling in verband met de bruikbaarheid van de rapportering was net iets minder gunstig, met name 6,4/10 (15% onvoldoendes) voor de leerwinstrapportering en 6,3/10 (24% onvoldoendes) voor de rapportering inzake toegevoegde

waarde. Bij de zeven scholen voor speciaal onderwijs in de steekproef waren de evaluatiescores lager dan bij de reguliere scholen. De gemiddelde scores voor de rapportering leerwinst-toegevoegde waarde waren er 5,9/10 en 5,4/10 voor inzichtelijkheid en 5,4/10 en 5,1/10 voor bruikbaarheid.

Verder werden in deze evaluatie door de projectgroep aan de directies en/of interne begeleiders stellingen voorgelegd in verband met de inschatting van de meerwaarde en van het mogelijke toekomstig gebruik van de rapporten inzake leerwinst en toegevoegde waarde. De resultaten hiervan zijn opgenomen in Tabel 12. Deze bevraging leidt tot de conclusie dat 60% tevreden is met de meerwaarde van de leerwinstrapportering en 53% met de meerwaarde van de toegevoegde waarde-meting. Respectievelijk 62% en 56% menen dat ze de rapporten inzake leerwinst en toegevoegde waarde ook daadwerkelijk zouden gebruiken wanneer deze in de toekomst beschikbaar zouden zijn voor scholen.

Tabel 12

Meningen van directeuren en interne begeleiders van reguliere pilotscholen over het nut en het toekomstig gebruik van rapportages over leerwinst (LW) en toegevoegde waarde (TW)

		Mee oneens	Neutraal	Mee eens
Leerwinst- schoolrapport	Ik vind het LW-rapport is een waardevolle aanvulling op wat er nu in LVS systemen zit	16%	24%	60%
	Ik ga het LW-rapport in toekomst gebruiken als het in ons LVS systeem beschikbaar komt	14%	24%	62%
Toegevoegde schoolrapport	Het TW-rapport is een waardevolle aanvulling op groepsrapportages die nu in LVS systemen zitten.	14%	33%	53%
	Ik zal het TW-rapport in toekomst gebruiken voor opbrengstgericht werken	14%	31%	56%

Verder blijkt uit deze analyse dat onder meer volgende zaken de bruikbaarheid en/of inzichtelijkheid van de rapporten zou kunnen versterken: de snellere beschikbaarheid van de rapporten; het beschikbaar maken van de rapporten in het leerlingvolgsysteem dat al door de scholen wordt gebruikt zodat alle data bij elkaar komen; het aanbieden van handvaten voor gebruik in de praktijk; het voorzien van een betere toelichting op de resultaten; en ten slotte het voorzien van een duidelijke handleiding met betrekking tot de interpretatie van de resultaten. Daarnaast moeten volgens de respondenten inspanningen worden geleverd om ook de methodiek achter het meetmodel zo doorzichtig mogelijk te maken voor de gebruikers.

Gebruik in een verantwoordingsperspectief

De meeste scholen die deelnamen aan het pilootproject stonden niet weigerachtig tegenover het idee om de resultaten van een berekening van de leerwinst van leerlingen door de onderwijsinspectie te laten gebruiken. Zo verklaart 61% van de bevroegde directies of interne begeleiders zich akkoord met de stelling 'ik denk dat het LW-rapport goed ingezet kan worden bij onderwijstoezicht door de Inspectie'. 17% van de respondenten gaat niet akkoord met deze stelling (22% bleef neutraal). Minder eensgezindheid is er evenwel omtrent de stelling inzake toegevoegde waarde 'ik vind dat het TW-

rapport wel door de Inspectie gebruikt mag worden'. Met deze stelling waren evenveel respondenten akkoord, als niet akkoord (beiden 31%; 39% bleef neutraal). Een vraag die hierbij wordt gesteld is evenwel of men het werkelijk wenselijk vindt dat deze rapporten door de onderwijsinspectie worden gebruikt, dan wel of men het ervaart als een verbetering ten aanzien van de huidige situatie (waarbij de inspectie enkel statusmodellen van toegevoegde waarde hanteert).

De projectgroep plaatst zelf ook enkele beperkingen bij het eventuele verantwoordingsgerichte gebruik van toegevoegde waarde-berekeningen: zo dient eerst verder onderzoek te worden verricht over de meest gepaste wijze om de toegevoegde waarde van scholen in te schatten; daarnaast dient dit ingebed te worden in een ruimer beleid, dat moet worden voorbereid in samenwerking met scholen, de onderwijsinspectie en academici.

3.3.4 *Kritiek en neveneffecten*

Zoals eerder vermeld, ondervond de projectgroep veel administratieve hinder in het verzamelen van de relevante gegevens om de toegevoegde waarde van scholen te berekenen. Ook de afstemming van verschillende systemen die in verschillende scholen werden gebruikt zorgden voor een onvoorziene werklast. Bovendien ervaren scholen ook dat ouders niet altijd eerlijk zijn bij het aangeven van gegevens over de eigen socio-economische status, wat een bedreiging is van de betrouwbaarheid van de resultaten.

Daarnaast blijft het in deze case onduidelijk of het doel van de pilot 'leerwinstmonitoring' om een valide en betrouwbare maatstaf te ontwikkelen voor leerwinst of toegevoegde waarde die ook praktisch bruikbaar is voor scholen. Er worden wel heel veel betrouwbare en valide data opgeleverd, maar toch blijkt de ervaring dat deze data niet altijd optimaal gebruikt kan worden in het kader van schoolontwikkeling:

"Daar komt heel veel informatie in terecht,... voor mij als onderzoeker is het een beetje hart-schokkend dat ze er zo weinig mee doen!" (interview, respondent NL7).

Ten slotte wordt kritiek geuit op de koppeling van ontwikkelingsgerichte en verantwoordingsgerichte doeleinden aan de leerwinst- en toegevoegde waarde-meting.

"We zien het graag als iets dat je voor scholen zelf kan gebruiken als een instrument om je eigen onderwijs te evalueren, om te zien in welke vakken doe ik het goed, doe ik het minder goed? Waar moeten we meer aandacht besteden? Daar willen we wel over waken dat het niet als zo'n soort afrekeninstrument wordt gebruikt. Een van de redenen daar voor is: je ziet, het model dat hier wordt gebruikt, en dat geldt voor alle modellen, dat is totaal afhankelijk van de beginmeting. En op het moment dat er zo'n heel hoog belang aan die outcome wordt gegeven, gaat natuurlijk de neiging ontstaan om die beginmeting te manipuleren: dan is het handig om daar heel laag op te scoren want dan is je groei natuurlijk ontzettend hoog." (interview, respondent NL6)

3.4. Leerwinst en toegevoegde waarde

Leerwinst en toegevoegde waarde staan centraal in deze case. Verschillende modellen worden uitgetoetst en naar waarde geschat. Uit de resultaten van de case studie blijkt dat, ondanks alle zorg die besteed wordt aan methodologische onderbouwing, men in scholen sterk geneigd is om voornamelijk de informatie te gebruiken die relatief eenvoudig kan worden voorgesteld. Het is pas wanneer de resultaten sterker vanuit een verantwoordingsgericht perspectief worden gehanteerd, dat de betrouwbaarheid van de toetsresultaten en een meer sluitende correcte interpretatie - in de ogen van de respondenten in scholen - aan belang winnen.

Ondanks de sterke focus in deze pilot op leerwinst en toegevoegde waarde, wordt er gewaarschuwd voor enkele nadelen, zoals het relativeren van zwakke leerprestaties door enkel te focussen op leerwinst, of een te weinig ambitieuze doelstelling ten aanzien van leerlingen met een minder gunstige achtergrond:

“Leerwinst is ook niet alles: een van de dingen die je nog altijd in de gaten moet houden, is van: ook al heb je veel leerwinst geboekt, je kunt het nog altijd dat je eindniveau zo laag is, dat je moet zeggen; jammer, mooi dat je zo ver vooruit bent gekomen, maar je kunt het nog steeds niet, dat is iets wat je nog altijd in de gaten moet houden. Een ander punt, of als je rekening gaat houden met achtergrondkenmerken van leerlingen is dat je eigenlijk minder hoge eisen gaat stellen aan mensen die op een laag niveau begonnen zijn. Je kunt de situatie krijgen dat iemand een klas met allemaal migrantenkinderen heeft en die zit dan op het niveau van, nationaal gezien, ‘goed gedaan jongens’, en je kunt een andere klas voorstellen waar allemaal hoogopgeleide ouders zijn, en zeg je van ‘jullie zitten allemaal gemiddeld, wat is dit nou?’ Dat is een heel vervelend dilemma als je met leerwinst of met toegevoegde waarde gaat werken.” (interview, respondent NL7)

Bovendien blijkt uit het pilootproject leerwinstmonitoring dat het moeilijk is om de toegevoegde waarde van scholen correct te meten: het al dan niet toevoegen van bepaalde factoren maakt dat de onverklaarde variantie (wat scholen dan zouden ‘toevoegen’) vergroot of verkleint.

4. CASE NOORWEGEN (INCLUSIEF FOCUS OP OSLO)

Ook bij de vierde case geven we eerst een korte inleiding in het lokale onderwijssysteem, met bijzondere aandacht voor enkele recente ontwikkelingen in het Noorse onderwijsbeleid. Deze case valt verder uiteen in twee delen, met name de gestandaardiseerde toetsen vanuit de nationale overheid en het gebruik van die toetsen, alsook enkele bijkomende gestandaardiseerde toetsen in scholen in de stad/provincie Oslo. Ook in deze case bespreken we nadien expliciet in welke mate er elementen van leerwinst en toegevoegde waarde worden geïdentificeerd.

4.1 Inleiding tot het onderwijssysteem in Noorwegen⁶³

Kinderen in Noorwegen zijn leerplichtig van de leeftijd van 6 tot 16 jaar. Het leerplichtonderwijs is gratis.

4.1.1 *Onderwijsstructuur*

Het leerplichtonderwijs wordt voorafgegaan door *'kindergartens'* vanaf de leeftijd van één jaar; er is geen formeel onderscheid tussen kinderopvang en kleuteronderwijs. Het leerplichtonderwijs zelf is zeer comprehensief: leerlingen volgen nagenoeg allemaal hetzelfde onderwijsprogramma en worden dus niet verdeeld overheen verschillende onderwijstypes of studierichtingen. Het leerplichtonderwijs omvat het lager onderwijs (leerjaar 1 tot en met 7) en het lager secundair onderwijs (leerjaar 8, 9 en 10).

Ruim 97% van de leerlingen stromen meteen na het leerplichtonderwijs door naar een vorm van hoger secundair onderwijs. Het algemene hoger secundair onderwijs duurt drie jaar (leerjaar 11, 12 en 13)⁶⁴, terwijl de beroepsgerichte opleidingen in het hoger secundair onderwijs afhankelijk van de opleidingsvorm drie of vier jaar duren. Figuur 29 geeft een overzicht van de onderwijsstructuur in Noorwegen.

4.1.2 *Leerlingen en scholen*

Het lager onderwijs en het lager secundair onderwijs wordt meestal ingericht door de stedelijke of gemeentelijke overheid. 2.697 van de 2.907 scholen (92,8%) voor lager en/of lager secundair onderwijs zijn ingericht door één van de 428 steden of gemeenten in Noorwegen⁶⁵. De resterende scholen zijn ingericht door de staat of de provincie ('fylke') (0,5%) of door privé-initiatief (6,7%). In totaal lopen 19.105 leerplichtige leerlingen (3,1% van het totaal aantal leerplichtige leerlingen) school in één van deze privéscholen. Bijna alle privéscholen worden gesubsidieerd door de overheid. 88 van

⁶³ Gebaseerd op: Elstad (2009); Norwegian Directorate for Education and Training (2011); Paulsen (2014); School Inspections Project (2016); Skedsmo (2011) + informatie tijdens gesprekken in case studie.

⁶⁴ Jaar 11, 12 en 13 worden ook vg1, vg2 en vg3 genoemd.

⁶⁵ Alle cijfermatige data in deze sectie zijn gebaseerd op Statistics Norway (2015). Het betreft cijfers uit 2013.

de 195 privéscholen zijn geënt op een specifieke onderwijsvisie, en 70 van deze scholen hebben een religieuze denominatie.

Leeftijd	Klas		Nationaal	Oslo
18	Leerj13 (vg3)	Hoger secundair		
17	Leerj 12 (vg2)			
16	Leerj 11 (vg1)		Leer-ondersteunende toets	Oslo-toetsen
15	Leerjaar 10	Lager secundair	Eindexamen	Beoordelingsonderst toets
14	Leerjaar 9		Nationale toetsen	
13	Leerjaar 8		Nationale toetsen	Leer-ondersteunende toets
12	Leerjaar 7	Lager onderwijs		Overgangstoetsen
11	Leerjaar 6			Leer-ondersteunende toets
10	Leerjaar 5		Nationale toetsen	Leer-ondersteunende toets
9	Leerjaar 4		Diagnostische toets	Overgangstoetsen
8	Leerjaar 3		Diagnostische toets	Oslo-toetsen
7	Leerjaar 2		Diagnostische toets	
6	Leerjaar 1		Diagnostische toets	
1-5	Kindergarten			

Figuur 29. Overzicht van het Noorse onderwijssysteem, inclusief gestandaardiseerde toetsen

Van de 2.697 gemeentelijke scholen zijn er 1.607 lagere scholen (*Barneskole*, 59,6%), 470 lager secundaire scholen (*Ungdomsskole*, 17,4%) en 620 scholen die beide onderwijsniveaus aanbieden (*Grunnskole*, 23,0%). Bij privéscholen ligt de verhouding lager onderwijs, lager secundair onderwijs, en combinatie van beide niveaus op respectievelijk 31,3%, 7,2% en 61,5%.

Er zijn in totaal 615.327 leerplichtige leerlingen in Noorwegen. In het leerplichtonderwijs zijn er 50.977 leerlingen met een erkende speciale onderwijsbehoefte. Dit is 8,3% van de totale leerlingenpopulatie.

Onder meer door de vele dun bevolkte gebieden zijn er veel kleine scholen in Noorwegen. De gemiddelde schoolgrootte bedraagt 211 leerlingen per school. 31,1% van de scholen lager en/of secundair onderwijs heeft minder dan 100 leerlingen. Respectievelijk 40,3% en 28,7% van de scholen hebben tussen 100 en 200, en meer dan 200 leerlingen. Er is een duidelijke trend doorheen de laatste tien jaar naar minder, maar grotere scholen.

Vaak is het zo dat waar een openbare school met een te klein leerlingaantal omwille van een rationalisatie gesloten wordt (in lijn met de tendens naar minder, maar grotere scholen), er een nieuw

lokaal privé-initiatief ontstaat zodat deze leerlingen zich niet over een grotere afstand hoeven te verplaatsen om naar school te gaan: de nieuwe privéscholen hebben dan ook maar een gemiddeld aantal leerlingen van 33 tijdens hun eerste werkjaar (terwijl de gesloten scholen gemiddeld 69 leerlingen telden). Volgens schattingen wordt een privéschool opgericht bij ongeveer 10% van de sluitingen van gemeentelijke scholen.

Het hoger secundair onderwijs wordt, in tegenstelling tot het leerplichtonderwijs, meestal ingericht door de 19 provincies ('fylker'): dit geldt voor 343 van de 435 scholen voor het hoger secundair onderwijs (78,9%). Daarnaast zijn er 3 staatsscholen (0,7%) en 89 scholen ingericht door een privé-initiatief (20,5%). Het onderwijsbureau binnen deze provincies is verantwoordelijk voor het toezicht op de onderwijskwaliteit.

4.1.3 Curriculum

Er is voor het leerplichtonderwijs, alsook voor het daaropvolgende hoger secundair onderwijs een uniform nationaal curriculum dat alle scholen dienen te volgen⁶⁶. Dit nationale curriculum bestaat uit drie delen: een kerncurriculum rond vijf fundamenteel geachte vaardigheden (mondelinge vaardigheden, schrijven, lezen, digitale vaardigheid en rekenvaardigheid), een set van principes en richtlijnen voor onderwijs, en subject-specifieke curricula. Hoewel de vijf fundamentele vaardigheden overal centraal staan, zijn er geen lessenspakketten of einddoelen aan verbonden - in tegenstelling tot bij de subject-specifieke curricula. Deze vaardigheden moeten, met andere woorden, doorheen de verschillende vakken worden verworven en zijn een gedeelde verantwoordelijkheid van alle (vak)leraren.

Het nationale curriculum bevat niet alles wat in scholen wordt aangeleerd. Het is de bedoeling dat dit curriculum wordt aangevuld door de scholen om zo maximaal mogelijk aangepast te zijn aan de lokale omstandigheden en aan individuele leerbehoeften van de leerlingen. Naarmate het onderwijsniveau vordert, is er evenwel minder ruimte voor vrije invulling van scholen: in het hoger secundair onderwijs worden bijna alle leerinhouden door het nationale curriculum vastgelegd.

Zowel in het lager als in het hoger secundair onderwijs worden leerlingen niet in verschillende studierichtingen opgedeeld: alle leerlingen (met uitzondering van buitengewoon onderwijs) volgen in principe hetzelfde curriculum, al wordt in grotere scholen soms wel voorzien in enige opdeling van leerlingen volgens interesse of capaciteit. De enige uitzondering op dit comprehensieve programma zijn de twee vormen van optionele vakken voor leerlingen in het hoger secundair onderwijs, die elk ongeveer 6,5% van het totale curriculum in de graden 8, 9 en 10 omvatten (gemiddeld elk ongeveer 2 lesuren per week). 73% van de leerlingen vullen het ene optionele vak in met een tweede vreemde taal (naast Engels). Populaire vakken bij het tweede optionele vak zijn 'fysieke activiteiten en gezondheid', 'muziek- en toneelproductie' en '(her)ontwerpen'. Daarnaast kunnen leerlingen er (in sommige scholen) ook voor kiezen om reeds vakken uit het hoger secundair onderwijs op te nemen.

⁶⁶ Er is een uitzondering voor scholen in bepaalde Sami Districten in Lapland. Het aparte Sami curriculum is bijkomend gericht op het ontwikkelen van een Sami identiteit en het aanleren van de Sami taal.

Pas in het hoger secundair onderwijs zijn er verschillende studierichtingen, al blijkt de diversificatie relatief beperkt. 48,2% van de leerlingen in leerjaar 11 kiest voor één van de drie programma's in 'algemene studies', met name 'specialisatie in algemene studies' (39,9%), 'sport en lichamelijke opvoeding' (5,3%), en 'muziek, dans en drama' (3,0%). De groep leerlingen in de 'specialisatie in algemene studies' wordt in leerjaar 12 opgesplitst in een groter aantal studierichtingen.

51,8% van de leerlingen in leerjaar 11 kiest voor één van de negen programma's in het beroepsgericht onderwijs. De meest populaire programma's hier zijn 'gezondheidszorg en ontwikkeling van kinderen en jongeren' (11,5%), 'technische en industriële productie' (9,3%) en 'elektriciteit en elektronica' (6,7%).

4.1.4 Centrale thema's: gelijkheid en comprehensiviteit

'Gelijkheid' is een bijzonder belangrijk thema in het Noorse onderwijs. De vele inspanningen daartoe blijken bijvoorbeeld in de verschillende PISA-studies, waarin kleine verschillen tussen scholen, en tussen leerlingen binnen scholen worden gerapporteerd.

Het comprehensieve karakter van het Noorse onderwijssysteem impliceert dat er weinig opdeling is van leerlingen, zoals reeds hierboven werd beschreven (4.1.3). Bovendien is zittenblijven in het lager onderwijs en in het lager secundair onderwijs onbestaande. Pas in het hoger secundair onderwijs kan de toegang tot een hoger leerjaar worden ontzegd op basis van onvoldoende leerprestaties. Zittenblijven is niet ongebruikelijk in het hoger secundair onderwijs: 75% van de leerlingen in het algemeen hoger secundair onderwijs behalen hun diploma binnen de vooropgestelde periode; in de beroepsgerichte opleidingen is dit slechts 39%. De drop-out is echter relatief laag: het gaat in het algemene en beroepsgerichte hoger secundair onderwijs om respectievelijk 8% en 26% van de leerlingen.

Om deze principes van gelijkheid en comprehensiviteit te verwezenlijken wordt een lage leerling-per-leraar-ratio noodzakelijk geacht. In het leerplichtonderwijs (enkel gemeentelijke scholen) is deze ratio 10,3 leerlingen per leraar (alle lerarenfuncties inclusief). In het hoger secundair onderwijs (enkel provincie-scholen) is de ratio 8,9.

4.1.5 Kwaliteitszorg

Alle scholen worden beschouwd als zelfsturende eenheden die de verantwoordelijkheid dragen voor de kwaliteit van hun onderwijs. Voor de openbare scholen in het leerplichtonderwijs dienen de gemeenten (als schoolbestuur) een systeem voor kwaliteitszorg te ontwikkelen om de output van scholen te evalueren en op te volgen, doch dat is zeker niet altijd het geval. In vele schoolbesturen beperkt de externe 'kwaliteitscontrole' zich tot een controle van de boekhouding⁶⁷, waarbij het stadsbestuur van Oslo (met een sterk uitgebouwd systeem van kwaliteitscontrole, zie sectie 4.3) de belangrijkste uitzondering vormt. Er is een toezicht op de werking van de gemeenten door het

⁶⁷ Deze stand van zaken werd vastgesteld in 2007 door Dosronningen (gerapporteerd in Elstad, 2009). Het is onduidelijk in welke mate dit heden ten dage veranderd is.

provinciale niveau. Algemeen kan echter worden gesteld dat inspecties van scholen slechts in een beperkt aantal gemeenten op systematische basis worden uitgevoerd.

“Oslo is more accountability-oriented than other municipalities. If you ask me about the small municipalities, we have really really small municipalities in Norway, they don’t have systems at all; they don’t care!” (interview, respondent NW10)

Er is ook geen instantie op het nationale niveau dat verantwoordelijk is voor het uitvoeren van inspecties van openbare scholen. Toch bestaat er een apart agentschap binnen het Ministerie van Onderwijs en Onderzoek verantwoordelijk voor inspecties van scholen. Dit agentschap richt zich echter enkel op private scholen.

4.2. Gestandaardiseerde toetsen in Noorwegen (federaal)⁶⁸

Het onderwijssysteem in Noorwegen kende verschillende ingrijpende aanpassingen in het eerste decennium van deze eeuw, als gevolg van een ‘PISA shock’ bij de resultaten van PISA 2000 (leesvaardigheid)⁶⁹. Daarin werd aangetoond dat de 15-jarige Noorse leerlingen, tegen de verwachtingen van de beleidsmakers in, nauwelijks betere prestaties konden voorleggen dan het gemiddelde uit de OESO-landen inzake lezen, wiskunde en wetenschappen. Ook in PISA 2003 (rekenvaardigheid) en PISA 2006 (wetenschappelijke geletterdheid) en in TIMSS 2003 (rekenvaardigheid en wetenschappelijke geletterdheid) werden de resultaten voor Noorwegen als teleurstellend ervaren. Sindsdien zijn de leerlingenresultaten in internationale vergelijkingen echter in stijgende lijn⁷⁰.

Tot voor 2000 waren er nauwelijks gestandaardiseerde toetsen in Noorwegen, al bestond er al wel een lange traditie inzake eindexamens aan het eind van het leerplichtonderwijs. Als gevolg van de PISA-shock in 2000 werd een grootschalig hervormingsprogramma ‘*The Knowledge Promotion*’ (‘*Kunnskapsløftet*’) opgezet met als doel om de leeruitkomsten van leerlingen te versterken. Enkele verwante ideeën werden evenwel reeds tevoren gelanceerd door de toenmalige regering, alsook door de toenmalige Minister van Onderwijs en Onderzoek Kristin Clemet. De resultaten van PISA werden dan voornamelijk gehanteerd als verantwoording voor de voorziene wijzigingen.

⁶⁸ Gebaseerd op: Elstad (2009); Elstad, Lejonberg, en Christophersen (2015); Elstad en Turmo (2011); Eurydice (2009); Fastling, Thygesen, Berge, Evensen, en Vagle (2009); Hopfenbeck, , Flórez Petour, & Tolo (2015); Norwegian Directorate for Education and Training (2008, 2011, 2014); Nusche, Earl, Maxwell, en Shewbridge (2011); Seland, Vibe, en Hovdhaugen (2013); Skedsmo (2011); Statistics Norway (2016); Tveit (2013); Utdanningsdirektoratet (z.j.) + informatie tijdens gesprekken in case studie.

⁶⁹ Ook de eerste comparatieve data, uit TIMSS 1995, toonden al aan dat de eigen hoge verwachtingen inzake effectiviteit, niet realistisch waren. Deze studie had echter een minder groot effect op het onderwijsdebat dan PISA 2000.

⁷⁰ De discussie over de kwaliteit van onderwijs gaat evenwel al langer terug. Een OESO-rapport uit 1988 stelde in vraag of Noorwegen voldoende middelen beschikbaar had voor het monitoren van kwaliteit van het onderwijssysteem. In het rapport stelde de OESO diverse strategieën voor om dit te bewerkstelligen. Onder meer als gevolg hiervan was er reeds in de jaren '90 van de vorige eeuw een levendige discussie op gang gekomen in verband met de wenselijkheid van het gebruik van gestandaardiseerde toetsen (zie onder meer Granheim en Lundgren (1991)).

4.2.1 *The Knowledge Promotion als kader voor de ontwikkeling van gestandaardiseerde toetsen*

Een basisidee van *The Knowledge Promotion* is dat scholen sterk(er) gericht moeten zijn op de uitkomsten van leerprocessen van leerlingen, waarbij goed gefundeerde kennis en vaardigheid inzake mondelinge communicatie, lezen, schrijven, digitale vaardigheden en rekenvaardigheden centraal dient te worden gesteld. *The Knowledge Promotion* gaf enerzijds meer autonomie aan de schoolbesturen (bijvoorbeeld inzake het differentiëren in de lonen van leraren, wat men kan doen onder meer op basis van resultaten van leerlingen op gestandaardiseerde toetsen), maar voerde anderzijds ook meer controle in op de output van scholen of schoolbesturen.

Binnen *The Knowledge Promotion* werd sinds 2005 een uitgebreid nationaal evaluatiesysteem uitgebouwd, met als doel te leiden tot een aangepast onderwijs voor alle individuele leerlingen en tot verbeterde leeruitkomsten. Bovendien moet het evaluatiesysteem een informatierijke basis opleveren over nationale en lokale evoluties om zo op een geïnformeerde basis het onderwijsbeleid te ontwikkelen. Ten slotte heeft het evaluatiesysteem tot doel om bij te dragen tot een versterkte openheid, transparantie en dialoog over de onderwijsprocessen in scholen (en in het onderwijsveld in het algemeen) en de mate waarin deze processen tot de verwachte doelstellingen leiden.

Aan dit evaluatiesysteem zijn zowel ontwikkelingsgerichte doeleinden, als verantwoordingsgerichte doelen verbonden. De focus van de verantwoordingsgerichte mechanismen ligt op het mesoniveau, en dan voornamelijk op het schoolbestuur, eerder dan op het niveau van de individuele scholen. De schoolbesturen worden ervoor verantwoordelijk geacht om de kwaliteit van scholen op te volgen en te ondersteunen. Toch blijven de *stakes* van het evaluatiesysteem in principe vrij laag: er zijn immers geen automatische gevolgen aan verbonden, en er is geen beleidsinstantie die scholen ter verantwoording roept of sanctioneert bij slechte resultaten. De perceptie in de realiteit is evenwel niet altijd in lijn met dit opzet van het *low-stakes* karakter. Uit een bevraging in 2005 - net nadat de eerste stappen richting meer gestandaardiseerde toetsen werden gezet - bleek dat ruim 90% van de schooldirecties (in zekere mate) akkoord ging met de stelling dat 'evaluatie leidt tot schoolontwikkeling', maar dat evenzeer 90% van de directies (in zekere mate) akkoord ging met de stelling dat 'evaluatie wordt gebruikt om het schoolleiderschap en leraren verantwoording te doen afleggen voor de processen in de school'⁷¹. Dit laatste duidt er op dat het evaluatiesysteem van bij aanvang niet enkel als ontwikkelingsgericht, maar ook als verantwoordingsgericht werd gepercipieerd.

Om de nodige relevante informatie te genereren die relevant is op de verschillende niveaus (micro, meso, en macro) werd het nodig geacht om meer dan één tool in te zetten, en een uitgebreid arsenaal aan instrumenten te ontwikkelen. In de afgelopen 15 jaar werd dan ook veel geïnvesteerd in de ontwikkeling van een kwaliteitsvol en omvattend nationaal toetssysteem voor de onderwijssector in Noorwegen. Dit toetssysteem bestaat uit:

- nationale toetsen (*Nasjonale Prøver*) die informatie opleveren over de basisvaardigheden van leerlingen en een basis vormen voor kwaliteitsontwikkeling in de school;

⁷¹ Onderzoek gerapporteerd door Skedsmo (2011).

- diagnostische toetsen (*kartleggingsmaterieel*), bedoeld om ondersteuning te bieden bij het identificeren van specifieke onderwijsleerbehoefte van bepaalde leerlingen;
- eindexamens die de competenties van elke leerling toetsen aan het eind van het leerplichtonderwijs;
- verschillende andere gestandaardiseerde toetsen die niet onder de vlag van *The Knowledge Promotion* zijn ontwikkeld, maar evenzeer deel uitmaken van het Noorse onderwijs.

Terwijl de eindexamens een traditie hebben die langer teruggaat, werden de nationale toetsen en de diagnostische toetsen specifiek ontwikkeld in het kader van *The Knowledge Promotion*. Naast de nationale toetsen, diagnostische toetsen en eindexamens, maken ook niet-gestandaardiseerde evaluaties deel uit van het evaluatiesysteem ontwikkeld in het kader van the Knowledge Promotion. Het betreft lokale toetsen (*'karakter- og- læringsstøttende prøver'*) met zowel formatieve als summatieve doelstellingen. Deze lokale toetsen worden ontwikkeld in opdracht van individuele schoolbesturen (gemeente of provincie), maar het gebruik ervan wordt door het federale beleid van *The Knowledge Promotion* sterk gepromoot. Daarnaast omvat het evaluatiesysteem verplichte tevredenheidsvragenlijsten van leerlingen en facultatieve tevredenheidsvragenlijsten van ouders en leraren. In hetzelfde beleidskader werd ook werk gemaakt van het ontwikkelen en beschikbaar stellen van materialen die ondersteunend zijn voor de interne kwaliteitszorg van de school. Ten slotte wordt ook de deelname aan internationale studies (bijvoorbeeld PISA en TIMSS) gerekend tot het programma van the Knowledge Promotion.

Tabel 13 geeft een overzicht van de verschillende toetsen in het Noorse onderwijssysteem, inclusief het leerdomein en het leerjaar waarin getoetst wordt. De rationale achter het evaluatiesysteem, en hoe de verschillende instrumenten verondersteld worden op elkaar in te spelen, is evenwel niet expliciet gemaakt.

De examens, nationale toetsen en diagnostische toetsen, de leerondersteunende toetsen en beoordelingsondersteunende toetsen worden ontwikkeld onder de auspiciën van de Norwegian Directorate for Education and Training (*'Utdanningsdirektoratet'*). Dit Directorate is een beleidsuitvoerend agentschap dat ressorteert onder het Ministerie van Onderwijs en Onderzoek en is verantwoordelijk voor het lager en secundair onderwijs (leerplichtonderwijs én hoger secundair onderwijs).

De kerntaken van dit Directorate bestaan uit toezicht op de inrichtende machten van scholen in het lager en secundair onderwijs (gemeenten, provincies en privé-initiatieven) en de ondersteuning van scholen door nascholing en ontwikkeling van educatieve materialen. Dit concretiseert zich in:

- de planning van het nationale curriculum;
- de organisatie van examens, nationale toetsen en diagnostische toetsen;
- het uitvoeren van onderzoek en bieden van feedback aan scholen, inrichtende machten en beleidsmakers, en daaropvolgend ontwikkelingswerk;
- de coördinatie van internationale studies.

Tabel 13

Overzicht van de leerjaren waarin gestandaardiseerde toetsen worden afgenomen, per leerdomein

	Leesvaardigheid (curriculum-onafhankelijk)	Leesvaardigheid (curriculum Noors)	Rekenvaardigheid (curriculum-onafhankelijk)	Rekenvaardigheid (curriculum wiskunde)	Engels	Digitale vaardigheid	Natuurwetenschappen	Sociale studies
Nationale toetsen	5, 8, 9		5, 8, 9		5, 8, 9			
Diagnostische toetsen	1, 2, 3		1, 2, 3		3 ⁷²	4 ⁷³		
Eindexamens		10		10	10			
Leerondersteunende toetsen		6, 11		6, 11	11	8		
Beoordelingsondersteunende toetsen							10	10
Oslo-toetsen⁷⁴		6, 11		11		7	4, 7	
Overgangstoetsen⁷⁵		4, 7		4, 7				

4.2.2 Nationale toetsen

Nationale toetsen (*Nasjonale Prøver*) maken sinds 2004 onderdeel uit van het nationale toetssysteem, en hebben in het eerste decennium van hun bestaan al enkele revisies ondergaan om te komen tot de huidige vorm.

De nationale toetsen zijn verplicht voor alle leerlingen van het vijfde leerjaar (lager onderwijs) en het achtste en negende leerjaar (lager secundair onderwijs) in alle scholen, met uitzondering van leerlingen in het buitengewoon onderwijs of leerlingen in een programma om de Noorse taal te leren

⁷² Tot en met 2015 ook in het elfde leerjaar.

⁷³ Tot en met 2015 ook in het elfde leerjaar.

⁷⁴ Enkel opgenomen ter vervollediging, want de Oslo-toetsen behoren niet tot het federale aanbod van gestandaardiseerde toetsen. Zie sectie 4.3.1 voor meer informatie in verband met de Oslo-toetsen.

⁷⁵ Enkel opgenomen ter vervollediging, want de overgangstoetsen behoren niet tot het federale aanbod van gestandaardiseerde toetsen. Zie sectie 4.3.2 voor meer informatie.

(anderstalige nieuwkomers)⁷⁶. De toetsen beogen een evaluatie van de vaardigheid van leerlingen in drie leerdomeinen, met name: leesvaardigheid in het Noors⁷⁷ (Sami voor *fykker* waar deze taal gesproken wordt), rekenvaardigheid en Engels⁷⁸. De eerste twee leerdomeinen zijn elk gerelateerd aan één van de vijf fundamentele vaardigheden die worden vastgelegd in het nationale curriculum en zijn bijgevolg niet gelinkt aan het subject-specifieke curriculum van bijvoorbeeld Noors of wiskunde. Voor Engels wordt wel het subject-specifieke curriculum van Engels gevolgd. Door de toetsen onafhankelijk van de leerdoelen in de verschillende leerdomeinen op te stellen (wat betreft lees- en rekenvaardigheid), wordt de kritiek vermeden dat bepaalde leerdomeinen belangrijker worden geacht dan andere, wat zou leiden tot een vernauwing van het curriculum tot de getoetste leerdomeinen:

"These fundamental skills, they are defined across all the other subjects, so you're testing reading in the natural sciences, reading in social sciences, reading in all the other things. So there is not any subject that gets pushed up (...) Even the music teacher for instance would see that reading is a basic skill in order to learn music because you have to read about it, to understand texts about music. I think it is a fantastic idea."(interview, respondent NW9)

De toetsen worden in het eerste trimester van het schooljaar afgenomen. Elk van de toetsen duurt maximaal 90 minuten. De toetsen bevatten zowel relatief moeilijke als relatief makkelijke items, om zo een accuraat beeld te schetsen van alle leerlingen. De toetsen voor het begin van het vijfde leerjaar zijn gebaseerd op de leerdoelstellingen voor het vierde leerjaar, en de toetsen voor het begin van het achtste en negende op de leerdoelstellingen voor het zevende leerjaar. De toetsen van het achtste en negende leerjaar zijn identiek, waardoor enerzijds kan nagegaan worden wat de voortgang van leerlingen is, en waardoor anderzijds aan scholen de mogelijkheid wordt gegeven om specifieke doelen te bepalen (op basis van de resultaten van het achtste leerjaar) die dan in het negende jaar kunnen geëvalueerd worden.

"We think that if we have the same tests in grade 8 and 9, we can see the difference in the results. They can see a difference in what we call 'learning outcome' of the reading in eighth grade and ... How much better can they read in the ninth grade, and can we set a proper goal on this?"(interview, respondent NW7)

De toetsen rekenvaardigheid en Engels worden digitaal afgenomen, waarbij zowel versies voor de PC als voor tablets beschikbaar zijn. De toetsen voor leesvaardigheid gebeuren op papier via testboekjes die twee weken op voorhand naar alle scholen worden verstuurd, al bestaat de ambitie om deze op korte termijn voor het secundair onderwijs ook digitaal te maken (en wellicht op langere termijn ook voor het lager onderwijs).

⁷⁶ De scholen zijn verantwoordelijk om de leerlingen aan te wijzen die om deze redenen een vrijstelling van toetsafname verdienen. Leerlingen of ouders kunnen hier evenwel beroep tegen aantekenen om alsnog deel te nemen aan de nationale toetsen.

⁷⁷ Gestandaardiseerde toetsen betreffende de vaardigheid van leerlingen in de Noorse taal, omvatten telkens de twee Noorse standaardtalen (Bokmål en Nynorsk).

⁷⁸ In een eerste versie werd ook een schrijfvaardigheidstoets afgenomen in de nationale toetsen. Deze werd na een eerste poging in 2005 echter niet meer opnieuw aangeboden.

Doelen

De nationale toetsen hebben een tweeledig doel. Enerzijds dienen ze om op gestandaardiseerde wijze een indicator te voorzien van de kennis en vaardigheden van leerlingen aan alle betrokkenen met name leerlingen, ouders, schoolleiders, gemeenten en nationale autoriteiten ten behoeve van het opvolgen, verbeteren en ontwikkelen van de kwaliteit van het onderwijssysteem. Anderzijds dienen de nationale toetsen op gestandaardiseerde wijze een indicator te voorzien van de sterktes en zwaktes van groepen van leerlingen zodat leraren en scholen hierop kunnen inspelen bij het lesgeven en het bepalen van het onderwijsaanbod.

Het eerste doel is verantwoordings- en ontwikkelingsgericht op micro-, meso- en macroniveau. Het tweede doel van de nationale toetsen is ontwikkelingsgericht op meso- en microniveau (ten behoeve van schoolbesturen, scholen, leraren en leerlingen), wat verklaart waarom de toetsen bij aanvang van het schooljaar worden afgenomen. De nationale toetsen zouden scholen een geobjectiveerd inzicht moeten geven in de prestaties van de leerlingen, en een benchmark met leerlingen in andere scholen.

De resultaten van de nationale toetsen hebben uitdrukkelijk geen gevolgen op de individuele schoolloopbaan van de leerlingen. Voor leerlingen gelden wel meer ontwikkelingsgerichte doelstellingen. Aanvullend wordt immers gesteld op de website van het Directorate for Education and Training dat men door de nationale testen wil bekomen dat leerlingen:

- een beter inzicht krijgen in wat er van hen wordt verwacht;
- feedback krijgen over de kwaliteit van hun vaardigheden;
- advies krijgen over hoe ze kunnen verbeteren;
- betrokken worden bij hun eigen leerproces en beoordeling.

De eerste twee doelen kunnen door de nationale toetsen rechtstreeks worden bereikt, de twee laatste doelen zijn afhankelijk van het gebruik van de resultaten door de klasleraren. Daartoe wordt door de beleidsmakers beklemtoond dat het belangrijk is dat leerlingen er in moeten kunnen vertrouwen dat deze resultaten enkel zijn bedoeld om hun leerproces te ondersteunen.

Het is bij wet verboden om de resultaten van de nationale toetsen te gebruiken in de evaluatie van leraren. Dit is opmerkelijk in het licht van de grote autonomie die aan schoolbesturen wordt gegeven.

Ontwikkeling

De toetsen worden ontwikkeld door academici en daartoe aangestelde toetsontwikkelaars. Dit proces gebeurt onder toezicht van het Directorate for Education and Training volgens voorafbepaalde kwaliteitsvereisten.

De toetsen worden jaarlijks volledig vernieuwd. Tot 2013 werden de toetsen van verschillende jaren niet aan elkaar gekoppeld. De cesuren tussen verschillende vaardigheidsniveaus werden elk jaar opnieuw bepaald. Sinds 2014 wordt bij de toetsen Engels en rekenvaardigheid elk van de items op een IRT-schaal ingeschaald, zodat de moeilijkheidsgraad van de toetsitems vergeleken kan worden. Hierdoor worden sinds 2015 vergelijkingen doorheen de tijd voor scholen (en voor het onderwijsbeleid) mogelijk gemaakt. Vanaf 2016 wordt dezelfde techniek ook toegepast voor de toetsen

leesvaardigheid, waardoor vanaf 2017 ook voor deze vaardigheid longitudinale gegevens beschikbaar zullen zijn. De vergelijkbaarheid van de moeilijkheidsgraad van toetsen wordt ook gegarandeerd door het werken met anker-items die in verschillende opeenvolgende toetsen worden opgenomen. Niet alle leerlingen krijgen de extra anker-items voorgeschoteld. De anker-items worden ad random aan de toetsen van 6% van de leerlingen toegevoegd. De items van verschillende toetsen (in verschillende leerjaren) worden evenwel niet op dezelfde schaal ingeschaald.

Het Directorate for Education and Training benadrukt dat de toetsontwikkeling een tijdrovend proces is. Gemiddeld is de ontwikkeling van een toets over twee jaar gespreid. Een idee voor de nabije toekomst is om de toetsen in de toekomst als adaptieve toetsen aan te bieden. De technische vereisten zijn hier al voor beschikbaar, maar naar de schatting van het Directorate for Education and Training is hiervoor een twintig- of dertigvoud van het aantal items vereist.

Voor de digitalisering van de toetsen doet het Directorate for Education and Training beroep op Inspira Assessment. Inspira is een bedrijf dat zich specifiek richt op de ontwikkeling van de technische aspecten van digitale toetsen. De nationale toetsen worden aangeboden via een *central hosting* via een *cloud-based platform*, waardoor er bij de afname veel minder problemen ontstaan door de uiteenlopende systemen die in verschillende scholen worden gebruikt. Via Inspira werden in 2014 door verschillende opdrachtgevers ongeveer 3.200.000 toetsen afgenomen.

Afname en correctie

De afname van de nationale toetsen valt onder de bevoegdheid van de schoolbesturen, die ervoor moeten zorgen dat er voldoende personele en technische middelen ter beschikking zijn. De schoolbesturen zijn ook verantwoordelijk om toezicht te houden op hun scholen opdat de afname op een eerlijke manier gebeurt.

In principe nemen alle leerlingen de toetsen op hetzelfde ogenblik af. Omwille van het low-stakes karakter op leerlingenniveau kan het echter dat bepaalde leerlingen (die bijvoorbeeld ziek waren op het moment van de toetsafname) de toetsen op een later tijdstip afleggen.

De digitale toetsen van Engels en rekenvaardigheid worden automatisch gecorrigeerd. De papieren testboekjes van leesvaardigheid worden door de leraren gecorrigeerd, aan de hand van een verbeter sleutel die werd opgesteld door de toetsontwikkelaars.

Tijdens de toetsafname wordt vanuit Inspira Assessment voortdurend opgevolgd of er technische problemen ontstaan. Men gaat er prat op dat men op elk moment kan zien waar er welke toetsen worden afgenomen, en in welke school er mogelijke problemen dreigen. Dit doet men onder meer wanneer er al enige tijd geen activiteit in een bepaalde opgestarte toets wordt opgetekend. Vanop afstand kan men meteen ingrijpen en backups maken van de afgelegde toetsen wanneer bijvoorbeeld browserproblemen voorkomen bij de scholen.

Daarnaast voorziet Inspira Assessment ook een telefonische technische ondersteuning tijdens de toetsafname. Deze hulplijn wordt door scholen sterk gewaardeerd.

Rapportage en publicatie

Bij de rapportage van de resultaten wordt elk van de leerlingen ingedeeld in een bepaald vaardigheidsniveau ('*mestringnivå*'). De leerlingen van het vijfde leerjaar worden ingedeeld in drie vaardigheidsniveaus, terwijl er vijf vaardigheidsniveaus zijn voor leerjaar 8 en 9.

Het beleid inzake openbaarheid van resultaten is in de korte bestaansperiode van de nationale toetsen al tweemaal gewijzigd. Aanvankelijk werden de geaggregeerde resultaten van leerlingen per school publiek bekend gemaakt. Van deze openbaarheid werd verwacht dat het een stimulus zou vormen voor scholen om de leerprestaties van leerlingen te versterken. De openbaarheid leidde ertoe dat de media jaarlijks ranglijsten opmaakten van de verschillende scholen op basis van de behaalde leerprestaties van leerlingen op het gebied van lezen, rekenen en Engels voor de jaren 5, 8 en 9. De focus van de media lag daarbij voornamelijk op de scholen die uitzonderlijk goede, of net uitzonderlijke slechte resultaten konden voorleggen.

Deze openbaarheid van de toetsresultaten op schoolniveau werd sterk bekritiseerd door verschillende stakeholders, en leidde in 2004 ook tot meerdere pogingen om de afname van de eerste reeks nationale toetsen te boycotten of saboteren door de vragen op voorhand online prijs te geven⁷⁹.

Onder meer omwille van deze kritiek werd de openbaarheid van de resultaten op schoolniveau na een eerste proeffase afgeschaft. Niet toevallig kwam deze aanpassing na een regeringswissel, waardoor het politieke draagvlak voor de openbaarheid van resultaten verdween. Enkel de resultaten op niveau van het schoolbestuur (in casu: de gemeente of de provincie) waren nog publiek beschikbaar. Ook van deze data werd door de media gebruik gemaakt om rankings op te maken, maar dan op het niveau van de gemeenten en provincies.

Door een recente beslissing in 2014, in lijn met een beleid van openbaarheid van bestuursdocumenten, werden ook de resultaten op schoolniveau opnieuw ter beschikking gesteld van het publiek. Deze resultaten worden via een daartoe ontwikkeld Schoolportaal ('*Skoleporten*') gecommuniceerd⁸⁰. Dit portaal biedt aan iedereen de mogelijkheid om de gemiddelde resultaten (incl. spreiding) van leerlingen in een bepaalde school te vergelijken ten opzichte van de gemeentelijke, provinciale en nationale parameters. Een voorbeeld rapportage is voorzien in Figuur 30⁸¹. De publieke rapportage bevat informatie over groepen van leerlingen, maar geen gedetailleerde informatie over individuele leerlingen.

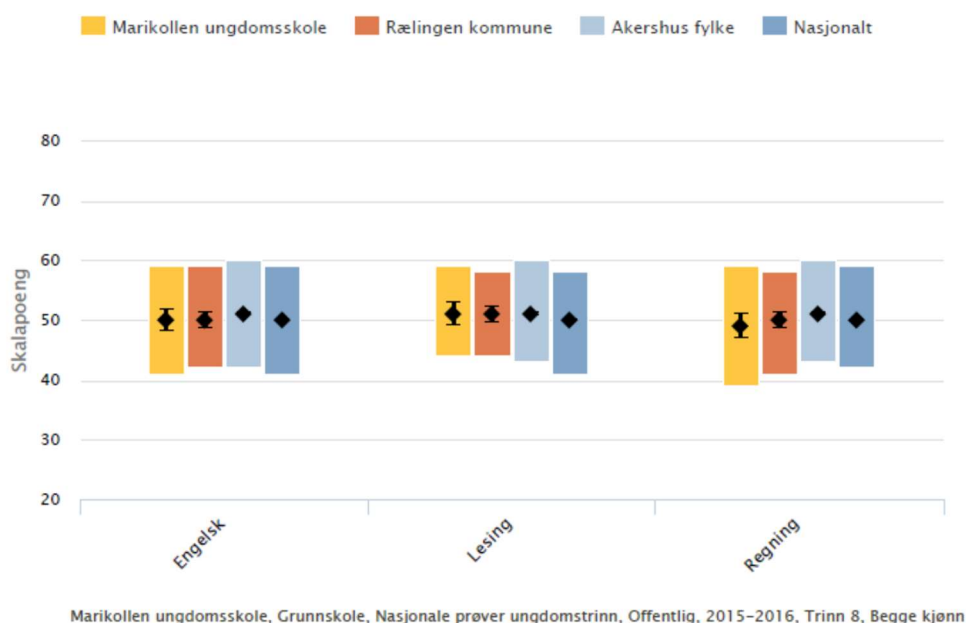
Het Directorate for Education and Training beklemtoont de verantwoordelijkheid van scholen om aan ouders uit te leggen dat de resultaten wel ter beschikking zijn, maar toch in de eerste plaats dienen om de interne kwaliteitszorg van de school en de leerprocessen van de leerlingen te ondersteunen.

⁷⁹ Hoewel de openbaarheid van de resultaten de belangrijkste reden was van het protest, was dit protest ook gericht op de gestandaardiseerde toetsen zelf. Bezwaren als het risico op vernauwing van het curriculum en de bedreiging voor de professionaliteit van leraren, werden genoemd als twistpunten die het protest in hand werkten.

⁸⁰ <https://skoleporten.udir.no/> (enkel beschikbaar in Noors)

⁸¹ met *Ungdomsskole*, *kommune*, *fylke* en *nasjonalt* als termen voor de school, de gemeente, de provincie en het nationale niveau; *Engelsk*, *Lesing* en *Regning* staan voor Engels, leesvaardigheid en rekenvaardigheid

De publiek bekende informatie leidt er nu ook toe dat er vergelijkingen en rankings van gemeenten/provincies en van individuele scholen worden opgemaakt in de media,⁸² ook al zeggen de resultaten door de afname van de toetsen in het begin van schooljaar op zich eigenlijk weinig over de



Figuur 30. Voorbeeld rapportage uit Skoleporten van een school in Rælingen kommune

kwaliteit van de school (zeker wat betreft het achtste jaar op schoolniveau, waar de leerlingen nog maar net in de secundaire scholen net zijn ingestroomd).

Naast deze publiek beschikbare informatie, krijgen scholen ook een overzicht van:

- de gedetailleerde prestaties van hun elk van eigen leerlingen (zie Figuur 31, waarbij de itemnummers in het groen zijn weergegeven bij een correct antwoord, en in het rood bij een foutief antwoord van de leerling);
- de globale prestaties van hun leerlingen per klasgroep volgens de vaardigheidsniveaus (zie Figuur 32);
- de prestaties van de verschillende klasgroepen in verhouding tot de prestaties op school-, gemeentelijk, provinciaal en nationaal niveau, volgens de vaardigheidsniveaus (zie Figuur 33);
- een detailanalyse op itemniveau, inclusief het percentage leerlingen dat het item in de school en op nationaal niveau correct heeft beantwoord, en de inschaling van het item op de IRT-schaal (*skalapoeng*) (zie Figuur 34).
- een vergelijking tussen de resultaten van dit jaar en de resultaten van het vorige jaar (vorige cohorte leerlingen)

Elk van deze deelrapporten is apart beschikbaar voor elk van de drie toetsdomeinen (Engels, leesvaardigheid, rekenvaardigheid).

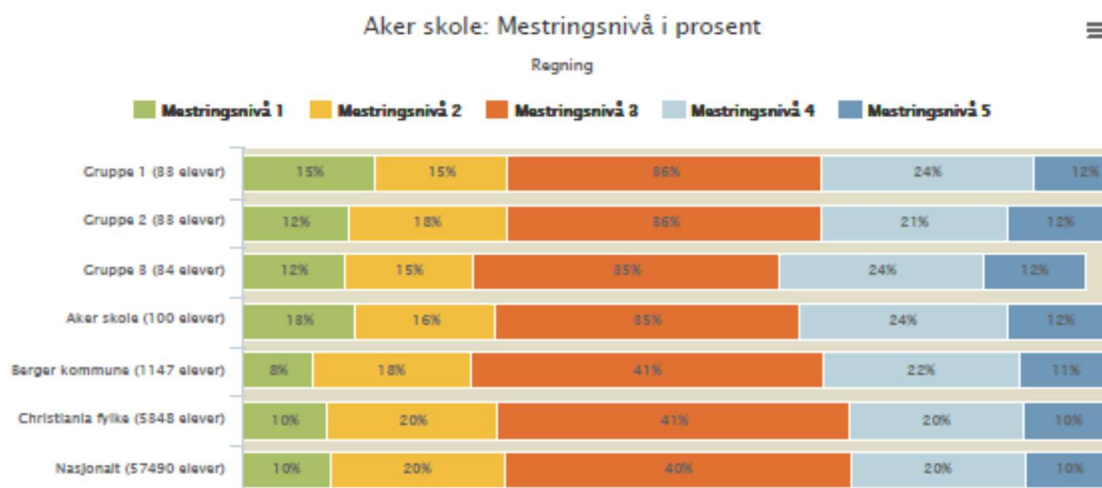
⁸² Bijvoorbeeld Berge (2015).



Figur 31. Voorbeeld rapportage uit digitaal feedbackrapport voor de school in verband met de antwoorden van een leerling op de items, volgens vaardigheidsniveaus.



Figur 32. Voorbeeld rapportage (fragment) uit digitaal feedbackrapport voor de school in verband met de prestaties van leerlingen in een klasgroep, volgens vaardigheidsniveaus.



Figur 33. Voorbeeld rapportage (fragment) uit digitaal feedbackrapport voor de school in verband met de prestaties van de verschillende klasgroepen, volgens vaardigheidsniveaus

Oppgave	Gruppe	Nasjon	Skalapoeng	Innhold	Område	Fag	Format
1	81 %	81 %	34	Dobling-/gjentatt addisjon desimaltall	Tall og algebra	ma	Åpen
2	70 %	72 %	39	Tolke informasjon i tabell	Statistikk og sannsynlighet	no, ma	Flervalg
3	59 %	50 %	49	Relatere negative tall til tallinja (temperatur), addisjon	Måling og geometri	ma, krls, sf	Åpen
4	78 %	76 %	40	Brek som del av en hel	Tall og algebra	m&h, ma, mu	Åpen
5	51 %	52 %	49	Brek som del av en mengde	Tall og algebra	m&h, ma	Åpen
6	63 %	60 %	45	Multiplikasjon av hele tall, sammenhengen måned og år	Måling og geometri	Sf, ma	Åpen
7	70 %	74 %	38	Beregne tidsintervall, analog og digital tid	Måling og geometri	ma, kro, m&h	Åpen
8	84 %	73 %	40	Geometriske egenskaper til kvadrat	Måling og geometri	ma, k&h	Åpen
9	72 %	66 %	43	Dobling/gjentatt addisjon/multiplikasjon og halvering/gjentatt subtraksjon/divisjon desimaltall	Tall og algebra	m&h, ma	Flervalg
10	75 %	76 %	34	Sammenligne størrelser	Måling og geometri	k&h, ma	Flervalg

Figur 34. Voorbeeld rapportage (fragment) uit digitaal feedbackrapport voor de school in verband met itemanalyse op schoolniveau

Het onderwijsbeleid motiveert scholen om deze resultaten op individuele basis met leerlingen te bespreken (met name de informatie in Figuur 31 en 32, maar mogelijk ook de informatie op school-, gemeentelijk, provinciaal en federaal niveau) om zo de leerlingen maximaal zicht te geven op hun eigen vaardigheden en hen te stimuleren om de verantwoordelijkheid voor hun eigen leerproces meer in handen te nemen.

Een algemene analyse van de resultaten op macroniveau verschijnt enkele maanden na de afname van de toetsen op de website van Statistics Norway⁸³. In deze analyse wordt gerapporteerd welk percentage van leerlingen op de verschillende vaardigheidsniveaus zijn ingeschaald. De resultaten kunnen bekeken worden van de volledige leerlingenpopulatie, of vergeleken tussen verschillende

⁸³ De resultaten van de nationale toetsen afgenomen eind 2015 worden besproken op <http://www.ssb.no/en/utdanning/statistikker/nasjprov>

groepen (bijvoorbeeld volgens geslacht, autochtone leerlingen versus leerlingen met een migratie-achtergrond, of volgens het opleidingsniveau van de ouders).

4.2.3 Diagnostische toetsen

Naast de nationale toetsen bestaan er ook diagnostische toetsen (*kartleggingsmateriell*, afwisselend vertaald als *screening tests*, *mapping tests* of *diagnostic tests*). De volgende toetsen zijn beschikbaar voor scholen: leesvaardigheid (eerste, tweede, derde leerjaar); rekenvaardigheid (eerste, tweede, derde leerjaar); Engels (derde leerjaar) en digitale vaardigheden (vierde leerjaar). Tot 2015 bestonden ook diagnostische toetsen in Engels en leesvaardigheid voor het elfde leerjaar, maar deze worden niet langer aangeboden.

De toetsen van Engels en digitale vaardigheden worden meestal afgenomen in februari; de toetsen reken- en leesvaardigheden in maart of april. De toetsen voor digitale vaardigheden en Engels zijn digitale toetsen. De toetsen inzake lees- en rekenvaardigheid worden op papier afgenomen. De toetsafname duurt gemiddeld 45 tot 60 minuten.

Doelen

Het uitgangspunt van deze toetsen is dat scholen ondersteuning nodig hebben bij het identificeren van de ongeveer 20% van de leerlingen die nood hebben aan extra ondersteuning. Diagnostische toetsen hebben als doel om de specifieke noden voor ondersteuning van individuele leerlingen te detecteren. De toetsen hebben dan ook een uitgesproken ontwikkelingsgericht doel op het microniveau.

Omwille van de gerichtheid van de diagnostische toetsen op de '*bottom 20%*' is er geen verplichting om deze toetsen van alle leerlingen af te nemen. De afname is in principe enkel zinvol van de leerlingen die mogelijk deze extra ondersteuning nodig hebben. Het diagnosticerende doel van deze toetsen maakt dat de toetsitems voor de doorsnee leerlingen relatief gemakkelijk zijn. De toetsresultaten differentiëren dan ook helemaal niet tussen de meer vaardige leerlingen. Scholen of schoolbesturen kiezen autonoom of ze de toetsen gebruiken, en bij welke leerlingen ze de toetsen afnemen.

Het ontwikkelingsgerichte karakter uit zich ook in het feit dat de diagnostische toetsen niet op zich staan, maar worden gekoppeld aan een uitgebreide database van materialen en richtlijnen die kunnen gebruikt worden bij specifieke leerlingenprofielen.

Ontwikkeling en correctie

Omwille van het ontwikkelingsgerichte karakter zouden de toetsen in principe jaar na jaar kunnen worden hergebruikt. Toch kiest men ervoor om vijfjaarlijks een nieuwe set van diagnostische toetsen te ontwikkelen. Er blijkt immers dat de diagnostische toetsen steevast vanaf het tweede jaar van ingebruikname een deel van hun diagnostische waarde verliezen. Waar bij de eerste afname in elke toets zoals voorzien ongeveer 20% van de leerlingen een score behaalt onder de cesaar voor 'nood aan extra ondersteuning', zakt dat vanaf het tweede jaar meestal naar ongeveer 15%, afhankelijk van de toets. Mogelijk is het onderwijs dus verbeterd waardoor minder leerlingen zwakker scoren, maar

aangezien bij de ingebruikname van een nieuwe toets telkens opnieuw 20% van de leerlingen onder deze cesuur scoort, is dat wellicht geen correcte interpretatie van de daling:

“When they set the score for the 20%-score, they did it the first year. So that means if the student average gets better, (...) it can also be that the teachers use questions from the test, so it is kind of a leak: the tests get ‘worn out’, they say.” (interview, respondent NW3)

De digitale toetsen voor Engels en digitale vaardigheden worden automatisch verbeterd. De resultaten worden automatisch weggeschreven naar het registratiesysteem waarin alle data van de leerlingen zijn opgeslagen (al stonden technische problemen deze automatische transfer in het verleden al enkele malen in de weg). De diagnostische toetsen van lees- en rekenvaardigheid moeten door de leraren worden gecorrigeerd.

Rapportering

Scholen krijgen een overzicht van de resultaten van alle deelnemende leerlingen, inclusief de identificatie van de leerlingen die volgens de diagnostische toetsen extra ondersteuning nodig hebben. De resultaten van de diagnostische toetsen worden niet gepubliceerd, waardoor er geen vergelijking tussen scholen mogelijk zijn. Ook analyses op macroniveau zijn aan de hand van de resultaten van de diagnostische toetsen niet mogelijk.

4.2.4 Eindexamens

Het eindcijfer van het lager secundair onderwijs (en daarmee het leerplichtonderwijs) wordt in de eerste plaats bepaald door de evaluaties die leraren maken van de individuele leerlingen. Voor één leerdomein echter wordt het eindcijfer samengesteld door enerzijds een schriftelijke gestandaardiseerd examen dat gelijk is voor alle leerlingen, en anderzijds door een mondeling examen dat onder de bevoegdheid van de schoolbesturen wordt afgenomen. De examens meten de kennis en vaardigheid van alle leerlingen in één van de leerdomeinen Noors, wiskunde, en Engels.

Doelen

Er wordt van elke leerling een eindscore berekend die wordt samengesteld op basis van zowel het eindexamen, als de evaluaties die in de school door de eigen leraren zijn uitgevoerd. Door het gewicht van dit examen op het eindcijfer van de leerlingen, is deze toets expliciet verantwoordingsgericht op het microniveau: de leerlingen moeten hier tonen in welke mate ze aan de einddoelen van het leerplichtonderwijs voldoen.

Toch wordt het belang van het eindcijfer ook gerelativeerd: zo blijkt quasi elke leerling het diploma van het leerplichtonderwijs te behalen (in de uitzonderlijke situaties waarin dat niet het geval is, is dat slechts uiterst zelden veroorzaakt door slechte leerresultaten); en hoewel het eindcijfer wordt geacht richtinggevend te zijn voor de keuze voor vervolgonderwijs, blijft deze beslissing de autonomie van de leerlingen en ouders. Er worden bijvoorbeeld geen leerlingen uitgesloten van de algemene studierichtingen in het hoger secundair onderwijs.

Uitzonderlijk stellen vervolgscholen (voornamelijk in Oslo, met een relatief competitieve onderwijsmarkt in het hoger secundair onderwijs) evenwel minimum-scores in als basis voor toelating van leerlingen.

Ontwikkeling en correctie

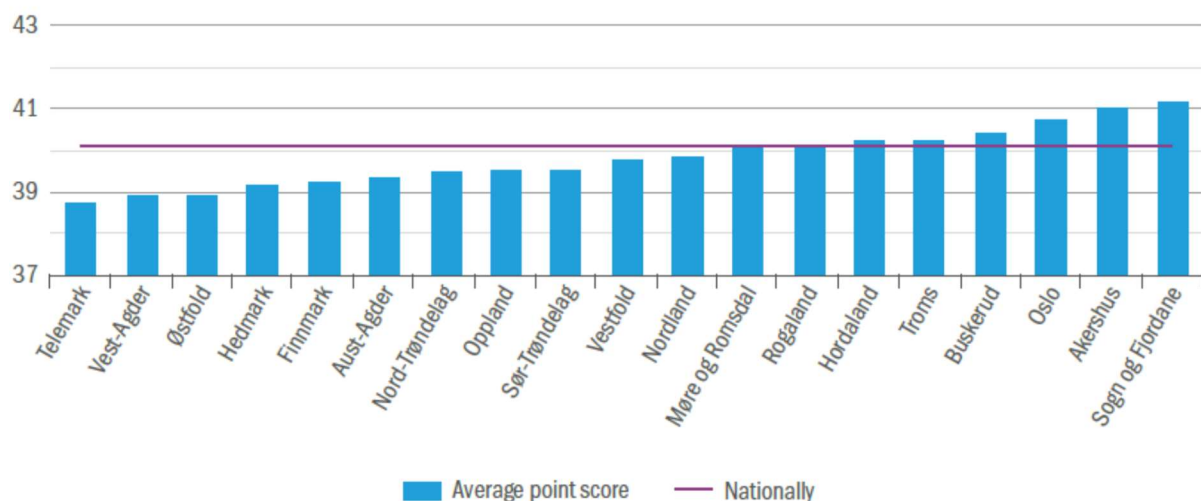
De toetsen worden ontwikkeld door een commissie van leraren en inhoudsexperten die wordt samengesteld door het Directorate for Education and Training. Dit gebeurt binnen dit Directorate evenwel in een dienst die volledig onafhankelijk werkt van de dienst die verantwoordelijk is voor de nationale toetsen, diagnostische toetsen, en andere gestandaardiseerde toetsen.

Omdat de *stakes* van het eindexamen relatief beperkt blijven (zeker in vergelijking met eindexamens in de meeste andere landen) acht men het geen probleem dat de correctie van de eindexamens overgelaten wordt aan de verantwoordelijkheid van de schoolbesturen. Gewoonlijk wordt elk examen verbeterd door twee onafhankelijke evaluatoren.

Elk van de leerlingen wordt in elk van de getoetste leerdomeinen ingedeeld in één van de vaardigheidsniveaus van 1 tot 6.

Rapportering

De individuele resultaten op leerling- of schoolniveau worden niet publiek bekend gemaakt. De gemiddelde resultaten van de leerlingen op het niveau van de provincies worden wel gepubliceerd in de jaarlijkse Education Mirror (uitgegeven door het Directorate, zie Figuur 35). In dit rapport worden ook opvallende resultaten gepubliceerd. Zo werd in de Education Mirror 2014 bijvoorbeeld het verschil in prestaties tussen jongens en meisjes aangekaart.



Figuur 35. Gemiddelde prestaties op eindexamen per provincie (Norwegian Directorate for Education and Training, 2014)

4.2.5 *Andere toetsen*

Drie andere toetssystemen worden niet tot *The Knowledge Promotion* gerekend, al betreft het evenzeer gestandaardiseerde toetsen die aan scholen worden aangeboden. Het gaat om leerondersteunende toetsen en beoordelingsondersteunende toetsen.

De leerondersteunende toetsen alsook de beoordelingsondersteunende toetsen zijn beide relatief recent ontwikkelde databanken van digitale toetsen die door scholen facultatief kunnen worden ingezet. De leerondersteunende toetsen bestaan uit toetsen inzake leesvaardigheid (zesde en elfde leerjaar) en rekenvaardigheid (zesde en elfde leerjaar), schrijfvaardigheid (vijfde en achtste leerjaar), digitale vaardigheid (achtste leerjaar) en Engels (elfde leerjaar). Deze toetsen hebben tot doel om leerlingen te ondersteunen door hen een beter zicht te geven op hun eigen vaardigheidsniveau. Daarnaast dienen ze om leraren extra ondersteuning te geven om gerichte feedback te geven aan leerlingen en om de validiteit van hun eigen evaluatie te bevestigen. De beoordelingsondersteunende toetsen hebben eerder tot doel om na te gaan of het eigen oordeel van leraren inderdaad de meest gepaste inschatting is van de vaardigheid van leerlingen. Bewust werd er dan ook voor gekozen om de focus te leggen op twee leerdomeinen die niet aan bod komen in het eindexamen aan het eind van het lager secundair onderwijs. Er zijn toetsen beschikbaar voor natuurwetenschappen en sociale studies, beiden voor het tiende leerjaar (wanneer de leraar een summatieve beoordeling van de leerlingen moet opmaken). Beide toetsdatabanken worden ter beschikking gesteld van scholen op een online systeem. Scholen kunnen zelf kiezen of, en wanneer ze deze toetsen gebruiken. De toetsen in beide toetsdatabanken zijn, in tegenstelling tot de nationale toetsen en diagnostische toetsen, gerelateerd aan het nationale curriculum in de betreffende leerdomeinen. Elk item wordt aan een specifiek doel in het curriculum gelinkt.

4.2.6 *Impact*

De nationale toetsen hebben tot doel dat het leren van leerlingen en de interne kwaliteitszorg van scholen wordt gestimuleerd. Het Directorate for Education and Training benadrukt de verantwoordelijkheid van scholen en schoolbesturen om de resultaten aan te wenden om te komen tot een versterking van de kwaliteit van het onderwijs. Daarnaast wordt beklemtoond dat scholen aan leerlingen en ouders de ontwikkelingsgerichte doelstelling van de nationale toetsen duidelijk moeten maken, en dat men de resultaten op leerlingenniveau dient te plaatsen binnen de ruimere context van informatie die men over de leerlingen via observatie en andere evaluatievormen verworven heeft. Of scholen en schoolbesturen effectief aan de slag gaan met de resultaten, wordt evenwel niet gecontroleerd. Wel voorziet men een nascholing waarin experts van het Directorate for Education and Training samen met toetsontwikkelaars ondersteuning bieden aan de aanwezige schoolleiders bij de interpretatie en het gebruik van de toetsresultaten.

Aanvaarding

Intussen menen de meeste respondenten in deze case studie wel dat de nationale toetsen algemeen aanvaard zijn bij de meeste leraren en directies, onder meer door de hoge kwaliteit van de toetsen. Bij

het Directorate for Education and Training is men ervan overtuigd dat, hoewel er zeker nog kritiek is op de nationale toetsen (zie sectie 4.2.7), er doorheen het eerste decennium van het bestaan van deze toetsen een algemeen inzicht is gegroeid in de meerwaarde van de nationale toetsen.

Nog steeds volgens het Directorate is de aanvaarding nog gestegen sinds men in 2015 bij de nationale toetsen longitudinale gegevens op schoolniveau ter beschikking kan stellen (door de introductie van IRT-inschaling en anker-items):

In December [2014] we published the new kind of results for the first time. Schools' reaction is very positive, because they got much more information than they had before. What they had before, was ... all tests were divided into benchmarks, in the fifth grade there were three levels, or benchmarks, with two cut scores; and in the eighth grade we had four or five levels. This was done over and over again each year. So it was new each year, it felt unfortunate, it was not possible to compare one year to another (interview, respondent NW9).

Verschillende respondenten geven aan dat de diagnostische toetsen intussen ruim aanvaard zijn door de meerderheid van leraren in het onderwijsveld, en ook in veel scholen worden gebruikt. Hiervoor worden globaal twee redenen gegeven: enerzijds de koppeling van de toetsresultaten aan ondersteunende materialen die kunnen worden ingezet bij de remediëring van leerlingen; en anderzijds het - in principe - louter ontwikkelingsgericht perspectief waarin de resultaten kunnen gebruikt worden. Toch geven verschillende respondenten aan dat sommige schoolbesturen ook de diagnostische toetsen een zekere verantwoordingsgerichte invulling geven, bijvoorbeeld door voor de afname op schoolniveau streefdoelen voorop te zetten (bijvoorbeeld: maximaal 10% van alle leerlingen onder de cesuur). De vaststelling is dat dit laatste voorbeeld *teaching to the test* in de hand werkt, waardoor de diagnostische waarde verloren gaat.

Bovendien geldt de aanvaarding van de diagnostische toetsen dan wel voor het leerplichtonderwijs, maar kan ze niet veralgemeend worden naar het hoger secundair onderwijs. Verschillende respondenten rapporteren dat er bij leraren hoger secundair onderwijs slechts in beperkte mate vertrouwen bestond in de diagnostische toetsen. Bijgevolg bleef ook het aantal gebruikers ervan beperkt, behalve in de zeldzame gevallen waar dit door een schoolbestuur werd verplicht (zoals onder meer in Oslo, zie sectie 4.3.3). Daarom werd besloten door het Directorate for Education and Training om geen investeringen (en dus geen aanpassingen) meer te doen in de tests van het elfde leerjaar. Omdat in 2015 werd vastgesteld dat de toetsen hun diagnostische waarde te zeer hadden verloren bij gebrek aan deze aanpassingen, werd beslist om de toetsen van het elfde leerjaar vanaf 2016 niet langer beschikbaar te stellen.

"I think it is a good thing that it is discontinued because people are very confused about what they were measuring. Because it's a completely different thing: measuring reading in the first to third grade and measuring reading when the kids are fifteen or sixteen, so it's a completely different thing!" (interview, respondent NW9)

Bij gebrek aan empirische evidentie blijft het moeilijk te zeggen in welke mate de hierboven besproken veronderstelling dat de aanvaarding van de nationale toetsen en diagnostische toetsen doorheen het voorbije decennium vergroot is, effectief in scholen waar te nemen is.

Zowel de leerondersteunende toetsen als de beoordelingsondersteunende toetsen zijn nog maar recent geïntroduceerd in het Noorse onderwijssysteem, en zijn nog in ontwikkeling. Momenteel is de bekendheid ervan nog relatief beperkt, waardoor ze ook nog niet zo vaak gebruikt worden door scholen.

Impact op microniveau

Volgens respondenten in het Directorate for Education and Training was het gebruik van de toetsen door leraren aanvankelijk relatief beperkt, maar komt daar stilaan verandering in. Men meent dat leraren zijn beginnen inzien dat er meer uit te leren valt dan men oorspronkelijk dacht:

“They didn’t know that they didn’t know” (interview, respondent NW7).

Toch wordt deze stellingname door andere respondenten tegengesproken: zo menen de bevroegde schooldirectie alsook de bevroegde academicus dat het leereffect voornamelijk voor individuele leraren beperkt is in termen van ‘nieuwe informatie over de kennis en vaardigheden van de leerlingen’. Dit is in lijn met de resultaten van een evaluatie door het Noorse Instituut voor Studies in Innovatie, Onderzoek en Onderwijs (NIFU). Deze evaluatie besloot dat hoewel de toetsen plichtsbewust en correct worden afgenomen door de scholen, het gebruik van de resultaten op leerling- en klasniveau relatief gering is⁸⁴. Een uitzondering gold voor de provincie Oslo, wat zal besproken worden in sectie 4.3.5. Uit deze studie blijkt ook dat het gebruik van de resultaten hoger was in het lager onderwijs dan in het lager secundair onderwijs. Voor leraren is de informatie die uit de nationale toetsen komt, te beperkt.

Een andere bron van empirische evidentie is de eerder vermelde bevraging van schooldirecties in 2005 (net nadat de nationale toetsen voor het eerst werden uitgetoetst)⁸⁵. Op dat ogenblik meende respectievelijk 14%, 36%, 38%, 12% en 1% van de directies dat de bijdrage van de nationale toetsen tot het verhogen van de leerresultaten van leerlingen respectievelijk onbestaande, klein, in zekere mate, groot, of zeer groot was. Met andere woorden: de meningen hierover waren sterk verdeeld. Ter vergelijking: 82% van de directies was ervan overtuigd dat de wettelijk bepaalde gesprekken tussen de scholen en leerlingen samen met hun ouders een (zeer) grote bijdrage hadden tot het verhogen van de leerprestaties, tegenover 13% bij de nationale toetsen.

Ook in de interviews blijkt dat weinig respondenten een inschatting kunnen maken van de mate waarin *The Knowledge Promotion* de leerresultaten van leerlingen heeft beïnvloed. Volgens de bevroegde directie draagt het hele evaluatiesysteem slechts in geringe mate bij tot veranderingen in de klaspraktijk. Daarvoor brengen de toetsen te weinig nieuwe informatie voor leraren, en is de periode tussen toetsafname en beschikbaarheid van de resultaten te groot. De diagnostische toetsen vormen hierbij de uitzondering, omdat wel kan gesteld worden dat in veel scholen specifieke initiatieven voor

⁸⁴ Seland et al., 2013.

⁸⁵ Onderzoek gerapporteerd door Skedsmo (2011).

extra ondersteuning gericht worden op de leerlingen die door deze volgens de resultaten van deze diagnostische toetsen nood aan hebben.

Impact op mesoniveau

Hierboven bespraken we dat volgens het eerder vermelde onderzoek van NUF⁸⁶ het gebruik op microniveau volgens leraren relatief beperkt was. Daar staat evenwel tegenover dat men in dezelfde studie vond dat meer dan 50% van bevroegde directies en schoolbesturen de resultaten van de nationale toetsen een zinvol gebruiksinstrument vonden voor verdere schoolontwikkeling. Het onderzoeksrapport stelt dat een belangrijk aspect van het effect op interne kwaliteitszorg er in bestaat dat er een meer geïnformeerde basis ontstaat voor schoolleiders om leraren te ondersteunen, maar ook om leraren te controleren en gericht aan te spreken bij eventuele tegenvallende resultaten. Dit effect wordt bevestigd door één van de respondenten:

“The national tests have really impact! Especially at the lower level, at the fifth level. This is the first time that teachers at the primary schools are controlled. When the national tests came, the school municipalities and the school headmasters were able to control the teachers: in your class is 3.2, in her class it is 3.8, but why?” (interview, respondent NW10)

Impact op macroniveau

Ook de impact van het evaluatiesysteem op macroniveaus tot op heden relatief beperkt. Hoewel de resultaten wel gebruikt worden binnen politieke discussies, zijn er (nog) geen aanpassingen aan het onderwijsbeleid die expliciet geënt zijn op de resultaten van de nationale toetsen. Wel merkt men dat meer en meer onderzoekers datasets opvragen om onderwijskundig onderzoek te verrichten.

4.2.7 Kritiek en neveneffecten

Er is enige empirische evidentie inzake enkele onwenselijke neveneffecten bij de afname van de nationale toetsen. Daarnaast worden in deze sectie verschillende punten van kritiek op het evaluatiesysteem aangehaald. Bij de introductie van The Knowledge Promotion en het bijbehorende uitgebreide evaluatiesysteem, was er protest van leraren, van ouders en van leerlingen. Hoewel de storm van protest intussen is geluwd, worden er nog heel wat kritische kanttekeningen bij de huidige praktijk inzake federale gestandaardiseerde toetsen.

Onwenselijke neveneffecten

De eerder vermelde evaluatie inzake de implementatie en het gebruik van nationale toetsen⁸⁷ rapporteert dat een minderheid van de scholen bijkomende tijd uittrekt om leerlingen voor te bereiden voor de leerdomeinen die in de nationale toetsen aan bod komen, waardoor andere leerdomeinen minder aandacht krijgen. Een belangrijk aandeel van de voorbereidingstijd gaan naar het trainen van leerlingen in het testformat, voornamelijk voor de digitale toetsen. Daarnaast geven leraren in het lager

⁸⁶ Seland et al., 2013.

⁸⁷ idem.

secundair onderwijs aan dat bepaalde leerlingen sterk tegen de toetsen op zien, wat wordt beaamd door ouders van de relatief zwakke leerlingen.

Gebrek aan leerwinstmeting

Een essentieel deel van de kritiek is gericht op het ontbreken van leerwinstmetingen van de groei van leerlingen. De verschillende toetsen zijn immers niet aan elkaar gekoppeld en laten dus geen betrouwbaar beeld over de groei van leerlingen doorheen de tijd door. Deze aanvulling zou - in tegenstelling tot wat nu het geval is - echt een nieuw licht kunnen laten schijnen op het leren van de verschillende leerlingen. De informatie die men wel krijgt, is naar aanvoelen van de bevraagde schooldirecteur, veel te weinig gedetailleerd. De vaardigheidsniveaus zijn daarvoor veel te breed. De leraren krijgen daardoor weinig specifieke informatie. De vele kleine scholen in Noorwegen maken het daarnaast moeilijk om betrouwbare uitspraken te doen en één of enkele uitzonderlijk sterke/zwakke leerlingen in een klasgroep kunnen de resultaten van één jaargroep sterk vertekenen.

"There is a willingness to look whether our work works, but this way, the testing is not motivating for the teachers, and it is not motivating for the pupils." (interview, respondent NW6)

Gebrek aan rationale

Het gebrek aan rationale van het evaluatiesysteem, de weinig transparante samenhang tussen de verschillende toetsen, en het feit dat aan zowat elke toets uiteenlopende doelstellingen worden toegeschreven, maakt bovendien dat leraren onvoldoende kunnen onderscheid maken tussen de geschiktheid van de verschillende toetsen voor verschillende doeleinden.

"As such, the system does not clearly differentiate between the needs of the individual student and the needs of the system, in terms of providing as well as using information for improvement purposes." (Skedsmo, 2011, p. 11)

Dit draagt bij tot een algemeen gevoel dat de toetsen vooral dienen afgenomen te worden omdat het nu eenmaal verplicht is, en niet omdat het de eigen school veel vooruithelpt:

"There are different test with different aims and different norms. That is very confusing for teachers (...) More education should be given to principals about what the purposes of these tests are. And how the tests can be used by the schools. Otherwise, the tests are only scaring people. And they are not helpful that way." (interview, respondent NW3)

Ten slotte wordt ook het bestaan van een evaluatiesysteem op zich nog steeds in zekere mate gecontesteerd. Een vraag die bijvoorbeeld in dit kader leeft is of het überhaupt mogelijk is dat leerresultaten de nodige informatie omvatten die vereist is om het onderwijssysteem en individuele leerprestaties te versterken.

"The aim of all these tests is not in a sufficient way oriented at the teacher and pupil level (...). An important condition is that is not an evaluation, that it is the teachers who

feel ownership (...) It is important for tests that, there is a need for a tool where the teachers can see themselves whether their teaching works. It requires a massive change in how one thinks about learning and teaching." (interview, respondent NW3)

Openbaarheid van de resultaten

Ook de vernieuwde publieke beschikbaarheid van de resultaten op schoolniveau blijft voor discussie zorgen. Deze openbaarheid voedt onder meer de vrees dat er verschillen zullen ontstaan in de criteria die scholen hanteren om leerlingen vrij te stellen voor deelname aan de nationale toetsen, wat een bedreiging inhoudt voor de betrouwbaarheid van de resultaten.

Kritiek op de openbaarheid van de resultaten op schoolniveau wordt nog sterker dan voor de andere toetsen geuit in verband met de nationale toetsen voor het achtste leerjaar. Op dat ogenblik, meteen na de start van het secundair onderwijs, zijn heel veel leerlingen nog maar enkele maanden in de school.

"For instance, in the eighth grade, those pupils came only one month before, so the schools can't be accountable for their achievement!" (interview, respondent NW10)

Een ander concreet punt van kritiek inzake de openbaarheid van de op schoolniveau geaggregeerde resultaten van leerlingen, is dat de ruwe scores worden gecommuniceerd: er wordt in de rapportering over nationale toetsen niet gecorrigeerd voor leerlingkenmerken.

Een onderzoek naar aanleiding van de openbaarheid van de eerste reeks nationale toetsen⁸⁸ stelde tot slot vast dat de tegenvallende resultaten (en de daaropvolgende negatieve media-aandacht) een katalysator kunnen zijn voor scholen om nooit meer in dezelfde situatie te verzeilen, zoals blijkt uit volgend citaat van een leraar: *"The issue with the national exams is in the back of my mind (poor results in 2004). And I think: 'Never again'"* (Elstad, 2009, p. 183). Dit leidde dan wel tot verbeteracties (soms paniecreacties), maar ook tot een sterke emotionele impact van stress, boosheid, schaamte en schuldgevoelens. Er zijn ook sterke reacties van ouders en lokale gemeenschap ten aanzien van de 'blamed'-scholen. De impact van de 'naming and shaming' voor scholen onderaan de ranking - zowel inzake mogelijke schoolontwikkeling, als inzake negatieve impact - hangt evenwel sterk samen met het leiderschap, de sociale normen en het innovatieve vermogen van scholen. Hierdoor is de impact op verschillende scholen erg uiteenlopend.

Voorwaarden voor goed gebruik onvoldoende ontwikkeld

Een andere kritiek omvat de vraag of de verschillende gebruikers - van micro tot macro - wel over de vereiste vaardigheden beschikken om de resultaten op een adequate manier te analyseren en te gebruiken in het licht van kwaliteitsverbetering, en of de huidige vorm van rapportering hen daarbij voldoende ondersteunt. De gebruiksvriendelijkheid van de huidige rapportering wordt onder meer door de geïnterviewde directie in vraag gesteld. Het wordt, aldus academisch onderzoek, te makkelijk verondersteld dat scholen over de nodige vaardigheden beschikken om de gegevens correct te interpreteren en om te zetten in verbeter-doeleinden:

⁸⁸ Zie Elstad (2009).

"In providing these data to the different actors; national authorities, municipalities, principals and teachers, it is presumed that the information automatically makes sense to them, and it is also taken for granted that they know how to use the data as well as what kind of actions they need to take in order to improve." (Skedsmo, 2011, p. 11)

Bovendien betekent de periode tussen de toetsafname en de beschikbaarheid van de resultaten bij de nationale toetsen (resultaten zijn niet automatisch beschikbaar) dat de waarde van de resultaten voor het gebruik op microniveau (door leraren, om bijvoorbeeld te differentiëren tussen leerlingen) sterk wordt aangetast. Een algemene kritiek is dan ook dat het design van het huidige evaluatiesysteem onvoldoende gericht is op het niveau van leerlingen en leraren.

Te weinig nieuwe informatie voor gebruik op microniveau

Zo blijft ook in verband met de diagnostische toetsen onduidelijk in welke mate deze toetsen volgens leraren werkelijk nieuwe informatie aanbrengen over de vaardigheid van leerlingen. Zowel bij het Directorate for Education and Training als bij het schoolbestuur van Oslo is men overtuigd dat dit wel het geval is:

"A lot of teachers would say that [that the tests do not bring new information about the pupils] Also some of our teachers [in Oslo] say that, but then ... it's probably not true. Because, of course, every good teacher will see what his students can, but in a class of maybe thirty pupils, you will probably not have enough information without good diagnostic tests." (interview, respondent NW3)

Om informatie over de kennis en vaardigheden van leerlingen in te winnen, geven leraren er vaker de voorkeur aan om andere toetsen te gebruiken. Het betreft toetsen die men zelf heeft ontwikkeld, of die school heeft aangekocht van educatieve uitgeverijen. Ook door de geïnterviewde directie wordt aangegeven dat leraren op dit ogenblik nog vaak meer waarde hechten aan de diagnostische waarde van hun eigen evaluaties, dan aan de waarde van de toetsen in het kader van *The Knowledge Promotion*. Ook de toetsen die leraren krijgen van educatieve uitgeverijen leveren volgens deze directie veel meer informatie op over de kennis en vaardigheden van de leerlingen.

Slechts beperkt aantal leerdomeinen

Een andere kritiek is gerelateerd aan het feit dat slechts drie leerdomeinen door de nationale toetsen worden opgenomen:

"I know some social studies teachers who say 'Why don't I have national tests in social studies? It is an important subject!' Some really demand a national test. This is not a general demand, but we had some 'noises' in that direction." (interview, respondent NW10)

Deze kritiek werd evenwel door geen enkele andere bron vernoemd. Ook de schooldirectie vond het op zich geen probleem dat de aandacht van de nationale toetsen voornamelijk gericht is op de huidige drie centrale leerdomeinen.

Afnameduur en -betrouwbaarheid

Specifiek in verband met de diagnostische toetsen is een ander punt van kritiek van zowel scholen als van ouders gerelateerd aan de afnameduur (45 tot 60 minuten). Volgens de critici is het voor kinderen in deze vroege fase van hun schoolloopbaan onwenselijk om toetsen van deze lengte af te leggen. Bovendien wordt gesteld dat de betrouwbaarheid van de resultaten er sterk negatief door zou worden beïnvloed.

Het eerder vermelde onderzoeksrapport van NUF⁸⁹ besluit dat er in sommige gevallen door verschillen in informatica-infrastructuur grote verschillen zijn in de toetsafnames tussen verschillende scholen, of zelfs tussen verschillende klassen. Deze verschillen worden als een bedreiging voor betrouwbaarheid van de toetsafname ervaren.

4.3. Gestandaardiseerde toetsen in de stad Oslo⁹⁰

De stad en de provincie Oslo zijn overlappend. De Stad Oslo, en met name het Bureau voor Onderwijs (*Utdanningsetaten*) is daardoor het schoolbestuur van bijna alle lagere en secundaire scholen in Oslo. Oslo is een stad die een sterke groei kent, waardoor het onderwijssysteem uit haar voegen barst. Elk jaar moeten scholen worden bijgebouwd of bestaande scholen worden uitgebreid. Tussen 2012 en 2014 steeg het aantal leerlingen in het leerplichtonderwijs met 5% van 56.476 tot 59.308 leerlingen, terwijl de leerling/(voltijds) leraar ratio bleef behouden op 12. Er is dus ook een grote aangroei van het onderwijspersoneel.

Sinds Noorwegen in het begin van deze eeuw een grotere autonomie aan schoolbesturen verleende, zijn er grote verschillen tussen hoe schoolbesturen aan kwaliteitszorg doen. De stad Oslo heeft, in tegenstelling tot de meeste schoolbesturen, een uitgebreid systeem van kwaliteitszorg uitgebouwd. In dit kwaliteitszorgsysteem staan de resultaten van gestandaardiseerde toetsen centraal, waarbij men extra de focus legt op kennis en vaardigheden die men cruciaal acht: wiskunde en Noors (met aanvullend natuurwetenschappen, zie sectie 4.3.1). Hiertoe heeft men eigen gestandaardiseerde toetsen ontwikkeld: de Oslo-toetsen (Oslo Prøven). Daarnaast maximaliseert men ook de inzet en het gebruik van de mogelijkheden van de diagnostische toetsen en de nationale toetsen.

De stad Oslo verleent als schoolbestuur op haar beurt een grote autonomie aan individuele scholen. Toch heeft men een intermediair niveau gecreëerd, waarbij een 'divisiedirecteur' toezicht houdt op ongeveer tien scholen. De divisiedirecteur dient jaarlijks de verschillende schooldirecteurs te evalueren, waarbij de resultaten van de school op de verschillende gestandaardiseerde toetsen een belangrijke rol (kunnen) spelen. Elke schooldirecteur krijgt een score op een schaal van 1 tot 5. In het verleden werd de score van elke directeur openbaar gemaakt, maar sinds 2013 is dat niet meer het geval. Op basis van de op schoolniveau geaggregeerde toetsresultaten kan onder meer overgaan worden tot extra bonussen voor schooldirecties (vorm van *performance-based payments*), alhoewel

⁸⁹ (Seland et al., 2013)

⁹⁰ Gebaseerd op Elstad (2009); Elstad, Christophersen, en Turmo (2012); Elstad et al. (2015); Elstad, Nordtvedt, en Turmo (2009); Norwegian Directorate for Education and Training (2014); Seland et al. (2013): + gesprekken tijdens case studie.

men hiermee niet makkelijk mee naar buiten komt. De manier waarop de toetsresultaten door het schoolbestuur in Oslo worden gebruikt, maakt dan ook dat directies van de scholen in Oslo aanvoelen onder een grote verantwoordingsdruk gebukt te gaan. Bovendien kan het schoolbestuur beslissen om op basis van de resultaten van gestandaardiseerde toetsen en de informatie van de divisiedirecteur scholen in te schakelen in een extern traject van intensieve begeleiding.

De verantwoordingsgerichte insteek van het schoolbestuur van Oslo is dan ook voornamelijk gericht op het mesoniveau:

“What I always say to anyone who wants to listen: it’s not really the kids who are tested; it’s their education.” (interview, respondent NW3)

4.3.1 Oslo-Toetsen

De Oslo-toetsen werden ontwikkeld parallel met de nationale toetsen. Niet toevallig was de persoon die mede verantwoordelijk was voor de ontwikkeling van de nationale toetsen bij het kabinet van de Minister van Onderwijs, in die periode ook deeltijds werkzaam in een leidinggevende functie bij het schoolbestuur van Oslo.

Er zijn toetsen in vier lesonderwerpen, met name lezen (Noors) in het zesde en elfde leerjaar, wiskunde (elfde leerjaar), digitale vaardigheid (zevende leerjaar) en natuurwetenschappen (vierde en zevende leerjaar). Vanaf 2016 zal digitale vaardigheid ook in het tiende leerjaar worden aangeboden, vanuit de vaststelling in scholen dat deze vaardigheid bij 15-jarige leerlingen vaak onvoldoende ontwikkeld is⁹¹.

Voor de toetsen voor het elfde leerjaar zijn er telkens twee versies ontwikkeld: één versie die aangepast is aan het curriculum voor ‘algemene studies’, en één versie opgemaakt volgens het curriculum van beroepsgericht onderwijs.

De Oslo-toetsen worden afgenomen in de periode januari-april, dus ongeveer in het midden van het schooljaar. De afname gebeurt in alle scholen op hetzelfde moment. Enkel de toetsen van het elfde leerjaar worden eind april - zo laat als mogelijk - gepland omwille van de meer verantwoordingsgerichte insteek op microniveau (zie sectie onder).

Doelen

De toetsen hadden van bij aanvang het uitdrukkelijke doel om complementair te zijn aan het aanbod op het federale niveau, waardoor er enkel toetsen moesten worden ontwikkeld voor die onderwerpen/leerjaren waarvoor men meende dat dit federale aanbod tekort schoot. Zo vond men het bijvoorbeeld een groot gemis in *The Knowledge Promotion* dat wetenschappen er onvoldoende in aanbod kwamen.

⁹¹ Men benadrukt dat het daarbij gaat om onder meer het gebruik van spreadsheets en tools voor wiskundige toepassingen; en niet om bijvoorbeeld het gebruik van sociale media, het aanmaken van filmpjes, en dergelijke meer, aangezien men ervaart dat deze laatste vaardigheden vaak wel zeer sterk ontwikkeld zijn bij de 15-jarige leerlingen.

De toetsen hebben volgens het schoolbestuur van Oslo in de eerste plaats een ontwikkelingsgericht doel op verschillende niveaus. Door de afname van de toetsen krijgen leraren specifieke informatie over de leerprestaties van leerlingen, krijgen de scholen een beeld van hun globale kwaliteit, en krijgt het schoolbeleid in Oslo een overzicht van de kwaliteit van de verschillende scholen in de hele stad. Toch wordt het ontwikkelingsgericht opzet aangevuld met een verantwoordingsgericht perspectief op mesoniveau: scholen die het niet goed doen, worden op hun resultaten aangesproken door het schoolbestuur. Bovendien bestaat er een vorm van verantwoording door de openbare beschikbaarheid van de resultaten (op schoolniveau) wat betreft het aandeel van de leerlingen dat bepaalde vaardigheidsniveaus heeft behaald op de Oslo-toetsen.

De toetsen in het elfde leerjaar worden relatief laat in het schooljaar afgenomen (meestal april), omdat deze toetsen in het elfde leerjaar deels een ander karakter hebben dan de andere Oslo-toetsen. Van de leraren wordt immers verwacht dat ze het finale cijfer dat ze voor Noors en wiskunde toekennen aan elk van de leerlingen, minimaal deels laten afhangen van de score van de leerlingen op de Oslo-toets. De Oslo-toetsen voor het elfde leerjaar hebben bijgevolg bijkomend een uitgesproken verantwoordingsgericht karakter op microniveau.

Ontwikkeling en correctie

De Oslo-toetsen zijn subject-gerelateerde toetsen die zijn opgesteld volgens het curriculum van een bepaald leerdomein. Bij de ontwikkeling van de toetsen van lezen en wiskunde heeft men zich dus, in tegenstelling tot bij de nationale toetsen, gebaseerd op het subject-specifieke curriculum van respectievelijk Noors en wiskunde.

De Oslo-toetsen zijn allemaal digitale toetsen. Men kiest voor digitale toetsen omwille van de voordelen van automatische correctie en directe feedback voor leraren en leerlingen.

De Oslo-toetsen worden jaarlijks vernieuwd, maar een aantal anker-items wordt behouden om na te gaan in welke mate het onderwijssysteem binnen Oslo erop vooruit of achteruit gaat.

“There are some exercises or questions that are the same each year. These are used to measure improvement from year to year within the schools. Most of the questions are new every year to avoid that teachers are gonna teach the items. If it were the same tests, we couldn’t publish the results and so on.” (interview, respondent NW2)

Deze anker-items worden wel jaarlijks op twee manieren gecontroleerd en mogelijk bijgeschaafd (maximaal 20% van de anker-items kan worden vervangen). Dit kan bij verschuivingen in het curriculum waardoor het item niet of minder relevant wordt, of indien uit de analyse blijkt dat het item te gemakkelijk is geworden als gevolg van *item shifting*. De anker-items worden evenwel niet vrijgegeven aan de leraren. In de rapportering over de resultaten van leerlingen kan men daarom niet het precieze item terugvinden, maar wel het thema/leerdoel waarop het item betrekking had.

Het systeem met anker-items is evenwel niet volledig sluitend, omdat ook deze anker-items ‘worn out’ raken na enige tijd, alsook omdat onvoorziene omstandigheden ertoe kunnen leiden dat de waarde van bepaalde anker-items ondermijnd wordt:

“An example: if an anchor question in a reading test is about a text about refugees from Africa, and if it has just this year been a lot of articles in the newspapers about refugees from Africa, the students will do very well. So we might have a result, and that result is too good, so we can't use it. It is a vulnerable system actually, but it is the only one we have.”(interview, respondent NW3)

Voor de ontwikkeling van de Oslo-toetsen inzake wiskunde heeft men binnen het schoolbestuur de nodige expertise. De andere toetsen worden door externe professionals ontwikkeld. Zo doet men voor de ontwikkeling van de toetsen inzake leesvaardigheid en natuurwetenschappen een beroep op de Universiteit van Oslo. Voor de toetsen van digitale vaardigheden wordt beroep gedaan op het Noorse ‘*Senter for IKT in Utdanningen*’ (Centrum voor ICT in onderwijs), een centrum onder de bevoegdheid van het federale Ministerie van Onderwijs en Onderzoek. Dit centrum voert voornamelijk onderzoek uit in opdracht van de Minister van Onderwijs inzake digitale vaardigheid van leerlingen, maar stelt ook gestandaardiseerde toetsen ter beschikking aan scholen en helpt scholen bij de ontwikkeling van strategieën om de ICT-investeringen optimaal in te zetten, alsook om de digitale vaardigheden van leerlingen te versterken.

Om de ‘afleiders’ bij de meerkeuzetoetsen zo realistisch mogelijk te maken, worden twee pre-tests uitgevoerd met telkens ongeveer 500 leerlingen. Bij de eerste pre-test worden geen antwoordcategorieën voorzien. De items die behouden blijven na de eerste pre-test, worden voor de tweede pre-test omgezet in meerkeuze-items met afleiders opgesteld op basis van de foutieve antwoorden van de leerlingen in de eerste pre-test. Op basis van de tweede pre-test worden items met een voldoende discriminerende kwaliteit worden behouden, en worden de resterende items ingeschaald op een IRT-schaal.

Eén van de uitdagingen voor de toekomst inzake het toetsen van digitale vaardigheden is om de toets een grotere inhoudsvaliditeit te geven, waarbij men minder afhankelijk is van items die de makkelijk meetbare vaardigheden in kaart brengen.

Bij de ontwikkeling van de Oslo-toetsen doet het schoolbestuur - net als het Directorate for Education and Training - beroep op de externe expertise van Inspera Assessment.

Kwaliteitszorg

Zoals hierboven gesteld, worden de toetsen op voorhand twee maal bij een groep van 500 leerlingen uitgeprobeerd.

Om de toetsafname zo betrouwbaar mogelijk te laten verlopen, wordt voor de afname van de Oslo-toetsen een infosessie georganiseerd waarin directies en leraren worden getraind in de procedures die moeten worden gevolgd. Op diezelfde infosessie wordt informatie gegeven over hoe de resultaten zullen worden gebruikt.

Deze infosessie is niet specifiek voor de Oslo-toetsen, maar omvat ook de procedures voor de afname van de nationale toetsen en Diagnostische toetsen.

Rapportering

De algemene resultaten op schoolniveau zijn voor het publiek beschikbaar. Er zijn rapporten per school, per district (van ongeveer 20 scholen elk) en voor geheel Oslo. Elk van deze rapporten wordt via een online platform (ontwikkeld en onderhouden door Inspira Assessment) ter beschikking gesteld⁹². Afhankelijk van de toegekende rechten kan men veel of weinig informatie raadplegen. Op dit platform worden onder meer ook de resultaten van de tevredenheidsvragenlijsten van leerlingen voor het geheel van de scholen uit Oslo gepubliceerd.

In het niet-openbare schoolspecifieke rapport kan de school per leerling nakijken wie welke items correct heeft beantwoord. Er wordt per leerling per leerdomein een z-score berekend, waarmee wordt aangegeven hoe de leerlingen hebben gepresteerd in vergelijking met alle leeftijdsgenoten in de scholen in Oslo. Op basis van deze vergelijking worden de leerlingen ook in drie categorieën (vaardigheidsniveaus) ingedeeld. Bij de toetsen in het secundair onderwijs worden, net als bij de nationale toetsen, vijf vaardigheidsniveaus onderscheiden.

Per item van de Oslo-toetsen kan gekeken worden in welke mate er correcte antwoorden waren op het niveau van individuele leerlingen, op schoolniveau alsook op het niveau van de gehele populatie in Oslo. Elk item in deze rapportering is ook meteen gelinkt aan het leerdoel uit het curriculum dat werd getoetst.

Daarnaast zijn er op populatieniveau ook longitudinale gegevens beschikbaar. Op basis van de anker-scores (op de jaarlijks terugkerende anker-items) kunnen scholen ook hun eigen absolute vooruitgang evalueren. Dit wordt evenwel niet automatisch in de rapportering voorzien. Bovendien wordt ervoor gewaarschuwd dat bij deze anker-items toch *item shifting* kan ontstaan waardoor longitudinale analyses van deze resultaten minder betrouwbaar worden.

Men is er zich bij het schoolbestuur van Oslo van bewust dat deze vorm van rapportage niet makkelijk te interpreteren is voor scholen. Om hieraan tegemoet te komen wordt er algemene ondersteuning voorzien voor scholen bij de interpretatie van data, en wordt de informatie zo grafisch mogelijk voorgesteld:

"We have a course after the tests where they will say mostly general things about how to read them [de feedbackrapporten] and also we publish something ourselves in the Inspira Platform called 'Infographic indicators' (...) for mainly graphs comparing the school to the Oslo level."(interview, respondent NW1)

Naast hulp bij de analyse wordt tijdens deze druk door leraren en directies bijgewoonde trainingssessies ook advies gegeven worden over welke acties men kan nemen op basis van informatie die blijkt uit de leerresultaten. Bij scholen die relatief zwak scoren, gaan adviseurs van het schoolbestuur zelf op bezoek om verbeteracties mee te helpen uitwerken.

⁹² <http://minosloskole.no/>

4.3.2 Overgangstoetsen

Daarnaast ontwikkelt het schoolbestuur van Oslo 'overgangstoetsen' die scholen verplicht dienen af te nemen van alle leerlingen aan het eind van het vierde leerjaar, en aan het eind van het zevende leerjaar. Deze toetsen worden jaarlijks ontworpen door experts binnen het schoolbestuur en zijn gebaseerd op het nationale curriculum. Beide toetsen worden gezien als toetsen met een ontwikkelingsgericht doel op zowel school- als leerlingniveau. Ze worden ingezet op overgangsmomenten: bij de overgang van het 'lager basisonderwijs' naar het 'hogere basisonderwijs' (tussen vierde en vijfde leerjaar) en bij de overgang van het lager onderwijs naar het secundair onderwijs (tussen zevende en achtste leerjaar). Op deze momenten worden zowel de rekenvaardigheid als de leesvaardigheid van leerlingen in kaart gebracht om na te gaan of minimale vereisten om optimaal te kunnen leren op het hogere niveau in voldoende mate aanwezig zijn. Deze toetsen dienen verder ook als informatiebron voor het systeem van interne kwaliteitszorg voor de lagere scholen. Tevens dienen ze als informatiebron voor de leraren van het 'hogere' niveau (respectievelijk vijfde en achtste leerjaar): de resultaten geven aan deze leraren aan wat het beginniveau van de leerlingen is bij de instroom in het hogere leerjaar.

Daarnaast tracht het schoolbestuur van Oslo het optimale gebruik van de overgangstoetsen te stimuleren door leraren van het zevende en achtste leerjaar samen te brengen om de resultaten van de leerlingen te bespreken.

4.3.3 Gebruik van de gestandaardiseerde toetsen van het federale niveau

In tegenstelling tot het federale beleid waarbij bepaalde nationale toetsen verplicht en de keuze voor deelname aan andere toetsen facultatief is, legt het schoolbestuur van Oslo de scholen op om alle leerlingen aan alle gestandaardiseerde toetsen deel te laten nemen die op het federale niveau werden ontwikkeld.

Dit verschil uit zich voornamelijk inzake de omgang van het schoolbestuur van Oslo ten aanzien van de diagnostische toetsen. Hoewel deze diagnostische toetsen als facultatieve toetsen zijn bedoeld, en eigenlijk er enkel op zijn gericht om de leerlingen te identificeren die extra ondersteuning nodig hebben (zie sectie 4.2.3), heeft het schoolbestuur van Oslo deze toetsen verplicht voor alle leerlingen in alle scholen.

Het schoolbestuur verzamelt de data van de verschillende scholen, maar zorgt ervoor dat deze niet gepubliceerd worden. In het verleden werden de resultaten van de scholen op de diagnostische toetsen door het schoolbestuur gebruikt om uitspraken te doen over de kwaliteit van de scholen. Hoewel het schoolbestuur deze praktijk momenteel afwijst, geeft men aan dat men tegelijkertijd de resultaten op een andere manier op een verantwoordingsgerichte manier hanteert:

"We anyway follow this very closely also in all our schools because we have all the results from these tests here and we have the number of students in every school who is below the line. So we know for instance one school that has in the first year twenty

pupils [below the line], then to be able to help the school, it is important for us to know if it is the same number of pupils who is below the line in the second year and in the third, or if they are getting more; because if they are, if they increase the number of pupils that are below the line, we know that they are doing something wrong. They are not managing this, they need intervention, they need help from us.” (interview, respondent NW3)

De standaardpraktijk houdt evenwel in dat het schoolbestuur verwacht dat de verschillende scholen kunnen aantonen op welke manier ze de achterstand bij de leerlingen ‘onder de cesuur’ trachten te remediëren.

4.3.4 Financiële impact van het evaluatiesysteem in Oslo

Het bleek moeilijk om zicht te krijgen op het totale financiële plaatje achter het evaluatiesysteem in Oslo. Wel kan gesteld worden dat drie personeelsleden voltijds bezig zijn met taken gerelateerd aan de gestandaardiseerde toetsen. Daarnaast zijn vier inhoudelijke experts (al dan niet voltijds) aangesteld voor de toetsontwikkeling. Ten slotte zijn heel veel toetsontwikkelaars extern aangesteld. Gevraagd naar een globale minimale inschatting, wordt aangegeven dat de ontwikkeling van één toets ongeveer de inzet van (equivalent) één voltijds personeelslid gedurende één jaar vergt. Door het comprehensieve karakter van het secundair onderwijs, waarbij alle leerlingen tot het tiende leerjaar hetzelfde programma volgen, blijft de financiële inspanning voor het schoolbestuur draagbaar.

“In Norway there is a very comprehensive system of secondary education, so we only need two different tests of digital skills in grade 11.” (interview, respondent NW4)

4.3.5 Impact

Het is moeilijk om veralgemenende uitspraken te doen over de impact van de Oslo-toetsen en het gebruik van de federale gestandaardiseerde toetsen in Oslo. De Oslo-toetsen en de publicatie van de resultaten hebben uiteenlopende effecten op verschillende scholen. Tussen de eerste en tweede afname van de toetsen was een opmerkelijke stijging van de resultaten, met name van het leesniveau van leerlingen in het tweede leerjaar. Dit kan echter ook te wijten zijn aan de bevinding dat leraren nu beter het toetsformat kenden en overgingen tot *teaching to the test* en zelfs onwenselijk strategisch gedrag.⁹³

Aanvaarding

Er zijn geen empirische gegevens beschikbaar over de aanvaarding van de Oslo-toetsen door de scholen. Bij de aanvankelijke implementatie, bleek er evenwel heel wat protest van leerlingen, ouders en leraren: immers, toen bij aanvang (2005) de Oslo-toetsen werden geïmplementeerd in het elfde leerjaar hadden de toetsen (net als andere Oslo-toetsen) enkel als doel om bij te dragen tot de ontwikkeling van leerlingen en van scholen, en om gestandaardiseerde wijze de kwaliteit van scholen

⁹³ Onderzoek van Elstad (2008) gepubliceerd in Noors, gerapporteerd in Engels door Elstad (2009)

in kaart te brengen. Het betrof dus ontwikkelingsgerichte doelen, gecombineerd met een verantwoordingsgericht doel op mesoniveau.

De toetsen werden evenwel massaal geboycot door de leerlingen. Tot deze staking werd opgeroepen door een belangenorganisatie (een leerlingenkoepel) die meende dat deze toetsen niet mochten gebruikt worden om conclusies over de scholen te trekken. Als gevolg hiervan besloot het schoolbestuur om leerlingen te verplichten en extrinsiek te motiveren tot deelname aan de toetsen door de toetsen later in het schooljaar te organiseren en de rechtstreekse link te leggen met het eindcijfer van leerlingen. Zo werd dus een verantwoordingsgericht doel op microniveau toegevoegd.

Bij het schoolbestuur gaat men er van uit dat deze Oslo-toetsen inmiddels aanvaard zijn door de meeste stakeholders. De 'Oslo-style' van omgaan met gestandaardiseerde toetsen heeft immers ook sterke voorstanders. Oslo komt bijvoorbeeld erg sterk uit de vergelijking qua leerresultaten in vergelijking met andere provincies. Dit wordt zowel door bepaalde politici als door verschillende academici toegeschreven aan de manier waarop het schoolbestuur van Oslo met de gestandaardiseerde toetsen omgaat.

"This style of targets and controls, it functions (...) it is an efficient style of government in Oslo." (interview, respondent NW10)

De aanvaarding van de diagnostische toetsen door leraren in Oslo is wellicht minder groot dan in andere gemeenten. De eerder geringe mate van aanvaarding heeft te maken met de verplichting die door het schoolbestuur wordt opgelegd om alle leerlingen elk van de diagnostische toetsen te laten afleggen, hoewel deze hier in principe niet voor zijn ontwikkeld. Daardoor leveren de diagnostische toetsen van de meeste leerlingen weinig bijkomende informatie op, maar moeten al deze toetsen wel worden afgenomen én (met name voor lees- en rekenvaardigheid in het eerste, tweede en derde leerjaar) manueel verbeterd. Hierdoor luidt de kritiek in scholen in Oslo dat de diagnostische toetsen voornamelijk hebben geleid tot een verhoogde bureaucratie en tijdsverlies.

Impact op microniveau

Een eerder genoemde studie van NUFI die de impact van de nationale toetsen naging in heel Noorwegen, stelde vast dat het gebruik van de resultaten van de nationale toetsen door leraren (alook de voorbereiding erop) in de scholen in Oslo hoger lag dan in andere provincies: het rapport concludeert dat in Oslo *"the use of national tests seems to be an integrated part of everyday schooling"* (Seland et al., 2013, p. 12)

Een andere bevinding die wijst op een impact van het evaluatiesysteem werd vastgesteld toen in 2014 de diagnostische toetsen werden uitgebreid met een toets voor digitale vaardigheden voor het vierde leerjaar. In Oslo bleken veel minder leerlingen dan normaal kon worden verwacht onder de cesuur te scoren die wijst op de nood aan extra ondersteuning. Dit wordt toegeschreven aan het feit dat de digitale vaardigheden van leerlingen in het vierde leerjaar in het kader van de Oslo-Toetsen reeds langere tijd op een gestandaardiseerde wijze in kaart worden gebracht. De hypothese is dan ook dat leraren sterker hebben ingezet op het aanleren van deze digitale vaardigheden, wat heeft geleid tot

een groter leereffect (inzake dit leerdomein). Dit sluit aan bij de sterke overtuiging in dit schoolbestuur dat het zichtbaar maken van leerprestaties bijna steeds leidt tot een verhoogde inzet en betere leerprestaties:

“When you shine a light on something, the quality will be almost automatically improved.” (interview, respondent NW3)

Door onderzoek uitgevoerd kort na de introductie van de Oslo-toetsen werd aangetoond dat deze mechanismen in sommige scholen met slechte resultaten op korte termijn leidden tot een verbetering van de leerprestaties, maar ook van de kwaliteit van de onderwijsleerprocessen. Dezelfde studie stelde echter vast dat er in andere scholen geen vooruitgang merkbaar was⁹⁴. Twee schooldirecties met een verschillende ervaring worden in het volgende tekstfragment geciteerd:

“This contributed to a stronger focus on the students’ learning at school 1: “We ... have not been very concerned with learning. When there is a stronger focus on learning outcomes my contact with the teachers, as their leader, is at a different level. We focus on results. I am able to monitor learning in the various grades in a completely new way”. But similar patterns are not evident in all schools. At another school we observed laughter when we used the words “strong leadership”. The principal said that “I cannot put pressure on the lower levels in the hierarchy. If I do so, and start wagging my finger ... saying ‘you should make sure the results are up to scratch’, then things get unhealthy around here.” (Elstad, 2009, p. 180)

Het schoolbestuur van Oslo geeft aan dat de verhoogde focus op leerprestaties mogelijk ook tot een vernauwing van het curriculum kan leiden tot deze leerdomeinen en thema’s die op gestandaardiseerde wijze kunnen gemeten worden. Dit ervaart het schoolbestuur echter niet als een probleem, omdat men het terecht vindt dat de meeste aandacht in het onderwijs gaat naar de kennis en vaardigheden die momenteel via de nationale toetsen, de Oslo-toetsen en de diagnostische toetsen worden geëvalueerd. Bovendien worden directies in de scholen in Oslo door het evaluatiesysteem aangemoedigd om ten aanzien van alle leerlingen hoge verwachtingen te hebben, ook ten aanzien van leerlingen met een minder gunstige socio-economische achtergrond.

Dat de diagnostische toetsen in Oslo verplicht worden afgenomen in het hoger secundair onderwijs⁹⁵, heeft ten slotte als voordeel dat de aandacht van scholen ook gevestigd wordt op de vaardigheden die in het hele onderwijsleerproces (volgens de visie van het nationale curriculum, zie sectie 4.1.3) centraal zouden moeten staan.

“We also see, for example in Oslo, it is necessary to have these tests when the people are seventeen or sixteen years old. In some schools, you can see that the students

⁹⁴ Studie uit 2005 door Andersen, Dellemir, & Eckhoff. Origineel gepubliceerd in Noors, gerapporteerd in Engels door Elstad (2009).

⁹⁵ Hoewel deze toetsen voor het hoger secundair onderwijs vanaf 2016 niet langer worden aangeboden, omdat ze spontaan te weinig worden gebruikt - zie sectie 5.2.3.

don't read good - even when they are in the eleventh grade and they need extra support to read properly." (interview, respondent NW8)

Door de diagnostische toetsen verplicht af te nemen in het hoger secundair onderwijs, werd in het leerproces van sommige leerlingen bijgevolg de nadruk gelegd op basisvaardigheden als lees- en rekenvaardigheid; zonder de diagnostische toetsen was dit wellicht niet het geval.

Dit effect wordt enerzijds bevestigd door de geïnterviewde schooldirectie, die meent dat er door het hervormde curriculum en de nationale toetsen er een groeiend besef is dat werken aan lees- en rekenvaardigheid van leerlingen de taak van alle leraren is. Anderzijds stelt ze evenwel dat deze vernauwing van het curriculum toch ook problematisch kan zijn:

"Teachers in science for instance will focus much more on how you can handle a written science task." (interview, respondent NW3)

Impact op mesoniveau

Net als in de sectie over de impact van de gestandaardiseerde toetsen op federaal niveau (zie sectie 4.2.6), zijn de meeste betrokkenen ervan overtuigd dat de impact van het evaluatiesysteem in Oslo zich voornamelijk op het niveau van de kwaliteitszorg van scholen bevindt. Schooldirecties krijgen onder meer door de jaarlijkse evaluatie een grote externe motivatie om met de resultaten van de verschillende toetsen aan de slag te gaan. Los van de vraag in welke mate men de resultaten kan gebruiken, wordt door verschillende directies benadrukt dat door de gestandaardiseerde toetsen de ambities van de school hoger worden gelegd. Deze toetsen maken immers dat er een externe druk ontstaat waardoor de school voor zichzelf hogere kwaliteitsnormen formuleert, zoals in deze studie wordt geïllustreerd door volgend citaat van een schooldirecteur:

"If we were allowed to (...) define our own standards, those standards would not be good enough for our students. Other schools have parents and other pressure groups that help improve the standards. We ... don't. It is essential that somebody external takes part in defining our standards, and that we push ourselves to deliver better results." (Elstad, 2009, p. 177)

4.3.6 Kritiek en neveneffecten

De manier waarop Oslo gestandaardiseerde toetsen inzet, verschilt sterk van het gebruik bij andere schoolbesturen. Daarom is er ook geregeld controverse rond het beleid in Oslo. Zo was het schoolbestuur zich enige tijd geleden bewust van het bestaan van een Facebook-groep 'Ouders die protesteren tegen het Oslo-schoolsysteem', die een groot aantal leden telde. De publieke verontwaardiging over de verantwoordingsgerichte insteek van het schoolbestuur in Oslo werd sterk gevoed door enkele kritische artikels in de media, onder meer in de periode waarin de diagnostische toetsen werden gebruikt als instrument om de kwaliteit van scholen in te schatten.

Emotionele impact op leraren en directies

De grote druk die wordt gelegd op schooldirecties is voor velen te hoog, wat zou leiden tot een zeer hoge turnover in de invulling van de functie van directies. Daarnaast is er een impliciete druk op leraren om het curriculum te vernauwen tot de leerstof die in de gestandaardiseerde toetsen aan bod komt. Diezelfde verantwoordingsgerichte invulling zorgt ook voor een emotionele impact op de leraren. Dit geldt met name voor de verplicht gestelde diagnostische toetsen:

“A teacher who gets his or her results, they get anxious. They feel like they are being evaluated (...) nothing is attached to it for the teachers, still many teachers experience it as an evaluation.” (interview, respondent NW3)

Bij het schoolbestuur van Oslo bevestigt men dat de toetsen leiden tot extra stress bij de leraren, hoewel dat uiteraard niet de bedoeling is. Er werden totnogtoe evenwel geen gevallen gerapporteerd van onwenselijke strategische gedragingen van scholen of individuele leraren.

Emotionele impact op leerlingen

Door het verplicht stellen van de diagnostische toetsen en de toevoeging van de eigen Oslo-toetsen leggen leerlingen in Oslo doorheen hun schoolloopbaan een groot aantal gestandaardiseerde toetsen af. Bovendien willen de meeste leraren ook hun eigen toetsen blijven behouden, omdat men meent dat eigen toetsen en toetsen van educatieve uitgeverijen (op microniveau) een grotere informatieve waarde hebben. Hierdoor is er toch enige bezorgdheid bij schoolbestuur over de ‘toetsdruk’ die op leerlingen worden gelegd. Bij de schooldirectie krijgen we een andere kijk op dit thema. Zij ervaart de toetsdruk niet als te groot - al pleit ze voor de afschaffing van de verplichting om alle toetsen van alle leerlingen af te nemen. Wel is er sprake van een emotionele impact op de leerlingen, voornamelijk bij de jonge leerlingen bij wie de diagnostische toetsen worden afgenomen. Het schoolbestuur van Oslo is zich hiervan bewust en raadt daarom de scholen aan om de leerlingen goed op de toetsomstandigheden voor te bereiden (door hen bijvoorbeeld al eens op de computers te laten werken) en het belang van de toetsresultaten te relativeren:

“You should not train the test - you should train them on the situation. What we try to emphasize to our schools, is that they have to prepare them [young pupils] for the test, they have to talk about it beforehand, they have to un-dramatize the tests.” (interview, respondent NW3)

Net als bij de nationale toetsen is een punt van kritiek op de rapportering van de Oslo-toetsen dat de ruwe scores worden gecommuniceerd. Met andere woorden: de resultaten worden niet gecorrigeerd voor input-kenmerken van de leerlingen in de verschillende scholen (hoewel uit de eigen analyse van het schoolbestuur blijkt dat er een sterke correlatie is).

Oneigenlijk gebruik van de diagnostische toetsen

De verplichting om alle leerlingen de diagnostische toetsen te laten afleggen, gaat in tegen het oorspronkelijke doel van deze toetsen. Hoewel in 4.3.5 een voorbeeld werd aangehaald waarin het

Directorate for Education and Training een voordeel van deze verplichting noemt, kan algemeen worden gesteld dat dit oneigenlijke gebruik niet kan rekenen op de goedkeuring van dit Directorate:

"Employees at the Directorate for Education and Training, they are angry about the Oslo practice. So we have some tensions between the Oslo style of government and the national authorities." (interview, respondent NW10)

Beperkte meerwaarde van de gestandaardiseerde toetsen

De beperkte meerwaarde van de resultaten op microniveau wordt deels toegeschreven aan de te lange periode tussen toetsafname en beschikbaarheid van de resultaten. In dezelfde lijn ligt de kritiek van leraren in het hoger secundair onderwijs (elfde leerjaar): omdat deze toetsen een impact hebben op het eindcijfer van de leerlingen, worden ze relatief laat in het schooljaar afgenomen. Deze late afname maakt evenwel dat er te weinig tijd over is om de resultaten te gebruiken in de voorbereiding van de leerlingen op het eigen schoolexamen⁹⁶. Ook hier blijken de ontwikkelingsgerichte en verantwoordingsgerichte doelen elkaar stokken in de wielen te steken:

"If we want this test to also be supportive to teachers' teaching, it's too late. Because the students, when they get the results back from us, it is the beginning of May, the middle of May, and the students are finished. So this is a big dilemma: the teachers say: 'this doesn't help us at all! We get the results, we see a student has challenges in this area, and I have only one week left to prepare for the exam. So we want this test earlier'. But then if it's earlier, they haven't been through the curriculum; so it can't be a support for assigning the grade." (interview, respondent NW3)

Geen leerwinstmeting

Net als bij de gestandaardiseerde toetsen op federaal niveau, hoort men in Oslo de kritiek dat de toetsen - zeker op microniveau- weinig nieuwe informatie aanleveren. Een belangrijke kritiek op de Oslo-toetsen en op het gebruik van gestandaardiseerde toetsen in Oslo is dat er ondanks alle metingen geen mogelijkheid bestaat om de leerwinst van leerlingen in kaart te brengen. De verschillende toetssystemen zijn daarvoor niet op elkaar afgestemd.

"It's not fair to the child not to keep track of the progress. It is far too important!"
(interview, respondent NW3)

Kwaliteit en gebruiksvriendelijkheid

Een bedreiging voor de kwaliteit van de vergelijking overheen toetsen, is dat verschillende toetsen door verschillende partners werden ontwikkeld, die elk hun eigen prioriteiten en accenten hebben gelegd. Het schoolbestuur van Oslo werkt daarom aan een set van verwachtingen en een 'common practice' waaraan elke toets moet voldoen en een omschrijving van wat kan vergeleken worden, en wat niet kan vergeleken worden.

⁹⁶ In de meeste provincies is er geen overkoepelend examen in het hoger secundair onderwijs. In Oslo wordt er evenwel een examen ontwikkeld dat gestandaardiseerd verloopt voor alle leerlingen in elke school. Aangezien het hier echter geen leerplichtonderwijs betreft, zijn deze examens niet opgenomen in de focus van dit rapport.

Ondanks de gegeven aandacht voor nascholing en trainingssessies omtrent hoe de rapportage kan worden geïnterpreteerd, blijkt dat de feedbackrapporten niet zo gebruiksvriendelijk zijn voor scholen, onder meer door het gebruik van z-scores. Bij het schoolbestuur van Oslo wordt inderdaad aangegeven dat dit een werkpunt is voor de nabije toekomst en een belangrijk aandachtspunt is bij de ontwikkeling van een nieuw platform om de resultaten te communiceren.

"For the future, we want the results to be more user-friendly. We want wider tools, and possibilities for the schools to use the assessments. More interactive as well".

(interview, respondent NW5)

Ook op het feit dat de anker-items niet worden vrijgegeven is er kritiek van leraren, omdat het gebrek aan detail een ontwikkelingsgericht gebruik (bijvoorbeeld voor adequate remediëring van bepaalde leerlingen) in de weg staat.

Daarnaast is de keuze van het schoolbestuur van Oslo voor digitale toetsen niet helemaal vrij van scepsis: zo blijkt het voor de ontwikkelaars van leesvaardigheidstoetsen niet makkelijk om digitale toetsen te ontwikkelen. Meer dan andere toetsen lijken de Oslo-toetsen van leesvaardigheid dan ook pen-en-papier toetsen die als het ware gekopieerd worden op een beeldscherm. Een ander probleem is dat men meent dat er door het digitaal aanbieden van de Oslo-toetsen naast leesvaardigheid wellicht ook andere vaardigheden worden geëvalueerd. Dit laatste is een bedreiging voor de validiteit van de toetsresultaten.

Een gelijkaardig probleem inzake validiteit (toetsen we wel wat we willen toetsen?) geldt voor de Oslo-toetsen inzake natuurwetenschappen, waaruit blijkt dat deze eigenlijk voornamelijk taalkundige kennis toetsen. Er bestaat dan ook een hoge correlatie tussen de resultaten van de leerlingen voor Noors en voor natuurwetenschappen.

"A lot of the challenges we see from the results on the science test is actually language. Science is above all actually a language subject: you have to know very specific terms and words, which sometimes are not the same as they mean in the every-day language. So this science test actually uncovers language challenges that are not uncovered in normal speaking language." (interview, respondent NW3)

In een academische analyse uitgevoerd op data van 2005 wordt een andere opmerkelijke correlatie vastgesteld: de resultaten op de digitaal aangeboden nationale toetsen Engels van het lager onderwijs (toen nog in het vierde leerjaar), bleken een significante samenhang te vertonen met de zelf-gerapporteerde vaardigheid van leerlingen in ICT. Dit suggereert dat minimaal voor een deel digitale vaardigheid wordt gemeten in de nationale toetsen Engels. De - eveneens digitale - nationale toetsen Engels in het secundair onderwijs (toen nog in het tiende leerjaar) vertoonden geen samenhang met de ICT-vaardigheid, maar bleken wel samen te hangen met de score van leerlingen op een non-verbale redeneervaardigheidstest⁹⁷.

⁹⁷ (Manger et al., 2009)

4.4. Leerwinst en toegevoegde waarde

Hoewel het Noorse onderwijssysteem – en bij uitstek het schoolbestuur van Oslo zich in verkennende gesprekken voorstelt als sterk ‘value-added-oriented’, wordt slechts in zeer beperkte mate de leerwinst en toegevoegde waarde door de gestandaardiseerde toetsen naar boven gebracht.

Er wordt voorlopig geen leerwinst op leerlingniveau gerapporteerd. Hoewel men vaak met IRT-inschaling van items werkt, zagen we geen voorbeelden van IRT-schalen die gelden voor meerdere toetsen (in verschillende leerjaren). De IRT-inschaling wordt enkel gebruikt om de vergelijkbaarheid van toetsen en toetsresultaten doorheen de tijd te analyseren. Hierdoor zijn *gain-based* modellen⁹⁸ van leerwinst niet mogelijk.

Het gebrek aan een duidelijke rationale, de zeer uiteenlopende en vaak dooreenlopende doelen van de verschillende toetsen zijn bovendien niet bevorderlijk voor het berekenen van leerwinst en toegevoegde waarde. Zoals eerder gesteld, worden toetsen door verschillende ontwikkelaars opgemaakt, elk vertrekkend vanuit eigen uitgangspunten en prioriteiten, waardoor het berekenen van leerwinst moeilijk wordt gemaakt.

Voor de scholen van Oslo is er echter een rapportagesysteem in ontwikkeling met de intentie om van elke leerling een overzicht te geven van alle toetsen die werden afgelegd, waarbij deze aan elkaar zouden worden gekoppeld. Het schoolbestuur acht het wenselijk om ook de resultaten van de nationale toetsen en diagnostische toetsen hierin te integreren, maar momenteel laten de verschillende systemen dit niet toe. Ook resultaten van de tevredenheidsvragenlijsten zouden dan kunnen worden opgenomen. Daarnaast heeft men de ambitie om op termijn ook de resultaten op andere toetsen (van private uitgevers) rechtstreeks in dit rapportagesysteem te integreren.

Het schoolbestuur van Oslo heeft verder de ambitie om in de nabije toekomst de resultaten van leerlingen op de nationale toetsen inzake leesvaardigheid van het vijfde leerjaar, te koppelen aan de resultaten op de Oslo-toetsen inzake leesvaardigheid van het zesde leerjaar, om zo leervooruitgang van leerlingen in kaart te brengen. Dit blijkt echter een methodologisch moeilijke uitdaging, omdat verschillende concepten worden gemeten met verschillende toetsen die niet aan elkaar gelinkt zijn.

De enige uitzondering inzake leerwinstmonitoring vormen de nationale toetsen in het lager secundair onderwijs, waarbij voor het achtste en het negende leerjaar dezelfde toets wordt afgenomen. Aangezien deze toetsen sinds 2014 aan elkaar verankerd zijn, kan hier in principe vanaf 2015 de leerwinst van leerlingen worden berekend. Deze leerwinst op leerlingniveau wordt evenwel niet automatisch gerapporteerd, en er zijn weinig aanwijzingen dat de lager secundaire scholen deze informatie op eigen initiatief uit de opeenvolgende rapporten halen.

Net omdat leerwinst wel een thema is dat ‘leeft’ binnen de scholen in Noorwegen, is het gebrek aan informatie over leerwinst van leerlingen volgens de bevraagde directie één van de grootste minpunten van het huidige toetsenarsenaal. Volgens haar is het gebrek aan leerwinstmonitoring één van de

⁹⁸ Zie Deel 3, sectie 3.2.2 van dit rapport.

belangrijkste redenen waarom de scholen relatief weinig aan de slag gaan met de data die uit deze gestandaardiseerde toetsen komen:

“Teachers see now much more that every pupil can make progress. But the tests are not sufficiently oriented at tracking the learning of the pupils.” (interview, respondent NW3)

Ook binnen de academische wereld is men ervan overtuigd dat leerwinstmetingen de logische volgende stap zouden moeten zijn:

“The schools should be able to compare the fifth level, the eighth level, and the ninth level.” (interview, respondent NW10)

De indicaties over de toegevoegde waarde van scholen blijven beperkt tot vergelijkingen die worden gemaakt tussen de gemiddelde leerprestaties van scholen. Bij het schoolbestuur van Oslo leeft evenwel de ambitie om in de toekomst hierop in te zetten.

“Maybe together with the bureau of Central Statistics, we can develop some kind of value added indicators. How much does the school add based on the national tests from year 5 to year 8?” (interview, respondent NW3).

5. CONCLUSIE: AANDACHTSPUNTEN VOOR VLAANDEREN VANUIT DE INTERNATIONALE STUDIE

In deze laatste sectie geven we een overzicht van aandachtspunten voor de ontwikkeling van een leer(winst)monitoringssysteem in de Vlaamse onderwijscontext, die afgeleid kunnen worden uit de vier verschillende case studies. Deze bespreking betreft een combinatie van ervaringen vanuit de verschillende case studies met concrete adviezen voor de Vlaamse onderwijscontext die werden uitgesproken door de verschillende respondenten tijdens de vier case studies.

5.1 Beslissing over implementatie

5.1.1 *Bepalen van een doel en rationale*

Uit de case Noorwegen komt de boodschap dat een duidelijke rationale waarin helder wordt gesteld welke doelen met de implementatie van een leer(winst)monitoringssysteem worden nagestreefd, op welke manier dit systeem geacht wordt te leiden tot de vooropgestelde doelen, en welke randvoorwaarden er ingevuld moeten zijn om deze doelen te bereiken, een belangrijke voorwaarde is die de ontwikkeling van gestandaardiseerde toetsen of een toetsensysteem moet voorafgaan. Zonder deze rationale bestaat het gevaar dat verschillende toetsen onvoldoende op elkaar zijn afgesteld of dat teveel doelen aan één toets worden gesteld. Men geeft in de case dan ook het advies om zeer duidelijk het doel van de toetsafname te omlijnen vooraleer men met de toetsontwikkeling aan de slag gaat. Dit doel zal immers - bij een goede toets - een grote impact hebben op verschillende keuzes die tijdens de toetsontwikkeling moeten worden gemaakt.

“Define what you want to get out of the test. The information you want to get out, that is important, and then you can think about specifications and methods.” (interview, respondent NW9)

Zo vinden respondenten in de Noorse case studie dat men, achteraf gezien, misschien te snel van start is gegaan bij de eerste nationale toetsen; waardoor men pas sinds 2014 de psychometrische expertise heeft aangewend die het mogelijk maakt om longitudinale gegevens te produceren op school- en provincie-niveau, alsook op het federale niveau.

Ook in de andere cases wordt door alle respondenten met ervaring in toetsontwikkeling benadrukt hoezeer het van belang is om eerst vast te stellen welk doel er aan de gestandaardiseerde toetsen wordt gesteld. Ook fervente voorstanders van gestandaardiseerde toetsen, menen dat men in Vlaanderen niet zonder meer op het gebruik van zulke toetsen moet inzetten. Het Vlaamse onderwijs wordt internationaal immers sterk gewaardeerd, en men ziet dus ook niet steeds op welke behoefte een leer(winst)monitoringssysteem een antwoord zou moeten bieden: *“Het onderwijs is daar [in Vlaanderen] toch wel goed en kan toch wel op een goede prestatie bogen”* (interview, respondent NL3). Vlaanderen kan in internationale vergelijkende studies (bijvoorbeeld PISA) goede resultaten voorleggen en dus is er - in de ogen van de respondenten van deze vier case studies - relatief weinig hefboom om ingrijpende veranderingen door te voeren. Gevraagd naar hun ervaringsdeskundig advies voor de Vlaamse context, is vaak het eerste antwoord een vraag: *“what is the problem you are hoping*

to fix?” (interview, respondent AU4): Toetsen dienen afgestemd zijn op een duidelijk onderwijskundig probleem, of om een mogelijk probleem in de toekomst te vermijden. De toetsontwikkeling moet sterk op dit doel worden gericht. Het expliciet definiëren van het ‘op te lossen probleem’ is bovendien ook een cruciale voorwaarde voor de aanvaarding van het leer(winst)monitoringssysteem door het onderwijsveld, wat dan weer een belangrijke voorwaarde is voor de effectiviteit ervan. Zo was het beleid van The Knowledge Promotion in Noorwegen een antwoord op de tegenvallende resultaten in internationale vergelijkende studies. Het evaluatiesysteem dat binnen dit beleid werd opgezet had het doel om de leerprestaties van leerlingen te verbeteren door leraren en scholen een beter inzicht te geven in deze leerprestaties. Door het ontbreken van een rationale (hoe dragen de verschillende toetsen bij aan het oplossen van het probleem) en het koppelen van zowel ontwikkelingsgerichte als verantwoordingsgerichte doelen aan de meeste toetsen (zie onder, sectie 5.1.2), is de impact van het evaluatiesysteem er momenteel evenwel beperkt. De drie andere leer(winst)monitoringsystemen onderzocht in de internationale studie zijn minder expliciet gericht op het ‘oplossen van een welbepaald probleem’ (hoewel er uiteraard specifieke doelen werden vooropgesteld); het gaat hier evenwel om drie onderwijscontexten waarin het gebruik van gestandaardiseerde toetsen door overheid en door scholen reeds een lange traditie kent.

5.1.2 Een duidelijk en eenduidig doel

In bovengenoemde rationale moet onder meer duidelijk blijken of het doel van de gestandaardiseerde toetsen ontwikkelingsgericht, dan wel verantwoordingsgericht is. Het samengaan van beide perspectieven bleek ook in verschillende cases ertoe bij te dragen dat de ontwikkelingsgerichte doelstelling slechts in beperkte mate werd bereikt.

“Het blijft altijd lastig, he. Naarmate je meer ter verantwoording wordt geroepen, heb je de neiging om je problemen te verbergen. En als je wil ontwikkelen, is het juist belangrijk dat je de problemen naar boven haalt om er wat aan te gaan doen. Dat is heel moeilijk, ik denk niet dat je’t allebei kunt doen. Met feedback is het zo dat als je mensen zegt dat het op een vertrouwelijke manier gaat, dan is het stimulerender om er iets mee te doen, dan iemand publiekelijk ter verantwoording te roepen, want dan gaat die in de defense modus zitten.” (interview, respondent NL7)

We zagen in de case studies verschillende voorbeelden waarin verantwoordingsgerichte doelen toegevoegd worden aan toetsen die werden ontwikkeld in een ontwikkelingsgericht perspectief. Het toevoegen van bijkomende doelen in de loop van het ontwikkelingsproces blijkt evenwel vaak ineffectief en te leiden tot onwenselijke neveneffecten. Voorbeelden daarvan zijn teaching to the test, oneigenlijk (en onwenselijk) gebruik van de toetsresultaten, exclusie van leerlingen uit scholen, en een emotionele impact op leerlingen en leraren.

Leer(winst)monitoringsystemen zonder duidelijk uitgesproken ontwikkelingsgerichte of verantwoordingsgerichte doelstelling, blijken zelden effectief. De kritiek op het gebrek aan succes van NAPLAN in de case Australië heeft vaak ook betrekking op deze mix van doelstellingen, de

neveneffecten die deze mix teweegbrengt en als gevolg daarvan de beperkte bruikbaarheid. De overtuiging leeft zeer sterk bij de respondenten in Australië dat het belangrijk is om nog voor de eerste stappen in toetsontwikkeling worden gezet, om duidelijk één doel te bepalen.

"I think you should be careful (...) that you don't get people using an assessment that is designed for one purpose to try and do other things, cause that is where things start to go wrong in all systems. You get tests that are designed for accountability and they start to be used to do diagnosis - they are the wrong instruments for that - and vice versa. (...) To me, it's really essential: sorting out what the purpose is and designing it for the purpose." (interview, respondent AU8)

Bovendien kunnen we uit de case studies ook concluderen dat bij de ontwikkeling van gestandaardiseerde toetsen best gericht wordt op één niveau (micro-, meso- of macroniveau). Het gebrek aan aanvaarding van de meeste gestandaardiseerde toetsen in Noorwegen kan onder meer worden verklaard door het feit dat de toetsen - voornamelijk gericht op het mesoniveau - te weinig opleveren op het microniveau. Het is een moeilijk evenwicht voor gestandaardiseerde toetsen tussen informatie vergaren die zinvol is op het mesoniveau, en informatie die zinvol is op het microniveau. De toetsen die leraren nieuwe informatie kunnen aanbrengen over hun leerlingen vereisen immers een grote mate van detail, waardoor er slechts een beperkt domein kan getoetst worden. Op schoolniveau is men echter meer gebaat bij toetsen die een iets ruimere focus hebben, en waar de schoolleiding en leraren een globaal beeld krijgen van het leren of de leerwinst van leerlingen in een volledig leerdomein.

5.1.3 Welke consequenties aan de toetsresultaten verbinden?

Uit het Australische voorbeeld kunnen we opmaken dat de formatieve toetsen door scholen sterk gewaardeerd worden en dat ze hen richting geven om hun kwaliteit verder uit te bouwen, zowel op micro- als op mesoniveau. In zowel Noorwegen als Australië zien we bevestigd dat hoe hoger de (gepercipieerde) inzet van een toets is, hoe meer onwenselijke neveneffecten verwacht kunnen worden. Bovendien wordt er in Nederland gewaarschuwd voor de indirecte kost van verantwoordingsgerichte toetsen, die ruimer is dan enkel de ontwikkeling, afname en correctie van de toetsen:

"Vaak wordt er niet goed nagedacht over alle nadelen of alle kosten die aan de controle verbonden zijn. Ik bedoel: er moeten mensen de controle uitvoeren, de mensen die verantwoording afleggen, die gaan ook veel tijd besteden aan papierwerk inleveren, en dat is allemaal tijd die ze niet kunnen besteden aan hun eigenlijke taak. En dan krijg je nog - meestal geeft het toch ergernis en gedoe, en vaak is het dan ook nog de vraag van 'ja, is alle informatie echt wel betrouwbaar?' Je kan het mooier voordoen dat het eigenlijk is, dat gevaar ligt ook heel erg op de loer." (interview, respondent NL7)

In elk van de case studies blijkt dat wanneer toetsen een verantwoordingsgericht karakter hebben, de gebruikte toetsitems meteen na de toetsafname worden neergeschreven en gebruikt worden bij de voorbereiding van volgende toetsen. Bij een verantwoordingsgericht gebruik of een competitief element verdient het dus aanbeveling om telkens een volledig nieuwe toets te ontwikkelen. Dit is in veel mindere mate het geval bij toetsen die enkel in een ontwikkelingsgericht perspectief worden gebruikt. Bij bijvoorbeeld de toetsen van CEM in Engeland⁹⁹, van de VCAA en van ACER in Australië, van CITO-LVS in Nederland alsook de diagnostische toetsen in Noorwegen is enkel een regelmatige evaluatie en herinschaling van de items nodig. De items op zich kunnen wel worden hergebruikt.

Toch komt uit de case studie in Australië ook een waarschuwing voor low-stakes toetsen die op geen enkele manier van belang zijn voor de leerlingen. Uit de resultaten van de leerlingen in het negende jaar in Western Australia voor de NAPLAN-toetsen leren we ook dat leerlingen - voornamelijk in het secundair onderwijs - niet altijd gemotiveerd zijn om op een toets te laten zien wat ze werkelijk kennen en kunnen. Dit is uiteraard een bedreiging voor de validiteit van de toetsresultaten. De motivatie wordt (extrinsiek) versterkt door ook voor de leerlingen consequenties aan de toetsresultaten te verbinden: zoals blijkt uit het aangehaalde voorbeeld, kan een hogere inzet van de toets voor de leerlingen dus een positief effect hebben op de validiteit van de toetsresultaten. Het roept de vraag op of louter formatieve toetsen, die geen enkele verdere betekenis hebben voor de leerlingen, een betrouwbare maatstaf kunnen vormen voor conclusies op meso- of macroniveau. Een soortgelijke ervaring in Noorwegen heeft ertoe geleid dat men aan de nationale toetsen in het elfde leerjaar naast het oorspronkelijke ontwikkelingsgerichte perspectief, tevens een verantwoordingsgericht perspectief op microniveau heeft toegevoegd.

5.1.4 Een ontwikkelingsgerichte impact op microniveau is niet makkelijk te bereiken

In Nederland heeft het CITO-LVS een sterke impact op de manier waarop leraren in de bezochte scholen de data in verband met leerprestaties en leerwinst gebruiken om het onderwijsaanbod bij te sturen en om individuele begeleidingstrajecten op te zetten. Een soortgelijke impact hebben bepaalde toetsen in het leer(winst)monitoringssysteem van CEM in Engeland. Dit gaat in tegen de gangbare stelling dat gestandaardiseerde toetsen nauwelijks nieuwe inzichten kunnen genereren voor leraren wat betreft hun leerlingen. Bovendien blijkt dat zelfs wanneer de toetsresultaten relatief weinig nieuwe informatie opleveren, ook de objectivering van een eigen oordeel door verschillende respondenten in het internationale onderzoeksluik als waardevol wordt omschreven.

Daar staat evenwel tegenover dat het grootste punt van kritiek op de gestandaardiseerde toetsen in Noorwegen (en meteen ook één van de hoofdredenen waarom de impact ervan relatief beperkt blijft) is dat scholen relatief weinig nieuwe informatie krijgen over het leren of de leerwinst van de leerlingen.

⁹⁹ Enkel in Noord-Ierland worden verschillende CEM-toetsen wel aangekocht en ter beschikking gesteld door de overheid, maar met de Noord-Ierse overheid werd op voorhand overeengekomen dat deze toetsen niet binnen een verantwoordingsgericht perspectief zouden worden ingezet. Men merkt negatieve reacties bij scholen waar deelname aan de gestandaardiseerde toetsen werd opgelegd (door een overheid, door een scholengroep, door een schoolbestuur). Waar er (zoals in Noord-Ierland) sterk op wordt gefocust dat de resultaten enkel voor ontwikkelingsgericht gebruik kunnen worden ingezet, en waar er sterk wordt ingezet op ondersteuning en training van leraren, zijn de problemen van aanvaarding minder groot.

5.1.5 Aantonen van een vaak nog ongekende behoefte

Bij een beslissing over een eventuele implementatie van ontwikkelingsgerichte toetsen kan niet enkel uitgegaan worden van wat de actuele behoeften zijn van leraren in Vlaamse scholen. Vaak zijn leraren er immers van overtuigd dat ze weinig kunnen leren van gestandaardiseerde toetsen. De resultaten lijken allen in het verlengde van hun verwachtingen te liggen. Om dit effect tegen te gaan (en aldus de aanvaarding/meerwaarde van gestandaardiseerde toetsen te vergroten) wordt door respondenten in Engeland aangeraden om leraren op voorhand een voorspelling te laten maken van de resultaten van de verschillende leerlingen en om in een latere fase deze voorspelling aan de hand van de resultaten te bespreken. Vaak ondervindt men pas de meerwaarde van gestandaardiseerde data wanneer men deze reeds heeft kunnen gebruiken. Dit past in het principe van *"Persuade people by doing"* (interview, respondent EN3). Door gestandaardiseerde toetsen kan nieuwe en verrassende informatie naar boven gebracht worden en het is net die meerwaarde, die moet worden getoond in scholen (*'creating the need'*):

"Just get started - if you collect data, you will definitely find things that people don't know." (interview, respondent EN3)

5.1.6 Bij aanvang op vrijwillige basis aanbieden

Bij implementatie van een leer(winst)monitoringssysteem kan verwacht worden dat er heel wat (terechte) bezorgdheden worden geuit. Het is dan ook belangrijk om bij de implementatie duidelijk te maken aan scholen en ouders wat het voordeel is voor hen als partners binnen het onderwijs om over betrouwbare informatie te beschikken inzake de leeruitkomsten van leerlingen. Het is daarbij noodzakelijk om sterk te investeren in het ondersteunen van een dialoog in scholen (of met scholen) op basis van de resultaten van gestandaardiseerde toetsen.

Bijgevolg zien de Australische respondenten weinig heil in verplichte toetsen, zeker bij aanvang: toetsen zijn maar effectief op het ogenblik dat ze aanvaard zijn door de scholen die ze moeten afnemen. De enige toetsen die er verplicht maar toch erg gecontesteerd zijn (NAPLAN), lijken hun ontwikkelingsgerichte doelstelling slechts in beperkte mate te realiseren.

Ook in andere onderzochte onderwijssystemen zien we dat de aanvaarding van elk leer(winst)monitoringssysteem doorheen de jaren gegroeid is - alhoewel er in elk systeem nog steeds fervente tegenstanders zijn. Initiële twijfelaars onder de stakeholders kunnen worden overtuigd van de meerwaarde van een leer(winst)monitoringssysteem, maar dan is het bereiken van een positieve impact, voornamelijk op micro- en mesoniveau, uiteraard een voorwaarde.

Om bovenstaande reden wordt door respondenten in CEM aangeraden om met de implementatie van een leer(winst)monitoringssysteem te beginnen in een relatief klein aantal scholen. Om reeds in de eerste scholen op te starten is het, aldus deze respondenten, niet noodzakelijk dat reeds het hele monitoringssysteem is ontwikkeld. Het toetsaanbod kan gradueel uitgebreid worden, net als de rapporteringsmogelijkheden. Een stapsgewijze ontwikkeling en implementatie reflecteert de eigen

ontwikkeling van het leer(winst)monitoringssysteem van CEM en biedt, aldus de ervaringen van deze respondenten, meer garantie op succes. Wellicht is het ook beter om wat leerdomeinen betreft kleinschalig te beginnen, en op basis van de aanvaarding ervan langzaam aan uit te breiden. Toch verdient het aanbeveling om niet enkel met de makkelijk meetbare zaken te beginnen, maar van bij aanvang reeds enkele moeilijker te toetsen eindtermen mee te integreren in de gestandaardiseerde toets. Zo waren bijvoorbeeld attitude-metingen ook reeds ontwikkeld bij de eerste stappen van CEM.

Daarnaast verdient het aanbeveling om in een eerste fase slechts op één versie (bijvoorbeeld een web-based versie voor PC, best via *Central Hosting* met een *cloud-based platform*) in te zetten, omdat de ontwikkeling van verschillende versies (lokale versie, tablet-versie) technisch bijzonder arbeidsintensief is.

5.2 Eigenschappen van het leer(winst)monitoringssysteem

5.2.1 *Digitale toetsen om snelle feedback te voorzien*

Verschillende respondenten geven de voorkeur aan digitale toetsen boven de klassieke pen-en-papier-toetsen. De snelheid van de feedback - die op grote schaal enkel mogelijk wordt geacht via digitale toetsen - wordt als essentieel ervaren om een impact te genereren, zeker wat betreft het microniveau. In Australië heeft NAPLAN momenteel nauwelijks een effect op het microniveau van leraren en leerlingen. Dit wordt sterk gerelateerd aan de lange periode tussen testafname en het opleveren van de resultaten. Ook bij het pilootproject in Nederland, de Nationale toetsen in Noorwegen en de Oslo-toetsen was een belangrijk onderdeel van de kritiek dat er een te lange periode was tussen toetsafname en beschikbaarheid van de resultaten. Daartegenover staat de directe feedback die wordt gegenereerd door leer(winst)monitoringssystemen als bij CEM, de VCAA, EAA en ACER. Digitale toetsen hebben bovendien twee extra stimulansen voor leraren, met name: dat een adaptieve toets kan ontwikkeld worden (zie sectie 5.2.2); en dat in de rapportering automatische linken kunnen worden voorzien naar delen van het curriculum (en eventueel naar concrete adviesmogelijkheden) (zie sectie 5.6.2).

Toch zijn digitale toetsen niet evident, zeker niet voor toetsen met een hogere inzet die door alle leerlingen simultaan moeten worden afgenomen. *High-stakes* toetsen worden dan ook voornamelijk op papier afgenomen. Er zijn immers heel wat logistieke problemen om een toets door alle leerlingen op hetzelfde moment te laten afnemen (vereiste van een goede IT-infrastructuur maakt het voor scholen heel wat duurder), en de verschillen tussen infrastructuur kunnen leiden tot oneerlijke of onbetrouwbare resultaten. Ook zijn er op die manier minder veiligheidsissues (inzake lekken van vragen). Enkel in Noorwegen zagen we een voorbeeld van digitale toetsen met een eerder high-stakes karakter op vrij grote schaal. Ook hier werd evenwel aangegeven dat een belangrijk thema in de beslissing van de vorm van gestandaardiseerde toetsen de beschikbaarheid van ICT-mogelijkheden in de verschillende scholen moet zijn. Aangezien Noorwegen sinds 2000 één van de leidende landen inzake de inclusie van ICT in onderwijs en de beschikbaarheid van laptops in scholen is (Krumsvik,

2006), werd het gebruik van digitale toetsen in een verantwoordingsgericht perspectief in deze context wel mogelijk geacht.

Los van welke keuze wordt gemaakt (digitaal of op papier), is het strategisch belangrijk om reeds bij aanvang van de toetsontwikkeling duidelijk deze beslissing te hebben genomen. Het gebruikte devies heeft immers een impact op uiteenlopende belangrijke parameters, zoals de soort vraagstelling, de context waarin items kunnen worden aangeboden en de hoeveelheid te verwerken informatie:

“Also, we didn’t realize that putting them on computer makes them completely different tests from on-paper. (...) Your decision whether it’s gonna be on computer or on paper is gonna be crucial.” (interview, respondent EN6).

5.2.2 *Adaptieve, of getrapt adaptieve toetsen*

Adaptieve toetsen worden door verschillende respondenten gezien als een grote meerwaarde, omdat ze meer gedetailleerde informatie van leerlingen kunnen weergeven, omdat ze motiverender zijn voor leerlingen die de toets afnemen, en omdat ze (zeker inzake hoog- en laagpresteerders) meer betrouwbare informatie voor leraren en scholen kunnen opleveren. Daar staat echter tegenover dat er een zeer ruime itembank vereist is. Zeker wanneer het curriculum-gebonden toetsen betreft, waarbij verschillende onderdelen van het curriculum evenredig dienen getoetst te worden, is een bijzonder grote itembasis noodzakelijk.

Adaptieve toetsing is volgens bepaalde respondenten dan weer minder aangewezen bij high-stakes toetsen met gevolgen op de schoolloopbaan van leerlingen of de evaluatie van leraren, omdat de hoge consequenties op microniveau een meting zouden vereisen waarbij alle leerlingen dezelfde items aangeboden krijgen (om de perceptie van een gelijke lat te behouden). De A-vision eindtoets voor het primair onderwijs in Nederland (als één van de alternatieven voor Eindtoets Primair Onderwijs van CITO en CvTE) is de enige high-stakes toets op microniveau die een adaptief opzet hanteert.

Naar het voorbeeld van bepaalde toetsen van ACER en van de plannen voor NAPLAN (alook, in zekere mate, de ‘General Adaptive Tests’ van EAA), kunnen toetsen ook getrapt adaptief worden aangeboden (‘*branching*’). Hierbij wordt niet na elk individueel item bijgestuurd naar een moeilijker of makkelijker vervolgitem, maar wordt de score op een set van items gebruikt om naar een vervolgset van een hogere of lagere moeilijkheidsgraad door te gaan. In tegenstelling tot de zeer arbeidsintensieve ontwikkeling van een volledig adaptieve toets, is een getrapt adaptieve toets veel economischer terwijl relatief weinig informatie verloren gaat. Bij CEM (bijvoorbeeld PIPS) wordt dan weer een andere invulling gegeven aan ‘adaptieve toetsing’, door een toets te laten overgaan naar een volgend onderdeel nadat de leerling een voorafbepaald aantal foutieve antwoorden heeft gegeven.

5.2.3 *Curriculum-gerelateerde toetsen (of net niet?)*

In Australië en in Nederland hecht men er sterk aan om een leer(winst)monitoringssysteem te baseren op het curriculum. Waar dat niet het geval is (bijvoorbeeld de meer algemene vaardigheidstoetsen)

worden er expliciet linken gelegd tussen de toetsitems en het curriculum. Ook deze link draagt sterk bij tot de mate waarin het leer(winst)monitoringssysteem door scholen wordt gebruikt voor de versterking van hun onderwijskwaliteit.

"It is important that it is linked to the curriculum. The teachers are teaching to the curriculum. The test should align with that, or it would be some sort of crazy thing that comes through. If it is not aligned to the curriculum it has no relevance whatsoever for what the teachers are doing and what the children are doing." (interview, respondent AU13)

Daartegenover staat dat zowel bij CEM in Engeland als in Noorwegen sterk wordt geïnvesteerd in toetsen die niet meteen gelinkt zijn aan het curriculum. In Noorwegen uit zich dat in toetsen rond vaardigheden die men ziet als essentiële einddoelen van het leerplichtonderwijs. Deze vaardigheden (leesvaardigheid en rekenvaardigheid) behoren tot de vijf vaardigheden die centraal staan in het Noorse nationale curriculum, maar niet aan de curriculum-leerdoelen van één bepaald leerdomein te koppelen zijn. Ze kunnen meer beschouwd worden als vakoverschrijdende doelen, die overheen de vakspecifieke curricula van alle leerdomeinen dienen te worden nagestreefd. Toch is er een belangrijke tegenbeweging aan de gang. Zo worden bijvoorbeeld in opdracht van het schoolbestuur van Oslo wel bijkomende curriculum-gerelateerde toetsen ontwikkeld om deze gepercipieerde lacune in te vullen. In CEM gaat men ervan uit dat het zinvol is voor scholen om een beeld op het algemene vaardigheidsniveau van leerlingen te krijgen dat niet aan het curriculum gelinkt is. Deze toetsen bieden scholen extra informatie over meer generieke capaciteiten van leerlingen aan en zijn daarom erg zinvol om geïndividualiseerde doelstellingen voor leerlingen voorop te stellen of om extra betekenis te geven aan leerwinstberekeningen. Daarnaast geven deze toetsen een overzicht van hoe leerlingen doorheen de tijd - in verschillende cohorten - presteren op algemene wiskundige vaardigheid en taalvaardigheid. Deze toetsen kunnen daarom ook bijzonder interessant zijn binnen een ontwikkelingsgericht gebruik op macroniveau. Een voorbeeld hiervan is de studie van Coe (2007), die op basis van verschillende cohorten aantoonde dat de gemiddelde 'algemene vaardigheid' van zestienjarige leerlingen is gestegen in de periode tussen 1997 en 2006.

De inzet van curriculum-gebaseerde toetsen leidt bovendien tot het probleem dat scholen niet allemaal op hetzelfde tempo dezelfde leerstof aanbieden. De doelen in het curriculum zijn immers in geen van de bestudeerde onderwijssystemen per trimester of leerjaar gedefinieerd (wat evenmin in Vlaanderen het geval is). Dit zou kunnen opgelost worden door een zekerheidsmarge in te bouwen, met name door te toetsen wat bijvoorbeeld gemiddeld een trimester tevoren werd aangeleerd, maar de ervaring bij CEM leert dat dit geen waardevol alternatief biedt omdat het leidt tot onwenselijk strategisch gedrag, zelfs bij de toetsen uit het low-stakes leer(winst)monitoringssysteem van CEM :

"The thing is with the curriculum testing, you have to be careful. We're faced with the problem, that we need to be prescriptive, if we're going to test the curriculum, then we have to make sure that the pupils have already been taught it. And if it's been taught last year and teachers know that we're going to test it, then teachers are going to give a refresher." (interview, respondent EN6)

De ontwikkeling van curriculum-gerelateerde toetsen houdt bovendien in dat de toetsitems bij elke mogelijke verandering in curricula moeten nagekeken worden op inhoudsvaliditeit. Ook moeten de aangeboden items en de inschaling van de items op een RASCH- of IRT-schaal telkens opnieuw nagekeken worden (wat bijvoorbeeld bij de VCAA in Australië gebeurt voor de *On Demand Testing*). Voortdurende wijzigingen aan een (low-stakes) toets worden evenwel niet altijd op prijs gesteld door leraren:

“Teachers are very conservative about assessment, and they sort of get very distrustful of the test if we say ‘oh, we have changed the test this year’, people don’t like that we change the test.” (interview, respondent EN6)

Bovendien bestaan er (in het secundair onderwijs) uiteenlopende curricula voor verschillende studierichtingen. De curriculum-gerelateerde gestandaardiseerde toetsen voor het secundair onderwijs in Noorwegen worden enkel mogelijk geacht door het sterk comprehensieve karakter van het Noorse onderwijs: wanneer er een uitgebreid scala van studierichtingen is, is dit minder haalbaar volgens de geïnterviewden.

Omwille van de stijgende vraag van scholen voor curriculum-gerelateerde toetsen, wordt sinds enkele jaren echter ook door CEM meer geïnvesteerd in de ontwikkeling van curriculum-gerelateerde toetsen. Ondanks de groeiende markt, heeft men enig voorbehoud bij deze toetsen, omdat ze zich meer lenen tot een verantwoordingsgericht gebruik (bijvoorbeeld om vakleraren van de leerdomeinen waarin leerlingen minder presteren ter verantwoording te roepen).

5.2.4 Welke leerdomeinen toetsen?

Uit deze internationale studie kan geen eenduidig advies worden gehaald over het vraagstuk of best alle domeinen, dan wel enkel de kernvakken getoetst worden. Bij een high-stakes toets (zeker wanneer deze effecten heeft op de studievoortgang van leerlingen, cf. het HSC-examen of de GCSE) lijkt het aangewezen om alle leerdomeinen als gelijkwaardig te beschouwen en dus voor elk vak een toets aan te bieden. Daar staat tegenover dat de meeste andere leer(winst)monitoringssystemen slechts enkele vakken omvatten, meestal reken- en leesvaardigheid. Bij ACARA is de rationale expliciet dat men vermoedt dat de prestaties in deze kernvaardigheden een goede proxy zijn voor de prestaties in andere leerdomeinen.

Een reden voor de beperkte focus van de meeste leer(winst)monitoringssystemen is dat het erg moeilijk is om nieuwe toetsen te ontwikkelen en uit te testen. Zo werd bij CEM een eerdere toets om *‘creative writing’* te evalueren sinds enkele jaren niet meer aangeboden omdat de correctie ervan (financieel) te omslachtig was. Wat betreft deze moeilijk toetsbare domeinen, gaat men er dan ook van uit dat deze misschien beter ongetoetst blijven in het bestaande leer(winst)monitoringssysteem:

“In some cases it is better not to have an assessment at all than to have one that is reporting something that it isn’t actually measuring.” (interview, respondent EN2)

Daarnaast gaan er meer en meer stemmen op om niet enkel cognitieve, maar ook niet-cognitieve vaardigheden en attitudes van leerlingen in kaart te brengen. Deze tendens zien we bijvoorbeeld in Australië (bijvoorbeeld de toetsen van ACER), en in Engeland (bij CEM).

5.2.5 Brede toetsenbatterij

Zowel bij CEM in Engeland als bij ACER in Australië luidt het dat het succes van de toetsen zeker deels kan worden verklaard door het feit dat het een leer(winst)monitoringssysteem betreft waarin uiteenlopende gestandaardiseerde toetsen worden aangeboden die één of meerdere malen kunnen worden afgenomen. Scholen kunnen, afhankelijk van hun eigen noden, kiezen welke gestandaardiseerde toetsen op welk ogenblik worden ingezet. In CEM wordt bovendien meer en meer geïnvesteerd in de koppeling van de verschillende gestandaardiseerde toetsen zodat de rapportering vanuit de uiteenlopende toetsinhouden op elkaar is afgestemd. In Noorwegen wil men in de toekomst meer duidelijke richtlijnen opstellen bij de toetsontwikkeling opdat de verschillende toetsen meer op elkaar zijn afgestemd; waardoor ook hier een evolutie kan ontstaan naar een echte toetsenbatterij.

5.2.6 Sterke psychometrische onderbouwing

In elke case studie wordt er uitvoerig gedocumenteerd welke procedures er worden gevolgd om de betrouwbaarheid van de toetsen te garanderen. In elke case vonden we dat gestandaardiseerde toetsen in een leer(winst)monitoringssysteem uitvoerig worden ge-pre-toetst en een stevige statistische onderbouw kennen. Deze statistische basis wordt door de scholen sterk gewaardeerd, helpt bij de aanvaarding van de toetsen door leraren, en draagt bijgevolg bij tot de mate waarin men in scholen van de toetsen gebruik maakt binnen hun eigen kwaliteitszorg op meso- of microniveau. RASCH-modellen of IRT-modellen vinden we in elk van de bezochte case studies.

Daarnaast wordt in elk van de systemen met veel zorg ingezet op cesuurbepaling - zeker waar de inzet op leerlingniveau erg hoog is, zoals in de HSC-examens in Australië. Er zijn verschillende manieren van cesuurbepaling. Vaak zien we dat de cesuren pas nadat de toetsresultaten zijn geanalyseerd definitief worden vastgelegd.

Toch zijn verschillende respondenten niet overtuigd van de noodzaak hiervan. Ondanks de lange traditie in landen als Australië en Engeland, is het in vele scholen nog steeds erg moeilijk om uit te leggen hoe een RASCH/IRT-model werkt en hoe het door de scholen kan gebruikt worden. Aangezien het een hele investering is om alle items op een gemeenschappelijke schaal te plaatsen, moet zorgvuldig worden nagegaan voor welke toetsen dit werkelijk noodzakelijk is om zinvolle informatie op te leveren.

5.2.7 Sterke procedures voor kwaliteitszorg bij afname en correctie

Het is opvallend dat voornamelijk bij high-stakes toetsen werk wordt gemaakt van strikte procedures voor de afname en de correctie van de gestandaardiseerde toetsen. De nauwgezet omschreven

procedures die worden gehanteerd bij toetsen met een hoge inzet worden ervaren als een cruciaal element binnen de wijd verspreide aanvaarding van deze gestandaardiseerde toetsen. Er gaan slechts zelden stemmen op die menen dat de toetsen een oneerlijk beeld geven van de kennis en vaardigheden van de leerlingen. Bij low-stakes toetsen zien we meestal dat de procedures voor afname en correctie minder sterk zijn omschreven. Nochtans is de betrouwbaarheid van de afname en correctie bij low-stakes toetsen op zich niet van minder groot belang om betrouwbare resultaten te bekomen.

5.2.8 *Samenwerking tussen verschillende actoren in toetsontwikkeling*

Het betrekken van veel actoren bij de ontwikkeling van een leer(winst)monitoringssysteem, kan helpen bij de aanvaarding van de toetsen (en de perceptie van betrouwbaarheid en validiteit). Dit betekent evenwel niet dat het proces van toetsontwikkeling daardoor vlotter verloopt. Het lijkt een vuistregel genoemd door verschillende respondenten in Australië dat hoe meer mensen betrokken zijn, hoe trager het proces verloopt, en hoe hoger de kostprijs is. Toch zien we in elke case studie dat bij zowat elke toets een samenwerking wordt tot stand gebracht tussen inhoudelijke experts, toetsexperten en (bij digitale toetsen) technische experts.

Bij de ontwikkeling van toetsen is een goede samenwerking tussen test-inhoudelijke ontwikkelaars en de technische ontwikkelaars een cruciale succesfactor. Dat beide groepen ontwikkelaars binnen dezelfde organisatie werken, wordt door CEM in Engeland, en door CITO in Nederland, als een enorm pluspunt ervaren. Het volledig ‘binnenshuis’ houden van het gehele ontwikkelingsproces past ook in het vorige principe om de ontwikkeling zo eenvoudig mogelijk te organiseren. Ook in Noorwegen is er een nauwe dialoog tussen de inhoudelijke toetsontwikkelaars en de technische ondersteuners van Inspera Assessment.

5.2.9 *Gebruik maken van bestaande (internationale) toetsen*

Verschiedende gevalideerde toetsen zijn reeds ter beschikking. Zo zijn voornamelijk niet-curriculum gebonden toetsen (CEM, nationale toetsen Noorwegen) relatief makkelijk te vertalen, aangezien deze systemen slechts in beperkte mate contextgebonden zijn. Een Nederlandse vertaling van PIPS Baseline (van CEM) is reeds beschikbaar (OnderBouw InformatieSysteem of OBIS, vertaling door Universiteit Nijmegen). Bij overname van buitenlandse toetsen moet evenwel zorg worden gedragen van een herstandaardisering die zeer afhankelijk is van de onderwijscontext en van het gebruikte curriculum. Van een standaardisering van toetsen ontwikkeld op basis van leerlingresultaten in het buitenland, kan niet zonder meer worden verondersteld dat deze ook voor Vlaanderen geschikt is.

5.3 Leerwinst

Verschiedende systemen stellen zich voor als leerwinstmonitoringsystemen. Toch zijn systemen *geïnitieerd door een overheid* die leerwinst meten uiterst zeldzaam. Dit zien we eigenlijk enkel bij het pilootproject leerwinst en toegevoegde waarde in Nederland - en zelfs daar maakt men gebruik van

eerder ontwikkelde toetsen (door CITO). NAPLAN in Australië laat dan weer toe dat leerwinst gemeten wordt, maar daar wordt in de rapportering in scholen niet op ingezet (op geaggregeerde scores van leerlingen na). Scholen kunnen wel zelf de leerwinst van elke leerling aan de hand van de opeenvolgende rapporten berekenen, indien twee opeenvolgende NAPLAN toetsen in de eigen school werden afgenomen. Het betreft dan evenwel leerwinst over een periode van twee schooljaren, wat een vrij lange periode is om te overspannen.

Toetssystemen van private aanbieders (buiten de overheden) voorzien wel vaak een leerwinstmonitoring, zoals we zagen in de case studies van Australië, Engeland en Nederland.

5.3.1 *Belang van het meten van leerwinst*

Uit de cases in Nederland, Noorwegen, en Australië blijkt duidelijk dat wanneer scholen informatie over de leerwinst van leerlingen kunnen verkrijgen, hier meestal gretig gebruik van wordt gemaakt. Kritiek op bestaande monitoringssystemen zonder deze mogelijkheid, is vaak gericht op het ontbreken van leerwinstinformatie wat door verschillende respondenten als een 'gemiste kans' wordt omschreven. Zo is bijvoorbeeld het grootste punt van kritiek op het evaluatiesysteem in Noorwegen (en meteen ook één van de hoofdredenen waarom de impact ervan relatief beperkt blijft) dat scholen relatief weinig nieuwe informatie krijgen. De meerwaarde zou zich voornamelijk manifesteren indien de toetsen zouden toelaten om het leerproces en de leerwinst van leerlingen op een betrouwbare manier in kaart te brengen. Ook doorheen het volledige toetsenarsenaal van CEM in Engeland wordt de meerwaarde gepropageerd van het vergelijken van de resultaten van leerlingen met hun eigen eerdere toetsresultaten. Deze leerwinstbenadering zet leerlingen en scholen aan tot een maximale ontwikkeling.

Uit de verschillende cases blijkt ten slotte ook dat leerwinstmeting met verschillende toetsen doorheen de schoolloopbaan neveneffecten als *'teaching the test'* kan voorkomen.

5.3.2 *Eigenschappen van een leerwinstmonitoringssysteem*

Het verdient verder aanbeveling om in de Vlaamse context goed na te gaan wat scholen wensen, en aan welke vereisten een leerwinstmonitoringssysteem moet voldoen om de kans op optimaal gebruik te maximaliseren. Daarbij moet onder meer worden ingezet op het vinden van een goede balans tussen enerzijds betrouwbaarheid van de berekening van leerwinst en anderzijds de relatieve eenvoud en transparantie van de concepten en berekeningswijzen en de gebruiksvriendelijkheid van de rapportage.

Leerwinst lijkt voornamelijk zinvolle informatie voor scholen op te leveren wanneer de periode tussen twee metingen niet langer is dan één jaar. De timing die in het pilootproject in Nederland wordt gehanteerd - met leerwinstberekening op basis van een halfjaarlijkse toetsing in het primair onderwijs - lijkt voor leraren en directies de meest zinvolle informatie op te leveren. Een nadeel hiervan is dat het grote aantal toetsen evenwel tot een bezorgdheid leidt inzake de toetsdruk die op leerlingen wordt gelegd.

Alle leerwinstmonitoringssystemen die werden beschreven, gebruiken een gemeenschappelijke vaardigheidsschaal die aan verschillende toetsen doorheen de schoolloopbaan ten grondslag ligt. Dit is overeenkomstig de eerste aanbeveling uit het Nederlandse pilootproject Leerwinst en Toegevoegde Waarde in het primair onderwijs.

Opvallend is dat leerwinstmonitoringssystemen zich bijna steeds richten tot het basisonderwijs, en dan nog meestal beperkt worden tot basisvaardigheden als lezen en rekenen. Voor meer complexe leerdomeinen is het moeilijker om een gemeenschappelijke vaardigheidsschaal te ontwikkelen. Dat gaat nog sterker op voor de leerdomeinen na het basisonderwijs, al zijn er toch enkele voorbeelden (bijvoorbeeld NAPLAN, of *On Demand Testing* van de VCAA) waar men ook een leerwinstmonitoringssysteem heeft voor het secundair onderwijs.

5.4 Toegevoegde waarde

De case studies resulteerden in geen voorbeelden van toegevoegde waarde-metingen die verder gaan dan het vergelijken van resultaten tussen scholen. Dit laatste gebeurt in Australië (bijvoorbeeld NAPLAN, HSC-examens), Engeland (Key Stage toetsen), en Noorwegen (bijvoorbeeld Oslo-toetsen, nationale toetsen). Naast statusmodellen, werd in Australië en Noorwegen ook een vergelijking gemaakt tussen de behaalde leerwinst in scholen. Dit zorgt voor een meer genuanceerd beeld tussen de scholen, waarbij relevante kenmerken van de leerlinginput van de school een minder grote impact hebben op de berekening van toegevoegde waarde.

Door de Nederlandse projectgroep Leerwinst en Toegevoegde Waarde in het primair onderwijs werd dan wel besloten dat toegevoegde waarde van scholen meten mogelijk is. Toch blijkt dat er nog onderzoek nodig is om de bestaande technieken te verfijnen. Ook de bezochte scholen in Nederland gaven aan dat het project hun verwachtingen inzake toegevoegde waarde nog niet had kunnen inlossen.

Bovendien blijken scholen (zowel in Nederland als in de andere case studies) voor hun interne kwaliteitszorg meest gebaat bij relatief eenvoudige vergelijkingen tussen de leerprestaties of leerwinst in de eigen school, en de vergelijkbare indicatoren in andere (vergelijkbare) scholen.

“Een probleem als je gaat werken met regressiemodellen is dat scholen niet precies begrijpen wat er gebeurt. Met die regressiemodellen, mensen wisten echt gewoon niet: wat gebeurt er nu? Een goed model is toch altijd vrij rechttoe rechtaan.” (interview, respondent NL6)

De berekening van de toegevoegde waarde is, aldus de projectgroep, het meest zinvol wanneer de periode waarover de toegevoegde waarde wordt berekend relatief beperkt is - bijvoorbeeld per twee leerjaren.

Bovendien concludeerde het pilootproject in Nederland dat het momenteel nog zeer onduidelijk is welke *fairness* kenmerken precies moeten worden gehanteerd om een eerlijke vergelijking tussen scholen mogelijk te maken. Men stuitte ook op veel praktische problemen bij het verzamelen van deze

gegevens. Niet alleen bleken er grote verschillen inzake de 'invoerdiscipline' van scholen - de bereidwilligheid van scholen om gegevens over leerlingen (bijvoorbeeld over achtergrondvariabelen) te verzamelen en (correct) in te voeren - ook de afstemming van verschillende systemen die in verschillende scholen werden gebruikt zorgden voor een onvoorziene werklast. Per deelnemende school moesten bestanden met toetsgegevens uit hun systemen worden geëxporteerd om de onderzoeksinstituten de berekeningen te laten uitvoeren.

Administratieve problemen kunnen tot op zekere hoogte vermeden worden indien met één systeem wordt gewerkt waarin de verschillende data beschikbaar zijn. Evenwel dienen ook in dat geval nog steeds afspraken met scholen te worden gemaakt over hoe (en welke) data in het systeem worden ingevoerd.

5.5 Rapportering

5.5.1 Gedetailleerde en interactieve rapportering

Een brede en gedetailleerde rapportering is belangrijk. De rapportering moet toelaten om informatie te geven over de prestatie op schoolniveau en van individuele leerlingen en klassen in vele subdomeinen van de toets. Daarnaast is het belangrijk dat de informatie snel ter beschikking staat van de leraren en de school.

Interactieve rapportering wordt in nagenoeg alle beschreven leer(winst)monitoringssystemen gebruikt. Door het interactieve karakter wordt de gebruiker in staat gesteld om de data op verschillende manieren te analyseren. De opgeleverde data sluiten op die manier zo dicht mogelijk aan bij de individuele doelstellingen van de gebruikers. Bovendien is dit in lijn met de zesde aanbeveling van het pilootproject in Nederland, met name om voldoende flexibiliteit in de rapportering te bewerkstelligen.

De bruikbaarheid en de inzichtelijkheid van de rapportering zijn belangrijk om ervoor te zorgen dat de gestandaardiseerde toetsen een impact kunnen hebben op micro- en mesoniveau, maar uit de cases kunnen we ook concluderen dat dit principe niet makkelijk te realiseren is. Dit blijkt onder meer uit de eigen evaluatie van het pilootproject in Nederland, die ondanks vele bijstellingen op basis van eerdere feedback nog steeds niet heeft geleid tot een overtuigend resultaat.

5.5.2 Vermijd publieke rapportering

In elk van de bezochte onderwijssystemen is er informatie over de geaggregeerde prestaties van leerlingen op gestandaardiseerde toetsen openbaar beschikbaar. Toch is deze publieke openbaarheid van resultaten in elk van de landen gecontesteerd, en werden verschillende onwenselijke neveneffecten aan deze openbaarheid toegeschreven. Ondanks de lange traditie in Australië om resultaten publiek bekend te maken, raden verschillende respondenten aan om de resultaten niet publiek te maken. Wel is iedereen in deze case studie het erover eens dat scholen én ouders informatie over hun school zouden moeten krijgen, waarbij de leerprestaties en/of leerwinst van de leerlingen in de eigen school in verhouding moet worden geplaatst met de prestaties en/of leerwinst

van leerlingen in andere (vergelijkbare) scholen (waarbij de vraag kan gesteld worden of dit niet tot een meer algemene openbaarheid zou leiden).

5.6 Ondersteuning van scholen

Een aanbod voor scholen doorheen het gehele proces van bij aanvang, afname, interpretatie van de opgeleverde data en de koppeling met actieplannen, wordt in elke van de case studies als een belangrijke succesvoorwaarde beschouwd. Dat aanbod gebeurt best geïndividualiseerd en op maat van de school, zoals uit de vermelde 'kritische kantelmomenten' in de Nederlandse case studie blijkt. Ook CEM in Engeland zet sterk in op individuele begeleiding van scholen. Of ook de meer algemene infosessies die door het schoolbestuur in Oslo worden georganiseerd een impact hebben op het effectieve gebruik van leraren en scholen van de resultaten van gestandaardiseerde toetsen, blijft onduidelijk.

Het is opvallend dat de noodzaak aan begeleiding op maat van scholen ook zo sterk wordt aanvoeld in de Nederlandse, Engelse en Australische onderwijscontext, waar scholen al decennia lang ervaring hebben met gestandaardiseerde toetsen. Dit begeleidingsaanbod wordt bij voorkeur aangeboden vanuit de organisatie die verantwoordelijk is voor de toetsontwikkeling. Zo geven verschillende respondenten in Australië aan dat er nog bijkomende professionele ontwikkeling voor leraren nodig is om hen voldoende vertrouwd te maken met de (interpretatie van) data. Dit blijkt ook uit de steeds sterkere vraag naar rechtstreekse koppelingen vanuit resultatenrapporten naar ondersteunende materialen en advies.

"If you want it to be an instrument of change, then I think you need professional development or that you need materials, or both, to do that follow-up..." (interview, respondent AU8)

Ondanks de jarenlange traditie in Engeland en de vele feedbackrapporten die scholen in de voorbije decennia reeds hebben ontvangen, wordt in alle geledingen van CEM benadrukt dat scholen nog steeds veel begeleiding nodig hebben (en vooral bij de interpretatie van de resultaten).

"It always worries me that they don't understand the feedback." (interview, respondent EN6)

Volgens de begeleiders heeft men in verschillende scholen evenwel vooral nood aan een externe instantie die hun eigen interpretatie ondersteunt ('*reassurance*').

Dat de informatie die uit de verschillende leer(winst)monitoringssystemen wordt opgeleverd, vaak niet ten volle leidt tot schoolontwikkeling, wordt ook in Nederland verklaard door de nood aan professionalisering en ondersteuning van professionals in scholen om de resultaten correct te interpreteren.

"Dat is een eerste les: je mag er echt niet van uitgaan dat alles goed aankomt!" (interview, respondent NL7)

5.6.1 *Ondersteuning bij (gebruiksvriendelijke) afname*

Zowel bij CEM als bij de nationale toetsen en de Oslo-toetsen in Noorwegen is er een voortdurend bemande helpdesk die telefonisch of via online kanalen zowel op inhoudelijke als technische vragen antwoord kan geven. In Noorwegen wordt in deze helpdesk zelfs actief naar mogelijke problemen gezocht tijdens de afname van de toetsen. Dit past in het kader van de overtuiging dat het noodzakelijk is om de werklast voor de school tot een minimum te beperken om de aanvaarding binnen scholen te verhogen.

5.6.2 *Investeren in vermogen van de scholen om informatie om te zetten in acties*

Naast technische ondersteuning bij de voorbereiding en afname, en de bovengenoemde ondersteuning bij de interpretatie van de resultaten, is er ondersteuning nodig bij het maken van de link tussen de geïnterpreteerde data en een actieplan (hoe kunnen we hier iets mee doen?).

Er lijkt in Australische scholen - algemeen gesproken - een relatief sterke cultuur van samenwerking en professionalisering aanwezig te zijn. Zo is er in de beide bezochte scholen '*professional learning time*' op elke woensdagnamiddag, geleid door interne '*professional leading teams*' bestaande uit leraren en middenkader in elke school. Deze sterke 'verbetercultuur' in scholen werd net als in Nederland genoemd als een belangrijke voorwaarde om een onderwijssysteem te doen groeien door middel van gestandaardiseerde toetsen. Het positieve effect van CITO-LVS werd in beide bevroegde scholen in Nederland voor een belangrijk deel toegeschreven aan de individuele besprekingen tussen leraren enerzijds en directie of interne begeleider anderzijds. Eén van de aanbevelingen in het Nederlands pilootproject was dan ook om een beleid te ontwikkelen waarin scholen de leerwinst-berekeningen maximaal zelf konden omzetten naar acties voor 'opbrengstgericht werken' en voor hun interne kwaliteitszorg.

Doorheen de verschillende cases worden uiteenlopende aspecten van het beleidsvoerend vermogen van scholen door verschillende respondenten vernoemd als cruciale randvoorwaarden. Het gaat daarbij onder meer om de gezamenlijke doelgerichtheid, het reflectief vermogen en het innovatief vermogen van scholen:

- De '*combined effort*' van de school: de toetsen hebben een grotere impact wanneer leraren en het management team samenwerken aan het verbeteren van de kwaliteit van de school. Intercollegiale consultatie wordt als een belangrijke factor genoemd in de impact van de toetsen op het onderwijs: mits feedback over de resultaten gekoppeld is aan een niet-bedreigende uitwisseling met een andere leraar (die door de eerste leraar sterk gerespecteerd wordt), wordt volgens de respondenten bij CEM in Engeland een sterkere impact op het lesgeven van leraren vastgesteld. Eenzelfde dynamiek geldt op schoolniveau bij directies, wanneer zij de schoolfeedback hebben kunnen bespreken met een collega-directeur.
- Daarnaast wordt gesteld dat de impact van de CEM-toetsen in Engeland groter is in scholen waar het duidelijk is welke persoon verantwoordelijk is om de data in de eerste plaats te analyseren en te bespreken met de verschillende geledingen binnen de school (in vele

gevallen is dat een *assistant headteacher*), maar er zijn ook voorbeelden van scholen waar een leraar of iemand in een andere functie deze rol op zich neemt. Bovendien blijkt in Engeland dat ook de Key Stage toetsen in de bezochte school een enorme impact hadden op het onderwijs door toedoen van de onderdirecteur die zich het gebruik van deze data sterk eigen had gemaakt.

- De toegang tot de resultaten voor alle leraren is ook een evidente voorwaarde, die evenwel niet in alle scholen aanwezig is:

"It's good for schools who are very open to the actual data, they will share it [with the teachers]." (interview, respondent EN1)

In de case studie van Noorwegen bepleiten respondenten dat de openheid over de leerprestaties van leerlingen zich idealiter niet beperkt tot de eigen school: de bereidheid van scholen om gegevens aan elkaar door te geven (zonder deze algemeen publiek te maken!) kan een grote invloed hebben op het onderwijsaanbod aan individuele leerlingen. Dit geldt zowel bij neveninstromers als bij de overgang van basis- naar secundair onderwijs. Het schoolbestuur van Oslo richt hiertoe overgangstoetsen in die door de leraren van lager en secundair onderwijs gezamenlijk besproken dienen te worden.

- Ook wordt in elke case studie het belang benadrukt dat het management in de school er niet op gericht is om leraren af te rekenen op de prestaties van leerlingen in hun leerdomein, maar dat de toetsresultaten binnen een ontwikkelingsgericht perspectief worden ingezet. Een context waarin leraren geen schrik hoeven te hebben om op de resultaten te worden (summatief) beoordeeld, wordt noodzakelijk geacht om een betrouwbare meting te garanderen. In Noorwegen werd bijvoorbeeld bij de nationale toetsen expliciet in de wet ingeschreven dat de resultaten ervan niet mogen worden gebruikt bij de evaluatie van leraren.

We zagen daarnaast enkele voorbeelden waarin het maken van de link tussen de geïnterpreteerde data en een actieplan een inherent deel uitmaakt van het leer(winst)monitoringssysteem. Verschillende toetsen in Australië en Noorwegen voorzien in de rapportering dat er automatische linken zijn tussen bepaalde items en het doel dat door middel van dit item werd getoetst. Daarnaast wordt de rapportering van de verschillende toetsen in Australië zo opgevat dat ze scholen en leraren automatische formatieve feedback geven, met ondersteunende informatie en inspiratie die kan worden ingezet wanneer eventuele problemen op klas-, of schoolniveau worden gedetecteerd. Ook bij CEM in Engeland blijkt dit één van de ambities voor de nabije toekomst: materialen ontwikkelen die bij bepaalde profielen van leerlingen (of van scholen) de leraar of de directeur inspiratie kunnen bieden door enkele model-acties voor te stellen.

Een voorbeeld waarbij dit reeds gerealiseerd is, zijn de diagnostische toetsen in Noorwegen. Een reden waarom de diagnostische toetsen relatief succesvol zijn, is de automatische koppeling van de toetsresultaten aan ondersteunende materialen die kunnen worden ingezet bij de remediëring van leerlingen:

“If there is any advice I would give to another government, it would be to emphasize the following up of the results, because it is very easy to be satisfied or dissatisfied: you get the results from the tests and then you say: I was down one tenth, or I was up one tenth, but that’s when the job starts.” (interview, respondent NW3)

5.7 Financiële inschatting is bijzonder moeilijk

Bij elk monitoringssysteem werd gepeild naar de financiële kost ervan. Geen enkele respondent kon een inzage geven in de totale kost. Zelfs in Australië met de ver doorgedreven cultuur van openbaarheid en transparantie, bleek het erg moeilijk om precieze cijfers te plakken op toetsontwikkeling, toetsafname, correctie en rapportering. In dit rapport zijn op basis van de beschikbare gegevens verschillende inschattingen gemaakt, die er allen op duiden dat de ontwikkeling van een leer(winst)monitoringssysteem een kost- en tijdsintensief werk is waarbij verschillende professionals gedurende een langere periode moeten worden betrokken.

LITERATUURLIJST

- ACARA. (2012). National Report on Schooling in Australia 2012 (pp. 101). Sydney: ACARA.
- ACARA. (2013). NAP National Assessment Program. Geconsulteerd op 11 augustus 2015, van <http://www.nap.edu.au/>
- ACARA. (2014). ACARA Annual Report 2013-14 (pp. 102). Sydney: ACARA.
- ACARA. (2015). MySchool. Geconsulteerd op 11 augustus 2015, van <http://myschool.edu.au/>
- ACARA. (s.d.). Report of 2014 NAPLAN test incidents (pp. 9). Sydney: ACARA.
- Acquah, D. (2013). Examining School Accountability (pp. 16). Manchester, UK: Centre for Education Research and Practice.
- Adler, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. London: Jessica Kingsley Publishers.
- AERA. (2000). AERA Position Statement on High-Stakes Testing in Pre-K - 12 Education. <http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx>
- Alberts, R. V. J., & Erens, B. J. M. (2015). Verslag van de examencampagne 2015 voortgezet onderwijs. Arnhem: Stichting CITO.
- Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, 20(1), 20-34.
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education policy analysis archives*, 10, 18.
- Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Educational Policy Analysis Archives*, 18(14), 1-36.
- Anagnostopoulos, D. (2003). The New Accountability, Student Failure, and Teachers' Work in Urban High Schools. *Educational Policy*, 17(3), 291-316. doi: 10.1177/0895904803017003001
- Anderman, E. M., Griesinger, T., & Westerfield, G. (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology*, 90(1), 84-93. doi: 10.1037/0022-0663.90.1.84
- Aronson, J., & Good, C. (2003). The development and consequences of stereotype vulnerability in adolescents. In F. Pajares & T. Urdan (Eds.), *Adolescence and education. Academic motivation of adolescents* (Vol. 2, pp. 299-330). Greenwich, CT: Information Age Publishing.

- Atkinson, R. C., & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher*, 38(9), 665-676. doi: 10.3102/0013189X09351981
- Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*(5), 258-267. doi: 10.3102/0013189X07306523
- Australian Bureau of Statistics. (2015). 4221.0 - Schools, Australia, 2014. Geconsulteerd op 12 augustus 2015, van <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4221.02014?OpenDocument>
- Australian Capital Territory Government. (2014). ACT students top the nation in NAPLAN. Canberra: ACT Government.
- Australian Council for Educational Research. (2015). Australian Council for Educational Research: Creating and promoting research-based knowledge, products and services to improve learning. Geconsulteerd op 12 augustus 2015, van <https://www.acer.edu.au/>
- Australian Government: Department of Education Employment and Workplace Relations. (2010). Country Background Report for Australia. In OECD (Ed.), *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes*. Paris: OECD.
- A-Vision. (2015). Welkom bij route8! Geconsulteerd op 09 november 2015, van <http://route8.nl/#home>
- Babo, G., Tienken, C. H., & Gencarelli, M. A. (2014). Interim Testing, Socio-Economic Status, and the Odds of Passing Grade 8 State Tests in New Jersey. *RMLE Online: Research in Middle Level Education*, 38(3), 1-9.
- Baird, J., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). Research evidence relating to proposals for reform of the GCSE (pp. 32). Oxford, UK: Oxford University Centre for Educational Assessment.
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.
- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of Frequent Classroom Testing. *The Journal of Educational Research*, 85(2), 89-99.
- Banta, T. W., Griffin, M., Flateby, T. L., & Kahn, S. (2009). Three promising alternatives for assessing college students' knowledge and skills. Champaign, IL: National Institute for Learning Outcomes Assessment (University of Illinois/Indiana University).
- Barber, M. (1998). The dark side of the moon: imagining an end to failure in urban education. In L. Stoll & K. Myers (Eds.), *No quick fixes. Perspectives on schools in difficulty* (pp. 17-33). Bristol, UK: Falmer Press.
- Barton, L., Willson, P., Langford, R., & Schreiner, B. (2014). Standardized Predictive Testing: Practices, Policies, and Outcomes. *Administrative Issues Journal: Education, Practice, and Research*, 4(2), 68-76.
- Barton, P. E. (1999). Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education. A Policy Information Perspective. Princeton N. J.: Educational Testing Service, Policy Information Center.
- Beaver, J. K., & Weinbaum, E. H. (2015). State Test Data and School Improvement Efforts. *Educational Policy*, 29(3), 478-503. doi: 10.1177/0895904813510774

- Belfield, C. R., & Levin, H. M. (2002). The Effects of Competition Between Schools on Educational Outcomes: A Review for the United States. *Review of Educational Research*, 72(2), 279-341. doi: 10.3102/00346543072002279
- Berge, J. (2015, 01 December). Her er de beste fylkene i nasjonale prøver, *Nettavisen*.
- Bishop, J. (1995). The impact of curriculum-based external examinations on school priorities and student learning. *International Journal of Educational Research*, 23, 653-752.
- Bishop, J. (1998a). The effect of curriculum-based external exit exams on student achievement. *Journal of Economic Education*, 29, 172-182.
- Bishop, J. (1998b). Do curriculum-based external exit exam systems enhance student achievement? University of Pennsylvania: Consortium for Policy Research in Education.
- Bishop, J., & Wössmann, L. (2004). Institutional effects in a simple model of educational production. *Education Economics*, 12, 17-38.
- Black, P., & Wiliam, D. (2005). Lessons from around the World: How Policies, Politics and Cultures Constrain and Afford Assessment Practices. *Curriculum Journal*, 16(2), 249-261.
- Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? . *Assessment in Education: Principles, Policy & Practice*, 18, 451-469.
- Block, T., Goeminne, G., Paredis, E., & Crivits, M. (2010). Toekomsten voorstellen. Over exploratieve scenario's voor Vlaamse steden en gemeenten. *Burger, Bestuur & Beleid*, 6(3), 241-252.
- Bolgatz, J. (2006). Using primary documents with fourth-grade students. In S. G. Grant (Ed.), *Measuring history* (pp. 133-156). Greenwich, CT: Information Age Publishing.
- Booher-Jennings, J. (2005). Below the bubble: 'Educational Triage' and the Texas accountability system. *American Educational Research Journal*, 42(2), 231-268.
- Börjeson, L., Höjer, M., Dreborg, K.-H., Ekvall, T., & Finnveden, G. (2006). Scenario types and scenario techniques: Towards a user's guide to scenarios. *Futures*, 38, 723-739.
- Bosker, R., Creemers, B. P. M., & Stringfield, S. (1999). *Enhancing educational excellence, equity and efficiency. Evidence from evaluations of systems and schools in change*. Dordrecht, Nederland: Kluwer.
- BOSTES. (2015a). How HSC results are determined. Geconsulteerd op 21 augustus 2015, van <http://www.boardofstudies.nsw.edu.au/hsc-results/determining-achievement.html>
- BOSTES. (2015b). HSC. Geconsulteerd op 21 augustus 2015, van <http://www.boardofstudies.nsw.edu.au/hsc/>
- Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, 17(3), 141-159.
- Boyd, B. T. (2008). Effects of State Tests on Classroom Test Items in Mathematics. *School Science and Mathematics*, 108(6), 251-262. doi: 10.1111/j.1949-8594.2008.tb17835.x
- Brennan, B. (2015). NAPLAN results: Annual numeracy and literacy report shows 'limited' significant improvement in students' skills. Geconsulteerd op 20 augustus 2015, van <http://www.abc.net.au/news/2015-08-05/naplan-results-show-limited-improvement-in-students-skills/6673118>
- Brennan, R., Kim, J., Wenz-Gross, M., & Siperstein, G. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review*, 71(2), 173-217.

- Brossman, B. G., & Guille, R. A. (2014). A Comparison of multi-stage and linear test designs for medium-size licensure and certification examinations. *Journal of Computerized Adaptive Testing*, 2(2), 18-36.
- Brozo, W. G., & Hargis, C. (2003). Using low-stakes reading assessment. *Educational Leadership*, 61(3), 60-64.
- Brunsdon, V., Davies, M., & Shevlin, M. (2006). Anxiety and stress in educational professionals in relation to Ofsted. *Education Today*, 56, 24-31.
- Buckendahl, C. W., & Hunt, R. (2005). Whose rules? The relation between the "rules" and "law" of testing. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 147-158). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Buisson-Fenet, H., & Pons, X. (2011). Les pratiques d'évaluation externe des établissements scolaires en France, au Royaume-Uni et en Suisse. Créteil, France: Université Paris-Est-Créteil.
- Bureau ICE. (s.d.). IEP eindtoets. Geconsulteerd op 09 november 2015, van <http://www.toets.nl/basisonderwijs/iep-eindtoets>
- Burgess, S., Wilson, D., & Worth, J. (2013). A natural experiment in school accountability: The impact of school performance information on pupil progress. *Journal of Public Economics*, 106(0), 57-67. doi: <http://dx.doi.org/10.1016/j.jpubeco.2013.06.005>
- Burton, K. D., Lydon, J. E., D'Alessandro, D. U., & Koestner, R. (2006). The differential effects of intrinsic and identified motivation on well-being and performance: Prospective, experimental, and implicit approaches to self-determination theory. *Journal of Personality and Social Psychology*, 91, 750-762. doi: 10.1037/0022-3514.91.4.750
- Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Case, P., Case, S., & Catling, S. (2000). Please show you're working; a critical assessment of the impact of Ofsted inspection on primary teachers. *British Journal of Sociology of Education*, 21(4), 605-621.
- Castellano, K. E., & Ho, A. D. (2013). A Practitioner's Guide to Growth Models. Geconsulteerd op from Council of Chief State School Officers website: <http://www.ccsso.org/Documents/2013GrowthModels.pdf>
- Chakrabarti, R., & Schwartz, N. (2013). Unintended consequences of school accountability policies. Evidence from Florida and implications for New York. *Economic Policy Review*, 19(1), 19-43.
- Chermack, T. J. (2005). Studying scenario planning: Theory, research suggestions, and hypotheses. *Technological Forecasting and Social Change*, 72(1), 59-73. doi: <http://dx.doi.org/10.1016/j.techfore.2003.11.003>
- Choi, K., & Seltzer, M. (2005). *Modelling heterogeneity in relationships between initial status and rates of change: latent variable regression in a three-level hierarchical model*. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.
- Chong, H., Adnan, H., & Zin, R. M. (2012). A feasible means of methodological advance from Delphi methods: a case study. *International Journal of Academic Research*, 4(2), 247-253.
- CITO. (2013a). *Hoe wordt een examen eigenlijk gemaakt?* CITO. Arnhem.
- CITO. (2013b). Toetsscore, vaardigheidsscore... en dan? Arnhem: CITO.
- CITO. (z.j.-a). LVS-toetsen primair en speciaal onderwijs Geconsulteerd op 12 maart 2016, van http://www.CITO.nl/onderwijs/primair%20onderwijs/lvs_toetsen
- CITO. (z.j.-b). Resultaten per examen 2015. Geconsulteerd op 14 maart 2016, van http://www.CITO.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/examenverslagen

- Cizek, G. J. (2005). High-stakes testing: contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 23-54). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Clarke, M., Haney, W., & Madaus, G. F. (2000). High Stakes Testing and High School Completion. *NBETPP Statements*, 1(3), 2-15.
- Cobb, F., & Russell, N. M. (2015). Meritocracy or complexity: problematizing racial disparities in mathematics assessment within the context of curricular structures, practices, and discourse. *Journal of Education Policy*, 30(5), 631-649.
- Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, 112(4), 469-495.
- College voor Toetsen en Examens. (2015a). De centrale eindtoets in 2016. Utrecht: College voor Toetsen en Examens.
- College voor Toetsen en Examens. (2015b). Interpretatie van het leerlingrapport bij de centrale eindtoets PO 2015. Utrecht: College voor Toetsen en Examens.
- College voor Toetsen en Examens. (s.d.-a). Centrale Eindtoets PO. Geconsulteerd op 16 maart 2015, van <https://www.centraleeindtoetspo.nl/>
- College voor Toetsen en Examens. (s.d.-b). Missie en taken. Geconsulteerd op 16 maart 2015, van https://www.hetcvte.nl/item/missie_en_taken
- Collins, S., Reiss, M., & Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales. *Assessment in Education: Principles, Policy & Practice*, 17(3), 273-286.
- Colmar Brunton. (2014). ACARA. Perspectives on the My School Website (pp. 68). Melbourne: Colmar Brunton.
- Colpin, M., Gysen, S., Jaspaert, K., Heymans, R., Van den Branden, K., & Verhelst, M. (2006). Studie naar wenselijkheid en haalbaarheid van de invoering van centrale taaltoetsen in Vlaanderen in functie van gelijke onderwijskansen. *Leuven: Centrum voor Taal en Onderwijs KU Leuven*.
- Cook, H. (2015, 9 augustus). Andrews government considers plan to tie NAPLAN results to school funding, *The Age*
- Corpus, J. H., McClintic-Gilbert, M. S., & Hayenga, A. (2009). Within-year changes in children's intrinsic and extrinsic motivational orientations: Contextual predictors and academic outcomes. *Contemporary Educational Psychology*, 34(2), 154-166. doi: 10.1016/j.cedpsych.2009.01.001
- Cosner, S. (2011). Supporting the initiation and early development of evidence-based grade-level collaboration in urban elementary schools key roles and strategies of principals and literacy coordinators. *Urban Education*, 46(4), 786-827.
- Costea, B., Crump, N., & Amiridis, K. (2007). Managerialism and Infinite Human Resourcefulness: A Commentary on the therapeutic Habitus, Derecognition of Finitude and the Modern Sense of Self. *Journal for Cultural Research*, 11(3), 245-264.
- Costigan, A. T. (2002). Teaching the Culture of High Stakes Testing: Listening to New Teachers. *Action in Teacher Education*, 23(4), 28-34. doi: 10.1080/01626620.2002.10463085
- Coughlan, S. (2016). Teachers' union calls for ballot on primary test boycott. Geconsulteerd op 30 maart 2016, van http://www.bbc.com/news/education-35904242?utm_source=twitterfeed&utm_medium=twitter
- Cunningham, G. K. (2005). Must high stakes mean low quality? Some testing program implementation issues. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 123-146). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.

- Custer, R. L., Scarcella, J. A., & Stewart, B. R. (1999). The Modified Delphi Technique--A Rotational Modification. *Journal of Vocational and Technical Education*, 15(2), 50-58.
- Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3), 458-467.
- Dammers, E. (2000). *Leren van de toekomst. Over de rol van scenario's bij strategische beleidsvorming*. Delft, Nederland: Eburon.
- Day, J., & Bobeva, M. (2005). A Generic Toolkit for the Successful Management of Delphi Studies. *Electronic Journal of Business Research Methods*, 3(2), 103-116.
- De Lange, M., & Dronkers, J. (2007). *Hoe gelijkwaardig blijft het eindexamen tussen scholen? Discrepanties tussen de cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005*. Nijmegen, The Netherlands/San Domenico di Fiesole, Italy: Radboud Universiteit Nijmegen/European University Institute.
- De Maeyer, S., & Rymenans, R. (2004). *Onderzoek naar kenmerken van effectieve scholen. Kritische factoren in een onderzoek naar schooleffectiviteit in het technisch en beroepssecundair onderwijs in Vlaanderen*. Gent: Academia Press.
- De Volder, I., & De Lee, L. (2009). *Bevraging van het welbevinden bij leerlingen in het basisonderwijs*. (Master), Universiteit Antwerpen, Antwerpen. Geconsulteerd op from <http://blogimages.bloggen.be/welbevindenleerlingen/attach/25023.pdf>
- de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education*, 33, 379-396.
- Deci, E., & Ryan, R. (2000). The "what" en "why" of goal pursuits: human needs and the self-determination of behavior. *Psychological Inquiry*, 11(227-268).
- Dembosky, J., Pane, J., Barney, H., & Christina, R. (2005). Data decision making in southwestern Pennsylvania school districts (pp. 59). Pittsburgh, PA: RAND Education.
- Denscombe, M. (2000). Social conditions for stress: Young people's experience of doing GCSEs. *British Educational Research Journal*, 26, 359-374.
- Dill, D. D., & Soo, M. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. *Higher education*, 49(4), 495-533.
- Dodd, T. (2015, 5 augustus). NAPLAN shows that something's wrong with high schools, *Financial Review*, p. 7.
- Doolaard, S. (2013). *Het streven naar kwaliteit in scholen voor primair onderwijs*. Groningen: Gronings Instituut voor onderzoek van het onderwijs.
- Driesler, S. D. (2001). Whiplash about backlash: The truth about public support for testing. *NCME Newsletter*, 9(3), 2-5.
- Driessen, G. (2002). School composition and achievement in primary education: a large-scale multi-level approach. *Studies in Educational Evaluation*, 28, 347-368.
- Durant, D. (2003). A comparative analysis of Key Stage tests and teacher assessments
- Eliot, L. (2010). The Myth of Pink and Blue Brains. *Educational Leadership*, 68(3), 32-36.
- Elmore, R. F. (2002). Bridging the gap between standards and achievement. Washington, DC: Albert Shanker Institute.
- Elstad, E. (2006). Understanding the nature of accountability failure in a technology-filled, laissez-faire classroom: disaffected students and teachers who give in. *Journal of Curriculum Studies*, 38, 459-481.

- Elstad, E. (2009). Schools which are named, shamed and blamed by the media: school accountability in Norway. *Educational Assessment Evaluation and Accountability*, 21(173-189).
- Elstad, E., & Turmo, A. (2011). Obeying the rules or gaming the system? Delegating random selection for examinations to head teachers within an accountability system. *Education, Knowledge and Economy*, 1-15.
- Elstad, E., Christophersen, K., & Turmo, A. (2012). The strength of accountability and teachers' organisational citizenship behaviour. *Journal of Educational Administration*, 50(5), 612-628.
- Elstad, E., Lejonberg, E., & Christophersen, K. (2015). Teacher evaluation as a contested practice: teacher resistance to teaching evaluation schemes in Norway. *Education Inquiry*.
- Elstad, E., Nordtvedt, G. A., & Turmo, A. (2009). The Norwegian assessment system: an accountability perspective. *CADMO*, 2, 89-103.
- EP-nuffic. (2014). Het onderwijssysteem van Nederland beschreven. Den Haag: EP-nuffic.
- Eurydice. (2007). School Autonomy in Europe: Policies and Measures. Brussels: Eurydice, EU.
- Eurydice. (2009). National testing of pupils in Europe: objectives, organisation and use of results. United Kingdom (England, Wales and Northern Ireland) (pp. 21): European Commission.
- Fan, X. (2001). Parental Involvement and Students' Academic Achievement: A Growth Modeling Analysis. *The Journal of Experimental Education*, 70(1), 27-61. doi: 10.1080/00220970109599497
- Fasting, R. B., Thygesen, R., Berge, K. L., Evensen, L. S., & Vagle, W. (2009). National assessment of writing proficiency among Norwegian students in compulsory schools. *Scandinavian Journal of Educational Research*, 53, 617-637.
- Faubert, V. (2009). School evaluation: current practices in OECD countries *OECD Working Papers*. Paris: OECD.
- Fielding, M. (1999). Target setting, policy pathology and student perspectives: learning to labour in new times. *Cambridge Journal of Education*, 29, 277-287.
- Figlio, D., & Ladd, H. F. (2008). School accountability and student achievement. In H. F. Ladd & E. Fiske (Eds.), *Handbook of Research in Education Finance and Policy*. London: Routledge.
- Firestone, W. A., Monfils, L., Camilli, G., Schorr, R. Y., Hicks, J., & Mayrowetz, D. (2002). The ambiguity of test preparation. A multimethod analysis in one state. *Teachers College Record*, 104, 1485-1523.
- Fleege, P. O. (1992). Stress Begins in Kindergarten: A Look at Behavior during Standardized Testing. *Journal of Research in Childhood Education*, 7(1), 20-26.
- Florestal, K., & Cooper, R. (1997). *Decentralization of Education*. Washington, DC: The World Bank.
- Fraillon, J., Schulz, W., Nixon, J., & Gebhardt, E. (2014). National Assessment Program - Civics and Citizenship Years 6 & 10 Report 2013 (pp. 198). Sydney: ACARA.
- Frey, A., & Seitz, N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, 35(2-3), 89-94. doi: <http://dx.doi.org/10.1016/j.stueduc.2009.10.007>
- Fryer, R. G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31, 373-407. doi: 10.1086/667757
- Fuchs, T., & Wößmann, L. (2008). *What accounts for international differences in student performance? A re-examination using PISA data* (pp. 209-240). Physica-Verlag HD.

- Gallagher, L., Means, B., & Padilla, C. (2008). Teachers' use of student data systems to improve instruction: 2005 to 2007. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.
- Gandal, M., & McGiffert, L. (2003). The power of testing. *Educational Leadership*, 60(5), 39-42.
- Gershenson, S., & Langbein, L. (2015). The Effect of Primary School Size on Academic Achievement. *Educational Evaluation and Policy Analysis*, 37(1 suppl), 135S-155S. doi: 10.3102/0162373715576075
- Gibbs, G. R. (2002). *Qualitative data analysis. Explorations with NVivo*. Berkshire, UK: Open University Press.
- Gijssel, M. A. R., & Bosman, A. M. T. (2010). Effects of a psycholinguistic intervention program for children with dyslexia. *Pedagogische Studien*, 87(2), 118-133.
- Gillborn, D. (1996). *Exclusions from school*. University of London, Institute of Education.
- Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 105-118). London: Springer.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. London: Sage Publications.
- Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). Growth model comparison study (pp. 20). Washington, DC: Council of Chief State School Officers.
- Goodman, D., & Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 91-110). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Goos, M., Van Damme, J., Onghena, P., Petry, K., & de Bilde, J. (2013). First-grade retention in the Flemish educational context: Effects on children's academic growth, psychosocial growth, and school career throughout primary education. *Journal of School Psychology*, 51(3), 323-347. doi: <http://dx.doi.org/10.1016/j.jsp.2013.03.002>
- Gradwell, J. M. (2006). Teaching in spite of, rather than because of, the test: a case of ambitious history teaching in New York state. In S. G. Grant (Ed.), *Measuring history* (pp. 157-176). Greenwich, CT: Information Age Publishing.
- Granheim, M. K., & Lundgren, U. P. (1991). Steering by Goals and Evaluation in the Norwegian Education System: A Report from the EMIL Project. *Journal of Curriculum Studies*, 23(6-), 481-505.
- Grant, S. G. (2003). *History lessons: teaching, learning, and testing in U.S. high school classrooms*. Mahwah, NJ: Lawrence Erlbaum.
- Grayson, L., & Gomersall, A. (2003). A difficult business: finding the evidence for social science reviews. London: ESRC UK Centre for Evidence Based Policy and Practice, University of London.
- Greene, J. P. (2001). Vouchers in Charlotte. *Education Next*, 1(2).
- Greene, J. P., Winters, M. A., & Forster, G. (2003). Testing High Stakes Tests: Can We Believe the Results of Accountability Tests? Civic Report.
- Grisham-Brown, J., Bagnato, S., & Pfeiffer-Fiala, C. "Applying Authentic Assessment Practices: How to be Critical Consumers across Assessment Purposes." Pre-Conference Session for International Division of Early Childhood Conference, Albuquerque, NM, October, 2009.
- Guldmond, H., & Bosker, R. J. (2009). School effects on students' progress - a dynamic perspective. *School Effectiveness and School Improvement*, 20(2), 255-268.

- Hadfield, M., & Chapman, C. (2009). *Leading School-based Networks*. Abingdon: Routledge.
- Haertel, E. H. (1999). Validity Arguments for High-stakes Testing: In Search of the Evidence. *Educational Measurement: Issues and Practice*, 18(4), 5-9. doi: 10.1111/j.1745-3992.1999.tb00276.x
- Haile, G. A., & Nguyen, A. N. (2008). Determinants of academic attainment in the United States: a quantile regression analysis of test scores. *Education Economics*, 16(1), 29-57.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.
- Hamilton, L. S., Berends, M., & Stecher, B. M. (2005). Teachers' responses to standards-based accountability. Santa Monica, CA: Rand.
- Haney, W. (2000). The myth of the Texas miracle in education. *Educational Policy Analysis Archives*, 8(41).
- Hanushek, E. A., & Kimko, D. D. (2000). Schooling, Labor-Force Quality, and the Growth of Nations. *The American Economic Review*, 90, 1184-1208.
- Hanushek, E. A., & Wössmann, L. (2010). The cost of low educational achievement in the European Union. Brussel: European Expert Network on Economics of Education.
- Harris, D. N. (2011). Value-Added Measures and the Future of Educational Accountability. *Science*, 333, 826-827.
- Harris, D. N. (2012). How Do Value-Added Indicators Compare to Other Measures of Teacher Effectiveness? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 5. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.
- Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment*, 7.
- Heyma, A., Bisschop, P., Van den Berg, E., Wartenbergh-Cras, F., Kurver, B., Muskens, M., & Spanjers, I. (2015). Effectmeting Innovatielimpuls Onderwijs. Eindrapport. Amsterdam: Seo Economisch Onderzoek / ResearchNed.
- Hirtt, N., Nicaise, I., & De Zutter, D. (2007). *De school van de ongelijkheid*. Antwerpen: EPO.
- Hopfenbeck, T. N., Flórez Petour, M. T., & Tolo, A. (2015). Balancing Tensions in Educational Policy Reforms: Large-Scale Implementation of Assessment for Learning in Norway. *Assessment in Education: Principles, Policy & Practice*, 22(1), 44-60.
- Hopkins, L. D., & Zapata, M. D. (2009). *Engaging the Future: Forecasts, Scenarios, Plans, and Projects*. Cambridge, UK: Lincoln Institute for Land Policy.
- Horn, C. (2005). Standardized Assessments and the Flow of Students Into the College Admission Pool. *Educational Policy*, 19(2), 331-348.
- Hornstra, L., Denessen, E., Bakker, J., van den Bergh, L., & Voeten, M. (2010). Teacher Attitudes Toward Dyslexia: Effects on Teacher Expectations and the Academic Achievement of Students With Dyslexia. *Journal of Learning Disabilities*, 43(6), 515-529. doi: 10.1177/0022219409355479
- Hoxby, C. M. (2003a). "School choice and school competition: Evidence from the United States," *Swedish Economic Policy Review*, 10, 11-67.
- Hoxby, C. M. (2012). The cost of accountability (pp. 27). Cambridge, MA: National Bureau of Economic Research.
- Hoxby, C.M. (2001). "How School Choice Affects the Achievement of Public School Students," Paper prepared for Koret Task Force meeting, Hoover Institution, Stanford CA.

- Hsu, C., & Sandford, B. A. (2007). The Delphi Technique: Making Sense Of Consensus. *Practical Assessment, Research & Evaluation*, 12(10).
- Inspectie van het Onderwijs. (2014a). Analyse en waarderings van opbrengsten. Utrecht: Ministerie van Onderwijs, Cultuur en Wetenschap.
- Inspectie van het Onderwijs. (2014b). Meerjaren opbrengsten VO 2014. Toelichting. Utrecht: Ministerie van Onderwijs, Cultuur en Wetenschap.
- Inspectie van het Onderwijs. (2015). De staat van het onderwijs. Onderwijsverslag 2013/14. Utrecht: Ministerie van Onderwijs, Cultuur en Wetenschap.
- Isenberg, E., & Hock, H. (2011). Design of value added models for IMPACT and TEAM in DC public school, 2010-2011 schoolyear. Washington DC: Mathematica Policy Research.
- Jacob, B. A. (2001). Getting Tough? The Impact of High School Graduation Exams. *Educational Evaluation and Policy Analysis*, 23(2), 99-121. doi: 10.3102/01623737023002099
- Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating* (No. w9413). National Bureau of Economic Research.
- Jäger, D. J., Maag Merki, K., Oerke, B., & Holmeier, M. (2012). Statewide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German states. *Assessment in Education: Principles, Policy & Practice*, 19(4), 451-467. doi: 10.1080/0969594X.2012.677803
- Janssen, R. (2009). Het Vlaamse onderwijs gepeild: achterliggende visies en analysemodellen. *ORD 2009*.
- Janssens, F. J. G., Rekers-Mombarg, L., & Lacor, E. (2014). Leerwinst en toegevoegde waarde in het primair onderwijs. Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschap; Inspectie van het Onderwijs; Rijksuniversiteit Groningen; CED Groep; Universiteit Twente; CITO.
- Jaradat, D. (1986). Assessment data obtained from the Jordanian GSECE: a critical look. *Studies in Educational Evaluation*, 12, 335-340.
- Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The Impact of High-Stakes Testing on Teachers and Students in North Carolina. *The Phi Delta Kappan*, 81(3), 199-203.
- Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics*, 23, 497-517. doi: 10.1007/s00148-008-0234-3
- Jürges, H., Büchel, F., & Schneider, K. (2005). The effect of central exit examinations on student achievement: quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association*, 3(5), 1134-1155.
- Jürges, H., Richter, W. F., & Schneider, K. (2005). Teacher Quality and Incentives: Theoretical and Empirical Effects of Standards on Teacher Quality. *FinanzArchiv / Public Finance Analysis*, 61, 298-326. doi: 10.2307/40913080
- Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2009). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1), 56-65. doi: <http://dx.doi.org/10.1016/j.econedurev.2011.08.007>
- Kall, L. B., Malmgren, H., Olsson, E., Linden, T., & Nilsson, M. (2015). Effects of a Curricular Physical Activity Intervention on Children's School Performance, Wellness, and Brain Development. *Journal of School Health*, 85(10), 704-713. doi: 10.1111/josh.12303
- Karsten, S., Visscher, A., & De Jong, T. (2001). Another side to the coin: the unintended effects of the publication of school performance data in England and France. *Comparative Education*, 37(2), 231-242.

- Katholiek Onderwijs Vlaanderen. (z.j.). Interdiocesane proeven. Geconsulteerd op 2 oktober 2015, van <http://info.dp.vvkbao.be/content/home>
- Keeves, J. P., Hungi, N., & Afrassa, T. (2005). Measuring Value Added Effects across Schools: Should Schools Be Compared in Performance? *Studies in Educational Evaluation*, 31(2-3), 247-266.
- Kelly, A. (2001). *Benchmarking for school improvement. A practical guide for comparing and improving effectiveness*. London, New York: Routledge Falmer.
- Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, 112, 496-520.
- Kesler, T. (2013). Unstandardized Measures: A Cross-Case Analysis of Test Prep in Two Urban High-Needs Fourth-Grade Classes. *The Elementary School Journal*, 113(4), 488-516. doi: 10.1086/669617
- Klein, E. D. (2013). *Statewide Exit Exams, Governance, and School Development*. Münster, Germany: Waxmann.
- Klein, E. D., & Van Ackeren, I. (2012). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation*, 37, 180-188.
- Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Educational Policy Analysis Archives*, 8(49), 1-22.
- Klenowski, V. (2014). Towards fairer assessment. *Australian Education Research*, 41, 445-470.
- Kobrin, J. L., Kim, Y. K., & Sackett, P. R. (2012). Modeling the Predictive Validity of SAT Mathematics Items Using Item Characteristics. *Educational and Psychological Measurement*, 72(1), 99-119.
- Koretz, D. M. (2008). A measured approach. Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*, 2008(fall), 18-39.
- Koretz, D. M., & Barron, S. I. (1998). The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: Rand Corp.scriven
- Koschoreck, J. W. (2001). Accountability and educational equity in the transformation of an urban district. *Education and Urban Society*, 33(3), 284-304.
- Kosow, H., & Gassner, R. (2008). Methods of future and scenario analysis: overview, assessment, and selection criteria. Bonn, Duitsland: Deutsches Institut für Entwicklungspolitik (DIE).
- Krumsvik, R. (2006). The digital challenges of school and teacher education in Norway: Some urgent questions and the search for answers. *Education and Information Technologies*, 11, 239-256.
- Kühn, S. M. (2011). Exploring the use of statewide exit exams to spread innovation—The example of Context in science tasks from an international comparative perspective. *Studies in Educational Evaluation*, 37(4), 189-195. doi: <http://dx.doi.org/10.1016/j.stueduc.2012.01.003>
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, 35, 64-70.
- Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, 29(3), 426-450. doi: 10.1002/pam.20504
- Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review*, 21, 1-17.

- Laevers, F., Bertrands, E., Declercq, B., & Daems, M. (2004). Instrumenten voor de screening van 6-12 jarigen en de observatie van de leeromgeving. Leuven: Centrum voor ErvaringsGericht Onderwijs.
- Lamprianou, I., & Christie, T. (2009). Why school based assessment is not a universal feature of high stakes assessment systems? *Educational Assessment, Evaluation and Accountability*, 21(4), 329-345. doi: 10.1007/s11092-009-9083-1
- Lange, C., Range, B., & Welsh, K. (2012). Conditions for Effective Data Use to Improve Schools: Recommendations for School Leaders. *International Journal of Educational Leadership Preparation*, 7(3), n3.
- Lee, J. (2008). Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies. *Review of Educational Research*, 78(3), 608-644. doi: 10.3102/0034654308324427
- Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal*, 37(1), 3-31. doi: 10.2307/1163470
- Leithwood, K. A., & Riehl, C. (2003). *What we know about successful school leadership*. Nottingham: National College for School Leadership.
- Levačić, R. (2004). Competition and the performance of english secondary schools: further evidence. *Education Economics*, 12(2), 177-193. doi: 10.1080/0964529042000239186
- Lin, C. Y., Hsieh, Y. H., & Chen, C. H. (2015). Use of latent growth curve modeling for assessing the effects of summer and after-school learning on adolescent students' achievement gap. *Asia Pacific Education Review*, 16(1), 49-61. doi: 10.1007/s12564-015-9356-y
- Linn, R. L. (2001). A Century of Standardized Testing: Controversies and Pendulum Swings. *Educational Assessment*, 7(1), 29-38. doi: 10.1207/S15326977EA0701_4
- Linstone, H. A., & Turoff, M. (1975). *The Delphi Method. Techniques and applications*. London, Amsterdam, Ontario, Sydney, Tokyo: Addison-Wesley Publishing Company.
- Loeb, S. (2013). How can value-added measures be used for teacher improvement? In Carnegie Knowledge Network (Ed.), *What we know series: value-added methods and applications*. Stanford, CA: Carnegie Knowledge Network.
- Longo, C. (2010). Fostering Creativity or Teaching to the Test? Implications of State Testing on the Delivery of Science Instruction. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 83(2), 54-57. doi: 10.1080/00098650903505399
- Looney, A. (2014). Assessment and the reform of education systems. From good news to policy technology. In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 233-248). London: Springer.
- Luna, C., & Turner, C. L. (2001). The Impact of the MCAS: Teachers Talk about High-Stakes Testing. *English Journal*, 91(1), 79-87.
- Luyten, H., & de Wolf, I. F. (2011). Changes in student populations and average test scores of Dutch primary schools. *School Effectiveness and School Improvement*, 22(4), 439-460. doi: 10.1080/09243453.2011.591614
- Lysne, A. (1984). Grading of students' attainment: Purposes and functions. *Scandinavian Journal of Educational Research*, 28, 149-165.
- Maag Merki, K. (2011). Effects of the implementation of state-wide exit exams on students' self-regulated learning. *Studies in Educational Evaluation*, 37(4), 196-205. doi: <http://dx.doi.org/10.1016/j.stueduc.2011.12.001>

- Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (pp. 83-121). Chicago, IL: University of Chicago Press.
- Magyar, A. (2015). *Comparing the measurement effectiveness of computer-based linear and adaptive tests*. (PhD), University of Szeged, Szeged.
- Manger, T., Vold, V., & Eikeland, O. (2009). A web-based national test of English reading as a foreign language: Does it test language ability, or computer competence? - A web-based measure of English reading and ICT competence. *Nordic Journal of Digital Literacy*, 4(03-04), 143-158.
- Marlow, R., Norwich, B., Ukoumunne, O. C., Hansford, L., Sharkey, S., & Ford, T. (2014). A comparison of teacher assessment (APP) with standardised tests in primary literacy and numeracy. *Assessment in Education: Principles, Policy & Practice*, 21(4), 412-426.
- Marchant, J. (2004). What is at Stake with High Stakes Testing? A Discussion of Issues and Research. *Ohio Journal of Science*, 104(2), 2-7.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280-295. doi: 10.1037/0022-0663.79.3.280
- Marsh, J. A., Pane, J. F., & Hamilton, L. A. (2005). Making sense of data-driven decision making in education. Santa Monica, CA: Rand Education.
- Massell, D. (2001). The theory and practice of using data to build capacity: State and local strategies and their effects. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (pp. 148-169). Chicago: University of Chicago Press.
- Masters, G. N., Rowley, G., Ainley, J., & Khoo, S. T. (2008). Reporting and comparing school performances. Melbourne: ACER.
- Mathis, W. (2012). Research-Based Options for Education Policymaking: Common Core State Standards. Boulder, CO: National Education Policy Center.
- McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, 95, 203-213.
- McNeil, L. M. (2000). *Contradictions of school reform: educational costs of standardized testing*. New York: Routledge.
- Meisels, S. J. (2000). On the side of the child. *Young Children*, 55(6), 16-19.
- Mietzner, D., & Reger, G. (2005). Advantages and disadvantages of scenario approaches for strategic foresight. *International Journal of Technology Intelligence and Planning*, 1, 220-239.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook (2nd ed.)*. Thousand Oaks, CA: Sage.
- Ministerial Council on Education Employment Training and Youth Affairs. (2008). Melbourne Declaration on Educational Goals for Young Australians. Melbourne: MCEETYA.
- Ministerie van Onderwijs Cultuur en Wetenschap. (s.d.). Referentiekader. Geconsulteerd op 09 november 2015, van <http://www.taalenrekenen.nl/referentiekader/>
- Moelands, A. H. J., Mommers, M. J. C., & Oud, J. H. L. (1990). Leerlingvolgsystemen verklaard en vergeleken. *School en Begeleiding*, 26, 19-28.
- Molnár, G., & Magyar, A. (2015). *Comparing the measurement effectiveness of computerised adaptive testing and fixed item testing*. Paper presented at the 16th EARLI Conference, Limassol, Cyprus.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters. The junior years*. Somerset, UK: Open Books.

- Muñoz, M. A., & Chang, F. C. (2007). The Elusive Relationship Between Teacher Characteristics and Student Academic Growth: A Longitudinal Multilevel Model for Change. *Journal of Personnel Evaluation in Education*, 20(3-4), 147-164. doi: 10.1007/s11092-008-9054-y
- Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants & Young Children*, 17(3), 198-212.
- Neisworth, John T., and Stephen J. Bagnato. "The mismeasure of young children: The authentic assessment alternative." *Infants & Young Children* 17.3 (2004): 198-212.
- Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. *Studies in Educational Evaluation*, 37(4), 206-217. doi: <http://dx.doi.org/10.1016/j.stueduc.2012.02.002>
- Norwegian Directorate for Education and Training. (2008). Our Responsibilities. Geconsulteerd op June 17, 2015, van <http://www.udir.no/Stottemeny/English/>
- Norwegian Directorate for Education and Training. (2011). OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Country Background Report for Norway (pp. 115). Oslo: Norwegian Directorate for Education and Training.
- Norwegian Directorate for Education and Training. (2014). The Education Mirror 2014. Facts and analysis of kindergarten, primary and secondary education in Norway. Oslo: Norwegian Directorate for Education and Training.
- Nusche, D., Earl, L., Maxwell, W., & Shewbridge, C. (2011). OECD Reviews of evaluation and assessment frameworks in education. Norway. Paris: OECD.
- OECD. (2013). Synergies for better learning. An international perspective on evaluation and assessment *OECD Reviews of Evaluation and Assessment in Education*. Paris: OECD.
- Oerke, B., Maag Merki, K., Holmeier, M., & Jäger, D. (2011). Changes in student attributions due to the implementation of central exit exams. *Educational Assessment, Evaluation and Accountability*, 23(3), 223-241. doi: 10.1007/s11092-011-9121-7
- Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1), 15-29. doi: 10.1016/j.im.2003.11.002
- O'Mara, F. (s.d.). *VCAA Assessment Online. On Demand Testing Program*. Victorian Curriculum and Assessment Authority. Melbourne.
- Onderwijskiezer. (2015). Welkom op onderwijskiezer. Geconsulteerd op 5 oktober 2015, van <http://www.onderwijskiezer.be/v2/index.php>
- Opdenakker, M., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education. Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, 11, 165-196.
- Oriëntatieproef. (2014). Over een Vlaamse oriëntatieproef of ijkingsstest. Geconsulteerd op 5 oktober 2015, van <http://www.orientatieproef.be/>
- Othuon, L., & Kishor, N. (1994). Hierarchical linear modelling of predictive validity: The case of Kenya certificate of primary education examination. *Studies in Educational Evaluation*, 20(2), 175-194. doi: [http://dx.doi.org/10.1016/0191-491X\(94\)90002-7](http://dx.doi.org/10.1016/0191-491X(94)90002-7)
- OVSG. (z.j.). Situering OVSG-toets. Geconsulteerd op 5 oktober 2015, van <http://www.ovsgtoetsen.net/ico/node/1>
- Paper presented at the British Educational Research Association Annual Conference, Heriot-Watt University, Edinburgh.

- Parkes, J., & Stevens, J. J. (2003). Legal issues in school accountability systems. *Applied Measurement in Education*, 16, 141-158.
- Paulsen, J. M. (2014). Norwegian Superintendents as Mediators of Change Initiatives. *Leadership and Policy in Schools*, 13(4), 407-423. doi: 10.1080/15700763.2014.945654
- Penninckx, M., Vanhoof, J., & Van Petegem, P. (2011). *Evaluatie in het Vlaamse onderwijs. Beleid en praktijk van leerling tot overheid*. Antwerpen, Belgium: Garant.
- Pentecost, T. C., & Barbera, J. (2013). Measuring Learning Gains in Chemical Education: A Comparison of Two Methods. *Journal of Chemical Education*, 90(7), 839-845. doi: 10.1021/ed400018v
- Perryman, J. (2007). Inspection and emotion. *Cambridge Journal of Education*, 37, 173-190.
- Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the Pressure Cooker - School League Tables and English and Mathematics Teachers' Responses to Accountability in a Results-Driven Era. *British Journal of Educational Studies*, 59(2), 179-195.
- phelps, R. L. (2000). Assessment and accountability. *Educational Researcher*, 29(2), 4-14.
- Phelps, R. P. (2000). Estimating the Cost of Standardized Student Testing in the United States. *Journal of Education Finance*, 25(3), 343-380. doi: 10.2307/40704103
- Phelps, R. P. (2005a). *Defending Standardized Testing*. Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Phelps, R. P. (2005b). The rich, robust research literature on testing's achievement benefits. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 55-90). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). Improving the quality of GCSE Assessment. *A Report Commissioned by the Qualifications and Curriculum Authority, Carrickfergus, UK*.
- Popham, W. J. (1999). Why Standardized Tests Don't Measure Educational Quality. *Educational Leadership*, 56(6), 8-15.
- Popham, W. J. (2002). Right Task, Wrong Tool. *American School Board Journal*, 189(2), 18-22.
- Popham, W. J. (2003). Trouble with Testing: Why Standards-based Assessment Doesn't Measure Up. *American School Board Journal*, 190(2), 14-17.
- Popham, W. J. (2005). Can Growth Ever Be Beside the Point? *Educational Leadership*, 63(3), 83-84.
- Popham, W. J., Cruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, R. L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan*, 66(628-634).
- Prokic-Breuer, T., & Dronkers, J. (2012). The high performance of Dutch and Flemish 15-year-old native pupils: explaining country differences in math scores between highly stratified educational systems. *Educational Research and Evaluation*, 18(8), 749-777. doi: 10.1080/13803611.2012.727359
- Purkey, S., & Smith, M. (1983). Effective schools. A review. *Elementary School Journal*, 83, 64-69.
- Putwain, D. (2008). Do Examinations Stakes Moderate the Test Anxiety-Examination Performance Relationship? *Educational Psychology*, 28(2), 109-118.
- Quist, J. (2007). *Backcasting for a sustainable future. The impact after 10 years*. Delft, Nederland: Eburon Academic Publishers.
- Ramsteck, C., Muslic, B., Graf, T., Maier, U., & Kuper, H. (2015). Data-based school improvement: The role of principals and school supervisory authorities within the context of low-stakes

- mandatory proficiency testing in four German states. *International Journal of Educational Management*, 29(6), 766-789.
- Randall, J., Cheong, Y. F., & Engelhard, G. J. (2011). Using Explanatory Item Response Theory Modeling to Investigate Context Effects of Differential Item Functioning for Students With Disabilities. *Educational and Psychological Measurement*, 71(1), 129-147.
- Ray, A. (2006). School value added measures in England: A paper for the OECD project on the development of value-added models in education systems. Londen: Department for Education and Skills.
- Reay, D. (2008). Tony Blair, the Promotion of the Active Educational Citizen, and Middle-Class Hegemony. *Oxford Review of Education*, 34(6), 639-650.
- Reback, R. (2008). Teaching to the Rating: School Accountability and Distribution of Student Achievement. *Journal of Public Economics*, 92, 1394-1415.
- Redden, G., & Low, R. (2012). My School, Education, and Cultures of Rating and Ranking. *Review of Education, Pedagogy & Cultural Studies*, 34(1-2), 35-48.
- Rijksoverheid. (s.d.-a). Basisonderwijs. Geconsulteerd op 16 februari 2015, van <https://www.rijksoverheid.nl/onderwerpen/basisonderwijs>
- Rijksoverheid. (s.d.-b). Voortgezet onderwijs. Geconsulteerd op 16 februari 2015, van <https://www.rijksoverheid.nl/onderwerpen/voortgezet-onderwijs>
- Rijksoverheid. (s.d.-c). Vrijheid van onderwijs. Geconsulteerd op 16 maart 2015, van <https://www.rijksoverheid.nl/onderwerpen/vrijheid-van-onderwijs>
- Ritchie, J., & Spencer, L. (1993). Qualitative data analysis for applied policy research. In A. Bryman & R. Burgess (Eds.), *Analysing qualitative data* (pp. 173-194). London: Routledge.
- Roeber, E. D. (1995). Using new forms of assessment to assist in achieving student equity: Experiences of the CCSSO State Collaborative on Assessment and Student Standards. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 265-288). Boston: Kluwer.
- Rogers-Ard, R., Knaus, C. B., Epstein, K. K., & Mayfield, K. (2013). Racial Diversity Sounds Nice; Systems Transformation? Not So Much: Developing Urban Teachers of Color. *Urban Education*, 48(3), 451-479.
- Rosenshine, B. (2003). High-stakes testing: Another analysis. *education policy analysis archives*, 11, 24.
- Rozenwajn, E. (2014). *Low-stakes testing and teacher practices: a mirror effect?* Paper presented at the First Bi-annual Conference of the Special Interest Group: Educational Evaluation, Accountability and School Improvement (SIG23) of the European Association for Research on Learning and Instruction (EARLI), Bielefeld, Germany.
- Russell, M., & Plati, T. (2000). Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight, and Ten. A Report of Findings Submitted to the Massachusetts Department of Education. NBETPP Statements World Wide Web Bulletin.
- Ryan, J. (2004). Perverse Incentives of the No Child Left Behind Act. *NYUL Review*, 79, 932-989.
- Sahlberg, P. (2011). *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* New York: Teachers College Press.
- Sammons, P., Hillman, J., & Mortimore, P. (1995). Key Characteristics of Effective Schools (London, Institute of Education). *SammonsKey Characteristics of Effective Schools 1995*.
- Santiago, P., Donaldson, G., Herman, J., & Shewbridge, C. (2011). OECD Reviews of Evaluation and Assessment Frameworks in Education. Australia. Paris: OESO.

- Saunders, L. (1999). A Brief History of Educational 'Value Added': How Did We Get To Where We Are? *School Effectiveness and School Improvement*, 10(2), 233-256.
- Saunders, L. (2000). Understanding schools' use of 'value added' data: the psychology and sociology of numbers. *Research Papers in Education*, 15(3), 241-258.
- Saye, J. (2013). Authentic Pedagogy: Its Presence in Social Studies Classrooms and Relationship to Student Performance on State-Mandated Tests. *Theory & Research in Social Education*, 41(1), 89-132. doi: 10.1080/00933104.2013.756785
- Scheerens, J. (2014). *The educational improvement potential of evaluation and accountability. Questions of effectiveness and efficiency*. Paper presented at the First Bi-annual Conference of the Special Interest Group: Educational Evaluation, Accountability and School Improvement (SIG23) of the European Association for Research on Learning and Instruction (EARLI), Bielefeld, Germany.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and teacher education*, 26(3), 482-496.
- Schildkamp, K., Rekers-Mombarg, L., & Harms, T. J. (2012). Student group differences in examination results and utilization for policy and school development. *School Effectiveness and School Improvement*, 23(2), 229-255.
- Schmidt, R. C. (1997). Managing Delphi Surveys Using Nonparametric Statistical Techniques. *Decision Sciences*, 28(3), 763-774.
- Schoenfeld, A. H. (2002). Making mathematics work for all children: issues of standards, testing, and equity. *Educational Researcher*, 31(1), 13-25.
- Scholastic, & Bill & Melinda Gates Foundation. (2012). Primary sources: 2012. http://www.scholastic.com/primarysources/pdfs/Gates2012_full.pdf
- School Inspections Project. (2016). Norway. Geconsulteerd op 20 Januari 2016, van <http://schoolinspections.eu/impact/norway/>
- Schoolfeedbackproject. (2015). Schoolfeedback. Geconsulteerd op 5 oktober 2015, van <http://www.schoolfeedback.be/schoolfeedback/>
- Schwartz, P. (1991). *The art of the long view*. London: Century Business.
- Scriven, M. (1995). A unified approach to teacher evaluation. *Studies in Educational Evaluation*, 21, 111-129.
- Segall, A. (2003). Teachers' Perceptions of the Impact of State-Mandated Standardized Testing: The Michigan Educational Assessment Program (MEAP) as a Case Study of Consequences. *Theory & Research in Social Education*, 31(3), 287-325. doi: 10.1080/00933104.2003.10473227
- Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened Test Anxiety among Young Children: Elementary School Students' Anxious Responses to High-Stakes Testing. *Psychology in the Schools*, 50, 489-499.
- Seland, I., Vibe, N., & Hovdhaugen, E. (2013). Evaluering av nasjonale prøver som system. Oslo: Nordisk institutt for studier av innovasjon, forskning og utdanning.
- Senate Standing Committee on Education and Employment. (2014). Effectiveness of the National Assessment Program - Literacy and Numeracy. Canberra: Senate Printing Unit, Parliament House.
- Senate Standing Committee on Education Employment and Workplace Relations. (2013). Questions on Notice: Budget estimates 2013-2014.

- Shepard, L. A. (1992). Will standardized test improve student learning? . Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Shewbridge, C., Hulshof, M., Nusche, D., & Stoll, L. (2011). School evaluation in the Flemish community of Belgium. In OECD (Ed.), *OECD Reviews of Evaluation and Assessment in Education*. Paris: OECD Publishing.
- Siguroardottir, A. K. (2010). Professional Learning Community in Relation to School Effectiveness. *Scandinavian Journal of Educational Research*, 54(5), 395-412. doi: 10.1080/00313831.2010.508904
- Sins, P. H. M., & Van der Zee, S. (2015). De toegevoegde waarde van traditioneel vernieuwingsonderwijs: Een studie naar de verschillen in cognitieve en niet-cognitieve opbrengsten tussen daltonscholen en traditionele scholen voor primair onderwijs. *Pedagogische Studiën*, 92, 254-274.
- Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending Standardized Testing* (pp. 111-121). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.
- Skedsmo, G. (2011). Formulation and realisation of evaluation policy: inconsistencies and problematic issues. *Educational Assessment, Evaluation and Accountability*, 23(1), 5-20. doi: 10.1007/s11092-010-9110-2
- Skorupski, W., & Egan, K. (2012). *A hierarchical linear modeling approach for detecting cheating and aberrance*. Paper presented at the Conference on the Statistical Detection of Potential Test Fraud, Lawrence, KS.
- Sloane, F. C., Oloff-Lewis, J., & Kim, S. H. (2013). Value-Added Models of Teacher and School Effectiveness in Ireland: Wise or Otherwise? *Irish Educational Studies*, 32(1), 37-67.
- Smagorinsky, P., Lakly, A., & Johnson, T. S. (2002). Acquiescence, Accommodation, and Resistance in Learning To Teach within a Prescribed Curriculum. *English Education*, 34(3), 187-213.
- Smith, A. (2015, 5 augustus). NAPLAN 2015: Education chiefs warn students are not improving, *The Sydney Morning Herald*, p. 6.
- Smith, M. L. (1991). Put to the test: the effects of external testing on teachers. *Educational Researcher*, 20(5), 8-11.
- Smith, M. S. (1995). Education reform in America public schools: The Clinton Agenda. In D. Ravitch (Ed.), *Debating the future of American education: Do we need national standards and assessment?* (pp. 9-32). Washington, DC: Brookings Institute.
- Smith, W. C. (2014). The Global Transformation Toward Testing for Accountability. 2014, 22. doi: 10.14507/epaa.v22.1571
- Soenens, B., & Vansteenkiste, M. (2005). Antecedents and outcomes of self-determination in life domains: The role of parents' and teachers' autonomy support. *Journal of Youth and Adolescence*, 34(6), 589-604. doi: 10.1007/s10964-005-8948-y
- Springer, M., & Winters, M. (2009). New York City's school-wide bonus pay program: Early evidence from a randomized trial Nashville, TN: Vanderbilt Peabody College, National Center on Performance Incentives.
- Srikantaiah, D., Zhang, Y., Swayhoeover, L., & Center on Education Policy, W. D. C. (2008). Lessons From the Classroom Level: Federal and State Accountability in Rhode Island: Center on Education Policy.
- Standaert, R. (2000). De verleiding van het simplificeren. Een pleidooi voor integratie van interne en externe kwaliteitsbewaking. In J. Letschert (Ed.), *Op de schouders van de voorgangers* (pp. 65-96). Enschede, Nederland: Stichting Leerplanontwikkeling.

- Standaert, R. (2014). *De becijferde school. Meetcultus en meetcultuur*. Leuven: Acco.
- Statistics Norway. (2015). Facts about education in Norway 2015 - key figures 2013. Oslo/Kongsvinger: Statistisk sentralbyrå.
- Statistics Norway. (2016). National tests, 2015. Geconsulteerd op 16 februari 2016, van <http://www.ssb.no/en/utdanning/statistikker/nasjprov/aar/2016-02-04>
- Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., . . . Naftel, S. (2008). Pain and gain: Implementing No Child Left Behind in three states, 2004-2006. Santa Monica, CA: Rand Corp.
- Steunpunt Toetsontwikkeling en Peilingen. (2015). Korte Omschrijving. Geconsulteerd op 2 oktober 2015, van <http://ppw.kuleuven.be/home/onderzoek/peilingsonderzoek/omschrijving>
- Strand, S., & Demie, F. (2007). Pupil Mobility, Attainment and Progress in Secondary School. *Educational Studies*, 33(3), 313-331.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Supovitz, J. A., & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.
- Taylor, C. S., & Lee, Y. (2009). Stability of Rasch Scales Over Time. *Applied Measurement in Education*, 23(1), 87-113. doi: 10.1080/08957340903423701
- Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer.
- Tertiary Institutions Service Centre. (s.d.). The Australian Tertiary Admission Rate (ATAR). East Perth: Tertiary Institutions Service Centre.
- Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, 33(3), 261-295. doi: 10.1080/03054980701366116
- Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology*, 100(4), 988-1001. doi: 10.1037/0022-0663.100.4.988
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98(4), 788-806. doi: 10.1037/0022-0663.98.4.788
- Tully, S. (2008). Interpreting Test Scores: More Complicated than You Think. *Chronicle of Higher Education*, 54(49), 23.
- Tveit, S. (2013). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice*, 21(2), 221-237. doi: 10.1080/0969594X.2013.830079
- Tymms, P. (1997). Responses of headteachers to value-added and the impact of feedback. London: School Curriculum and Assessment Authority.
- Universities Admissions Centre. (2015). Australian Tertiary Admission Rank (ATAR). Geconsulteerd op 25 augustus 2015, van <http://www.uac.edu.au/atar/>
- UNSW Global. (2015). Educational Assessment Australia. from <https://www.eaa.unsw.edu.au/>
- Upton, J., & Supovitz, J. (1996). Measuring Student Impact in the Context of Statewide Education Reform. Paper presented at the Annual Meeting of the American Education Research Association. New York.

- Utdanningsdirektoratet. (z.j.). Nasjonale prøver. Geconsulteerd op 02 juli 2015, van <http://www.udir.no/Vurdering/Nasjonale-prover/>
- Van Ackeren, I., Block, R., Klein, E. D., & Kühn, S. M. (2012). The Impact of Statewide Exit Exams: A Descriptive Case Study of Three German States with Differing Low Stakes Exam Regimes. *Education Policy Analysis Archives*, 20(8), 1-32.
- van de Grift, W. (2009). Reliability in measuring the value added of schools. *School Effectiveness and School Improvement*, 20(2), 269-285.
- Van de Ven, A. L., & Delbecq, A. H. (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of Management Journal*, 17(4), 605-621.
- Van den Bergh, H. (1993). Rapportcijfers in onderwijsonderzoek. Rapportcijfers Duits versus leesprestaties Duits. In P. Van den Hoven, L. Lentz, J. Sturm & F. Vonk (Eds.), *Taal in onderwijs en organisatie. Een vloerschouw van onderzoek* (pp. 1-6). Utrecht, Nederland: Centre for Language and Communication.
- Van der Heijden, K. (1996). *Scenarios: the Art of Strategic Conversation*. Chichester, UK: Wiley.
- Van Eck, E., & Boogaerd, M. (2007). Interne en externe kwaliteitszorg in het basisonderwijs en de rol van het bovenschools management Amsterdam: Universiteit van Amsterdam: SCO-Kohnstamm Instituut.
- Van Keer, H., & Verhaeghe, J. P. (2005). Comparing two teacher development programs for innovating reading comprehension instruction with regard to teachers' experiences and student outcomes. *Teaching and Teacher Education*, 21(5), 543-562. doi: <http://dx.doi.org/10.1016/j.tate.2005.03.002>
- Van Petegem, P., & Imbrecht, I. (2012). *Wegwijs in het Vlaamse onderwijs : onderwijsorganisatie en -beleid in kaart gebracht*. Mechelen: Plantyn.
- Van Petegem, P., Devos, G., Mahieu, P., Dang Kim, T., & Warmoes, V. (2006). *Hoe sterk is mijn school? Het beleidsvoerend vermogen van Vlaamse scholen*. Mechelen: Wolters-Plantyn.
- Van Vliet, M., Kok, K., & Veldkamp, T. (2010). Linking stakeholders and modellers in scenario studies: The use of Fuzzy Cognitive Maps as a communication and learning tool. *Futures*, 42(1), 1-14. doi: <http://dx.doi.org/10.1016/j.futures.2009.08.005>
- Vanhoof, J., & Van Petegem, P. (2007). Matching internal and external evaluation in an era of accountability and school development: lessons from a Flemish perspective. *Studies in Educational Evaluation*, 33, 101-119.
- Vanhoof, J., Deneire, A., & Van Petegem, P. (2011). Waar zit het beleidsvoerend vermogen in (ver) scholen?[Where are the policy-making capacities concealed in schools?]. *Antwerpen: Wolters Plantyn*.
- Vanhoof, J., Verhaeghe, G., Van Petegem, P., & Valcke, M. (2011). Datageletterdheid versterken bij scholen: lessen uit het schoolfeedbackproject. *Kwaliteitszorg in het onderwijs*.
- Ver Eecke, E. (2004). Leerwinst als kwaliteitsindicator: een haalbare kaart of een brug te ver? *Impuls*, 34(3), 149-163.
- Verhaeghe, J. P. (2010). Schoolfeedback als input voor interne kwaliteitszorg. *status: published*.
- Verhelst, M. (2010). Taaltoetsen aan het begin van het schooljaar. *Nieuwsbrief Taal & Onderwijs*, 1(3).
- Verhoeven, J., Devos, G., Bruylant, W., & Warmoes, V. (2002). *Evalueren in Vlaamse secundaire scholen: een onderzoek bij leraren en directeurs*. Garant.
- Victorian Curriculum and Assessment Authority. (2013). *VCAA Assessment Online. On Demand General Information*. Victorian Curriculum and Assessment Authority: Melbourne.

- Vlaams Ministerie van Onderwijs en Vorming, Onderwijsinspectie (2015). Onderwijsspiegel 2015. Geraadpleegd op: <https://www.vlaanderen.be/nl/publicaties/detail/onderwijsspiegel-2015-jaarlijks-rapport-van-de-onderwijsinspectie-1>.
- Vlaams Ministerie van Onderwijs en Vorming. (2010). Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for the Flemish Community of Belgium OECD.
- Vlaams Ministerie van Onderwijs en Vorming. (z.j.). Inschrijven in het basisonderwijs. Geconsulteerd op 5 oktober 2015, van <http://www.ond.vlaanderen.be/wetwijs/thema.aspx?id=37&fid=5>
- Vlaams Ministerie van Onderwijs en Vorming. (z.j.). Wat is een leerlingvolgsysteem? Geconsulteerd op 5 oktober 2015, van <http://www.ond.vlaanderen.be/toetsenvoorscholen/LVS/infoleerlingvolgsysteem.asp>
- Vlaamse Overheid. (2005). Verkennen van de toekomst met scenario's. Brussel: Studiedienst van de Vlaamse Regering.
- Vlaamse Regering (2009). Decreet betreffende de Onderwijskwaliteit]. Geconsulteerd op 5 oktober 2015, van <http://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=14129>
- Vlaamse Regering. (2014). Regeerakkoord Vlaamse Regering 2014-2019. Brussel: Vlaamse Regering.
- Voeten, M. (1990). Longitudinaal onderzoek van de leesontwikkeling. In C. Aarnoutse & M. Voeten (Eds.), *Gaat en onderwijst. Liber amicorum voor dr. M.J.C. Mommers bij gelegenheid van zijn afscheid van de Katholieke Universiteit Nijmegen*. Tilburg: Zwijssen.
- Vogler, K. E. (2006a). Impact of a High School Graduation Examination on Tennessee Science Teachers' Instructional Practices. *American Secondary Education*, 35(1), 33-57.
- Vogler, K. E. (2006b). Impact of an exit examination on English teachers' instructional practices. *Essays in Education*, 16(Spring). <http://www.usca.edu/essays/vol162006/vogler.pdf>
- Wallace, M. (2002). Modelling distributed leadership and management effectiveness: Primary school senior management teams in England and Wales. *School effectiveness and school improvement*, 13(2), 163-186.
- Wang, L., Beckett, G. H., & Brown, L. (2006). Controversies of Standardized Assessment in School Accountability Reform: A Critical Synthesis of Multidisciplinary Research Evidence. *Applied Measurement in Education*, 19(4), 305-328.
- Wang, T., & Kolen, M. J. (2001). Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. *Journal of Educational Measurement*, 38(1), 19-49.
- Wayman, J. G., & Stringfield, S. (2006). Data use for school improvement: School practices and research perspectives. *American Journal of Education*, 112, 463-468.
- White, J. (2013). Trust in Testers. *London Review of Education*, 11(1), 1-6.
- Wiggins, A., & Tymms, P. B. (2000, September). Dysfunctional effects of public performance indicator systems: a comparison between English and Scottish Primary Schools. In *Paper presented at the European Conference on Educational Research* (Vol. 20, p. 23).
- Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research & Evaluation*, 2(2), 1-6.
- Wikeley, F. (1998). Dissemination of research as a tool for school improvement? *School Leadership & Management*, 18(1), 59-73.
- Wikeley, F. (1998). Dissemination of research as a tool for school improvement?. *School Leadership & Management*, 18(1), 59-73.

- Wilcox, B., & Gray, J. (1996). *Inspecting Schools: Holding Schools to Account and Helping Schools To Improve*. Taylor & Francis, Inc., Orders and Customer Service, 7625 Empire Drive, Florence, KY 41042 (paperback: ISBN-0-335-19674-8, \$30.95; hardback: ISBN-0-335-19675-6, \$102.95).
- William, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, 45(2), 107-122.
- Williamson, J. (1990). The lack of impact of information: performance indicators for A levels. *Educational Management & Administration*, 18(1), 37-45.
- Williamson, J., Tymms, P., & Haddow, M. (1992). ALIS through the looking glass: changing perceptions of performance indicators. *Educational Management & Administration*, 20(3), 179-187.
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39, 1-37. *Journal of Educational Measurement*, 39, 1-37.
- Woods, P. A., Bagley, C., & Glatter, R. (1998). School responsiveness in a competitive climate: The public market in England. *Educational Administration Quarterly*, 34(1 suppl), 650-676.
- Wormington, S. V., Corpus, J. H., & Anderson, K. G. (2012). A person-centered investigation of academic motivation and its correlates in high school. *Learning and Individual Differences*, 22(4), 429-438. doi: 10.1016/j.lindif.2012.03.004
- Wössmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, 13(2), 143-169.
- Wu, M. (2010). Measurement, Sampling, and Equating Errors in Large-Scale Assessments. *Educational Measurement: Issues and Practice*, 29(4), 15-27. doi: 10.1111/j.1745-3992.2010.00190.x
- Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, 20(1), 35-52.
- Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education*, 25, 469-483.
- Zenisky, A., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15, 337-362.

