

MAARTEN PENNINGCKX, JAN VANHOOF, AMY QUINTELIER,
SVEN DE MAEYER & PETER VAN PETEGEM

ZICHT OP LEERWINST

SCENARIO'S VOOR **GESTANDAARDISEERD TOETSEN**



Deze publicatie kwam tot stand in het kader van OBPWO-project 13.03
in opdracht van het Ministerie van de Vlaamse Gemeenschap.

Al het mogelijke werd gedaan om de informatie in dit boek zo juist en actueel te maken. Auteurs of uitgever kunnen niet verantwoordelijk gesteld worden voor mogelijke nadelen die lezers door eventuele onvolkomenheden in het boek zouden ondervinden.

WOORD VOORAF

De Vlaamse onderwijscontext is traditioneel gekenmerkt door een hoge mate van autonomie van scholen. Het grondwettelijke principe van 'vrijheid van onderwijs' bepaalt dat scholen over het recht beschikken om een eigen visie op de kwaliteit van onderwijs en een schoolspecifiek pedagogisch project na te streven. In het Vlaamse Regeerakkoord staat de ambitie centraal om als overheid een terughoudende rol te spelen, en bevestigt daarbij de scholen de nodige autonomie toe te kennen.

Over die autonomie van scholen werd de laatste jaren heel wat onderzoek gevoerd. Op dit moment legt de overheid eindtermen en ontwikkelingsdoelen op aan scholen. De eindtermen en ontwikkelingsdoelen vertolken maatschappelijk vooropgestelde doelen die het onderwijs respectievelijk moet bereiken en nastreven. Daarnaast zijn de eindtermen en ontwikkelingsdoelen ook de weergave van een door het parlement gelegitimeerde opvatting over kwaliteit en kwaliteitsbewaking. De onderwijsinspectie staat garant voor de controle hierop en maakt tijdens een schooldoorlichting een inschatting van de mate waarin de eindtermen en ontwikkelingsdoelen worden bereikt/nagestreefd en van de interne kwaliteitszorg van een school.

In het jaarlijks verslag van de Onderwijsinspectie van 2015, de Onderwijs Spiegel, blijkt dat scholen op één van vier kwaliteitsaspecten, namelijk 'doeltreffendheid' jaar na jaar het laagst scoren (Onderwijsinspectie, 2015). Onder doeltreffendheid wordt de mate verstaan waarin scholen onderzoeken of de vooropgestelde doelen bereikt worden op zowel school- als klasniveau. Uit het verslag blijkt dat de evaluatiepraktijk in Vlaamse scholen onvoldoende is afgestemd op de onderwijsdoelen en dat heel wat leraren er niet in slagen om hun onderwijs te richten op de onderwijsdoelen, die in toenemende mate geformuleerd zijn in termen van competenties. Daarmee samenhangend blijkt het zichtbaar maken van leren (en leerwinst) een uitdaging te blijven.

Mede door bovenstaande vaststellingen, maar ook door internationale tendensen en een toenemende vraag naar transparantie gaan steeds meer stemmen op om ook in Vlaanderen een systeem van leer(winst)monitoring te ontwikkelen om leerresultaten van leerlingen en scholen in kaart te brengen. Boven rijst de vraag of bij de evaluatie van de opbrengsten van leerlingen of scholen ook gegevens met betrekking tot de leerwinst of de toegevoegde waarde betrokken moet worden.

Er bestaan goede redenen om een (meer) systematisch zicht te krijgen op leerresultaten, leerwinst en toegevoegde waarde in het Vlaamse leerplichtonderwijs. Daar staat tegenover dat leer(winst)monitoringssystemen bekritiseerbaar zijn en dat ongewenste effecten zich kunnen voordoen. Het monitoren van leerresultaten/leerwinst is bovendien vanuit diverse perspectieven bekeken een complexe aangelegenheid (zowel politiek-maatschappelijk als onderwijskundig en wetenschappelijk-methodologisch). Als de Vlaamse overheid dergelijke initiatieven overweegt in te voeren, is het aangewezen om bestaande (buitenlandse) monitoringssystemen met hun positieve en negatieve ervaringen te bestuderen. Omwille van die redenen werden naast, een uitgebreide literatuurstudie, ook vier gevalsstudies uitgevoerd in Australië, Engeland, Nederland en Noorwegen.

Tot slot werd ook een Delphi-studie uitgevoerd om de haalbaarheid en wenselijkheid van leer(winst)monitoringssystemen in Vlaanderen te evalueren. Deze onderzoeksmethode is geschikt om een gestructureerd communicatieproces tussen onderwijsexperts op gang te brengen en alzo verschillende standpunten en argumenten te exploreren.

Dit boek bestaat uit negen delen. In Deel 1 worden de probleemstelling en de onderzoeksvragen voorgesteld. Deel 2 schetst de onderzoeksmethodiek van de diverse luiken. De daaropvolgende delen rapporteren de resultaten van de drie te onderscheiden onderzoeksactiviteiten, namelijk de literatuurstudie (Deel 3), de eerste en tweede ronde van de Delphi-studie (Deel 4) en de internationale studie (Deel 5). In Deel 6 wordt de ontwikkeling van acht afzonderlijke scenario's met een concreet uitgewerkt leer(winst)monitoringssysteem beschreven. Deel 7 beschrijft de resultaten afkomstig uit de derde ronde van de Delphi-studie en geeft een inschatting van de wenselijkheid en haalbaarheid van elk van de scenario's aan de hand van de standpunten van het expertenpanel. Deel 8 en Deel 9 sluiten ten slotte dit boek af met een algemeen besluit en enkele beleidsaanbevelingen.

Aan het einde van dit project willen wij enkele personen en instanties danken omdat zij een grote betekenis gehad hebben voor de realisatie van het project. Dat zijn in de eerste plaats de respondenten die hebben meegewerkt aan de Delphi-studie. De gedrevenheid en de genuanceerdheid waarmee de verschillende vragenrondes beantwoord werden, waren essentieel voor de succesvolle realisatie van het onderzoeksproject. Daarnaast willen wij een speciaal woord van dank richten aan de respondenten in de vier cases van het internationale onderzoeksluik. Hen willen we bedanken voor de intensieve correspondentie en hun bereidwilligheid om tijdens een interview dieper in te gaan op specifieke aspecten van het publicatiesysteem in hun land.

Een woord van dank is ook op zijn plaats voor de stuurgroep opgezet vanuit de dienst Strategische Beleidsondersteuning van het Departement Onderwijs en Vorming. Wij hebben onder meer hun kritische blik en enthousiasme voor dit project geapprecieerd.

Wij hopen dat de bevindingen van dit onderzoek een bijdrage kunnen leveren aan de kwaliteit van het (beleids)debat over onderwijskwaliteit in het algemeen en over leer(winst)monitoringssystemen in het bijzonder. Dat het kleur moge brengen in het zwart-wit denken dat de discussie over de betreffende problematiek vaak kenmerkt.

Antwerpen, december 2016

De auteurs

INHOUDSTAFEL

Woord Vooraf	i
Inhoudstafel.....	iii
Deel 1 Inleidend.....	1
1 AANLEIDING EN SITUERING	1
2 BEGRIPSAFBAKENING	3
2.1 Gestandaardiseerde toetsen	3
2.2 Leerwinst en toegevoegde waarde	5
2.3 Uiteenlopende doelen van gestandaardiseerde toetsen.....	5
2.4 High-stakes en low-stakes leer(winst)monitoringssystemen.....	8
2.5 Verschillende criteria.....	9
3 HISTORIEK	10
3.1 Historiek in Europees perspectief	10
3.2 Historiek in Vlaanderen	11
4 DE HUIDIGE SITUATIE: GESTANDAARDISEERDE TOETSEN IN HET VLAAMSE ONDERWIJS	12
4.1 Peilingen en paralleltoetsen	Fout! Bladwijzer niet gedefinieerd.
4.2 Toetsen van Katholiek Onderwijs Vlaanderen en OVSG.....	14
4.3 Leerlingvolgsystemen	15
4.4 Examencommissie van de Vlaamse Gemeenschap	16
4.5 Schoolfeedbacksysteem	16
4.6 Toetsen voor oriëntering van leerlingen.....	17
4.7 Deelname aan internationaal vergelijkende studies van leerprestaties.....	17
4.8 Andere toetsen	18
Deel 2 Onderzoeksaanpak	21
1 ONDERZOEKSDESIGN	21
2 LITERATUURONDERZOEK	22
3 INTERNATIONALE CASESTUDIES	23
3.1 Australië	25
3.2 Engeland (Verenigd Koninkrijk) - Centre of Evaluation and Monitoring (CEM).....	25
3.3 Nederland - pilootproject Leerwinstmonitoring en Toegevoegde Waarde	26
3.4 Noorwegen - met focus op Oslo	26
4 DELPHI-STUDIE.....	27
4.1 Definitie en doeleinden van een Delphi-studie	27
4.2 Verloop van de Delphi-studie in dit onderzoek	28
5 SCENARIO-ONTWIKKELING	30
5.1 Definitie en doeleinden van scenario's	30
5.2 De meerwaarde van scenario-onderzoek in deze context	31
5.3 Scenariotypologieën	31
5.4 Scenario-ontwikkeling en Delphi-studies	33
5.5 Scenario-ontwikkeling in dit onderzoek.....	33
5.6 Het gebruik van de scenario's in dit onderzoek	34

Deel 3 Resultaten literatuurstudie.....	35
1 IMPACT VAN GESTANDAARDISEERDE TOETSEN ALS BASIS VAN LEER(WINST)MONITORING	35
1.1 Positieve effecten op microniveau	37
1.2 Onwenselijke effecten op microniveau	47
1.3 Positieve effecten op mesoniveau	54
1.4 Onwenselijke effecten op mesoniveau	55
1.5 Positieve effecten op macroniveau	57
1.6 Onwenselijke effecten op macroniveau	58
2 BOUWSTENEN VAN LEER(WINST)MONITORINGSSYSTEMEN	59
2.1 Mogelijkheid tot berekening van leerwinst en/of toegevoegde waarde	59
2.2 Opzet van de gestandaardiseerde toetsen	61
2.3 Individuele kenmerken van leraren	75
2.4 Kenmerken van scholen	76
2.5 Kenmerken van het onderwijssysteem	78
3 CONCEPTUALISERING VAN LEERWINST EN TOEGEVOEGDE WAARDE	79
3.1 Leerwinst	79
3.2 Toegevoegde waarde	85
4 CONCLUSIES UIT DE LITERATUURSTUDIE	90
Deel 4 Resultaten Delphi-studie (eerste en tweede onderzoeksrunde).....	95
1 INSCHATTING VAN BEHOEFTE AAN EEN LEER(WINST)MONITORINGS-SYSTEEM	95
2 IMPACT VAN EEN LEER(WINST)MONITORINGSSYSTEEM	98
2.1 Microniveau	98
2.2 Mesoniveau	104
2.3 Macroniveau	109
3 LEERWINST EN TOEGEVOEGDE WAARDE	113
3.1 Leerwinst	114
3.2 Toegevoegde waarde	116
4 ANDERE AANDACHTSPUNTEN UIT DE DELPHI-STUDIE	118
4.1 Psychometrisch sterk onderbouwde toetsen	118
4.2 Toetsformat	119
4.3 Vrijwillig of verplicht karakter	119
4.4 Andere voorwaarden	120
Deel 5 Resultaten internationale studie	121
1 BESLISSING OVER IMPLEMENTATIE	121
1.1 Bepalen van een doel en rationale	121
1.2 Een duidelijk en eenduidig doel	122
1.3 Welke consequenties aan de toetsresultaten verbinden?	123
1.4 Een ontwikkelingsgerichte impact op microniveau is niet makkelijk te bereiken	124
1.5 Aantonen van een vaak nog ongekende behoefte	124
1.6 Bij aanvang op vrijwillige basis aanbieden	125
2 EIGENSCHAPPEN VAN HET LEER(WINST)MONITORINGSSYSTEEM	126
2.1 Digitale toetsen om snelle feedback te voorzien	126
2.2 Adaptieve, of getrapt adaptieve toetsen	126
2.3 Al dan niet curriculum-gerelateerd toetsen	127
2.4 Welke leerdomeinen toetsen?	129
2.5 Brede toetsenbatterij	129
2.6 Sterke psychometrische onderbouwing	129
2.7 Sterke procedures voor kwaliteitszorg bij afname en correctie	130
2.8 Samenwerking tussen verschillende actoren in toetsontwikkeling	130

2.9	Gebruik maken van bestaande (internationale) toetsen	131
3	LEERWINST	131
3.1	Belang van het meten van leerwinst	131
3.2	Eigenschappen van een leerwinstmonitoringssysteem	132
4	TOEGEVOEGDE WAARDE	132
5	RAPPORTERING	133
5.1	Gedetailleerde en interactieve rapportering	133
5.2	Vermijd publieke rapportering	134
6	ONDERSTEUNING VAN SCHOLEN	134
6.1	Ondersteuning bij (gebruiksvriendelijke) afname	135
6.2	Investeren in vermogen van de scholen om informatie om te zetten in acties	135
7	FINANCIËLE INSCHATTING IS BIJZONDER MOEILIJK	137
Deel 6 Bouwstenen van de scenario's.....		139
1	ONTWIKKELING VAN DE ACHT SCENARIO'S AAN DE HAND VAN BOUWSTENEN	139
2	OVERZICHT VAN DE BOUWSTENEN	140
2.1	Primaire bouwstenen	140
2.2	Secundaire bouwstenen	144
3	SCENARIO'S	146
3.1	Scenario 1 'High-stakes eindtoetsen voor leerlingen'	146
3.2	Scenario 2 'Adaptief leerwinstmonitoringssysteem'	151
3.3	Scenario 3 'Diagnostisch meetinstrumentarium & toetsenbank'	154
3.4	Scenario 4 'High-stakes eindtoetsen voor leerlingen en scholen'	158
3.5	Scenario 5 'Leerwinstmeting op schoolniveau met openbaar karakter'	162
3.6	Scenario 6 'Leerwinstmeting op schoolniveau met afgeschermd karakter'	166
3.7	Scenario 7 'Facultatieve toetsenbank'	170
3.8	Scenario 8 'Huidige trend wordt verdergezet'	173
Deel 7 Inschatting van wenselijkheid en haalbaarheid van de scenario's.....		175
1	ALGEMENE BESPREKING	175
1.1	Wenselijkheid	175
1.2	Haalbaarheid	177
1.3	Verschillen tussen categorieën van respondenten	178
2	BESPREKING PER SCENARIO	178
2.1	High-stakes eindtoetsen voor leerlingen	178
2.2	Adaptief leerwinstmonitoringssysteem	183
2.3	Diagnostisch meetinstrumentarium & toetsenbank	187
2.4	High-stakes eindtoetsen voor leerlingen en scholen	191
2.5	Leerwinstmeting op schoolniveau met openbaar karakter	195
2.6	Leerwinstmeting op schoolniveau met afgeschermd karakter'	198
2.7	Facultatieve toetsenbank	201
2.8	Huidige trend wordt verdergezet.....	204
Deel 8 Conclusie.....		209
1	DOELEN VAN HET LEER(WINST)MONITORINGSSYSTEEM	209
2	LEERWINST EN TOEGEVOEGDE WAARDE	210
2.1	Leerwinst monitoren.....	210
2.2	Toegevoegde waarde monitoren	211
3	OPENBAARHEID VAN DE RESULTATEN	212
4	EIGENSCHAPPEN VAN DE TOETSEN IN EEN LEER(WINST)MONITORINGSSYSTEEM	213
4.1	Toetsenbank	213

4.2	Sterke psychometrische onderbouwing	213
4.3	Link met de eindtermen en ontwikkelingsdoelen	214
4.4	Brede toetsing.....	215
5	ONDERSTEUNING.....	216
6	DRAAGVLAK CREËREN.....	218
Deel 9 Beleidsaanbevelingen.....		219
Literatuurlijst.....		226

DEEL 1 INLEIDEND

1 AANLEIDING EN SITUERING

In verschillende onderwijssystemen woedt een aanhoudende discussie over de implementatie van systemen voor het monitoren van leerresultaten van leerlingen. Hoewel sommige landen doelbewust afzien van dergelijke initiatieven is toch een groeiende tendens waar te nemen van het aantal landen dat wél stappen in die richting onderneemt of waar het gestandaardiseerd in kaart brengen van leerresultaten en leerwinst van leerlingen reeds gemeengoed is geworden. Bovendien worden aan de uitkomsten van gestandaardiseerde toetsen steeds meer gevolgen gekoppeld (OECD, 2013). Naast het uitreiken van diploma's, die vaak de enige mogelijkheid zijn om leerlingen toegang tot hoger onderwijs te verschaffen, zijn in diverse regio's en landen een aantal andere consequenties verbonden aan de resultaten van een leer(winst)monitoringssysteem. Zo kunnen toetsresultaten de basis vormen voor het evalueren (en borgen of verbeteren) van de kwaliteit van het onderwijs, zowel op klas- en schoolniveau als op landelijk niveau (Cizek, 2005; Loeb, 2013). Deze resultaten kunnen ook leiden tot het opstellen van rankings van scholen, waarbij de toets gebruikt wordt als meetinstrument voor de kwaliteit van de scholen (de Wolf & Janssens, 2007; Dill & Soo, 2005). In sommige landen worden de uitkomsten van een leer(winst)monitoringssysteem eveneens gebruikt om leraren te evalueren en worden er financiële stimuli toegekend aan goed presterende scholen en/of leraren (Baker et al., 2010; Loeb, 2013). In de meest uiteenlopende onderwijscontexten geven gestandaardiseerde toetsen die deel uitmaken van een leer(winst)monitoringssysteem aanleiding tot reeds decennialang aanslepende discussies (Linn, 2001; Wang, Beckett, & Brown, 2006). Het aanhouden van deze discussie wordt verklaard door de inbedding ervan in een ruimere '*clash of philosophical positions*', waarbij gestandaardiseerde toetsen de uiting zijn van twee tegengestelde visies op onderwijs en kwaliteit in het algemeen. Aan de ene zijde van het continuüm staat een visie die pleit voor competitie en het belonen van verdiensten en prestaties. Deze visie, waarin gestandaardiseerde toetsen een belangrijke plaats innemen, omarmt het idee van individuele verantwoordelijkheid en erkent verschillen tussen individuen. Aan de andere zijde van het continuüm staat een visie waarbij competitie wordt ontmoedigd, waarbij inzet en goede processen worden beloond, waarbij collectieve verantwoordelijkheid de voorkeur krijgt boven individuele verantwoordelijkheid en waarbij gestreefd wordt naar gelijkheid. Het betreft daarnaast een '*clash of interests*', waarbij aan de ene kant wordt gepleit voor een transparante meting van de prestatie van elke openbare dienst, terwijl aan de andere kant wordt beweerd dat meting en transparantie enkel tot onwenselijke neveneffecten leiden (Phelps, 2005a).

Door de sterke geladenheid van het thema is er internationaal al bijzonder veel gepubliceerd omtrent de (veronderstelde) effecten en neveneffecten van een leer(winst)monitoringssysteem of van losstaande gestandaardiseerde toetsen. Sinds de opkomst van de eerste studies, in de jaren dertig van de vorige eeuw, is er een continue stroom van publicaties geweest rond dit thema (Phelps, 2005b). Het betreft evenwel talrijke opiniërende publicaties en slechts zelden goed onderbouwde

empirische studies. Vaak wordt er door beleidsmakers veel aangenomen inzake effectiviteit van gestandaardiseerde toetsen en zijn de onderliggende assumpties (hoe moeten de verschillende toetsen in het leer(winst)monitoringssysteem leiden tot een verbetering van de onderwijskwaliteit?) slechts matig geëxpliciteerd (Skedsmo, 2011).

Ook in Vlaanderen is het idee van het opzetten van systemen die leerprestaties van leerlingen op een gestandaardiseerde manier in kaart brengen steeds gecontesteerd (Shewbridge, Hulshof, Nusche, & Stoll, 2011). Er gaan verschillende stemmen op van stakeholders die het wenselijk achten een (meer) systematisch en gestandaardiseerd zicht te krijgen op de leerresultaten van leerlingen in het leerplichtonderwijs, waarbij deze resultaten zowel op micro-, meso- én macroniveau kunnen dienen. Daar staat tegenover dat er ook stakeholders van mening zijn dat op bestaande (buitenlandse) systemen om leerresultaten te monitoren gefundeerde kritieken te geven zijn en dat ongewenste neveneffecten zich (kunnen) voordoen (Hirtt, Nicaise, & De Zutter, 2007; Standaert, 2014). De discussie in Vlaanderen moet gezien worden in de context van het grondwettelijk vastgelegde principe van 'vrijheid van onderwijs' waarin bepaald wordt dat scholen een grote vrijheid hebben om een eigen invulling aan onderwijskwaliteit te geven (gegeven de decretale bepalingen). De visie wordt dan ook vaak gehuldigd dat verschillen tussen scholen niet enkel moeten worden toegelaten, maar zelfs moeten worden aangemoedigd. Uit onderzoek in het kader van een voorbereidende evaluatieschets voor een review van de OESO over evaluatie in het Vlaamse onderwijs blijkt dat de argumenten die door verschillende tegenstanders van centrale examens werden genoemd erg talrijk zijn (Vlaams Ministerie van Onderwijs en Vorming, 2010). Het betreft pedagogisch/onderwijskundige bezwaren (bijvoorbeeld dat onderwijskwaliteit niet mag worden verengd tot meetbare output), normatieve bezwaren (bijvoorbeeld dat een examen een statisch beeld weergeeft van leerlingen, en dat rankings kunnen worden opgemaakt van scholen), bezwaren tegen de examens zelf (bijvoorbeeld over beperkingen inzake de metingsbetrouwbaarheid), en praktische bezwaren (zoals de organisatorische en financiële kost). Desalniettemin is er toch vanuit het onderwijsveld alsook vanuit de Vlaamse Scholierenkoepel vraag naar het aanreiken van meer gestandaardiseerde tools die helpen bij beslissingen over de loopbaan van de leerlingen (Penninckx, Vanhoof, & Van Petegem, 2011).

Ondertussen kan men niet om de vaststelling heen dat in het Vlaamse onderwijs steeds meer aandacht wordt geschonken aan de resultaten van de onderwijsprocessen (op micro-, meso- én macroniveau). De tendens naar het aanspreken van onderwijsaanbieders (leraren, scholen en intermediairen) op de wijze waarop ze de autonomie die ze genieten, succesvol weten aan te wenden, is daar uiteraard niet vreemd aan. Bovendien groeit bij diverse stakeholders de overtuiging dat door het creëren van synergieën inzake het monitoren van leerresultaten op micro-, meso- en macroniveau krachtige instrumenten voor het bewaken en ontwikkelen van (aspecten van) onderwijskwaliteit kunnen opgezet worden (OECD, 2013). In het Vlaamse Regeerakkoord staat onder meer het voornemen om van een soort 'toolkit' die kan begrepen worden als een aanzet tot een leer(winst)monitoringssysteem werk te maken:

“Op het einde van het basisonderwijs worden alle leerlingen getoetst. De doelstelling is zowel feedback te kunnen geven in het kader van de interne kwaliteitscontrole van de scholen als gegevens te verzamelen op systeemniveau. Met de onderwijsverstrekkers wordt overeengekomen om de huidige OVSG-toetsen en de interdiocesane toetsen te valideren in samenspraak met het Steunpunt

Toetsontwikkeling, dat de peilingsproeven en paralleltoetsen ontwikkelt. Op die manier kan een toets toolkit ontstaan. Met de realisatie van dezelfde doelstelling voor ogen voeren we gesprekken met de onderwijsverstrekkers van het secundair onderwijs en bekijken we hoe de bestaande peilingsproeven daartoe een basis kunnen vormen” (Vlaamse Regering, 2014, p.96).

Het monitoren van leerresultaten/leerwinst is vanuit diverse perspectieven bekeken een complexe aangelegenheid (zowel politiek-maatschappelijk als onderwijskundig en wetenschappelijk-methodologisch) (Janssens, Rekers-Mombarg, & Lacor, 2014). Als de Vlaamse overheid initiatieven tot monitoring van leerresultaten overweegt in te richten, te faciliteren of te stimuleren, dan is het aangewezen te leren uit bestaande (buitenlandse) systemen en uit bestaande literatuur. Heldere en gedragen doelen dienen vooropgesteld te worden, net zoals ook de haalbaarheid op voorhand moet worden afgetoetst. Bestaande monitoringsystemen zijn immers binnen een lokale traditie en cultuur gegroeid, waardoor hun determinanten niet automatisch gegeneraliseerd kunnen worden naar andere onderwijscontexten. Onderzoek naar de haalbaarheid en de mogelijkheden van leerwinstmeting in de Vlaamse context dringt zich dus op.

Deze studie onderbouwt en ontwikkelt mogelijke scenario's voor het monitoren van leerresultaten/leerwinst en gaat na of deze haalbaar en wenselijk zijn in de Vlaamse context. Het betreft dus een beleidsevaluatie ex ante met als doel de beleidsvoorbereiding wetenschappelijk te ondersteunen. De focus ligt op het leerplichtonderwijs (lager en secundair onderwijs) en op de leerdomeinen wiskunde, wetenschappen en Nederlands. De volgende onderzoeksvragen worden vooropgesteld:

- OV1 - Welke zijn (volgens Vlaamse stakeholders) de karakteristieken van een optimaal leer(winst)monitoringsysteem in het Vlaamse leerplichtonderwijs?
- OV2 - Welke zijn de (volgens Vlaamse stakeholders) karakteristieken van een optimale conceptualisering en operationalisering van het concept 'leerwinst' in het Vlaamse leerplichtonderwijs?
- OV3 - Tot welke beleidsscenario's geven de ontwerpprincipes die uit OV1 en OV2 afgeleid kunnen worden aanleiding?
- OV4 - Hoe beoordelen Vlaamse stakeholders de haalbaarheid en wenselijkheid van deze beleidsscenario's?

2 BEGRIPSAFBAKENING

Deze sectie heeft tot doel om enkele centrale concepten binnen dit boek af te bakenen. Zo worden de termen 'gestandaardiseerde toetsen', 'leerwinst' en 'toegevoegde waarde' kort besproken. Daarna worden functies en vormen van gebruik van gestandaardiseerde toetsen ingeleid.

2.1 Gestandaardiseerde toetsen

De focus van deze studie ligt op een leer(winst)monitoringssysteem. Een leer(winst)-monitoringssysteem bestaat per definitie uit één of meerdere gestandaardiseerde toetsen. De literatuur rond gestandaardiseerde toetsen richt zich zeer sterk op 'state tests': toetsen die door alle leerlingen op een van bovenaf opgelegd moment dienen te worden afgelegd. Vaak worden er ook

directe consequenties aan de resultaten van deze toetsen verbonden voor leerlingen, leraren of scholen.

De invulling van gestandaardiseerde toetsen in een leer(winst)monitoringssysteem, en dus ook in dit rapport, is evenwel breder. Volgens Sireci (2005, p. 113) betekent een gestandaardiseerde toets dat: *“the test content is equivalent across administrations and that the conditions under which the test is administered are the same for all the test takers”*.

Tabel 1 - Kenmerken en mate van standaardisatie

	Licht gestandaardiseerd	Medium gestandaardiseerd	Sterk gestandaardiseerd
Organisatie			
Verspreiding	Regionaal		Nationaal
Soort examen	Schriftelijk overheen hele gebied in aanvulling van door scholen ontwikkelde toetsen	Schriftelijke toetsen overheen het hele gebied, mondelinge toetsen door scholen	Schriftelijke en mondelinge toetsen overheen het hele gebied
Aantal leerdomeinen	1	4-5	Alle leerdomeinen
Belangrijkste bron van informatie	Toetsen van school meest belangrijk	Even belangrijk	Enkel eindexamens
Leerdomeinen	Enkel hoofdvakken	Hoofdvakken en specialisatie-vakken	Alle leerdomeinen
Keuze van leerdomeinen door leerlingen	Vrije keuze	Deels verplicht, deels vrije keuze	Alles verplicht
Keuze van moeilijkheidsgraad/ beheersingsniveau door leerlingen	Vrije keuze	Keuze automatisch gerelateerd aan keuze voor vakken	Geen keuze: moeilijkheidsgraad identiek voor alle leerlingen
Verbetering van toetsen			
Aantal fases in verbetering	1	2	3
Evaluatoren	Van de school zelf	Van andere scholen	Centrale commissie
Anonimiteit	Niet anoniem	Deels anoniem	Anoniem
Standaarden	Achteraf bepaald		Vooraf bepaald
Gebruik van de resultaten			
Aandeel in eindresultaat van de leerling	Minder dan 30%	50% of meer	100%
Publicatie van geaggregeerde resultaten op schoolniveau	Niet mogelijk	Mogelijk, maar wordt niet openbaar gemaakt	Openbaar gemaakt

Deze definitie houdt in dat een bepaalde inhoud overheen leerlingen en scholen dezelfde is en dat de toets in identieke condities door alle deelnemende leerlingen wordt afgelegd. Er zijn binnen deze

definitie geen vooropgezette bepalingen inzake verplichting of vrijwilligheid van afname noch inzake het doel van de toets, de consequenties, de frequentie van afname, de getoetste leerdomeinen, de leeftijd van leerlingen en de toetsvorm. Een gestandaardiseerde toets kan dus op zeer uiteenlopende manieren vorm worden gegeven.

Klein en Van Ackeren (2012) stelden een inventaris op om het concept gestandaardiseerde toetsen te operationaliseren en deelden toetsen op in 'licht gestandaardiseerd', 'medium gestandaardiseerd' en 'sterk gestandaardiseerd'. Om tot deze conceptualisering te komen, vertrokken zij vanuit drie kenmerken, namelijk (1) organisatorische kenmerken, (2) kenmerken inzake het verbeteren, en (3) kenmerken inzake het gebruik van de toetsresultaten (waarbij weliswaar vooral wordt uitgegaan van eerder summatieve eindtoetsen, die worden ingezet op een verantwoordingsgerichte manier op microniveau). Tabel 1 is een vertaling van de tabel uit deze studie (Klein & Van Ackeren, 2012, p. 182).

2.2 Leerwinst en toegevoegde waarde

De termen 'leerwinst' en 'toegevoegde waarde' (in de Angelsaksische literatuur worden termen als *learning gains*, *student growth* en *added value* aangewend) worden heel vaak door elkaar gebruikt. Uiteenlopende en soms zelfs tegenstrijdige invullingen - vooral wat de term 'added value' betreft - maken dat het vaak moeilijk is om correct te interpreteren wat door verschillende auteurs precies wordt bedoeld (Saunders, 1999).

In overeenstemming met een recent rapport over een pilootproject rond leerwinstmonitoring in Nederland (Janssens et al., 2014), wordt in deze studie over leerwinst gesproken wanneer het individuele leerlingen betreft (microniveau). De term toegevoegde waarde wordt gereserveerd voor de bijdrage van scholen aan het leren of aan de leerwinst van leerlingen. Toegevoegde waarde situeert zich dus op het mesoniveau (zie 2.3.2 voor een definitie van micro- en mesoniveau).

2.3 Uiteenlopende doelen van gestandaardiseerde toetsen

Een leer(winst)monitoringssysteem wordt vaak vanuit uiteenlopende doelen opgezet. Het gebruik van de resultaten van een leer(winst)monitoringssysteem kan gebeuren in ontwikkelingsgericht of een verantwoordingsgericht perspectief, en gericht zijn op micro-, meso- of macroniveau.

2.3.1 Ontwikkelingsgerichte en verantwoordingsgerichte doelen

Er kan onderscheid worden gemaakt tussen het ontwikkelingsgericht of verantwoordingsgericht perspectief wat het doel van een leer(winst)monitoringssysteem betreft. Overheden nemen immers initiatieven om scholen te ondersteunen bij hun ontwikkelingen, maar organiseren ook mechanismen waarbij scholen verantwoording dienen af te leggen. Dit kan op drie manieren: via onderzoek door experts (bijvoorbeeld inspectie), via administratieve monitoring of via gestandaardiseerde toetsen (Scheerens, 2014).

In het ontwikkelingsgericht perspectief dient een gestandaardiseerde toets een formatief doel. Of de resultaten ervan leiden tot consequenties hangt in de eerste plaats af van degene waarop de evaluatie betrokken is: een score van een leerling heeft in dit perspectief gevolgen die ofwel door de leerlingen zelf ofwel door de leraar na een dialoog met de betrokken leerling worden geïnitieerd.

Doelen in een verantwoordingsgericht perspectief zijn daarentegen gericht op het maken van een summatieve beoordeling. De resultaten van het leer(winst)monitoringssysteem leiden tot consequenties die worden opgelegd aan degene die het onderwerp is van de evaluatie (Vanhoof & Van Petegem, 2007; Wang et al., 2006).

In recente decennia is er een geleidelijke verschuiving in vele onderwijssystemen geweest van een 'toetscultuur' naar een meer brede 'assessmentcultuur' waardoor meer toetsen met een eerder ontwikkelingsgericht doel worden afgenomen. Tegelijkertijd betekent de verhoogde autonomie van scholen in vele onderwijssystemen echter dat scholen ook meer verantwoording dienen af te leggen, wat onder meer heeft geresulteerd in een toename van gestandaardiseerde toetsen met een verantwoordingsgericht doel (OECD, 2013; Smith, 2014). Vanuit economisch standpunt kan inderdaad worden geredeneerd dat autonome scholen, bij gebrek aan rigoreuze doelstellingen en competitie, hun eigen belangen zullen nastreven (bijvoorbeeld het beperken van de werklast van personeelsleden) ten koste van de belangen van leerlingen (Bishop & Wössmann, 2004; Wössmann, 2005).

Vaak worden aan toetsen zowel ontwikkelingsgerichte als verantwoordingsgerichte doelen toegeschreven (Linn, 2000; Wang et al., 2006). Tevens geldt dikwijls de assumptie dat een toets met verantwoordingsgerichte doelen indirect zal leiden tot ontwikkeling (Looney, 2014). Andere auteurs wijzen evenwel op de gevaren van het combineren van ontwikkelingsgerichte en verantwoordingsgerichte doelen. De verantwoording zal immers maken dat er een *incentive* bestaat om resultaten op allerlei oneigenlijke manieren te verhogen, waardoor ze niet meer bruikbaar zijn in een ontwikkelingsperspectief (Saunders, 1999).

2.3.2 Macro-, meso- en/of microniveau

Een leer(winst)monitoringssysteem kan ingezet worden om gegevens te verzamelen op verschillende niveaus. Binnen elk van deze niveaus kan bovendien zowel het ontwikkelingsgerichte als het verantwoordingsgerichte perspectief worden gehanteerd. Van elk van de mogelijke combinaties worden in Tabel 2 enkele voorbeelden gegeven.

Zo kan op het microniveau het gebruik op niveau van de klas en de individuele leraar enerzijds en het gebruik door de individuele leerling anderzijds onderscheiden worden. Wanneer de resultaten worden weergegeven op klasniveau, krijgt elke leraar gegevens over zijn/haar professionele effectiviteit wat de getoetste leerdoelen betreft (Baker et al., 2010). De gegevens over de prestaties van verschillende leerlingen kunnen worden aangewend om op een meer efficiënte wijze differentiatie in te zetten (Collins, Reiss, & Stobart, 2010). Gestandaardiseerde toetsen kunnen ook worden gebruikt om leerlingen inzicht te helpen krijgen op hun eigen voortgang, en kunnen op die manier leiden tot toename van de motivatie (Wang et al., 2006). Daarnaast kunnen de resultaten van gestandaardiseerde toetsen in een leer(winst)monitoringssysteem ook ingezet worden binnen een verantwoordingsgericht perspectief. Voornamelijk in de Verenigde Staten is al geëxperimenteerd met verschillende vormen van *performance-based payments* waarbij de resultaten van gestandaardiseerde toetsen een sterke invloed hebben op de beoordeling van leraren (Fryer, 2013, Harris, 2011; Wiliam, 2010). Ten slotte kunnen gestandaardiseerde toetsen ook leiden tot summatieve oordelen over leerlingen, waarbij bijvoorbeeld de resultaten worden gebruikt om leerlingen al dan niet toe te laten tot bepaalde vormen van vervolgonderwijs (Acquah, 2013; Atkinson & Geiser, 2009).

Het mesoniveau betreft de scholen, aangezien resultaten van gestandaardiseerde toetsen in een leer(winst)monitoringssysteem op schoolniveau geaggregeerd kunnen worden. Op die manier kan informatie over sterktes en zwaktes in verschillende scholen worden gegenereerd (Eurydice, 2009). Deze informatie kan de school hulp bieden bij haar processen van interne kwaliteitszorg en kan een bron zijn om bijvoorbeeld professionele ontwikkeling van schoolteams effectiever in te zetten (Cizek, 2005; Loeb, 2013). Schooleffectiviteitsonderzoek wees uit dat scholen die veel data gebruiken, inderdaad op verschillende parameters meer succesvol zijn (Teddle & Reynolds, 2000; Wayman & Stringfield, 2006). Daarnaast kunnen de resultaten van een leer(winst)monitoringssysteem ook worden gebruikt om scholen verantwoording te laten afleggen over de mate waarin ze met het aan hen toegekende overheidsgeld efficiënt zijn omgegaan. Zo kunnen het leer(winst)monitoringssysteem één van de bronnen zijn bij de beoordeling van de kwaliteit van scholen. Het is een internationale trend dat steeds vaker de resultaten van gestandaardiseerde toetsen worden gebruikt in een verantwoordingsgerichte functie bij het evalueren van scholen (Smith, 2014), of openbaar worden gemaakt om marktmechanismen toe te laten waarbij ouders de scholen kiezen op basis van, onder meer, outputinformatie (Bishop & Wössmann, 2004). Scholen kunnen ten slotte ook de resultaten van een leer(winst)monitoringssysteem op een strategische manier gebruiken, bijvoorbeeld om zich aan ouders te verantwoorden voor de kwaliteit van hun onderwijs (Saunders, 1999).

Tabel 2 - Voorbeelden van doelen van een leer(winst)monitoringssysteem

Niveau		Ontwikkeling	Verantwoording
MICRO	Leraren	Zelfevaluatie van de eigen prestatie Beter inzetten van curriculum en van didactische methoden Differentiatie	Evaluatie van leraren Beslissingen inzake benoemingen en loopbaan
	Leerlingen	Zelfevaluatie, zicht op voortgang Leergerichte feedback	Behalen diploma Beslissingen inzake voortgang of zittenblijven
MESO	Scholen	Zelfevaluatie van scholen: interne kwaliteitszorg Betere inzet van pedagogische begeleidingsdiensten	Gebruik door onderwijsinspectie Gebruik voor schoolbesturen of scholengemeenschappen
MACRO	Onderwijssysteem	Beleidsontwikkeling	Beleidsevaluatie Maatschappelijke verantwoording

Op macroniveau kan een onderwijsoverheid van de resultaten van een leer(winst)monitoringssysteem gebruik maken om het onderwijsbeleid van een informatierijke omgeving te voorzien en om haar eigen effectiviteit op een systematische manier te onderzoeken (Cizek, 2005; Loeb, 2013). Bovendien kan het onderwijsbeleid zich aan de hand van deze objectieve gegevens ook verantwoorden voor haar beleid of voor wijzigingen in haar beleid. Metingen en cijfers “maken een wezenlijk en zelfs erg bepalend deel uit van het ontwerpen, plannen, bijsturen en beheren van een onderwijssysteem” (Standaert, 2014, p. 16).

2.4 High-stakes en low-stakes leer(winst)monitoringssystemen

De impact van een leer(winst)monitoringssysteem hangt in zekere mate samen met de grootte van de eraan verbonden inzet (de '*stakes*' van een toets). Toetsen binnen zulk systeem hebben een hoge inzet (*high-stakes*) wanneer er rechtstreeks gevolgen aan de resultaten zijn gekoppeld, zoals beslissingen over voortgang/zittenblijven van leerlingen, salarissen of bonussen voor personeelsleden of beloningen of sancties voor scholen (Au, 2007; Cizek, 2005; Madaus, 1988). Aanvullend dient te worden gesteld dat toetsen ook als high-stakes worden beschouwd indien een bepaald gebruik van de resultaten, bedoeld of onbedoeld, kan leiden tot beslissingen met ernstige gevolgen voor leerlingen, leraren of scholen (AERA, 2000). Voorbeelden daarvan zijn al dan niet doorstroom in de schoolloopbaan of al dan niet toegang verlenen tot vervolgonderwijs.

Typische voorbeelden van high-stakes toetsen zijn eindexamens die bepalen of leerlingen al dan niet een diploma behalen. Ook leer(winst)monitoringssystemen waarbij de resultaten gebruikt worden om werkingsmiddelen toe te kennen of waarvan de resultaten openbaar gepubliceerd worden, zijn hier voorbeelden van. Dergelijke systemen floreren vooral in onderwijscontexten waar er een grote competitie tussen scholen bestaat en het marktmechanisme bij schoolkeuze wordt bevorderd (Sahlberg, 2011). High-stakes toetsen zijn dan ook meestal toetsen die door alle leerlingen in een heel onderwijssysteem moeten worden afgelegd, en die in een verantwoordingsgericht perspectief worden gebruikt (Marchant, 2007).

Een leer(winst)monitoringssysteem hoeft niet noodzakelijk voor elke betrokkene dezelfde inzet te hebben. Zo bestaat er in de Verenigde Staten een lange traditie van gestandaardiseerde toetsen die *low-stakes* zijn voor leerlingen, maar high-stakes voor leraren (Wiliam, 2010).

Wanneer de resultaten van een leer(winst)monitoringssysteem openbaar worden gemaakt, is er steeds een zekere inzet aan de toetsen verbonden (Jürges, Büchel, & Schneider, 2005). Daarom worden ze in dat geval door sommige auteurs per definitie als high-stakes bestempeld (Au, 2007; McNeil, 2000). Een leer(winst)monitoringssysteem heeft daarentegen een lage inzet (*low-stakes*) als er nauwelijks consequenties aan de resultaten zijn verbonden. High-stakes toetsen zijn vaak verbonden aan een verantwoordingsgericht perspectief (zie 3.1.1), terwijl *low-stakes* gestandaardiseerde toetsen eerder voor een ontwikkelingsgericht doel worden gebruikt. In de hoogte van de inzet bij verantwoordingsgerichte toetsen is nog een grote variatie mogelijk. Zo bestaan er bijvoorbeeld verschillende praktijken van gebruik van gestandaardiseerde toetsen om een diploma toe te kennen aan leerlingen: in sommige landen tellen de resultaten op het examen voor 100% (bijvoorbeeld Frankrijk), in andere landen 50% (bijvoorbeeld Nederland) of minder dan 50% (bijvoorbeeld Portugal), terwijl de rest van het eindcijfer dan wordt ingevuld door resultaten op interne evaluaties van de school (Klein & Van Ackeren, 2012).

De inzet van een leer(winst)monitoringssysteem is niet enkel sterk gerelateerd aan de impact of aan het gebruik van de toetsresultaten, maar hangt ook nauw samen met de betrouwbaarheid van de toetsafname. Bij toetsen met een hoge inzet zijn scholen, leraren en leerlingen immers meer geneigd om de resultaten op oneigenlijke of onwenselijke manieren te verhogen, waardoor bias ontstaat en de toetsresultaten geen getrouwe weergave van de werkelijkheid zijn (Harris, 2012). Tegelijk wordt geargumenteerd dat de toetsbetrouwbaarheid net extra zorgvuldig moet worden bewaakt wanneer het high-stakes toetsen betreft (Elstad, 2009; Wiliam, 2010).

2.5 Verschillende criteria

Een leer(winst)monitoringssysteem levert gegevens op aan degene voor wie de resultaten beschikbaar zijn. Mogelijke gebruikers zijn de: overheid, schooldirecties, leraren, leerlingen, ouders, inspectie, andere onderwijsbetrokkenen of het ruimere publiek. Deze gegevens worden evenwel maar zinnvolle informatie op het moment dat ze binnen een bepaald referentiekader kunnen worden afgetoetst. Dat referentiekader wordt ook het 'criterium' genoemd dat wordt gebruikt om aan de gegevens betekenis te geven (Neumann, Trautwein, & Nagy, 2011). Er zijn drie soorten van criteria mogelijk: een groepsnorm-criterium, een standaard-criterium en een zelf-criterium (Lysne, 1984).

Bij een groepsnorm-criterium wordt een bepaalde score afgewogen ten opzichte van de scores van anderen. De score van een leerling voor een wiskundetoets wordt bijvoorbeeld in perspectief geplaatst door een overzicht van de prestaties van zijn klasgenoten of door te vergelijken met het klasgemiddelde. In klassituaties is dit het meest gebruikte criterium (Neumann et al., 2011). Hoewel dit criterium vaak een goed kader geeft voor de interpretatie van een bepaald resultaat kunnen ongelijkheden in de groep waarmee wordt vergeleken tot foutieve conclusies leiden: zo zal een leerling in een 'zwakke klas' schijnbaar een beter resultaat halen dan een leerling met dezelfde prestatie in een 'sterke klas' wanneer de gegevens enkel binnen de eigen klas worden vergeleken (Willingham, Pollack, & Lewis, 2002). Onderzoek wees zelfs uit dat identieke prestaties in een sterkere context een lager waardeoordeel krijgen dan in een minder sterke context (Südkamp & Möller, 2009 in Neumann et al., 2011). Gestandaardiseerde toetsen hanteren vaak een groepsnorm-criterium, maar breiden de vergelijkingsbasis uit naar alle leerlingen in een bepaald land. Daardoor wordt duidelijk gemaakt hoe een leerling presteert in vergelijking met alle andere leeftijdsgenoten, wat een beoordeling van zijn/haar prestaties kan informeren (Bishop & Wössmann, 2004). Het groepsnorm-criterium kan uiteraard ook op mesoniveau worden gehanteerd, waarbij het vaak *benchmarking* wordt genoemd (Kelly, 2001). In deze methode wordt de gemiddelde leerprestatie van leerlingen in scholen vergeleken met de gemiddelde leerprestatie in andere scholen.

Bij een standaard-criterium wordt een score vergeleken met een absoluut doel, bijvoorbeeld (voor leerlingen) een eindterm of ontwikkelingsdoel. Meer dan bij het groepsnorm-criterium, krijgt men hier een duidelijk beeld van wat en hoeveel iemand geleerd heeft, en of dit als voldoende kan worden beschouwd (Sireci, 2005). Dit criterium wordt best gebruikt wanneer een evaluatie moet worden gemaakt over de vraag of een doelstelling van een bepaald proces al dan niet werd bereikt (Neumann et al., 2011).

Bij een zelf-criterium wordt de score van een leerling, klas, school of onderwijssysteem vergeleken met een score in het verleden. Dit criterium lijkt het meest effectief wanneer een motiverend effect wordt beoogd (Neumann et al., 2011). Bij de berekening van 'leerwinst als groei' of 'leerwinst op basis van eerder leerprestaties' wordt per definitie een zelf-criterium gebruikt.

Vaak worden bij de interpretatie van gestandaardiseerde toetsen in een leer(winst)-monitoringssysteem verschillende van deze criteria gecombineerd (McMillan, Myran, & Workman, 2002; Willingham et al., 2002). Bij grootschalige gestandaardiseerde toetsen met een verantwoordingsgericht doel gaat men meestal in twee fasen te werk. Eerst bepaalt men een absoluut doel (standaard) dat bereikt moet zijn om te 'slagen'. Na de toetsafname kan dit doel worden bijgesteld, op basis van de gemiddelde scores op verschillende items (Klein & Van Ackeren, 2012). Daarbij wordt dus een groepsnorm-criterium toegevoegd aan het standaard-criterium.

3 HISTORIEK

De aard en de bestaansredenen van leer(winst)monitoringssystemen en gestandaardiseerde toetsen zijn zo uiteenlopend, waardoor het moeilijk is om een globaal beeld te krijgen van het totale aanbod (Standaert, 2014). In wat volgt, wordt een beknopt chronologisch overzicht gegeven van het ontstaan van gestandaardiseerde toetsen binnen diverse Europese onderwijssystemen. Vervolgens wordt ook geschetst waarom gestandaardiseerde evaluatiemechanismen in Vlaanderen tot op heden beperkt zijn gebleven.

3.1 Historiek in Europees perspectief

In verschillende buitenlandse onderwijssystemen, en in het bijzonder in Angelsaksische landen, is er reeds een lange traditie van gestandaardiseerde toetsen. Zo opperde de *Royal Commission* in 1861 bij het overzien van het onderwijs in Engeland en Wales dat alle leerlingen in lagere scholen aan een mondeling examen zouden worden onderworpen, dat door een inspecteur zou worden afgenomen. Dit examen werd geïmplementeerd om de gelijkheid tussen scholen te borgen met betrekking tot zowel het geboden onderwijs als de evaluatiepraktijken van scholen. Officieel diende dit initiatief om de goede besteding van overheidsmiddelen aan scholen te verzekeren, al is het mogelijk dat ook een politiek motief (met name het uitbouwen van de zaak van publiek onderwijs als tegengewicht voor confessioneel onderwijs) ten grondslag lag aan deze eerste initiatieven rond gestandaardiseerde toetsen (Wiliam, 2010). In Ierland volgde reeds in 1879 een verplicht eindexamen geïmplementeerd om veralgemeende standaarden te verzekeren in het secundair onderwijs (Klein & Van Ackeren, 2012).

Aanvankelijk werden nationale examens ingezet om beslissingen te kunnen nemen over individuele schoolloopbanen van leerlingen (Eurydice, 2009). Zo is IJsland het eerste land dat in 1946 nationale examens introduceerde om beslissingen te nemen over de doorstroom van leerlingen naar het volgende schooljaar. Kort daarna volgden Portugal (1947) en het Verenigd Koninkrijk (1947) waar ook certificering op basis van de resultaten mogelijk werd gemaakt. In de jaren '60 en '70 van de vorige eeuw volgden ook nog onder meer Luxemburg, Nederland, Malta en Denemarken. In deze periode ging het bijna uitsluitend om eindexamens (Eurydice, 2009). In landen met een afsluitend eindexamen hebben scholen een duidelijk idee van het eindniveau per onderwijsvorm- en richting. Dat scholen hierdoor een deel aan autonomie inboeten om zelf inhoud en doelen te bepalen, leek in deze onderwijssystemen niet op te wegen tegen de gepercipieerde voordelen (Fuchs & Wössmann, 2008).

Toch waren er reeds in die periode onderwijssystemen waarbij de gestandaardiseerde toetsen verder werden uitgebreid met achterliggende doelstellingen (Eurydice, 2009). Zweden, Ierland, Frankrijk, het Verenigd Koninkrijk en Hongarije maakten het mogelijk om meer te analyseren dan louter de individuele schoolloopbanen van leerlingen. Zweedse leraren konden in 1962 de resultaten van hun leerlingen vergelijken met de nationale standaard. Ook in Ierland (1972) werden gestandaardiseerde leestoetsen ontwikkeld als antwoord op een maatschappelijk debat omtrent dit topic. Frankrijk implementeerde in 1977 een gestandaardiseerd meetinstrument dat het mogelijk maakte om verbeterpunten van het onderwijssysteem te identificeren. In het Verenigd Koninkrijk bracht men de leeromstandigheden van leerlingen in kaart en kende men op basis van de resultaten extra werkingsmiddelen toe aan bepaalde scholen. Tot slot implementeerde Hongarije in 1986 een

meetinstrument dat het monitoren, evalueren en controleren van de onderwijskwaliteit mogelijk maakte (Eurydice, 2009).

In de 21^e eeuw nam het aantal landen of regio's toe waarin gestandaardiseerde toetsen werden gebruikt. Tegelijk met de internationale tendens naar meer autonomie voor scholen, hebben overheden zich steeds vaker toegelegd op het bepalen van minimumvereisten (*core curriculum of education standards*) waaraan (leerlingen in) scholen dienen te voldoen (Buisson-Fenet & Pons, 2011; Eurydice, 2007; OECD, 2013). Door het vooropstellen van minimumdoelen - een vorm van *outcomes-based governance* (E. D. Klein, 2013) - worden uiteenlopende doelstellingen beoogd, zoals de gelijkwaardigheid tussen scholen (Mathis, 2012), de gerichtheid van het onderwijs op minimale maatschappelijke verwachtingen (Florestal & Cooper, 1997; Schoenfeld, 2002), de geïnformeerde ontwikkeling en verantwoording van het onderwijssysteem (Elmore, 2002), en de contractuele verantwoording van scholen ten aanzien van de overheid (Faubert, 2009; OECD, 2013).

De toetsen die in de 21^e eeuw bijkomend worden ontwikkeld, zijn vaak gebaseerd op deze kerncurricula. Ze zijn tevens vaak van een andere aard dan de traditionele examens aan het eind van een onderwijsfase. Een voorbeeld van dit soort toetsen zijn de grootschalige toetsafnames bij een representatieve groep leerlingen in een steekproef van scholen. Deze toetsen hebben geen summatieve functie op systeemniveau en worden afgenomen om een overzicht te krijgen van de onderwijskwaliteit (op macroniveau). In verschillende landen/regio's, zoals Frankrijk, Estland, Finland, Schotland en Spanje, is deze vorm van gestandaardiseerde toetsen intussen geïntegreerd in het onderwijssysteem. Meer en meer zijn leer(winst)monitoringssystemen ontstaan met toetsen die ook ontwikkeld zijn om scholen te ondersteunen in hun interne kwaliteitszorg en beleidsvoerend vermogen. Daarnaast zijn leer(winst)monitoringssystemen met een meer diagnostische doelstelling, op het niveau van de individuele leerlingen, in gebruik in verschillende Europese onderwijssystemen, zoals Denemarken, Ierland, Frankrijk, Engeland en Schotland. In Cyprus, bijvoorbeeld, bestaat sedert 2007 een landelijke toets die als doel heeft om leerlingen met verhoogde kans op analfabetisme zorg op maat te kunnen aanbieden (Eurydice, 2009). Tegelijkertijd zijn er evenwel onderwijssystemen die ook recent formele nationale examens hebben geïmplementeerd. Zo zijn verschillende deelstaten in Duitsland (als gevolg van de 'PISA-shock' aan het begin van de 21^e eeuw) begonnen met een centraal eindexamen (E. D. Klein, 2013).

Tevens hebben technologische ontwikkelingen de tendens tot gestandaardiseerd toetsen versterkt, maar ook de vorm van toetsen sterk beïnvloed. De grotere beschikbaarheid van computers in scholen leidde niet alleen tot een toename van gestandaardiseerde toetsing, maar ook tot een toename van toetsvormen die geschikt zijn voor de computer (Wang et al., 2006; Zenisky & Sireci, 2002).

Momenteel zijn er slechts een beperkt aantal Europese onderwijssystemen waarin geen enkele vorm van gestandaardiseerde toetsen op grote schaal hebben geïmplementeerd. Voorbeelden hiervan zijn Duitstalig België, Tsjechië, Griekenland en Liechtenstein (Eurydice, 2009).

3.2 Historiek in Vlaanderen

In Vlaanderen is de traditie inzake leer(winst)monitoringssystemen of gestandaardiseerde toetsen relatief klein. Eén van de redenen hiervoor is het grondwettelijke principe van vrijheid van onderwijs.

Deze vrijheid van onderwijs impliceert immers onder meer een pedagogische vrijheid. Dat wil zeggen dat een school recht heeft om eigen waarden, attitudes en overtuigingen bij te brengen. Toch moeten scholen die erkende diploma's willen uitreiken en overheidssubsidies willen ontvangen, voldoen aan wettelijke en statutaire voorschriften (Van Petegem & Imbrecht, 2012). Tegenstanders van gestandaardiseerde toetsen hebben evenwel altijd volgehouden dat die toetsen een inbreuk zouden betekenen op het principe van vrijheid van onderwijs. Dit is een opmerkelijke situatie, want in andere landen waar scholen een vrij grote autonomie toebedeeld krijgen, gaat deze autonomie vaak hand in hand met externe gestandaardiseerde toetsen die aan scholen worden opgelegd (OECD, 2013). Zoals vermeld in de probleemstelling zijn de bezwaren ten aanzien van gestandaardiseerde toetsen evenwel steeds talrijk en gevarieerd geweest. De terughoudendheid ten aanzien van gestandaardiseerde toetsen in het algemeen, en verplichte centrale examens in het bijzonder, stoelt op een lange traditie. In tegenstelling tot de economische benadering van onderwijs, waarin scholen worden beschouwd als instellingen die allemaal dezelfde (meetbare) resultaten dienen te produceren, wordt in Vlaanderen doorgaans de visie gehuldigd dat verschillen tussen scholen niet enkel mogelijk maar ook wenselijk zijn, bijvoorbeeld in het creëren van een rijk onderwijsaanbod (Penninckx et al., 2011; Standaert, 2000).

Als gevolg hiervan is het steeds zo geweest in het Vlaamse onderwijs dat, hoewel de overheid scholen verplicht om hun leerlingen op regelmatige basis te evalueren, nauwelijks eisen worden gesteld aan de leerlingenevaluatie. Leraren in Vlaamse scholen ontwikkelen hun eigen evaluatiemateriaal en nemen deze ook zelf af bij hun leerlingen. De beoordeling van het werk van de leerlingen is in handen van de leraar waardoor de rollen van leraar en evaluator in Vlaanderen volledig samenvallen (Van Petegem & Imbrecht, 2012).

Toch werd met de implementatie van de ontwikkelingsdoelen en eindtermen sinds 1997 een eerste stap richting standaardisatie van de minimale leerresultaten gezet. Bovendien hebben de voornoemde bezwaren tegen gestandaardiseerde toetsen verschillende stakeholders, waaronder beleidsmakers, niet weerhouden om voorstellen uit te werken om leerprestaties en leerwinst in kaart te brengen. Zo schreef een werkgroep binnen de toenmalige Dienst voor Onderwijsontwikkeling (DVO) reeds in 1998 en 1999 twee opeenvolgende nota's waarin een meting van leerprestaties en leerwinst werd voorgesteld met zowel ontwikkelings- als verantwoordingsgerichte doelstellingen (Ver Eecke, 2004). Ook recentelijk laaide de discussie weer sterk op, onder meer door het rapport van de OESO over de evaluatiemechanismen in het Vlaamse onderwijs (Shewbridge et al., 2011).

In tegenstelling tot de beperkte ontplooiing van eigen initiatieven om de leerprestaties van leerlingen te monitoren, heeft Vlaanderen wel een lange traditie van deelname aan internationaal vergelijkend onderwijsonderzoek waarbij deze prestaties worden nagegaan. België (onderwijs was toen nog een federale bevoegdheid) was één van de eerste deelnemers in de First International Mathematics Study (FIMS), die in opdracht van de OESO werd uitgevoerd tussen 1963 en 1969 (Penninckx et al., 2011).

4 DE HUIDIGE SITUATIE: GESTANDAARDISEERDE TOETSEN IN HET VLAAMSE ONDERWIJS

Kwaliteitszorg in het Vlaamse onderwijssysteem wordt traditioneel sterk toevertrouwd aan de individuele scholen. Met het Decreet betreffende de kwaliteit van onderwijs werd dit in 2009 nog

sterker uitgesproken. De Inspectie heeft de taak om de interne kwaliteitszorg te controleren en de pedagogische begeleidingsdiensten ondersteunen scholen bij de ontwikkeling ervan. Meer en meer worden gestandaardiseerde toetsen aan scholen ter beschikking gesteld als bron van informatie bij deze interne kwaliteitszorg op microniveau (individuele leerlingen en leraren) en mesoniveau (scholen). Om de kwaliteit van het onderwijssysteem systematisch na te gaan, zijn sinds 2002 peilingen ontwikkeld (Janssen, 2009). Elk van deze gestandaardiseerde toetsen wordt hieronder besproken. Tabel 3 geeft een overzicht van de onderstaande toetsen en hun doelstellingen.

Tabel 3 - Overzicht van gestandaardiseerde toetsen volgens doelstelling

	Niveau	Ontwikkeling	Verantwoording
MACRO	Onderwijssysteem	Peilingen	
MESO	Scholen	Parelleltoetsen Toetsen van OVSG en Katholiek Onderwijs Vlaanderen Schoolfeedback systeem	
MICRO	Leraren ----- Leerlingen	Leerlingvolgsystemen Oriënteringstoetsen SALTO	----- Centrale Examencommissie Ingangstoets lager onderwijs

4.1 Peilingen en paralleltoetsen

Tot in 2002 was Vlaanderen één van de weinige regio's binnen de Europese Unie die niet over landelijke data beschikte inzake leerlingenprestaties. Hoewel Vlaanderen wel deelnam aan internationaal vergelijkend onderzoek naar de kwaliteit van het onderwijs richtten deze meetinstrumenten zich niet specifiek op het Vlaamse curriculum. Hierdoor kon geen antwoord gegeven worden op de vraag of Vlaamse leerlingen de vooropgestelde minimumdoelen behaalden (Penninckx et al., 2011).

Door jaarlijks grootschalige toetsafnames bij een representatieve steekproef van Vlaamse scholen en leerlingen te organiseren, kan onderzocht worden in welke mate leerlingen bepaalde eindtermen of ontwikkelingsdoelen bereiken. De resultaten van deze peilingen, die telkens afgenomen worden bij ongeveer 1500 leerlingen, kunnen belangrijke parameters opleveren voor het Vlaamse Ministerie van Onderwijs en Vorming om de kwaliteit van het onderwijs na te gaan en te verbeteren (Standaert, 2000).

De peilingen hebben als doel om op een betrouwbare en valide manier zowel sterktes als zwaktes in het Vlaams onderwijs in kaart te brengen. Via IRT-analyse (Itemresponstheorie) worden items op een gemeenschappelijke ladder ingeschaald, zodat vergelijkingen doorheen jaren mogelijk worden gemaakt. Niet alleen wordt nagegaan hoeveel procent leerlingen de eindtermen behalen, maar ook kan de haalbaarheid en wenselijkheid van de eindtermen worden nagegaan (Steunpunt Toetsontwikkeling en Peilingen, 2015). Daarnaast maken peilingen het eveneens mogelijk om te onderzoeken of er systematische verschillen zijn in prestaties tussen scholen, klassen en leerlingen en of deze verschillen samenhangen met bepaalde school-, klas- of leerlingenkenmerken (Standaert, 2014). Er worden immers ook achtergrondkenmerken verzameld bij leerlingen, ouders,

leraren en directies. Deze informatie stelt beleidsmakers in staat om de kwaliteit van het onderwijs gericht te evalueren, en indien nodig te verbeteren. Deelnemende scholen ontvangen ook een individueel, niet publiekelijk feedbackrapport over de gemiddelde prestaties van hun leerlingen (Steunpunt Toetsontwikkeling en Peilingen, 2015).

Om tegemoet te komen aan de vraag van scholen die niet deelnemen aan de peilingen worden evenwaardige parallelversies van de peilingen ontwikkeld. Deze toetsen hebben een niet-verplichtend karakter en worden gratis ter beschikking gesteld aan alle basisscholen en secundaire scholen, die de resultaten ervan dan zelf kunnen interpreteren, rekening houdend met de context van de school (Steunpunt Toetsontwikkeling en Peilingen, 2015). Scholen kunnen dan informatie verzamelen over zowel de prestaties van hun leerlingen op eindtermen en ontwikkelingsdoelen als over de positie van de school in vergelijking met vergelijkbare scholen. De paralleltoetsen zijn dus een instrument voor de interne kwaliteitszorg van scholen (Standaert, 2014; Verhaeghe, 2010).

4.2 Toetsen van Katholiek Onderwijs Vlaanderen en OVSG

Ook de koepelorganisaties Katholiek Onderwijs Vlaanderen en OVSG stellen gestandaardiseerde toetsen ter beschikking aan scholen die de netgebonden leerplannen volgen. Deze toetsen worden jaarlijks vernieuwd en hebben - in tegenstelling tot de peilingen en paralleltoetsen - geen gemeenschappelijke schaal zodat resultaten niet over de jaren heen vergeleken kunnen worden. Bij beide vormen van toetsen ligt het doel voornamelijk op het mesoniveau: scholen bijkomende informatie verschaffen over de kwaliteit van hun onderwijsaanbod door de leerprestaties van leerlingen in kaart te brengen.

Bij het ontwikkelen van interdiocesane proeven baseert het Katholiek Onderwijs Vlaanderen zich op het eigen curriculum. Deze schooloverstijgende toetsen worden afgenomen in het vierde en/of zesde leerjaar. Naast de toetsen voor wiskunde en Nederlands wordt ook het leerplan wereldoriëntatie geëvalueerd. Doordat elk jaar een andere focus gekozen wordt, komen niet alle eindtermen jaarlijks aan bod. De toetsen gaan na in welke mate de leerplandoelen bij de leerlingen zijn gerealiseerd. Het inrichten van interdiocesane proeven verschaft de school in de eerste plaats informatie op schoolniveau, en heeft niet als doel om individuele leerlingen te beoordelen. Het feedbackrapport kan vooral worden aangewend om de interne werking van de school te beoordelen. De schoolresultaten kunnen vergeleken worden met de gemiddelde resultaten van de Vlaamse katholieke basisscholen en kunnen worden gespiegeld aan referentiegroepen op basis van SES-kenmerken (Katholiek Onderwijs Vlaanderen, z.j.; Verhaeghe, 2010).

De OVSG-toets baseert zich inhoudelijk op het curriculum van het zesde leerjaar van het stedelijk en gemeentelijk onderwijs, hoewel ook andere scholen (waaronder GO! basisscholen) van deze proeven gebruik maken. De OVSG-toets toetst alle leergebieden en domeinen van de basisschool op schriftelijke wijze, zoals wiskunde, Nederlands, Frans, wereldoriëntatie en artistieke opvoeding, en richt daarnaast ook zes praktische toetsen in: Nederlandse en Franse spraakvaardigheid, lichamelijke opvoeding, technologie, artistieke opvoeding en verkeersveiligheid. De toets levert niet enkel betrouwbare informatie over de competenties van hun leerlingen, maar brengt ook informatie om de onderwijskwaliteit te bewaken (OVSG, z.j.).

4.3 Leerlingvolgsystemen

Bij een leerlingvolgsysteem worden verschillende toetsen voorzien die op vlak van de getoetste inhoud en moeilijkheidsgraad van elkaar kunnen onderscheiden worden en waarbij de voortgang van individuele leerlingen in kaart kan worden gebracht. De definitie die werd gegeven door het pilootproject Leerwinstmonitoring en Toegevoegde waarde in primair onderwijs in Nederland wordt hier gevolgd:

“Onder een leerlingvolgsysteem wordt een systeem verstaan dat bestaat uit:

- *een serie toetsen die gebaseerd is op longitudinale leerlijnen uitgezet over meerdere aaneensluitende leerjaren;*
- *een registratiesysteem waarin de vorderingen per leerling en per groep kunnen worden vastgelegd;*
- *normen voor de beoordeling van de vorderingen”* (Janssens et al., 2014, p.9).

Een leerlingvolgsysteem is een instrument dat zowel voor de leraar als voor de school nut heeft (Colpin et al., 2006). Enerzijds brengt een leerlingvolgsysteem de leervorderingen en leeropbrengst van de individuele leerling in beeld. De resultaten van deze metingen kunnen worden aangewend om zorg op maat te bieden door zowel de pedagogische aanpak en didactische werkvormen toe te spitsen op de behoefte van de leerling als door het leer- en onderwijsproces van de leerling bij te sturen (Vlaams Ministerie van Onderwijs en Vorming, z.j.).

Anderzijds ondersteunt een leerlingvolgsysteem ook de school zelf in het coördineren van een planmatige aanpak. Een inkijk in de resultaten op schoolniveau geeft schoolleiders en leraren een beeld van de leer- en onderwijsprocessen en onderwijskwaliteit van de eigen school. Een goed leerlingvolgsysteem houdt rekening met de ontwikkelingsdoelen en eindtermen en kan daardoor worden ingezet als hulpinstrument voor een zelfevaluatie. Doordat het systeem de vorderingen van alle leerlingen in kaart brengt, rekening houdend met de achtergrondgegevens van de leerlingen, kunnen scholen nagaan of het geboden onderwijs voor alle leerlingen even effectief is (Vlaams Ministerie van Onderwijs en Vorming, z.j.).

Er zijn verschillende voorbeelden van leerlingvolgsystemen beschikbaar voor de scholen. De website Toetsen voor Scholen van het Vlaamse Ministerie van Onderwijs en Vorming biedt een Leerlingvolgsysteem voor Vlaanderen (LVS) aan dat verschillende toetsen omvat die de kennis van spelling en de vaardigheid technisch lezen registreren. Ook de Vrije Centra voor Leerlingenbegeleiding in Vlaanderen hebben een eigen leerlingvolgsysteem ontwikkeld bestaande uit toetsen en analyse-instrumenten die specifieke problemen helpen achterhalen.

Omdat een leerlingvolgsysteem het doel heeft om vaststellingen te doen over mogelijke vorderingen tussen twee meetmomenten, is het noodzakelijk dat de resultaten die op die verschillende momenten verzameld zijn wel met elkaar vergeleken kunnen worden. Hierdoor is het noodzakelijk om met meetinstrumenten te werken die over de hele periode hetzelfde blijven (Voeten, 1990). Hoewel enkele leerdomeinen zich hiertoe makkelijk lenen, blijkt dit niet voor alle toetsen het geval te zijn. Vele opeenvolgende toetsen zijn immers vaak zowel onvergelijkbaar inzake inhoud, moeilijkheidsgraad als aantal opgaven.

Om aan dit probleem tegemoet te komen, zijn leerlingvolgsystemen vaak gebaseerd op de itemresponsetheorie (IRT) die middels een uni-dimensionale schaal de ontwikkeling van leerlingen

over meerdere leerjaren kan volgen. Deze schaal kan worden voorgesteld als een meetlat met een constante meeteenheid waarop zowel de vaardigheid van de leerlingen als de moeilijkheidsgraad van de opgaven kunnen worden afgezet. Door de resultaten van leerlingen op verschillende meetmomenten te vertalen naar een positie op de schaal wordt hun vooruitgang in kaart gebracht (Moelands, Mommers, & Oud, 1990).

4.4 Examencommissie van de Vlaamse Gemeenschap

Leerlingen die niet in het bezit zijn van een getuigschrift basisonderwijs, getuigschrift eerste en tweede graad of geen diploma secundair onderwijs hebben, kunnen terecht bij de Examencommissie van de Vlaamse Gemeenschap (EVG).

Het getuigschrift basisonderwijs kan behaald worden door alle leerlingen vanaf het jaar waarin de leerling de leeftijd van negen jaar bereikt. In elke Vlaamse provincie wordt minstens één school van het officieel onderwijs en één school van het vrij onderwijs aangeduid door de Vlaamse regering. Deze scholen, waar de kandidaat zelf uit kan kiezen, hebben de taak om de examens zowel te organiseren als op te stellen (Vlaams Ministerie van Onderwijs en Vorming, z.j.). Wel kunnen kandidaten voor een getuigschrift of diploma secundair onderwijs kiezen tussen examens die gebaseerd zijn op het curriculum van de EVG zelf of op het curriculum van een bepaalde school. Anders dan bij de examens voor het getuigschrift basisonderwijs, is de EVG zelf verantwoordelijk voor de organisatie en het opstellen van de examens voor secundair onderwijs (Penninckx et al., 2011).

Om de kwaliteit van de examens te kunnen borgen, worden de toetsitems voor een bepaald domein in het secundair onderwijs binnen een vakgroep ontwikkeld. Afhankelijk van de leerinhoud bestaan er op dit onderwijsniveau mondelinge en schriftelijke examens, maar ook praktijkexamens en assessments kunnen een vorm van evaluatie zijn. Per vak wordt nagegaan in welke mate de leerling de eindtermen of ontwikkelingsdoelen beheerst. Tevens is de Examencommissie een examenvragendatabank aan het uitbouwen, waarop toetsvragen digitaal ter beschikking worden gesteld (Vlaams Ministerie van Onderwijs en Vorming, z.j.).

4.5 Schoolfeedbacksysteem

Naar aanleiding van IWT-Strategisch Basisonderzoek 'Schoolfeedbackproject' van drie Vlaamse universiteiten (KULeuven, Universiteit Antwerpen en Universiteit Gent) werd in 2011 het Centrum voor Schoolfeedback opgericht (Schoolfeedbackproject, 2015).

Deze instelling biedt scholen ondersteuning aan voor hun interne kwaliteitszorg aan de hand van toetsen die scholen kunnen afnemen bij hun leerlingen. Op basis van de resultaten wordt dan een feedbackrapport gegenereerd dat scholen kunnen aanwenden voor een zelfevaluatie van de eigen werking en het verbeteren van de interne kwaliteitszorg.

Scholen krijgen zowel resultaten op schoolniveau als op individueel leerlingniveau te zien. Zo wordt in eerste instantie informatie gegeven over de toegevoegde waarde van een school. Indien mogelijk wordt er ook rekening gehouden met de toegevoegde waarde betreffende diverse leerlingencategorieën (moedertaal, welvaart gezinnen, ...). Scholen die hun leerlingen op meerdere momenten bevragen, krijgen feedback over zowel de gemiddelde leerwinst als individuele

vaardigheidsscores. Doordat ook achtergrondkenmerken van scholen mee worden opgenomen in de analyses kunnen resultaten van de school vergeleken worden met de resultaten van scholen met een vergelijkbaar leerlingenpubliek (Schoolfeedbackproject, 2015).

4.6 Toetsen voor oriëntering van leerlingen

In het basisonderwijs zijn geen gestandaardiseerde oriënteringstoetsen terug te vinden. Wel biedt de website www.onderwijskiezer.be, een overkoepelend initiatief van centra voor leerlingenbegeleiding van het gemeenschapsonderwijs, het vrije, het provinciaal, het stedelijk en het gemeentelijk onderwijs, de belangstellingsproef *I-like* aan.

Een belangstellingsproef is een hulpmiddel om de belangstelling voor bepaalde studiedomeinen in kaart te brengen. De *I-like* belangstellingsproef richt zich voornamelijk naar leerlingen vanaf het zesde leerjaar basisonderwijs tot en met de tweede graad van het secundair onderwijs. De vragen toetsen de mate van interesse in het uitoefenen van beroepen uit verschillende beroepssectoren op een zevenpuntschaal. Nadat alle vragen beantwoord zijn, verschijnt een grafiek met de resultaten per belangstellingsdomein en bijhorende studierichtingen (Onderwijskiezer, 2015).

In het secundair onderwijs bestaat de *I-Prefer* belangstellingsproef voor laatstejaarsstudenten. Dit meetinstrument tracht de interesse voor opleidingen van het hoger onderwijs weer te geven. De *I-Prefer* is een genormeerde proef waardoor de student zijn of haar resultaten kan vergelijken met die van leeftijdsgenoten. Hij houdt evenwel geen rekening met de competenties van de leerlingen.

In het huidige regeerakkoord is sprake van het implementeren van een verplichte niet-bindende oriënteringsproef voor leerlingen aan het einde van het secundair onderwijs. Deze toelatingsproef beoogt een betere ondersteuning van leerlingen bij hun studiekeuze in het hoger onderwijs. Doordat de slaagcijfers van eerstejaarsstudenten redelijk laag zijn, hoopt de Vlaamse regering dat de resultaten van deze proef jongeren bewustmaken voor de gevolgen van een verkeerde studiekeuze (Vlaamse Regering, 2014). In het academiejaar 2016-2017 zou gestart worden met een toelatingsproef in alle STEM-richtingen en in de lerarenopleiding (Oriëntatieproef, 2014).

Een aanzienlijk deel van de informatie die beschikbaar is voor systeemevaluatie wordt verzameld door het vergelijken ('benchmarken') van voornamelijk leerlingenprestaties. Vergelijkende data van internationale studies creëren een raamwerk voor het Vlaams Ministerie van Onderwijs en Vorming, voor verschillende onderwijsactoren en voor de maatschappij om de sterktes, zwaktes, items en kansen voor verdere ontwikkeling te bepalen.

4.7 Deelname aan internationaal vergelijkende studies van leerprestaties

De Vlaamse Gemeenschap neemt daarnaast deel aan verschillende internationale vergelijkende studies die gericht zijn op de meting van leerlingenprestaties: het Programme for International Student Assessment (PISA; meest recente onderzoek in 2015), de Progress in International Reading Literacy Study (PIRLS; 2016 (lopend)), de Trends in International Mathematics and Science Study (TIMSS, 2015), de International Civic and Citizenship in Education Study (ICCS, 2016 (lopend)) en de European Survey on Language Competences (ESLC 2011). PIRLS, TIMSS en ICCS worden georganiseerd door de International Association for the Evaluation of Educational Achievement (IEA). ESLC tenslotte is een onderzoeksproject van de Europese Commissie.

4.8 Andere toetsen

4.8.1 Taalvaardigheidstoetsen

Het Vlaams Ministerie van Onderwijs en Vorming biedt op haar website toetsinstrumenten aan die de taalvaardigheid van leerlingen aan het begin van het lager onderwijs (SALTO-toets) of het secundair onderwijs (TAS, TASAN en TIST) evalueren. Deze toetsen hebben geen summatief karakter, maar geven de leraren de mogelijkheid om de schoolse taalvaardigheid bij instromende leerlingen te beoordelen (Verhelst, 2010). Deze normgerichte screening zorgt ervoor dat leraren de prestatie van hun leerlingen kunnen inschatten en vergelijken met de prestaties van een referentiegroep. Hierdoor worden niet enkel de leerlingen met een taalachterstand geïdentificeerd, maar kan ook duidelijkheid worden geschapen over de omvang van de achterstand door middel van een berekening van de kloof tussen de norm (verwachte prestaties) en de prestatie van de individuele leerling. Op die manier kunnen zowel de individuele begeleiding als de stimuleringsmaatregelen efficiënter gestuurd worden zodat de taalontwikkeling van deze leerlingen kan bevorderd worden.

4.8.2 Aafgeschafte taalproef - instap lagere school

De verplichte taalproef Nederlands die in het schooljaar 2010-2011 zijn intrede deed in het basisonderwijs, had als doel om kinderen die onvoldoende naar school waren gegaan in het kleuteronderwijs, toch de mogelijkheid te geven om in te stromen in het lager onderwijs bij aanvang van de leerplicht. Het onderliggende doel was om ouders te stimuleren hun kinderen ook voor de leerplicht zo vaak mogelijk naar school te sturen. Deze taalproef was verplicht voor leerlingen die minder dan 220 halve dagen aanwezig waren in de derde kleuterklas en die de overstap naar het eerste leerjaar in een Nederlandstalige school wilden maken. De resultaten van deze taalproef wezen uit of de leerlingen over voldoende kennis van de Nederlandse taal beschikken om de doorstart naar het lager onderwijs te kunnen maken. Kinderen die vaak thuisblijven, kampen mogelijk met een taalachterstand wat kan resulteren in een leerachterstand. Tegenstanders van de taalproef waren van mening dat het overdoen van de derde kleuterklas enkel en alleen omwille van taalachterstand de ontwikkeling van andere domeinen in het gedrang brengt. Mede door dit debat werd deze taalproef in het schooljaar 2014 - 2015 afgeschaft.

Momenteel is er wel een taalscreening beschikbaar voor scholen om de beginsituatie van leerlingen in kaart te brengen bij instroom in het eerste leerjaar. Scholen kunnen op basis daarvan beslissen om leerlingen met een grote taalachterstand een remediëringstraject aan te bieden (Vlaams Ministerie van Onderwijs en Vorming, z.j.).

4.8.3 Betrokkenheid en welbevinden van leerlingen

Daarnaast zijn er een aantal meetinstrumenten beschikbaar die de betrokkenheid en het welbevinden van leerlingen in de klas meten. Het Expertisecentrum Ervaringsgericht Onderwijs (ECEGO) ontwikkelde een meetinstrument dat zowel de betrokkenheid als het welbevinden van de leerling situeert op een vijfpuntenschaal op basis van observatie door de leraar (Laevers, Bertrands, Declercq, & Daems, 2004). Een alternatief instrument om het welbevinden van leerlingen te meten is beschikbaar op de website van de onderwijsinspectie (De Volder & De Lee, 2009).

4.8.4 Prodiagnostiek

De website www.prodiagnostiek.be is een initiatief van de onderwijskoepels en CLB-centrumnetten met de steun van de Vlaamse overheid. Het project heeft als uitgangspunt het realiseren van gelijkgerichte, kwaliteitsvolle en wetenschappelijk verantwoorde diagnostiek in leerlingenbegeleiding. Protocollen vormen de leidraad hoe men diagnostische praktijken moet structureren en welke meetinstrumenten moeten gebruikt worden. Ook de interpretatie- en beslissingscriteria werden op voorhand vastgelegd.

4.8.5 Toetsen van educatieve uitgevers

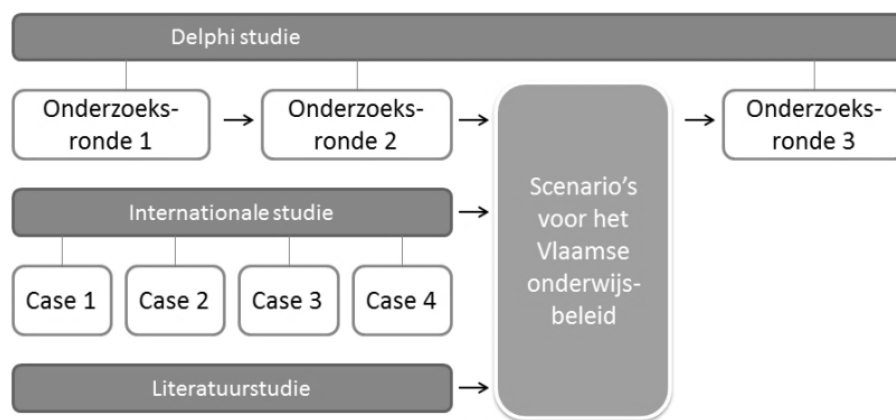
Tal van uitgeverijen springen op de kar van gestandaardiseerde toetsen. Zo beschikt Uitgeverij Van In over twee digitale leerplatformen 'Bingel' voor leerlingen uit het basisonderwijs en 'Diddit' voor leerlingen uit het secundair onderwijs. Beide systemen richten zich op differentiatie en worden aangestuurd door middel van een digitale verbeteringstool die het niveau en de prestaties van de leerlingen evalueert. Onder- en bovenpresteerders krijgen automatisch oefeningen met de juiste moeilijkheidsgraad.

Uitgeverij Abimo is in Vlaanderen vooral gekend om zijn publicaties inzake de technische leesontwikkelingen bij leerlingen. Met behulp van genormeerde toetsen 'Technisch Lezen' kan de leraar het technisch leesniveau van zijn of haar leerlingen bepalen. Wie nood heeft aan meer diagnostisch informatie kan leerlingen toetsen met behulp van de 'Drie-Minuten-Toets' en de 'AVI-leeskaarten'. AVI staat voor 'Analyse van Individualiseringsvormen'. Daarna kan gestart worden met de leesboeken die het juiste AVI-niveau hanteren.

DEEL 2 ONDERZOEKSAANPAK

1 ONDERZOEKSDESIGN

Vanuit methodologisch oogpunt omvat deze studie meerdere deelstudies: een literatuuronderzoek, een internationale studie en een Delphi-studie (zie Figuur 1). Elke deelstudie draagt op een unieke manier bij tot de resultaten van dit onderzoek. Als gevolg hiervan wordt over elk van de studies individueel gerapporteerd in de resultatensectie.



Figuur 1 - Onderzoeksdiseign

Het literatuuronderzoek had als voornaamste doel om relevante nationale en internationale opiniestukken en empirische evidentie in het kader van gestandaardiseerde toetsen en leerwinstmeting, op systematische wijze bijeen te brengen. Daarnaast verleenden de internationale casestudies zich ertoe om informatie in te winnen over de kwaliteiten en valkuilen van buitenlandse monitoringssystemen. De beschrijving is gebaseerd op zowel schriftelijke bronnen als op gesprekken ter plaatse met betrokken actoren uit het onderwijsveld. Tot slot zorgden een eerste en tweede ronde van de Delphi-studie ervoor dat ook Vlaamse onderwijsexperten informatie en standpunten betreffende gestandaardiseerde toetsen konden meedelen.

Vanuit deze drie luiken werd vervolgens toegewerkt naar acht werksenario's voor het Vlaamse onderwijsbeleid. De wenselijkheid en de haalbaarheid van elk van deze scenario's werd nadien afgetoetst aan de opinies van de personen in het expertenpanel in een derde onderzoeksrunde van de Delphi-studie.

In de volgende secties wordt ingegaan op de gehanteerde methodologie in elk van de verschillende deelstudies.

2 LITERATUURONDERZOEK

Door middel van een systematisch literatuuronderzoek werd een inventaris gemaakt van de bestaande (inter)nationale kennisbasis omtrent effecten, processen en randvoorwaarden van leerwinst-monitoringsystemen. In tegenstelling tot bij een systematische review studie, werd zowel empirische als non-empirische (opiniërende en feitelijke) literatuur weerhouden met betrekking tot de thema's die centraal staan in dit rapport. Dit gebeurde op een zodanige manier dat de inbreng van beide soorten bronnen in de analyse gescheiden kon blijven. Er werd evenwel voorkeur gegeven aan *peer-reviewed* tekstmateriaal.

Het literatuuronderzoek startte in oktober 2014 met een zoektocht in de databases ERIC en Web of Science (specifiek het onderzoeksdomein *education - educational research*). Daarbij werd de zoekterm *effects* gecombineerd met de volgende zoektermen: *standardized tests*, *national tests*, *added value*, *learning progress* en *learning gain*. Deze oefening werd herhaald in oktober 2015 om recente publicaties op te sporen. De resultaten (in termen van aantal publicaties) zijn gerapporteerd in Tabel 4.

Tabel 4 - Aantal publicaties per zoekterm en database in oktober 2015

Zoektermen	Database	Totaal	Peer-reviewed	Voor 2000	2000 - 2009	Sinds 2010
"Standardized tests"+ effects	ERIC	1178	740	175	266	299
	Web of Science	107	107	13	41	53
"national tests" + effects	ERIC	29	13	1	6	6
	Web of Science	10	10	1	3	6
"added value" + effects	ERIC	49	39	1	17	21
	Web of Science	55	55	1	23	31
"learning progress" + effects	ERIC	22	5	0	0	5
	Web of Science	18	18	0	5	13
"learning gain" + effects	ERIC	18	16	0	6	10
	Web of Science	20	20	2	5	13

Door het grote aantal publicaties rond de (vermeende) effecten en neveneffecten van gestandaardiseerde toetsen diende een keuze te worden gemaakt inzake de te behandelen literatuur. Daarbij werd in de eerste plaats uitgegaan van het abstract van de artikels. Publicaties met de hoogste relevantie qua inhoud werden geselecteerd vanaf publicatiejaar 1995; bij een vermoede 'hoge relevantie' werd de tekst geselecteerd vanaf publicatiejaar 2005; bij een 'medium relevantie' vanaf 2014. Enkele meta-analyses van empirische studies gaven bovendien een snel overzicht van bestaande literatuur. Daarnaast werd aan themanummers van academische tijdschriften inzake gestandaardiseerde toetsen, leerwinst en toegevoegde waarde prioriteit toegekend. Ten slotte werd ook de datum van publicatie als een criterium vooropgesteld om de prioriteit van een publicatie te bepalen. Na de lectuur van de geselecteerde publicaties werd, op basis van de tekst, een hoge of lage prioriteit toegekend aan de literatuur in de referentielijsten. De peer-reviewed materialen gaven daarbij toegang tot een ruim arsenaal van boeken, boekhoofdstukken, academische tijdschriften, beleids- en onderzoeksrapporten, en papers van congrespresentaties. Zoals algemeen wordt aanbevolen in het veld van sociale wetenschappen, blijft het literatuuronderzoek niet enkel beperkt tot peer-reviewed studies (Grayson & Gomersall, 2003). Het omvattend karakter van de zoektocht maakt dat bias in de onderzoeksresultaten op een adequate wijze werd vermeden (Glass, McGaw, & Smith, 1981).

3 INTERNATIONALE CASESTUDIES

In het internationale onderzoeksluik werd nagegaan of en op welke manier ervaringen met vier buitenlandse leer(winst)monitoringssystemen onderbouwend kunnen zijn en inspirerend kunnen werken voor Vlaanderen. De beschrijving van de verschillende leer(winst)monitoringssystemen vertrok allereerst vanuit een analyse van geschreven academische bronnen en werd daarna aangevuld met aan het systeem gelieerde praktijk- en beleidsgeoriënteerde literatuur. Vervolgens werden ook de resultaten verwerkt van de interviews die ter plaatse werden afgenomen met een aantal bevoorrechte getuigen (beleidsmakers, ontwikkelaars en gebruikers).

Een leer(winst)monitoringssysteem kan bestaan uit één enkele toets, maar zal vaker berusten op een combinatie van meerdere toetsen of toetssystemen, mogelijks met uiteenlopende doelen. Een leer(winst)monitoringssysteem omvat evenwel niet noodzakelijk alle gestandaardiseerde toetsen binnen een bepaald onderwijssysteem. Het was van cruciaal belang om een goede selectie te maken uit het rijk palet aan buitenlandse monitoringssystemen, waarbij de combinatie het toelaat om de voorliggende thema's vanuit verschillende invalshoeken te benaderen.

Bij de selectie werd in twee fasen gewerkt. In een eerste fase werd op basis van beschikbare informatie via de OESO (met name de Country Background Reports (CBR) en OECD Review on Evaluation and Assessment Frameworks) en via een rapport in opdracht van de Europese Commissie (Eurydice, 2009) een eerste screening gemaakt van mogelijk interessante leer(winst)-monitoringssystemen. Tabel 5 biedt een overzicht van de 39 onderwijssystemen waarvan op deze manier informatie werd verzameld.

Tabel 5 - Overzicht van informatiebronnen per onderwijssysteem

Informatiebronnen	Aantal	Landen
Enkel OECD CBR	1	Zuid-Korea
OECD CBR + OECD Review	4	Australië Chili Mexico Nieuw-Zeeland
Enkel Eurydice	20	België (Duitstalige Gemeenschap) Bulgarije Cyprus Duitsland Engeland Estland Finland Griekenland IJsland Italië Letland Liechtenstein Litouwen Malta Oostenrijk Polen Roemenië Schotland Spanje Wales
OECD CBR + Eurydice	6	België (Franstalige Gemeenschap) Frankrijk Hongarije Ierland Slovenië Tsjechië
OECD CBR + OECD Review + Eurydice	8	Denemarken Luxemburg Nederland Noord-Ierland Noorwegen Portugal Slovakije Zweden

Een moeilijkheid bij de lectuur en analyse van deze rapporten, bleek het non-conforme gebruik van termen als *'standardized tests'* en voornamelijk *'value added'* in de beleids- en praktijkliteratuur. Deze termen krijgen in verschillende rapporten zeer uiteenlopende betekenissen en worden zelfs binnen rapporten vaak weinig consistent gebruikt, waardoor het proces bemoeilijkt werd om een

precieze inschatting te maken van doelstellingen en operationalisering in de verschillende onderwijscontexten. Deze onduidelijkheid noopte het onderzoeksteam om een ruime selectie van 15 leer(winst)monitoringssystemen te weerhouden na de eerste screening. Het betroffen monitoringssystemen in Australië, Denemarken, Duitsland, Engeland, Finland, Frankrijk, Ierland, IJsland, Nederland, Noord-Ierland, Noorwegen, Slovenië, Verenigde Staten, Wales, en Zweden. Aan deze selectie werd ook Canada toegevoegd, aangezien dit onderwijssysteem ook een traditie kent inzake gestandaardiseerde toetsen, maar waar noch via Eurydice, noch via de betreffende OESO-studie informatie over kon worden verzameld in de eerste screening.

Van elk van de overblijvende 16 leer(winst)monitoringssystemen werd in een tweede fase via verschillende kanalen op zoek gegaan naar bijkomende informatie om een zo duidelijk mogelijk beeld te krijgen van de huidige praktijk en recente ontwikkelingen. Deze zoektocht naar bijkomende informatie werd geïnitieerd met behulp van openbaar beschikbare beleidsdocumenten en gericht onderzoek in academische literatuur. Indien nodig, werden sleutelpersonen binnen het monitoringssysteem ingeschakeld om ongepubliceerde literatuur te verkrijgen of om onduidelijke punten uit te klaren.

Daarnaast werden criteria omschreven waaraan de selectie van vier leer(winst)-monitoringssystemen diende te voldoen. Elk van de leer(winst)monitoringssystemen moest uiteraard gestandaardiseerde toetsen aanbieden en moest gericht te zijn op het leerplichtonderwijs. Op basis van de in deze fase verzamelde data moest bovendien een redelijk adequaat beeld worden gevormd van het leer(winst)monitoringssysteem en de verschillende gestandaardiseerde toetsen, inclusief van de initiatiefnemers, de doelstellingen, de inzet van de toetsen, de getoetste leerdomeinen, en van de wijze van toetsafname.

Voor de combinatie van monitoringssystemen werden volgende criteria gehanteerd:

- elk van de leerdomeinen rekenen, (moeder)taal en wetenschappen worden door minimaal drie monitoringssystemen getoetst;
- zowel de overheid als niet-overheidspartners (bijvoorbeeld dienstenbedrijven) zijn bij minimaal twee monitoringssystemen de initiatiefnemers;
- zowel formatieve doeleinden als summatieve doeleinden worden door ten minste twee monitoringssystemen nagestreefd;
- doelen op macro-, meso- en microniveau worden elk door minimaal twee monitoringssystemen nagestreefd;
- zowel verplichte toetsen als facultatieve toetsen worden door minstens twee monitoringssystemen gebruikt;
- zowel leerwinst als toegevoegde waarde worden door ten minste drie monitoringssystemen aangeboden;
- de selectie bevat maximaal twee onderwijssystemen van een zelfde 'type' van onderwijscontext (bijvoorbeeld Angelsaksisch, Scandinavisch).

Bovendien werd bepaald dat, indien bovenstaande criteria tot meerdere combinatiemogelijkheden zouden leiden, voorkeur gegeven wordt aan leer(winst)monitoringssystemen waarbij taalbarrières de dataverzameling (zowel voor document-analyse als interviews met respondenten) zo min mogelijk zou hinderen.

Op basis van de informatie die de hierboven omschreven methode opleverde, en mede na advies van de stuurgroep, werd besloten dat de volgende selectie de meeste kansen op een maximaal informatie-effect voor de Vlaamse onderwijscontext biedt: Australië, Engeland (met focus op het Centre of Evaluation and Monitoring), Nederland (met focus op het pilootproject 'Leerwinstmonitoring en toegevoegde waarde in het primair onderwijs') en Noorwegen (met focus op de stad Oslo).

In elk van de cases werden ter plaatse diepte-interviews afgenomen van minimaal acht en maximaal veertien respondenten. Elk van de interviews werd opgenomen om het analyseren te faciliteren. Net als bij de tweede en derde ronde van de Delphi-studie (zie 4.2), werden ook in deze analyses de principes van de kaderanalyse toegepast (Miles & Huberman, 1994; Ritchie & Spencer, 1993), waarbij in de interviews zowel vanuit de theorie als vanuit de inbreng van de respondenten gezocht wordt naar patronen. Hieronder wordt de selectie van elk van de vier monitoringssystemen gemotiveerd. In Deel 5 van dit werk zullen de resultaten inhoudelijk worden toegelicht.

3.1 Australië

De case 'Australië' is interessant omwille van de lange traditie in het monitoren van leerprestaties. Uit de literatuur blijkt bovendien dat het gebruik van gestandaardiseerde toetsen ingebed is in een breed debat tussen uiteenlopende stakeholders.

Het geheel van toetsen die worden aangeboden door de federale overheid, verschillende staten en private aanbieders, vormt een breed scala aan mogelijkheden met verantwoordingsgerichte en ontwikkelingsgerichte doeleinden op zowel microniveau als meso- en macroniveau. Er zijn zowel verplichte als facultatieve toetsen. Bovendien is er een lange traditie inzake de berekening van zowel leerwinst als toegevoegde waarde.

Bij het in kaart brengen van het Australische monitoringssysteem lag de focus op een selecte steekproef die de verschillende onderdelen van dit brede aanbod goed weergeeft, met name het geheel van (verplichte) toetsen op federaal niveau, de verplichte eindexamens op staatsniveau (focus op New South Wales omwille van de sterke traditie van deze staat inzake verantwoordingsgerichte high-stakes toetsing) en de vrijwillige toetsen op staatsniveau (focus op Victoria wegens het ruime aanbod aan On Demand Testing, één van de vroegste computer-adaptieve systemen). Ten slotte werden ook de gestandaardiseerde toetsen van de twee grootste private toetsontwikkelaars bestudeerd. De respondenten werden geselecteerd op basis van (1) hun bereidwilligheid om mee te werken, (2) een contextanalyse en (3) hun functie/deskundigheid betreffende het thema. Aanvullend werden ook twee schoolleiders geselecteerd om deel te nemen aan het onderzoek.

3.2 Engeland (Verenigd Koninkrijk) - Centre of Evaluation and Monitoring (CEM)

Deze case, waarbij gefocust wordt op het *Centre of Evaluation and Monitoring* (CEM), dat verbonden is aan de Universiteit van Durham, wordt geselecteerd omwille van de uitgebreide ervaring van dit centrum inzake gestandaardiseerde toetsen. Bovendien heeft de instelling een ruime ervaring en deskundigheid opgebouwd inzake leerwinstmonitoring van leerlingen. De toetsen van het

monitoringssysteem dat door CEM ontwikkeld en voortdurend verbeterd wordt, worden op het ogenblik van de data-verzameling in 70 verschillende landen gebruikt.

De ontwikkeling van de toetsen in CEM staat niet los van de Engelse context. De onderwijscontext bepaalt immers mede het succes van toetsen in het standaard-aanbod van CEM die - in tegenstelling tot de verplichte verantwoordingsgerichte toetsen - sterk ontwikkelingsgericht zijn. Ten slotte investeert CEM ook in het aanbieden van nascholing aan scholen die de gestandaardiseerde toetsen gebruiken.

Om deze case gedetailleerd in kaart te brengen, werd ervoor geopteerd om verschillende respondenten binnen CEM met uiteenlopende functies in het onderzoek te betrekken. Daarnaast werd door middel van een schoolbezoek een licht geworpen op het gebruik van de toetsen van CEM binnen de context van het Engelse onderwijssysteem.

3.3 Nederland - pilootproject Leerwinstmonitoring en Toegevoegde Waarde

De keuze voor Nederland als casestudie is niet enkel gebaseerd op de nabijheid van deze case en de lange traditie van het Nederlandse onderwijs inzake gestandaardiseerde toetsen, maar steunt vooral op de vaststelling dat zeer recent een pilootproject inzake leerwinstmonitoring en toegevoegde waarde in het primair onderwijs werd afgerond in opdracht van het Ministerie van Onderwijs, Cultuur en Wetenschap en de Onderwijsinspectie. Binnen dit pilootproject worden zowel leerwinst van leerlingen als toegevoegde waarde van scholen bestudeerd en in kaart gebracht.

In deze case komen zowel ontwikkelingsgerichte als verantwoordingsgerichte doelen bij elkaar en zijn deze vaak nauw vervlochten. Bovendien is het toetsontwikkelingsinstituut CITO, nauw betrokken bij het pilootproject, een internationaal bekende naam inzake de ontwikkeling van gestandaardiseerde toetsen.

Binnen deze case werd ervoor gekozen om respondenten te selecteren bij de verschillende betrokkenen bij het pilootproject Leerwinstmonitoring. Dit houdt in dat een bezoek werd gebracht aan de Onderwijsinspectie, aan de Universiteit Twente, aan CITO en aan verschillende scholen die aan het pilootproject hebben deelgenomen.

3.4 Noorwegen - met focus op Oslo

De case Noorwegen wordt geselecteerd omwille van de relatief recente ontwikkelingen van het onderwijssysteem inzake gestandaardiseerde toetsen, waarmee het een aanvulling vormt ten opzichte van de reeds langere tradities in de andere casestudies.

Binnen de case werd ervoor gekozen om de focus te richten op een combinatie van overheidsinitiatieven op verschillende niveaus, met name het centrale (nationale) niveau, en het lokale (stedelijke/provinciale) niveau. Voor dit laatste viel de keuze op Oslo, omdat uit de verzamelde informatie blijkt dat in de hoofdstad veel meer beroep wordt gedaan op gestandaardiseerde toetsen dan in andere steden of provincies. Bovendien stelt deze case zich ook voor als een voorbeeld van leerwinstmonitoring en lijken de digitale toetsen met grote flexibiliteit in rapportering een interessante inspiratiebron voor Vlaanderen.

4 DELPHI-STUDIE

4.1 Definitie en doeleinden van een Delphi-studie

De opinies van verschillende actoren in het Vlaamse onderwijs omtrent de wenselijkheid en haalbaarheid en mogelijkheden van gestandaardiseerde toetsen, leerwinstmonitoring en toegevoegde waarde, waren bij aanvang van deze studie nauwelijks op een systematische manier onderzocht. Deze opinies identificeren en tot één geheel kneden, is een werk van exploratieve aard. Omwille van dit exploratieve karakter werd gekozen voor een Delphi-studie. In deze methode worden middels een feedback-techniek doorheen verschillende onderzoeksrondes de standpunten van experts eerst verkend om ze vervolgens met elkaar te confronteren (Dalkey & Helmer, 1963; Day & Bobeva, 2005; Okoli & Pawlowski, 2004). De Delphi-methode wordt gebruikt om een proces van groepscommunicatie zo te structureren dat een groep van individuele experts zich als geheel over een complexe vraag kunnen buigen (Day & Bobeva, 2005; Linstone & Turoff, 1975). Het gaat bij een Delphi-methode steeds om het verzamelen van opinies eerder dan objectieve feiten (Schmidt, 1997). Het doel is tweeledig: enerzijds beoogt men een onderbouwde en creatieve verkenning van diverse standpunten, terwijl men anderzijds gepaste informatie met het oog op beleidskeuzes wil produceren (Adler & Ziglio, 1996). In de meeste toepassingen van de Delphi-methodiek gaat het voornamelijk om kwalitatief onderzoek (Linstone & Turoff, 1975), zoals ook in deze studie het geval was. Daarbij stonden de verwachtingen en bekommernissen inzake functies/doelen, haalbaarheid en wenselijkheid van gestandaardiseerde toetsen, en monitoring van leerwinst en toegevoegde waarde in de context van het Vlaamse leerplichtonderwijs centraal.

Eén van de meest kritische vereisten voor de validiteit van de resultaten van een Delphi-studie is de selectie van geschikte experts in het expertenpanel (Chong, Adnan, & Zin, 2012; Custer, Scarcella, & Stewart, 1999; Okoli & Pawlowski, 2004). Elk van de geselecteerde stakeholders dient niet enkel te beschikken over substantiële kennis over leer(winst)monitoring, maar dient ook gemotiveerd te zijn om zich doorheen het volledige proces te engageren voor deelname aan de Delphi-studie. Tot slot is het aangewezen dat hij of zij in staat is om op autonome basis opinies te vormen en te expliciteren (Day & Bobeva, 2005). Het is daarbij aangewezen om niet enkel experts op beleidsniveau te selecteren, maar om bij het samenstellen van het expertenpanel rekening te houden met personen die zelf rechtstreeks in aanraking komen met het onderwerp van de studie, zoals mogelijke gebruikers (leraren, directies) en mogelijke ‘*evaluees*’ (in dit geval de leerlingen die getoetst zullen worden) (Hsu & Sandford, 2007). Rekening houdend met de functie en de expertise, en in overleg met de stuurgroep, werden vier categorieën van stakeholders bepaald (zie Tabel 6).

Tabel 6 - Verdeling van het expertenpanel over negen categorieën

Categorie	Aantal experts
Beleidsmakers	4
Onderwijsverstrekkers	6
Scholen	8
Inhoudsdeskundigen	6
Totaal	24

De categorie ‘beleidsmakers’ bevat ook een respondent uit de onderwijsinspectie. De categorie ‘onderwijsverstrekkers’ bestaat uit respondenten van de koepelorganisaties en onderwijsnetten,

inclusief pedagogische begeleidingsdiensten. De categorie 'scholen' bestaat uit directies, leraren en een vertegenwoordiger van leerlingen, waarbij respondenten evenredig zijn verdeeld over basis- en secundair onderwijs. De categorie 'inhoudsdeskundigen' bestaat uit respondenten met expertise in de ontwikkeling van gestandaardiseerde toetsen en academici met een belangrijke stem in dit veld.

In een volgende stap werden deze categorieën geconcretiseerd naar 24 personen die op basis van hun expertise en beschikbaarheid het meest geschikt werden geacht voor deelname aan deze Delphi-studie. Verder werd ook gestreefd naar een evenwicht inzake onderwijsnetten (gemeenschapsonderwijs, officieel gesubsidieerd onderwijs, vrij onderwijs - met aandacht voor inclusie van minimaal één expert gelinkt aan Freinet- of Steineronderwijs), onderwijsniveaus (lager en secundair onderwijs) en gender van de respondenten. Waar mogelijk, werd een vervangend persoon geselecteerd indien de eerste persoon zich niet kon engageren tot deelname aan dit onderzoek.

Hoewel elke expert in het panel deel uitmaakt van één van bovenstaande categorieën, en in vele gevallen ook van een belangenorganisatie in het onderwijs, wordt nadrukkelijk gesteld dat ieder vanuit zijn/haar persoonlijke expertise wordt aangesproken. Het is immers niet de bedoeling van de Delphi-studie om de officiële standpunten van belangenorganisaties te verzamelen, maar wel om in kaart te brengen wat leeft in het gedachtengoed van experts die beroepshalve met deze materie in aanraking komen.

Eén van de belangrijkste voordelen van de Delphi-studie is dat er geen rechtstreekse confrontatie ontstaat tussen de verschillende experts. Hierdoor wordt de zelfstandige reflectie door de experts versterkt (Dalkey & Helmer, 1963; Okoli & Pawlowski, 2004) en wordt vermeden dat experts zich te gemakkelijk aansluiten bij een eerder ingenomen standpunt (Dalkey & Helmer, 1963). Ook ervaren de deelnemers geen druk om zich te conformeren naar gezichtspunten van andere experts, wat vooral het geval kan zijn wanneer er een (impliciete of expliciete) vorm van hiërarchie tussen de experts bestaat (Van de Ven & Delbecq, 1974). Om deze voordelen te vrijwaren, wordt de anonimiteit van de deelnemers aan de Delphi-studie doorheen en na afloop van het onderzoek behouden, zowel ten aanzien van elkaar als ten aanzien van de opdrachtgever (Dalkey & Helmer, 1963).

4.2 Verloop van de Delphi-studie in dit onderzoek

Concreet werd in deze studie gekozen om de Delphi-methode te operationaliseren middels drie onderzoeksrondes. Na definitieve bevestiging van alle leden van het expertenpanel werd zo snel als mogelijk aan elk van de 24 respondenten een verkennende open vragenlijst bezorgd. Van de uitgestuurde vragenlijsten in de eerste ronde van de Delphi-studie werden 22 ingevulde vragenlijsten terugbezorgd aan het onderzoeksteam. De data van de eerste ronde werden geanalyseerd op deductieve wijze volgens de principes van Grounded Theory: er werd bij de codering en analyse niet uitgegaan van vooraf bepaalde categorieën (Strauss & Corbin, 1990). In een eerste fase werden bij elk van de data codes aangemaakt. In een tweede fase werden de codes met elkaar vergeleken en werden ze, waar mogelijk, samengevoegd. Daarna werden de data gehercodeerd naar de resulterende codes (Gibbs, 2002). De eerste onderzoeksronde resulteerde in een op data 'gegronde theorie' omtrent effecten en neveneffecten van het monitoren van

leerprestaties en leerwinst van leerlingen en de toegevoegde waarde van scholen, alsook omtrent de conceptualisering van de experten van deze begrippen.

In de tweede onderzoeksrunde werden de standpunten van elk van de respondenten verder verkend en verfijnd middels kwalitatieve diepte-interviews. Tevens werd de houding ten aanzien van stellingnames van andere experten afgetoetst. Op die manier werd de 'gegronde theorie' uit de eerste onderzoeksrunde verder uitgediept en werden belangrijke principes en twistpunten geïdentificeerd. Daarnaast werd in elk van de diepte-interviews toegewerkt naar een ideaal scenario inzake gestandaardiseerde toetsen, leerwinst en toegevoegde waarde voor het Vlaamse leerplichtonderwijs, volgens de opinies van de betreffende respondent. De interviews in de tweede ronde van de Delphi-studie namen tussen 53 minuten en 118 minuten in beslag.

In een volgende stap werden de 24 scenario's geanalyseerd en werden gemeenschappelijke kenmerken en twistpunten in kaart gebracht, die de bouwstenen van acht verdere werkscenario's vormen. Op basis van het literatuuronderzoek en de resultaten van de internationale casestudies werden deze bouwstenen verder afgebakend tot er uiteindelijk acht volwaardige werkscenario's ontwikkeld werden. Tot slot werden ook mogelijke implicaties per werkscenario in kaart gebracht (zie sectie 5 over scenario-ontwikkeling).

In de derde onderzoeksrunde van de Delphi-studie werden de acht werkscenario's ter beoordeling aan de 24 leden van het expertenpanel voorgelegd. De respondenten ontvingen deze werkscenario's enkele weken voor het eigenlijke interview, zodat iedere respondent de tijd had om zich inhoudelijk voor te bereiden. Daarnaast werd een schriftelijke vragenlijst bijgevoegd, met het verzoek om deze vragenlijst ingevuld te retourneren enkele dagen voorafgaand aan het gesprek. Deze vragenlijst peilde op een kwantitatieve wijze naar de wenselijkheid en haalbaarheid van de diverse werkscenario's. Er werd geopteerd voor een Likertschaal met vijf antwoordmogelijkheden gaande van (1) 'Helemaal oneens' tot (5) 'Helemaal eens'. Tijdens het diepte-interview kregen de respondenten de kans om hun standpunten verder te verduidelijken en de gegeven scores te motiveren.

Omwille van langdurige ziekte konden twee respondenten uit het expertenpanel niet deelnemen aan de derde onderzoeksrunde. In beide gevallen van overmacht werd in samenspraak met de betreffende respondenten een vervangende respondent uit dezelfde organisatie aan het expertenpanel toegevoegd, zodat ook in deze derde ronde gegevens werden verzameld bij 24 respondenten. Eén van beide originele respondenten kon de vragen op voorhand samen bespreken met de vervangende respondent; in het ander geval was dit niet mogelijk. Bij één andere respondent was het in de derde onderzoeksrunde enkel mogelijk om een telefonisch interview af te nemen. De interviews in de derde ronde van de Delphi-studie namen tussen 50 minuten en 117 minuten in beslag.

Alle interviews van de tweede en derde onderzoeksrunde werden opgenomen en ad verbatim getranscribeerd met het oog op analyse. De gehanteerde analysetechniek verschilt van die in de eerste onderzoeksrunde, aangezien bij de verwerking van de diepte-interviews een eerder deductieve wijze wordt gehanteerd volgens de principes van een kaderanalyse (Miles & Huberman, 1994; Ritchie & Spencer, 1993). Het voorliggende kader bij de analyse van de data van de tweede en derde onderzoeksrunde werd gevormd door respectievelijk de resultaten van de eerste

onderzoeksrunde, en de voorliggende werkscenario's. Bij aanvang van de studie waren uiteraard geen van beide kaders reeds gekend.

5 SCENARIO-ONTWIKKELING

5.1 Definitie en doeleinden van scenario's

In een steeds complexer wordende samenleving neemt de invloed van maatschappelijke, economische, politieke en ecologische aspecten steeds meer toe. Ook in onderwijskundige context zijn deze invloeden voelbaar en bepalen ze in sterke mate de klijtlijnen van hoe veranderingsprocessen gepland en gestuurd moeten worden (Chermack, 2005). De onzekerheid betreffende toekomstige ontwikkelingen neemt toe en wordt versterkt door gebeurtenissen op vlak van schaalvergroting, kennisverruiming en de steeds groter wordende autonomie van scholen (Van Petegem, Devos, Mahieu, Dang Kim, & Warmoes, 2006). Bovendien staan de ontwikkelingen in het onderwijsveld niet los van evoluties in andere sectoren.

Zoals eerder vermeld in de probleemstelling is het debat omtrent het monitoren van leerresultaten en/of leerwinst in Vlaanderen niet nieuw, en evenmin afgesloten. Om het draagvlak van en de kennis over deze initiatieven te vergroten, is het aangewezen om de mogelijkheden van leer(winst)monitoring in de Vlaamse onderwijswereld te onderzoeken. Om tijdig te kunnen anticiperen op mogelijke gebeurtenissen in de toekomst is het de taak van de overheid om de toekomst zoveel mogelijk richting te geven in plaats van ze te ondergaan (Van Vliet, Kok, & Veldkamp, 2010). Prospecties zijn in dat opzicht essentieel om vooruit te kunnen. Scenarioanalyse is een onderzoeksmethode die zich ertoe verleent om een aantal mogelijke toekomstige situaties te verkennen.

Hopkins en Zapata (2009, p.9) omschrijven scenario's als "*stories about how the world changes and how it will be changing at some future time*". Hoewel vaak verondersteld wordt dat toekomstonderzoek moet leiden tot het voorspellen van de toekomst en het kunnen nemen van eenduidige beslissingen, benadrukken ze dat scenario's 'verhalen' zijn die uitgedrukt worden in aspecten van mogelijke of wenselijke situaties. Het doel van een scenario is om een beeld te krijgen van deze alternatieve toekomsten. Het ontwikkelen van scenario's kan in dat opzicht dan ook beschouwd worden als een hulpmiddel ter ondersteuning van strategische besluitvormingsprocessen over mogelijke evoluties. Als onderzoeksteam nemen wij geen a priori positie in ten opzichte van deze scenario's. Elk scenario heeft in principe evenveel kans om de werkelijke ontwikkeling in de toekomst te beschrijven.

In dit onderzoek betreffende leer(winst)monitoring in de Vlaamse onderwijscontext worden verschillende stakeholders aangezet om te reflecteren over de toekomst en om buiten de lijnen van de bestaande kaders te gaan denken (Block, Goeminne, Paredis, & Crivits, 2010). Een Delphi-studie helpt hier om de input op een gestructureerde wijze te verzamelen. Samen met een analyse van enkele bestaande (buitenlandse) systemen en een uitgebreid literatuuronderzoek betreffende dit onderwerp, zorgt de input van de Delphi-studie dan voor de registratie van sleutelfactoren en haalbaarheid van een aantal mogelijke toekomstige scenario's.

5.2 De meerwaarde van scenario-onderzoek in deze context

In onderwijskundige context is het gebruik van scenario's beperkt. Daarnaast vinden toekomstonderzoekers vaak ook dat het beleid weinig gebruik maakt van hun resultaten (Dammers, 2000). Het maken van scenario's vergt veel inzicht van de onderzoekers in de beleidscontext. Zo moet niet enkel de kernvraag van het onderzoek aansluiten bij actuele beleidsvragen, ook wordt de kernvraag onderbouwd op basis van de kennis, waarden, normen en verwachtingen van de betrokken actoren.

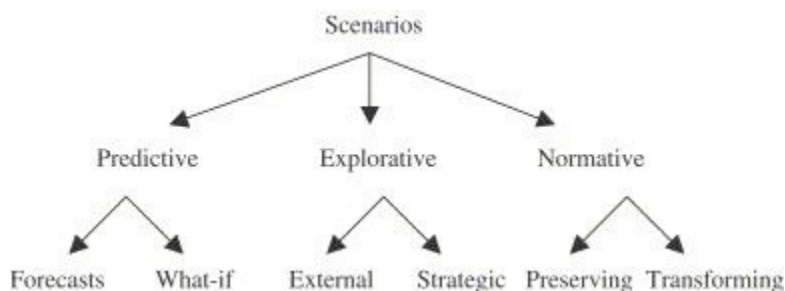
De waarde van scenario-onderzoek blijkt uit een aantal voordelen. Allereerst ondersteunen scenario's beleidvoeringsprocessen doordat verschillende strategische standpunten ten opzichte van elkaar worden afgewogen (Vlaamse Overheid, 2005). Daarnaast benadrukken Mietzner en Reger (2005) dat het ontwikkelen van scenario's bijdraagt tot kennisontwikkeling. Door uit te gaan van verschillende mogelijkheden komen ook toekomstbeelden tot stand waarover voordien niet werd nagedacht. Ook worden scenario's gebruikt om de communicatie omtrent bepaalde onderwerpen te verbeteren. Door de diverse stakeholders te stimuleren om zelf na te denken over beleidsproblemen en oplossingen, wordt niet alleen de betrokkenheid van de participanten groter, maar vergroot ook het draagvlak voor de beleidskeuzes. Tevens zorgen de verhaallijnen en het perspectief binnen de scenario's voor het uitlokken van een gestructureerd maatschappelijk debat. Tot slot zorgen scenario's voor het identificeren van kritische indicatoren of sleutelfactoren (Van der Heijden, 1996).

Ondanks de vele voordelen zijn er ook een aantal nadelen en valkuilen verbonden aan deze onderzoeksmethode. Het uitvoeren van een scenario-onderzoek is een tijdrovende bezigheid. De kwalitatieve aanpak leent zich ertoe om de respondenten nauwkeurig te selecteren. Tot slot is het belangrijk om bij de keuze van sleutelfactoren en scenario's geen zwart-witverhaal neer te zetten als meest relevant scenario (Mietzner & Reger, 2005).

5.3 Scenariotypologieën

In dit onderzoek maken we gebruik van de scenariotypologie van Börjeson, Höjer, Dreborg, Ekvall, en Finnveden (2006) omdat deze typologie gericht is op zowel de ontwikkelaars als de gebruikers van scenario's. Volgens Börjeson et al. (2006) bestaan er drie types van scenario's (zie Figuur 2) die zich van elkaar onderscheiden middels de vraag die men wil beantwoorden.

Predictieve scenario's trachten een antwoord te bieden op de vraag 'wat zal er gebeuren?'. Binnen dit type scenario's wordt een onderscheid gemaakt tussen 'forecasts' en 'wat als' scenario's. Het eerste type scenario bouwt verder op trends en gangbare ontwikkelingen, terwijl 'wat als' scenario's zich focussen op de impact van specifieke gebeurtenissen en condities. Aan de hand van kwantitatieve data en causale modellen richten de ontwikkelaars zich vooral op korte termijn. Ontwikkelaars van predictieve scenario's streven naar resultaten die zeer waarschijnlijk zijn. Ze worden dan ook vooral gebruikt om investeerders en beleidsmakers de kans te geven tijdig te anticiperen op mogelijke gebeurtenissen in de toekomst en om het bewustzijn rond mogelijke toekomstige problemen en opportuniteiten te vergroten (Block et al., 2010; Börjeson et al., 2006).



Figuur 2 - Scenariotypologieën (Börjeson et al., 2006).

Normatieve scenario's richten zich op lange termijn doelstellingen of gewenste toekomstbeelden en vertrekken daarbij van de vraag 'wat moet er gebeuren?' (Block et al., 2010; Börjeson et al., 2006). Dit type scenario wordt opgesplitst in scenario's die vertrekken vanuit het behouden van bestaande structuren en in scenario's die verandering vergen van die structuren. In een behoudscenario wordt onderzocht hoe men vanuit de bestaande structuren efficiënt en met minimale inbreuk op die bestaande situatie een gewenst doel kan bereiken. In een transformatiescenario is een trendbreuk onafwendbaar om een doel te verwezenlijken. Vaak werkt men eerst een toekomstbeeld uit dat 25 tot 50 jaar vooruitloopt, waarna men achteruit gaat nadenken over de afstemming tussen de te bereiken en huidige structuren (Block et al., 2010; Börjeson et al., 2006).

Tot slot zijn er de exploratieve toekomstscenario's die verschillende perspectieven kunnen bevatten en die zich richten op de vraag 'wat kan er gebeuren?'. Hoewel dit type scenario inzicht wil verschaffen op lange termijn (gaande van 20 tot 100 jaar), vertrekken ontwikkelaars uit het heden en niet vanuit een gewenst toekomstbeeld. Door het ontwikkelen van verschillende hypothesen over mogelijke toekomstbeelden en het beredeneren van bedreigingen, opportuniteiten en implicaties streeft men ernaar om meerdere scenario's in kaart te brengen. Ook hier maken Börjeson et al. (2006) een onderscheid tussen twee vormen. Enerzijds focussen externe scenario's zich op de impact van externe, moeilijk beïnvloedbare factoren die niet door de betrokken actoren te controleren zijn. Anderzijds zijn er de strategische scenario's die ontwikkeld worden om de gevolgen van zelfgenomen strategische beslissingen te onderzoeken. Hoewel men rekening houdt met externe factoren, zijn de belangrijkste elementen in deze ontwikkelingsprocessen toch de interne factoren waarop een organisatie invloed kan hebben (Block et al., 2010).

In dit onderzoek worden hoofdzakelijk exploratieve scenario's ontwikkeld. Exploratieve scenario's hebben een belangrijke meerwaarde voor zowel de beleidsmakers als de betrokken actoren. Wie de focus richt op het eindresultaat zal merken dat scenario's een bron van kennis zijn die kunnen bijdragen aan besluitvormingsprocessen. Daarnaast is het niet onbelangrijk om te weten dat het betrekken van diverse actoren in het ontwikkelen van de scenario's een proces op gang brengt waarin de betrokken actoren gestimuleerd worden om grondiger te gaan nadenken over de problemen en mogelijke oplossingen (Quist, 2007). Block et al. (2010) stellen dat het transitiedenken in grote mate samenhangt met het ontwikkelen van normatieve scenario's. Doordat transitie management zich richt op het duurzaam (her)oriënteren van maatschappelijke systemen spelen toekomstvisies, die vaak worden opgesteld met behulp van een delegatie stakeholders, hier een rol in. Door bestaande problemen anders te benaderen en oplossingsmogelijkheden aan te reiken zorgen deze scenario's voor het vernieuwen en herstructureren van huidige regelgevingen. Het coöperatieve karakter van scenario's leidt vaak tot gedeelde toekomstbeelden waardoor het

draagvlak kan vergroot worden. Omdat instellingen tijdig moeten anticiperen op toekomstige ontwikkelingen, kan het interessant zijn om met behulp van predictieve scenario's een voorsprong op te bouwen. Doordat de scenario's voor conceptuele onderbouwing zorgen, kunnen strategische beslissingen snel en gefundeerd gemaakt worden. Het werken met een bepaald type scenario sluit de anderen niet uit. In werkelijkheid lopen de verschillende types door elkaar en werkt men vaak met hybride scenario's (Block et al., 2010).

5.4 Scenario-ontwikkeling en Delphi-studies

De Delphi-studie en de methode van scenario-ontwikkeling kunnen elkaar op verschillende manier verrijken. Enerzijds kunnen scenario's geïntegreerd worden als input voor een Delphi-studie, maar evengoed bestaat onderzoek waarbij de resultaten van een Delphi-studie geïntegreerd worden in ontwikkelde scenario's (Kosow & Gassner, 2008). In dit onderzoek werden de resultaten van de eerste en tweede ronde van de Delphi-studie als één van de bronnen van input gebruikt om de sleutelfactoren of bouwstenen te identificeren. Op basis van de resultaten van het literatuuronderzoek, landenstudies en de Delphi-studie werden acht scenario's ontwikkeld die een gestructureerd en onderbouwd debat mogelijk maken.

5.5 Scenario-ontwikkeling in dit onderzoek

Zoals eerder gesteld, onderbouwde en ontwikkelde deze studie mogelijke scenario's voor het monitoren van leerresultaten en/of leerwinst. Het scenario-ontwikkelingsproces werd gestructureerd aan de hand van de klassieke methode voor scenario-ontwikkeling van Schwartz (1991). In vier stappen werden exploratieve scenario's ontwikkeld omtrent het monitoren van leerresultaten/leerwinst in de Vlaamse context:

- (1) Relevante kenmerken en trends in kaart brengen
- (2) Bouwstenen bepalen
- (3) Scenario's ontwikkelen rond de verschillende bouwstenen
- (4) Mogelijke implicaties oplijsten per scenario

In de eerste stap werd de omgeving geanalyseerd aan de hand van de eerste en tweede ronde van de Delphi-studie (zie sectie 4). Er werden in totaal 24 individuele scenario's voor leer(winst)monitoringssystemen uitgeschreven, die daarna horizontaal en verticaal geanalyseerd werden. Op basis van de resultaten van het literatuuronderzoek en de internationale casestudies werden in deze scenario's relevante en mogelijke toekomstige ontwikkelingen geïdentificeerd, geïnventariseerd en gegroepeerd per geselecteerd topic. In een volgende fase werd de impact van elk van deze topics naar waarde geschat. Op die manier werden de drijvende factoren of bouwstenen in kaart gebracht. In de derde stap werden acht scenario's uitgewerkt die gebaseerd zijn op de bouwstenen uit de eerdere fases: doelen op micro-, meso- en macroniveau, frequentie, verplichtend karakter, ontwikkelings- of verantwoordingsgericht perspectief, beschikbaarheid van de gegevens, leerwinst, toegevoegde waarde, toetsvorm, link met curriculum, inhoud, ontwikkeling en correctie. Nadat de scenario's vorm kregen, werden de implicaties (positieve effecten en neveneffecten) per scenario opgelijst. Eveneens werd een raming van de kostprijs gemaakt.

5.6 Het gebruik van de scenario's in dit onderzoek

In de derde ronde van de Delphi-studie werd elk gesprek ingeleid met een korte introductie in het ontwikkelingsproces en de samenhang van de scenario's. Doordat de respondenten op voorhand de mogelijkheid gekregen hadden om de scenario's grondig door te nemen en in te schatten naar wenselijkheid en haalbaarheid (zie 4.2), was er tijdens de gesprekken voldoende tijd om de standpunten van de respondenten diepgaand te verkennen.

Elke respondent had de mogelijkheid om per scenario zijn of haar standpunt betreffende de wenselijkheid en haalbaarheid te geven. Er werd daarbij vooral gepeild naar oorzaken of kenmerken die ervoor zorgden dat een scenario meer of minder wenselijk en/of haalbaar werd geacht. Respondenten hadden de mogelijkheid om bijkomende suggesties te geven die ervoor zorgden dat de wenselijkheid en/of haalbaarheid van de scenario's kenterden.

DEEL 3 RESULTATEN LITERATUURSTUDIE

De literatuur inzake de effectiviteit van leer(winst)monitoringssystemen met gestandaardiseerde toetsen is zeer uitgebreid. Uit de eerdere definitie (zie Deel 1) blijkt dat het concept 'gestandaardiseerde toetsen' op diverse wijzen geïnterpreteerd kan worden. Enkele niet door empirie ondersteunde normatieve theorieën buiten beschouwing gelaten, is er echter geen overkoepelend theoretisch kader dat uitlegt tot welke effecten gestandaardiseerde toetsen kunnen leiden of hoe gestandaardiseerde toetsen deze veronderstelde effecten zou kunnen genereren (Rozenwajn, 2014). Ook een review studie kwam tot een dergelijke conclusie inzake de impact van gestandaardiseerde toetsen op prestatie-indicatoren (J. Lee, 2008). Volgens deze review ontbrak in het merendeel van de studies een theoretisch kader dat duiding geeft over de effecten van verantwoordingsgerichte initiatieven betreffende de leerprestaties van leerlingen.

In wat volgt, worden veronderstelde positieve en negatieve effecten van leer(winst)monitoringssystemen en individuele gestandaardiseerde toetsen op zowel microniveau als meso- en macroniveau in kaart gebracht. Daarbij worden ook meteen de impactstudies verwerkt die het effect van deze diverse systemen (centrale eindexamens, schooltoetsen, toetsenbank, ...) hebben bestudeerd. Vaak bevestigen deze impactstudies het bestaan van het veronderstelde positieve of negatieve effect, maar soms worden ook andere bevindingen gerapporteerd.

Uiteraard zijn de effecten van leer(winst)monitoringssystemen niet in elke situatie gelijk. Daarom gaat in de tweede sectie aandacht uit naar een theoretische onderbouwing betreffende een aantal relevante bouwstenen van leer(winst)monitoringssystemen. Van elk van die bouwstenen wordt een impact op de effectiviteit van de leer(winst)monitoringssystemen verwacht. In de derde sectie wordt ingezoomd op leerwinst en toegevoegde waarde. De literatuur die vaak uiteenlopende betekenissen toekent aan deze beide begrippen wordt hier samengebracht.

De vierde sectie brengt ten slotte de belangrijkste bevindingen van de eerdere drie secties kort samen.

1 IMPACT VAN GESTANDAARDISEERDE TOETSEN ALS BASIS VAN LEER(WINST)-MONITORING

In elk onderwijssysteem waarin gestandaardiseerde toetsen - individueel of als onderdeel van een leer(winst)monitoringssysteem - worden geïmplementeerd, gaat men ervan uit dat dit zal leiden tot een verbetering van het onderwijs. Hoewel het uiteindelijke gebruik daarbij vaak een verbetering van de leerprestaties van leerlingen beoogt, wordt de nadruk meestal op tussenliggende doelen (op meso- of macroniveau) gelegd. Voornamelijk de introductie van gestandaardiseerde eindexamens, waar de toekenning van een diploma of de toegang tot een bepaalde vorm van vervolgonderwijs geheel of gedeeltelijk samenhangt met de resultaten op het examen, gaat gepaard met expliciete en impliciete assumpties over onderliggende mechanismen die zouden leiden tot betere leerprestaties, met name:

- leraren zullen nauwgezet het volledige curriculum onderwijzen (ook die onderdelen waar ze zelf minder vertrouwd mee zijn) (*'coverage of content standards'*);
- examens hebben het potentieel om de kwaliteit van het curriculum en de gebruikte evaluatie-opdrachten te verhogen (*'innovation potential'*);
- daaruit kan volgen dat leraren ook meer kwaliteitsvolle opdrachten gebruiken in de klas om tegemoet te komen aan de sturende opdrachten gehanteerd in de examens (*'alignment of testing and instruction'*);
- eveneens wordt de (extrinsieke) motivatie van zowel leraren als leerlingen verhoogd, alsook hun wederzijdse verbondenheid als partners in het proces, omdat ze gezamenlijk een uitdagende situatie dienen te beheersen (*'extrinsic motivation'*);
- gestandaardiseerde examens kunnen, ten slotte, de diagnostische vaardigheden van leraren versterken en meer richten op de mate waarin leerlingen bepaalde doelen al dan niet hebben bereikt (*'criterion-referenced rigor'*) (Klein & Van Ackeren, 2012; Van Ackeren, Block, Klein, & Kühn, 2012).

Een moeilijkheid bij de bespreking van de literatuur is het feit dat opinie en empirie niet strikt van elkaar gescheiden worden. Daarbij komt nog dat de empirie moeilijk te generaliseren is door de uiteenlopende vormen van gestandaardiseerde examens die in verschillende onderwijscontexten worden georganiseerd (Klein & Van Ackeren, 2012).

Zoals zal blijken uit de onderstaande bespreking, komt een groot deel van de meer recente empirische evidentie omtrent de effecten van centrale examens uit Duitsland en Noord-Amerika. De verklaring hiervoor is dat er in de verschillende regio's (deelstaten in Duitsland, provincies in Canada, steden in de Verenigde Staten) andere beleidslijnen betreffende eindexamens gelden, wat uitnodigt tot effectstudies die de verschillen tussen deelstaten voor bepaalde variabelen nagaan. De andere kenmerken van de onderwijssystemen in verschillende deelstaten worden immers geacht om relatief gelijk te zijn, waardoor eventuele verschillen (in bijvoorbeeld leerprestaties) aan de implementatie van gestandaardiseerde toetsen zouden kunnen worden toegeschreven (Jürges, Richter, & Schneider, 2005). Het is evenwel niet altijd duidelijk in welke mate deze verschillen echt causaal te verklaren zijn, waardoor het dan ook een bijzonder moeilijke opdracht is om de impact van de implementatie van gestandaardiseerde toetsen op veronderstelde effecten na te gaan. Leerprestaties kunnen immers ook door heel andere variabelen, of door algemene maatschappelijke trends worden beïnvloed (Carnoy & Loeb, 2002; Jürges, Büchel, et al., 2005). Gestandaardiseerde toetsen worden bovendien meestal geïmplementeerd als onderdeel van een ruimere onderwijsvernieuwing, waardoor het effect van één element (met name gestandaardiseerde toetsen) moeilijk te isoleren is (Figlio & Ladd, 2008). Tot slot is experimenteel onderzoek zo goed als onmogelijk, aangezien beslissingen over de invoering van gestandaardiseerde toetsen meestal voor alle scholen binnen een bepaald gebied genomen worden, zodat quasi-experimenteel onderzoek met een controlegroep onmogelijk is (Figlio & Ladd, 2008; Jürges, Richter, et al., 2005).

In het onderstaande worden de veronderstelde effecten op micro-, meso-, en macroniveau beschreven, die wordt aangevuld met data uit empirisch onderzoek (op voorwaarde dat deze beschikbaar is). Er wordt telkens onderscheid gemaakt tussen effecten die men als positief beschouwt en neveneffecten die men onwenselijk acht.

1.1 Positieve effecten op microniveau

1.1.1 Beter beeld van leerprestaties door een meer betrouwbare en valide evaluatie

Er woedt al lange tijd een discussie over de vraag of de resultaten van gestandaardiseerde toetsen een betere weergave zijn van de capaciteiten van de leerling dan de resultaten van toetsen die leraren zelf ontwerpen. Voorstanders menen dat een gestandaardiseerde toets garant staat voor een meer nauwkeurige en objectieve beoordeling (Marlow et al., 2014). Zo zouden leraren onvoldoende kennis en vaardigheden bezitten om een valide en betrouwbare evaluatie uit te voeren (Black, Harrison, Hodgen, Marshall, & Serret, 2011). Wyatt-Smith en Klenowski (2013) menen dat de kennis en opvattingen van de leraar inzake waarden, normen en sociaal-culturele factoren een grote invloed uitoefent op de manier waarop de capaciteiten van een leerling beoordeeld worden.

Empirische resultaten

Onderzoek wijst uit dat leraren soms onbewust een bias hanteren in hun lesgeven en hun evaluatie, waarbij het oordeel van leraren beïnvloed wordt door 'irrelevante' factoren zoals gender, socio-economische achtergrond, inzet en gedrag in de klas (Baird, Ahmed, Hopfenbeck, Brown, & Elliott, 2013). Toetsen die door leraren zelf worden opgesteld, houden in vele gevallen een onvoldoende valide beoordeling van de kennis en vaardigheden van leerlingen in (Van den Bergh, 1993; Verhoeven, Devos, Bruylant, & Warmoes, 2002). Leraren zouden vaak ook rekening houden met de omstandigheden waarin leerlingen een bepaalde taak afwerken: zo kan een prestatie van een leerling met een anderstalige achtergrond onbewust hoger worden ingeschat dan een identieke prestatie van een autochtone leerling (De Lange & Dronkers, 2007). Bovendien zou de invloed van leerlingkenmerken, zoals de culturele bagage van het thuismilieu, een grotere impact kunnen hebben op taken die vaker op schoolniveau worden aangeboden (bijvoorbeeld het maken van een werkstuk) dan op taken die vaker via gestandaardiseerde toetsen worden aangeboden (De Lange & Dronkers, 2007). Andere onderzoekers maken ook gewag van een discrepantie tussen *resultaten* op high-stakes testen in het nadeel van etnische minderheden en vrouwen (Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Cobb & Russell, 2015).

Een analyse in Noorwegen leert dat de verschillen tussen jongens en meisjes kleiner zijn op de gestandaardiseerde toetsen in vergelijking met de evaluaties die door de scholen zelf worden opgesteld (Norwegian Directorate for Education and Training, 2014). Dit gegeven wordt bevestigd door een analyse op de data van de eindexamens in Nederland, en de eindscores die door leraren zelf worden bepaald: de verschillen tussen jongens en meisjes zijn ook hier kleiner bij de gestandaardiseerde toets. De gemiddelde achterstand van etnische minoriteiten op autochtone leerlingen is evenwel net groter op het gestandaardiseerde eindexamen dan op de evaluaties van scholen (Schildkamp, Rekers-Mombarg, & Harms, 2012).

Onderzoek van Eliot (2010) wijst uit dat verschillen in prestaties tussen jongens en meisjes niet kunnen worden verklaard door biologische of neurologische factoren en besluit dat verschillen tussen jongens en meisjes inzake leren en toetsen niet 'aangeboren' zijn, maar voornamelijk het gevolg lijken te zijn van andere interesses én van een andere behandeling (door ouders, leraren) op basis van genderspecifieke stereotypes (Eliot, 2010). Door de evaluatie te baseren op gestandaardiseerde toetsen, kan de bias inzake evaluatie worden uitgeschakeld. Op die manier dragen gestandaardiseerde toetsen bij tot gelijke onderwijskansen voor alle leerlingen (Koschoreck, 2001).

Kritische kanttekeningen

Door Bowers (2011) wordt dan weer gesteld dat punten die door leraren worden toegekend, goede voorspellers zijn voor de latere schoolresultaten. Wikeley (1998) stelt dat leraren soms menen dat de resultaten van gestandaardiseerde toetsen weinig nieuwe informatie aanleveren ten opzichte van de eigen evaluaties.

Tegenover de evidentie inzake het positieve objectiverend effect staat verder ook de bewering dat gestandaardiseerde toetsen een bias vertonen, net omdat ze afwijken van de evaluaties van leraren (Gipps & Stobart, 2009; Klenowski, 2014). Zo zouden jongens het op gestandaardiseerde toetsen systematisch beter doen dan meisjes omdat ze meer durven gokken tussen verschillende antwoordalternatieven. Voor deze bewering bestaat echter geen neurologisch bewijs (Eliot, 2010). Naast bovenstaande positieve gevolgen van gestandaardiseerde toetsen, blijkt evengoed dat ze bepaalde groepen van leerlingen kunnen benadelen. Sectie 1.2.6 verduidelijkt de mogelijke ongelijkheid die gestandaardiseerde toetsen veroorzaken.

1.1.2 Garantie op een 'gelijke lat' voor alle leerlingen in alle scholen

Naast het voordeel dat gestandaardiseerde toetsen in een leer(winst)monitoringssysteem een meer betrouwbare en valide meting zouden kunnen inhouden, zouden deze toetsen - zeker wanneer het gestandaardiseerde eindexamens betreft - het voordeel hebben dat voor alle leerlingen dezelfde lat wordt gehanteerd bij de toekenning van diploma's, getuigschriften of certificaten. Gelijkheid inzake de verwachtingen ten aanzien van leerlingen en de evaluatie van leerlingen is niet enkel in het belang van leerlingen zelf. Ook instellingen voor hoger onderwijs en potentiële werkgevers kunnen baat hebben bij een gestandaardiseerde meting van kennis en competenties aan het eind van het leerplichtonderwijs (Bishop & Wössmann, 2004; Van Ackeren et al., 2012). Aangezien de arbeidsmarkt meer belang hecht aan de resultaten van leerlingen wanneer ze gebaseerd zijn op gestandaardiseerde toetsen (in plaats van op evaluaties die verschillen van school tot school), kan worden verwacht dat leerlingen meer gemotiveerd zullen zijn om betere punten te behalen indien er een centraal eindexamen wordt georganiseerd (Bishop & Wössmann, 2004; De Lange & Dronkers, 2007; Wössmann, 2005), alsook dat scholen een grotere incentive ervaren om kwaliteitsvol onderwijs aan te bieden (Bishop & Wössmann, 2004).

Empirische resultaten

Empirische evidentie die dit effect bevestigt, is afkomstig van een vergelijkende analyse van de resultaten op het eindexamen secundair onderwijs in Duitsland en de resultaten die leerlingen behaalden op basis van de evaluaties van scholen zelf. Neumann et al. (2011) bevestigden empirisch dat gestandaardiseerde toetsen leiden tot meer vergelijkbare (lees: eerlijke) eindresultaten: leerlingen in sterke scholen krijgen gemiddeld minder hoge punten dan even sterke leerlingen in scholen met een gemiddeld zwakker leerlingenpubliek. Toch vond deze studie dat ook dit '*frame-of-reference*' effect ook in zekere mate een invloed had op de eindexamens, al was het effect daar veel kleiner (Neumann et al., 2011).

In het verlengde hiervan blijkt uit een analyse op de verschilcores tussen de eindexamens van scholen, en de gestandaardiseerde eindexamens aan het eind van het secundair onderwijs in Nederland, dat er inderdaad discrepantie zijn. Bovendien blijkt dat er systematische verschillen zijn tussen scholen inzake de toekenning van punten voor het eindexamen en dat deze verschillen samenhangen met de mate van onderlinge afstemming tussen leraren binnen een school in de

becijfering (waarbij scholen met een hoge mate van onderlinge afstemming gemiddeld lagere punten toekennen) (De Lange & Dronkers, 2007). Bij gestandaardiseerde toetsen is de beoordeling van de prestaties van leerlingen onafhankelijk van de prestaties van andere leerlingen in dezelfde klasgroep, wat vaak het geval is bij toetsen die enkel binnen een enkele school worden afgenomen (cf. het groepsnorm-criterium) (Neumann et al., 2011; Willingham et al., 2002). In het bovenstaande onderzoek van De Lange en Dronkers (2007) blijkt inderdaad dat een hogere verschilscore tussen gemiddelde score schooltoets en gestandaardiseerd eindexamen correleert met een lager gemiddeld cijfer op het gestandaardiseerde eindexamen. Eveneens wijst het onderzoek uit dat de discrepantie tussen de gemiddelde cijfers van school-gebaseerde eindtoetsen en gestandaardiseerde eindexamens groeit naarmate een school een hoger percentage allochtone leerlingen heeft (De Lange & Dronkers, 2007).

Dat ook scholen vragende partij kunnen zijn voor een instrument dat deze gelijke lat garandeert, bewijst het voorbeeld van Wales. Toen Wales in het begin van deze eeuw meer autonomie kreeg over haar eigen onderwijssysteem werden de verplichte *Key Stage 2* toetsen afgeschaft. Deze toetsen aan het eind van het lager onderwijs werden gebruikt om leerlingen (en ouders) een objectief beeld te geven van de eigen kennis en vaardigheden. Meteen na de afschaffing van deze toetsen werden verschillende gestandaardiseerde toetsen ter beschikking gesteld met als doel om de summatieve beoordeling van leraren over leerlingen te ondersteunen. Deze toetsenbank kon facultatief gebruikt worden door leraren of scholen. In de focusgroepen bestaande uit directies en leraren van lagere scholen én leraren uit secundaire scholen in een onderzoek van Collins et al. (2010) bleek er een consensus te zijn over het feit dat deze toetsen waardevolle instrumenten zijn om een summatieve beoordeling te ondersteunen: er bleek nog steeds een nood aan 'harde evidentie' over wat leerlingen kennen en kunnen. In sommige gevallen legde het schoolbestuur een gestandaardiseerde toets op aan al hun scholen.

1.1.3 Betere leerprestaties

Voorstanders van leer(winst)monitoringssystemen menen dat de implementatie ervan op korte of lange termijn een positief effect kan hebben op de leerprestaties van leerlingen (Bishop, 1998a; Chakrabarti & Schwartz, 2013; Linn, 2000). Uit review studies blijkt dat er inderdaad evidentie bestaat om de bovenstaande positieve bewering inzake het effect van gestandaardiseerde toetsen op leerprestaties te staven (Phelps, 2005b; Wiliam, 2010). Deze studies zijn voornamelijk gericht op de impact van gestandaardiseerde eindexamens die door alle leerlingen van een bepaalde leeftijd (of in een bepaald leerjaar) in alle scholen in een bepaald gebied moeten worden afgelegd, met als primair doel om een diploma of toegang tot bepaalde vervolgstudies te verkrijgen. Het betreft dus de impact van high-stakes verantwoordingsgerichte toetsen op het microniveau. Daarnaast zijn er evenwel een aantal studies die ook de impact van andere vormen van gestandaardiseerde toetsen in kaart hebben gebracht.

Onderliggende mechanismen

Volgens Bishop en Wössmann (2004) worden leerprestaties versterkt door drie mechanismen: (1) beloningen, (2) een verminderde negatieve invloed van klasgenoten en (3) een verhoogde monitoring van leraren en scholen. Een leerling of leraar die een (im)materiële beloning ontvangt bij een goede prestatie (en eventueel bestraft wordt bij een slechte prestatie), wordt geacht meer gemotiveerd te zijn om te leren of om kwaliteitsvol les te geven (Hanushek & Wössmann, 2010).

Tegenstanders van gestandaardiseerde toetsen argumenteren echter dat hierdoor enkel extrinsieke motivatie wordt gestimuleerd, wat ten koste gaat van de (meer kwaliteitsvolle) intrinsieke motivatie (Wang et al., 2006). Daarnaast wijzen Bishop en Wössmann (2004) op de verminderde negatieve invloed van leerlingen op het leren van klasgenoten en op het lesgeven van leraren: bij gestandaardiseerde toetsen heeft men er immers baat bij dat de eigen groep zo goed mogelijk presteert, terwijl men er bij toetsen beperkt tot de eigen school (waarbij vaak een groepsnorm-criterium op klasniveau wordt gehanteerd) baat bij heeft dat medeleerlingen zo laag mogelijk scoren. Ook de verhoogde monitoring van leraren en scholen, waardoor ouders informatie krijgen over de prestatie van de school in vergelijking met andere scholen en op basis daarvan actie ondernemen, zou bijdragen tot het bevorderen van leerprestaties. Tegenstanders, zoals Harris (2011), menen echter dat aangezien de toetsresultaten meestal enkel op schoolniveau worden vrijgegeven, de onderpresterende leraren zich kunnen verschuilen achter collega's en niet de verantwoordelijkheid zullen opnemen voor eventuele slechte resultaten op schoolniveau. Een andere, meer fundamentele kritiek komt van tegenstanders die menen dat geen enkele vorm van 'verantwoordingsgericht' toetsen kan leiden tot echte kwaliteitsverbetering, omdat dit steeds van binnen in de school moet komen (Smith, 1995).

Empirische resultaten uit Duitsland

Verschillende studies in Duitsland onderzochten de leerresultaten van leerlingen in verschillende deelstaten, waarbij een vergelijking wordt gemaakt tussen deelstaten mét en deelstaten zonder gestandaardiseerde eindexamens. Jürges, Büchel, et al. (2005) besloten op die manier dat leerlingen beter scoren op wiskunde in de deelstaten met gestandaardiseerde toetsen. Het verschil wordt in deze studie conservatief ingeschat op één vijfde van de normale groei van een leerling tijdens een volledig schooljaar (Jürges, Büchel, et al., 2005). In een andere analyse van dezelfde onderzoeksgroep op de resultaten van PISA-E (een uitbreiding op de PISA-studie met een veel grotere steekproef) en bijkomende vragenlijsten voor ouders, leraren en leerlingen bleek dat slechts twee van de vele variabelen in alle drie de onderzochte schooltypes significant verschilden van elkaar tussen deelstaten met centrale eindexamens en deelstaten zonder centrale examens. Het betreft de leerprestaties in wiskunde en het oordeel van ouders over het academisch niveau van de school: beiden waren hoger in de deelstaten met centrale examens. De verschillscore voor wiskunde was het grootst in *Hauptschule*, die meestal de minst sterke leerlingen herbergen. Vermoedelijk komt dit doordat de meting in deze scholen plaatsvond op een tijdstip dicht bij de afname van de eindexamens, wat wijst op een 'nabijheidseffect': de *Hauptschule* wordt immers afgerond na het negende leerjaar, terwijl *Realschule* en *Gymnasia* respectievelijk ook het tiende en twaalfde leerjaar aanbieden (Jürges, Richter, et al., 2005).

Een latere analyse op dezelfde data, aangevuld met de resultaten van een gestandaardiseerde toets in het tiende leerjaar, toont aan dat er na controle voor eerdere prestaties, cognitieve vaardigheden, zittenblijven en verschillende variabelen inzake socio-economische status, geen effect is van de centrale eindexamens op de algemene wiskundige vaardigheid van leerlingen. Toch werd een effect gevonden op de curriculumgebonden kennis en vaardigheid inzake wiskunde, aangezien leerlingen in deelstaten met een centraal eindexamen beter presteren. Dit effect beperkt zich evenwel tot de *Hauptschule* (mogelijk speelt opnieuw het bovengenoemde nabijheidseffect een rol) (Jürges, Schneider, Senkbeil, & Carstensen, 2009).

Later onderzoek van dezelfde onderzoeksgroep gaf dan weer aan dat leerlingen in de deelstaten met gestandaardiseerde eindexamens betere resultaten konden voorleggen inzake zowel wiskunde als wetenschappen (gemeten via TIMSS). Bovendien was ook de zelf-effectiviteit (vertrouwen in de eigen bekwaamheid) van deze leerlingen in beide leerdomeinen hoger, terwijl dat in de vorige studie niet het geval was (Jürges & Schneider, 2010).

Een paarsgewijze vergelijking tussen de geaggregeerde resultaten van leerlingen op PISA-E 2003 concludeerde dat er nauwelijks verschillen zijn tussen de leerlingen in deelstaten met en deelstaten zonder centrale eindexamens, al vond men dat, na controle voor relevant geachte leerlingkenmerken, leerlingen in drie van de zeven deelstaten met centrale examens statistisch significant (met kleine of medium effectgrootte) beter scoorden dan leerlingen in vijf van de zeven deelstaten zonder centrale examens wat wiskunde betreft (maar niet voor leesvaardigheid of wetenschappelijke vaardigheid) (Block, Klein, Van Ackeren, & Kühn (2011) in Van Ackeren et al., 2012).

Andere onderzoekers (Baumert & Watermann (2000), in Klein & Van Ackeren, 2012) stellen dat het effect van gestandaardiseerde examens aan het einde van het secundair onderwijs sterk afhankelijk is van uiteenlopende factoren, zoals het leerdomein, het gevolgde studieprogramma en het behaalde resultaat van de leerlingen. Zo vonden de onderzoekers bij de laagst presterende groep wel een positief effect van de *Abiturprüfung*, het verplichte centraal eindexamen aan het einde van het Gymnasium, voor wiskunde, maar niet voor fysica. Al komt dit mogelijk omdat fysica in de betreffende onderzoekscontext een keuzevak is dat vaak enkel door de sterkst presterende leerlingen wordt gekozen (Klein & Van Ackeren, 2012).

Empirische resultaten uit Noord-Amerika

Ook in de Verenigde Staten werd aan het eind van vorige eeuw een positief effect van gestandaardiseerde eindexamens op de leerprestaties gedocumenteerd. Bishop (1998a) vergeleek de prestaties van leerlingen in de staat New York (op dat ogenblik de enige staat met centrale eindexamens) met die van leerlingen in andere staten. In een andere publicatie werden de resultaten van leerlingen uit verschillende Canadese provincies met elkaar vergeleken. Uit deze resultaten kon eveneens geconcludeerd worden dat leerlingen significant beter presteren op wiskunde en wetenschappen in scholen met een centraal examen (Bishop, 1998b).

Empirische resultaten uit comparatieve studies

Daarnaast zijn er ook andere comparatieve studies in het kader van centrale eindexamens uitgevoerd. Zo werd op basis van de resultaten van TIMSS 1994-1995 en IAEP 1991 (International Assessment of Educational Progress) aangetoond dat leerlingen in onderwijssystemen met een gestandaardiseerd centraal examen betere studieresultaten inzake wiskunde en wetenschappen behalen (Bishop, 1998a). Andere mogelijke verklaringen voor deze verschillen tussen uiteenlopende landen worden door de auteur evenwel slechts zeer beperkt besproken. Net als in andere comparatieve studies kan ook hier kritiek worden geuit dat eventuele significante verbanden niet noodzakelijk een causale aard hebben. Deze kritiek gaat slechts in mindere mate op voor de bovengenoemde studies in Duitsland of Noord-Amerika omdat de verschillende (deel)staten er een algemene politieke en historische achtergrond delen (Klein & Van Ackeren, 2012).

Het onderzoek van Wössmann (2005) ging uit van de veronderstelling dat centrale eindexamens een differentiële impact kunnen hebben op leerlingen, rekening houdende met zowel

leerlingfactoren (zoals prestaties en de achtergrond van ouders) als schoolfactoren (zoals de mate van autonomie van de school en het gebruik van de centrale examens). Op basis van de resultaten van twee cohorten op TIMSS (1995 en 1999) en van PISA 2000 werd geconcludeerd dat leerlingen in landen met centrale eindexamens gemiddeld betere leerprestaties behalen (in wiskunde en wetenschappen) dan leerlingen in landen zonder centrale eindexamens. Bovendien blijkt uit de analyse van de TIMSS-data dat volgende groepen van leerlingen meer baat hebben van centrale eindexamens: '*higher-ability*' leerlingen (versus *lower-ability* leerlingen), leerlingen met een migratie-achtergrond (versus autochtone leerlingen), en leerlingen die bij samenwonende ouders leven (versus leerlingen met alleenstaande ouders). Daarnaast stelt Wössmann (2005) dat de invloed van de achtergrond van ouders op de leerprestaties van leerlingen minder groot is in landen met centrale eindexamens, al blijkt ook dat het positieve effect van centrale eindexamens op de prestaties in wetenschappen net weer groter is bij leerlingen met veel boeken in huis. Deze verschillen werden evenwel niet bevestigd door de PISA-data. Een andere bevinding van deze studie wijst op een differentieel effect op leerprestaties van autonomie van scholen (inzake lonen van leraren, inzake het aankopen van goederen, en inzake het bepalen van leerinhouden). In landen met centrale eindexamens behalen leerlingen in sterk autonome scholen betere leerresultaten dan leerlingen in minder autonome scholen, terwijl deze laatste categorie van leerlingen het in landen zonder centrale eindexamens dan weer beter doet dan de andere leerlingen. Het positieve effect van centrale eindexamens is groter in het achtste leerjaar dan in het zevende leerjaar. Tot slot onderzocht deze studie ook het effect van het frequent gebruiken van gestandaardiseerde toetsen (naast eindexamens). Uit de resultaten blijkt dat deze toetsen enkel een positief effect hebben op leerprestaties in landen met centrale eindexamens, terwijl ze de leerprestaties negatief beïnvloeden indien ze niet gekoppeld zijn aan centrale eindexamens (Wössmann, 2005).

Kritische kanttekeningen

Dat leer(winst)monitoringssystemen in het algemeen, of gestandaardiseerde toetsen in het bijzonder, leiden tot hogere leerprestaties wordt evenwel ook gecontesteerd door andere studies. Een Amerikaanse studie stelde bijvoorbeeld vast dat er geen verschil is inzake de leerwinst van leerlingen (van instroom tot eindtoets secundair onderwijs) wat betreft hun leesvaardigheid en wiskundeprestaties tussen staten mét en staten zonder high-stakes eindexamens (Jacob, 2001).

Critici argumenteren dat verbeterde leerprestaties enkel gemeten worden op de betrokken gestandaardiseerde toetsen en dat dit ten koste gaat van niet-getoetste kennis en vaardigheden (Boyd, 2008; Wiliam, 2010). Er zijn inderdaad aanwijzingen dat het positieve effect van gestandaardiseerde eindexamens op leerprestaties beperkt blijft tot die onderdelen van het curriculum die door gestandaardiseerde toetsen in kaart worden gebracht. In Duitsland werd in een uitvoerige studie vastgesteld dat op de PISA-E voor wiskunde een stijging was in de behaalde scores voor de onderdelen die ook door de lokale eindexamens werden getest terwijl er geen impact was op andere onderdelen van wiskunde (met name *mathematical literacy* of de capaciteit om rekenvaardigheden in alledaagse contexten toe te passen) (Jürges et al., 2009). De resultaten van deze studie zetten de onderzoekers aan tot de aanbeveling om meer strategieën van kennistransfer in het curriculum te integreren.

De resultaten van bepaalde studies in de Verenigde Staten liggen in dezelfde kritische lijn. Hieruit blijkt evenwel dat er, ondanks hogere prestaties op nieuw geïmplementeerde gestandaardiseerde toetsen, geen transfer ontstaat naar andere toetsen die een gelijkaardig construct meten (onderzoek

gereviewd door Wiliam, 2010). In deze studies werd nagegaan of een betere prestatie op een gestandaardiseerde toets in een bepaalde staat ook leidde tot het behalen van betere prestaties op andere (nationale) toetsen. Het onderzoek toonde aan dat de leerresultaten verbeterden door een nieuwe gestandaardiseerde toets in Texas (Texas Assessment of Academic Skills: TAAS). Deze toets deed niet enkel dienst als eindexamen voor leerlingen (slagen is een vereiste voor het behalen van een diploma), maar werd tevens gebruikt bij de evaluatie en mogelijk verloning van leraren en directies. Wanneer echter dezelfde constructen werden getoetst door het reeds bestaande nationale NAEP (National Assessment of Educational Progress) stelt men vast dat de 'geboekte winst' van leerlingen in het vierde leerjaar veel kleiner was dan men gemeten had met de TAAS, terwijl de geboekte winst voor leerlingen in het achtste leerjaar volledig uitgevaagd was (Klein, Hamilton, McCaffrey, & Stecher, 2000). Ook bij een eerdere analyse in Kentucky werd het positieve effect van de eindexamens op staatsniveau niet vertaald in hogere scores op de NAEP (Koretz & Barron, 1998). Onderzoek van Wiliam (2010) toont daarentegen aan dat er groeiend bewijs is dat de effecten van high-stakes toetsen niet beperkt blijven tot de leerprestaties zelf, maar dat het gebruik ervan ook niet-gerelateerde leerlingprestaties zou verhogen.

Deze kritische onderzoeksresultaten leiden, volgens sommige onderzoekers, tot de conclusie dat betere prestaties niet te wijten zijn aan 'echte' of gewenste leereffecten, maar eerder het gevolg zijn van specifieke 'teaching to the test' (S. P. Klein et al., 2000; Koretz & Barron, 1998; Wang et al., 2006). Dit neveneffect wordt dan ook uitgebreid besproken in 1.2.2.

Evaluatie van leraren?

Hoewel er enkele modellen bekend zijn - voornamelijk in de Verenigde Staten - waarbij een leer(winst)monitoringssysteem geacht wordt bij te dragen tot betere leerprestaties doordat de resultaten van leerlingen een formele plaats krijgen bij de evaluatie (en zelfs verloning) van leraren, is er in de literatuur weinig steun voor dit idee. Leraren evalueren op basis van leerlingprestaties is niet wenselijk omwille van de te grote abstractie, de onvolledigheid van de criteria, de irrelevantie van andere criteria, de variabiliteit tussen verschillende klasgroepen, meetfouten en de hoge waarschijnlijkheid van onwenselijke manipulatie van data (Scriven, 1995). Ook onderzoek met een experimenteel opzet waarbij sommige leraren konden beloond worden mits goede prestaties van leerlingen, terwijl dat bij een controlegroep niet het geval was, toont aan dat *performance-based payments* weinig effectief zijn (Fryer, 2013; Springer & Winters, 2009).

1.1.4 Beter lesgeven

Volgens Klein en Van Ackeren (2012) wordt vaak verondersteld dat gestandaardiseerde toetsen in een leer(winst)monitoringssysteem tot beter lesgeven leiden. Het lesgeven zou in dit geval ondersteund worden door het gebruik van gestandaardiseerde informatie en de toenemende druk op scholen. Daarnaast zou dit positieve effect eveneens tot stand komen door de explicitering van minimumstandaarden, het motiveren van de implementatie van nieuwe syllabi, de ondersteuning van het lesgeven en het verhogen van de motivatie. Dat gestandaardiseerde toetsen zo leiden tot een betere selectie inzake instructiemethoden en aangeboden inhouden wordt sterk gepropageerd door onder meer Popham, Cruse, Rankin, Sandifer, en Williams (1985) onder de noemer '*measurement-driven instruction*'. Een andere verwachting inzake 'beter lesgeven als gevolg van gestandaardiseerde toetsen' is dat leraren meer opbrengstgericht gaan werken (Janssens et al., 2014).

Daarbij aansluitend zijn verschillende auteurs van mening dat het gebruik van gestandaardiseerde toetsen ertoe bijdraagt dat de verwachtingen van leraren beter aansluiten bij de vereiste condities dan het geval is bij een lijst doelstellingen (zoals eindtermen of leerplannen) (Popham, 2003; Shepard, 1992). Leraren worden door (de resultaten van) gestandaardiseerde toetsen aangezet om in hun lesgeven nauwer aan te sluiten bij het bestaand curriculum (J. A. Marsh, Pane, & Hamilton, 2005). Daarnaast wordt ook verwacht dat het gebruik van gestandaardiseerde toetsen leraren stimuleert om het curriculum volledig aan te bieden, zelfs de delen waar leraren minder vertrouwd met zijn doordat men niet weet welke examenvragen er gesteld zullen worden (Van Ackeren et al., 2012). Dit veronderstelde positieve effect is eigenlijk een vorm van *teaching to the test* dat bij de veronderstelde onwenselijke neveneffecten als *'teaching tot he standard'* zal worden besproken (zie 1.2.2).

Eveneens wordt verwacht dat de informatie die de gestandaardiseerde toetsen opleveren over de kennis en vaardigheden van elke leerling, leraren in staat zullen stellen om zowel op een meer geïnformeerde basis (en dus beter) te differentiëren in de klas (Cunningham, 2005) als om specifieke leerlingen gericht te remediëren (J. A. Marsh et al., 2005).

Game theory

Dit veronderstelde positieve effect wordt verder uitgewerkt door Elstad (2006). Hij gebruikt de *game theory* om te argumenteren dat het - zonder toetsen van buitenaf - zowel voor leerlingen als voor leraren loont om de ambitie zo laag mogelijk te leggen, hoewel de meest wenselijke situatie eruit bestaat dat de leraren hoge (academische) eisen stellen en dat leerlingen zich maximaal inspannen om de leerstof te verwerken en verwerven. Echter, zowel in geval een leraar hoge eisen stelt als in het geval een leraar lage eisen stelt, is de leerling meer gebaat bij een 'laissez-faire'-houding. Ook voor de individuele leraar geldt dat een 'laissez-faire'-houding in elk geval minder inspanning vraagt dan een veeleisende houding. Daardoor leidt de interactie tussen leerlingen en leraren van nature naar een evenwicht dat beide groepen in een comfortzone brengt waarin de ambities relatief laag zijn (zie Figuur 3). Een externe invloed is nodig om deze preferentiestructuur te wijzigen, waarbij (onder meer) kan gekeken worden naar gestandaardiseerde toetsen, aldus Elstad (2006, 2009).

		TEACHER	
		High academic ambitions	Low academic ambitions
STUDENT	High intensity	Student: 3 Teacher: 3	Student: 1 Teacher: 5
	Low intensity	Student: 5 Teacher: 1	Student: 2 Teacher: 2

Zone of comfort

←

Figuur 3 - Pay-off matrix (Elstad, 2009)

Empirische resultaten

De empirische evidentie rond dit effect komt voor een groot deel uit Duitsland. In een eerder vermeld onderzoek op basis van de resultaten van PISA-E (Jürges, Richter, et al., 2005) blijkt dat

wiskundeleraren van de *Hauptschule* in deelstaten met eindexamens van mening zijn dat ze een meer gedisciplineerde leeromgeving creëren en dat ze meer druk zetten op leerlingen om goed te presteren. Bovendien blijkt ook dat deze leraren meer frequent het leerproces van alle leerlingen maximaal trachten te ondersteunen en meer innovatieve oefeningen aanbieden. In *Gymnasia* zijn de resultaten minder gunstig in de deelstaten met centrale examens: daar blijkt dat wiskundeleraren minder vaak extra ondersteuning aanbieden aan leerlingen die daar nood aan hebben. Ook zijn deze leerlingen vaker van oordeel dat er te veel van hen gevraagd wordt. Daartegenover staat dat ouders van leerlingen in zowel het *Gymnasium* als in de *Realschule* in deelstaten met centrale examens een meer positief beeld hebben van de school en minder vaak hun ontevredenheid uiten over de school. Ze gaan ook vaker akkoord met de stelling dat leraren voldoende inspanningen leveren. Tot slot rapporteren leerlingen in de *Realschule* minder herhalingen tijdens de lessen in de deelstaten met centrale eindexamens (Jürges, Richter, et al., 2005).

In een andere eerder vermelde studie van dezelfde onderzoeksgroep (Jürges et al., 2009) wordt aangetoond dat wiskundeleraren in de *Hauptschule* meer actief en prestatiegericht zijn in deelstaten met centrale eindexamens. Deze verschillen tussen aanpak van leraren in *Hauptschule* worden evenwel niet gevonden bij leraren wetenschappen noch bij leraren in het *Gymnasia*. Zowel wiskunde- als wetenschapsleraren in verschillende schooltypes geven in deelstaten met centrale eindexamens vaker aan veel zorg te dragen voor het leerklimaat en voor discipline (Jürges et al., 2009).

Een opmerkelijk resultaat uit een andere Duitse studie is dat leraren en leerlingen in deelstaten met centrale eindexamens vaker 'partners' waren die een gezamenlijke uitdaging tegemoet gingen in het leerproces. Dit effect was nog sterker in de deelstaat waarin centrale eindexamens nog maar net waren geïmplementeerd (ten koste van een *school-based* summatieve beoordeling) (Van Ackeren et al., 2012). In de studie van Jäger et al. (2012) werd vastgesteld dat leraren de examenresultaten aanwendden als een persoonlijke vorm van feedback.

Een studie in de staat New York beschreef een bepaalde case waarbij een verplichte toets voor sociale wetenschappen in *grade 5* met een low-stakes karakter een positief effect had op het lesgeven van leraren. Door leerlingen via de gestandaardiseerde toets te evalueren op de materialen die ze aanbrachten om een bepaald argument in een discussie naar voren te brengen, gaf deze toets aanleiding tot het aanleren van onderzoeksvaardigheden en andere 'hogere orde'-vaardigheden (Bolgatz, 2006). Een andere studie keek naar de reacties van leraren naar aanleiding van een beleidsverandering in Pennsylvania. Deze kwalitatieve studie onderzocht hoe 97 leraren in 11 scholen - waarvan negen scholen door de betreffende toets werden gediagnosticeerd als '*in need of improvement*' - reageerden op de implementatie van een verplichte toets. Eén van de meest voorkomende 'categorieën van reacties' was het inzetten van *state tests* bij leerlinggerichte activiteiten, zoals het identificeren van leerlingen die bijzondere remediëring nodig hebben, het vergelijken van huidige scores met vroegere toetsscores (cf. leerwinst), het her-samenstellen van klassen volgens prestatieniveau (zodat homogene groepen ontstaan), en de resultaten over hun individuele prestatie delen met de leerlingen (Beaver & Weinbaum, 2015).

Bishop (1998b) onderzocht de resultaten van TIMSS, IAEP en de New Yorks States Regents Examinations en stelt dat in regio's met curriculumgebonden eindexamens de leerprestaties van leerlingen significant hoger waren dan in andere regio's. Regio's met centrale eindexamens die gebaseerd zijn op het curriculum worden gekenmerkt door de aanwezigheid van hogere

standaarden voor het leraarschap, hogere lonen voor leraren en door een betere scholing van leraren in het secundair onderwijs in het vak dat zij onderwijzen. Daarnaast wordt in de scholen in deze regio's meer voorbereiding gewijd aan wiskunde en wetenschappen en wordt er meer geïnvesteerd in infrastructuur voor wetenschappen (labo's en vaklokalen). Tot slot geven de leerlingen in scholen in deze regio's vaker aan dat het memoriseren van de leerstof niet de beste manier is om de leerstof onder de knie te krijgen, alsook dat wetenschappelijke experimenten worden uitgevoerd (Bishop, 1998b). Het gaat dus veelal om indirecte effecten en samenhangen zonder aangetoond causaal verband.

1.1.5 Impact op doelmatigheidsbeleving en motivatie van leerlingen

Maag Merki (2011) onderzocht of het inrichten van eindexamens voor wiskunde en Engels in het hoger onderwijs gevolgen met zich meebracht in de mate van zelfregulerend leren van studenten. Voor het onderzoek werden data verzameld bij studenten in twee Duitse deelstaten in de periode tussen 2007 en 2009, vlak voordat voor de eerste maal schriftelijke eindexamens werden afgenomen. Deze eindexamens tellen respectievelijk voor slechts 16% en 23% mee in het eindresultaat van de leerlingen, terwijl er geen publicatie van de resultaten op schoolniveau plaatsvindt: ze werden daarom beschouwd als low-stakes eindexamens. Doordat de examens in de deelstaat Hessen reeds in 2007 werden ingevoerd, terwijl Bremen pas in 2008 aan de beurt was, kon men in deze studie de gevolgen van de overgang van het klassieke systeem naar het centrale systeem onderzoeken. Uit de resultaten blijkt dat het implementeren van gestandaardiseerde eindexamens in wiskunde het zelfregulerend leren van studenten niet significant beïnvloed heeft. Wel werd opgemerkt dat gedurende de eerste drie jaar van deze implementatie niet enkel de interesse in wiskunde toenam bij de studenten in Hessen, maar dat ook de schoolse zelf-effectiviteit en het gebruik van elaboratiestrategieën (bijvoorbeeld samenvatting maken of in eigen woorden herhalen) bij deze studenten aangewakkerd werden. De resultaten op het examen Engels in Bremen tonen eveneens een significant positief effect aan van het invoeren van centrale eindexamen op het gebruik van elaboratiestrategieën door studenten om zich voor te bereiden op de examens. Daarnaast werd ook hier een toegenomen interesse voor Engels vastgesteld. Opvallend is dat studenten aangeven onzekerder te zijn over het slagen op het centraal eindexamen Engels dan op het klassieke examen in het verleden.

Daarnaast voerden Oerke, Maag Merki, Holmeier, en Jäger (2011) onderzoek naar de gevolgen die de implementatie van bovenstaande eindexamens met zich meebrengt op de succesbeleving van studenten en leraren. Uit de resultaten kon worden vastgesteld dat leraren succesvolle leerprestaties sneller toeschrijven aan het gegeven onderwijs en de geleverde inspanningen. In het jaar waarop studenten het eindexamen aflegden, kon worden opgemerkt dat leraren minder geneigd waren om de resultaten aan het gegeven onderwijs toe te schrijven. Men hield vaker rekening met de invloed van stress als verklaring voor de ondermaatse resultaten van hun leerlingen. De onderzoekers concludeerden dat studenten die zichzelf als 'succesvol' ervaren, een nog hogere succesbeleving ervaren als ze zware inspanningen hebben geleverd voor het examen. Een mogelijke verklaring hiervoor is dat studenten met goede resultaten op de schoolexamens eveneens hoge scores willen halen op de centrale eindexamens, waardoor ze zich extra goed zullen voorbereiden. Oerke et al. (2011) zijn dan ook van mening dat het invoeren van centrale examens voor een positieve bijstelling van attributiepatronen ('studeren' in plaats van 'niets doen') zorgt bij succesvolle studenten.

Er is evenwel ook evidentie die het tegendeel beweert, met name dat leerlingen door gestandaardiseerde toetsen eerder motivatie verliezen dan winnen. Deze tegenevidentie wordt besproken bij de onwenselijke neveneffecten op microniveau (1.2.3).

1.1.6 Betere evaluatie door leraren zelf

Bepaalde auteurs expliciteren de verwachting een leer(winst)monitoringssysteem een effect kan hebben op de manier waarop leraren hun eigen toetsen organiseren. Leraren zouden zich meer bewust kunnen worden van wat kwaliteitsvolle evaluatie is, met name in verband met selectie van inhoud, maar ook in verband met de constructie van toetsvragen die aan psychometrische standaarden voldoen (Cizek, 2005). Dit positieve effect zou zowel op de manier van toetsing (inzake betrouwbaarheid en validiteit), maar ook op de moeilijkheidsgraad van de toetsing (met name de gebruikte standaarden) van toepassing kunnen zijn (Neumann et al., 2011). Daarnaast kunnen gestandaardiseerde toetsen ook aanleiding geven tot het evalueren van meer complexe 'hogere orde'-vaardigheden als leerlingen deze ook moeten gebruiken bij de gestandaardiseerde toets (Shepard, 1992).

Uit empirische onderzoeksresultaten blijkt dat de gestandaardiseerde toetsen de eigen evaluaties van leraren zeker beïnvloeden. Het eerder vermelde onderzoek van Collins et al. (2010), waarbij men drie jaar na het wegvallen van Key Stage 2 toetsen in Wales het lesgeven (inzake wetenschappen) in het laatste jaar van het lager onderwijs vergeleek tussen Wales en Engeland, kwam tot enkele opvallende resultaten. Zo vond men dat de eigen toetsen van scholen in Wales (die de Key Stage 2 toetsen hadden vervangen) erg gelijkend waren aan de eerdere evaluatiepraktijk.

1.1.7 Betere onderbouwing schoolkeuze

Wanneer de resultaten van gestandaardiseerde toetsen openbaar worden gemaakt, biedt dit aan ouders de mogelijkheid om op een geïnformeerde basis een schoolkeuze te maken. Zeker in een context van een grote vrijheid van onderwijs, is vergelijkende informatie over de prestaties van de diverse scholen volgens bepaalde auteurs van cruciaal belang (Cizek, 2005; Hanushek & Wössmann, 2010).

1.1.8 Betere onderbouwing studiekeuze

Een leer(winst)monitoringssysteem verkleint de kans dat er mindere eisen worden gesteld aan lager presterende leerlingen, waardoor ze leerlingen een betere inschatting van hun eigen mogelijkheden geven. Hierdoor zijn leerlingen beter in staat om de ingangs- en studievereisten van studierichtingen in zowel secundair als tertiair onderwijs af te wegen tegenover hun eigen capaciteiten (de Lange & Dronkers, 2007).

1.2 Onwenselijke effecten op microniveau

1.2.1 Emotionele impact

Door veel auteurs wordt benadrukt dat gestandaardiseerde toetsen stress en angstgevoelens bij leerlingen in de hand werken, zeker wanneer ze door hen als high-stakes worden ervaren (met name wanneer ze een invloed hebben op de verdere studieloopbaan van de leerlingen) (Denscombe, 2000; White, 2013). Denscombe (2013) beschreef hoe leerlingen die hun eindexamen secundair

onderwijs dienden af te leggen deze toetsen niet enkel zagen als een beoordeling van hun leerprestaties, maar ook als een beoordeling van hun hele waarde als persoon. Hierdoor zouden toetsen ook nefast kunnen zijn voor het zelfvertrouwen van leerlingen (Meisels, 2000).

Deze veronderstelde negatieve emotionele impact is niet enkel onwenselijk op zich. Sterke stress en angst bedreigen eveneens de betrouwbaarheid van de meting van de toets: door deze negatieve emoties kan de leerling mogelijk niet ten volle zijn/haar kennis, vaardigheden of leerwinst tonen (Baird et al., 2013). Daarnaast leggen gestandaardiseerde toetsen ook een grote druk op leraren en directies in scholen (Jones et al., 1999; M. L. Smith, 1991). Deze druk komt niet enkel rechtstreeks vanuit de gestandaardiseerde toetsen, maar ook vanuit de eraan gekoppelde assumptie dat scholen altijd nog meer effectief kunnen worden met een gelijk aantal middelen. Zo wordt de vraag gesteld of scholen blijvend verbeterplannen kunnen opstellen en uitvoeren (Costea, Crump, & Amiridis, 2007).

In een eerder vermelde studie in Duitsland werden ook de emoties en attitudes van leerlingen ten opzichte van de getoetste leerdomeinen onderzocht. Men stelde vast dat leerlingen in deelstaten met centrale eindexamens significant vaker negatieve emoties zoals angst, boosheid en wanhoop rapporteren (in *Hauptschule* bijkomend prestatiedruk) (Jürges et al., 2009).

Verder geven leerlingen (in het laatste jaar secundair onderwijs) in de deelstaat met eindexamens in het onderzoek van Van Ackeren et al. (2012) aan dat er een grotere druk op hen wordt uitgeoefend om goed te presteren in het getoetste leerdomein wiskunde (door meer huiswerk en hogere les-intensiviteit). Leerlingen schatten zichzelf tegelijk ook hoger in op het vlak van zelfeffectiviteit en verwachten betere resultaten op wiskunde. Deze leerlingen geven ook vaker aan dat ze buitenschoolse, private bijlessen wiskunde volgen (Van Ackeren et al., 2012).

Hoewel toetsen in een oudere studie in de Verenigde Staten low-stakes waren, stelde men bij de afname van de gestandaardiseerde toetsen ook bij jonge leerlingen (in het kleuteronderwijs) reeds een verhoging van stress vast tijdens de toetsafname (Fleege, 1992). Een meer recente studie stelde een verschil vast tussen de angst die werd ervaren door kinderen in grade 3 tot grade 5 (leeftijd 6 tot 8 jaar) bij verschillende toetsen: de ervaren angst ligt significant hoger bij de high-stakes gestandaardiseerde toets (NCLB-toets) dan bij de gewone *classroom assessment*, maar slechts met een kleine effectgrootte. Een mogelijke alternatieve verklaring voor dit kleine verschil is dat deze leerlingen voor het eerst worden geconfronteerd met high-stakes testing, terwijl ze gewone classroom assessments al meer gewoon zijn (Segool, Carlson, Goforth, von der Embse, & Barterian, 2013). In Engeland leverde Putwain (2008) dan weer een tegengesteld bewijs: hier vond men de grootste toetsgerelateerde angst in de conditie waar leerlingen een low-stakes toets aflegden (opnieuw significant met een kleine effectgrootte). Deze onderzoeker geeft evenwel zelf aan dat andere verklaringen aan de basis van dit verschil kunnen liggen, met name de methodologische beperking dat leerlingen niet ad random aan de drie condities (low-stakes/mid-stakes/high-stakes) werden toegewezen: verschillende ervaringen in de scholen met eerdere toetsen kunnen dus ook een invloed hebben op de ervaring van angst (Putwain, 2008).

1.2.2 Teaching to the test

Teaching to the test wordt door verschillende auteurs genoemd als één van de onwenselijke neveneffecten van leer(winst)monitoringssystemen. Algemeen gesteld, houdt *teaching to the test* in dat leraren hun onderwijs- of evaluatiepraktijk aanpassen aan de toets om leerlingen een beter

toetsresultaat te laten bekomen. Hierdoor kan de aandacht voor andere belangrijke zaken verslappen. *Teaching to the test* is doorheen de vele literatuur evenwel een soort containerbegrip geworden, waar verschillende invullingen aan gegeven worden.

Verskillende invullingen van teaching to the test

Een vaak gehanteerde invulling is een vernauwing van het curriculum tot enkel die inhouden die getoetst worden of een overmatige beklemtoning van deze leerinhouden die ten koste gaat van niet-getoetste inhouden (bijvoorbeeld Au, 2007; Baird et al., 2013; Janssens et al., 2014; Srikantaiah, Zhang, & Swayhoover, 2008). Het kan gaan om een verschuiving van aandacht tijdens het lesgeven van het ene (niet-getoetste) leerdomein naar een ander (getoetst) leerdomein of om een verschuiving binnen een bepaald leerdomein naar de vaak getoetste kennis (ten koste van dieperliggende vaardigheden) (Au, 2007). Soms wordt deze vorm van *teaching to the test* niet als nadelig ervaren, bijvoorbeeld wanneer het betekent dat leraren meer focussen op dat wat de kern van het onderwijs behoort te zijn: *“Ironically, when curriculum and assessment are in alignment, teaching to the test is exactly what teachers should be doing because the test or assessment will contain a sampling of questions from the curriculum”* (Goodman & Hambleton, 2005, p. 94). Teaching to the test wordt in dit geval *‘teaching to the standard’* en wordt als een voordeel van gestandaardiseerde toetsen beschouwd (Cizek, 2005; Ver Eecke, 2004).

Een andere vorm van vernauwing van het curriculum - en meteen de tweede invulling van *teaching to the test* - houdt in dat niet zozeer andere leerdomeinen, maar wel andere vaardigheden binnen deze leerdomeinen worden beklemtoond. Zo kunnen leer(winst)monitoringssystemen ertoe leiden dat leraren vooral inzetten op het aanleren van feiten en procedures, eerder dan op meer fundamentele vaardigheden zoals conceptueel denken of complexe probleemoplossende vaardigheden (Au, 2007; Shepard, 1992). Daartegenover staat de bewering van Cizek (2005) dat zogenaamde ‘hogere orde’-vaardigheden in de Verenigde Staten net omwille van gestandaardiseerde toetsen te vroeg worden aangeboden, omdat de toetsen in verschillende staten op dit vlak te uitdagend en te moeilijk zijn.

Bij een derde invulling van *teaching to the test* wordt bedoeld dat de manier waarop leerinhouden worden aangeboden, verengd wordt. Dit betekent dat leraren inhouden enkel op die manier aanbieden zoals ze ook in de toetsen worden gepresenteerd (bijvoorbeeld eerder korte teksten met een beperkt aantal korte vragen, dan langere teksten met meer beschouwende vragen binnen ‘begrijpend lezen’) (Shepard, 1992). Gestandaardiseerde toetsen met meerkeuzevragen kunnen ertoe leiden dat ook het onderwijs zich vernauwt tot enkel korte en vaak gedetailleerde vragen tijdens het lesgeven (Boyd, 2008; M. L. Smith, 1991).

Een vierde invulling van *teaching to the test* houdt in dat leraren met veel zorg de leerlingen voorbereiden op de toetscondities. Leerlingen worden ‘strategisch gecoacht’ in bijvoorbeeld het invullen van meerkeuzevragen en het beantwoorden van type-vragen waardoor dit ten koste gaat van onderwijstijd die aan andere zaken kan besteed worden (Elstad, 2009; Redden & Low, 2012). Het gaat hier om een oneigenlijke vorm van toetsvoorbereiding. Bovendien gebeurt dit niet enkel tijdens de schooluren, ook ouders spelen hierop in: er is immers een bloeiende industrie van toetsboekjes die door ambitieuze ouders worden aangekocht om leerlingen goed voor te bereiden (Standaert, 2014).

Een vijfde invulling van *teaching to the test* houdt een vernauwing in van de leerlingen die in de focus worden geplaatst: het onderwijs concentreert zich dan voornamelijk op die leerlingen die zich net boven of onder een bepaalde standaard of een cut-off point bevinden (Booher-Jennings, 2005; Reback, 2008). Er worden daardoor weinig inspanningen geleverd om het leerproces te stimuleren van leerlingen die geacht worden sowieso niet in aanmerking te komen voor een goede score (Kesler, 2013). Verschillende auteurs spreken in dit geval van een focus van het onderwijs op zogenaamde '*bubble kids*'; ten koste van andere leerlingen (Booher-Jennings, 2005; Srikantiah et al., 2008).

Ten slotte wordt *teaching to the test* ook gebruikt om aan te geven dat leraren de items die in de toets zullen voorkomen (of worden verwacht) identiek overnemen in hun lessen. Hier gaat het eerder om '*teaching the test*' dan om '*teaching to the test*'. Aangezien dit neveneffect voornamelijk op schoolniveau wordt geobserveerd, wordt dit besproken in de sectie over onwenselijk strategisch gedrag op mesoniveau (zie 1.4.1).

Empirische resultaten

De ruime basis aan empirie rond dit thema komt voor een belangrijk deel uit de context van high-stakes toetsen in de Verenigde Staten. Een grootschalige kwalitatieve meta-analyse identificeerde 45 studies uitgevoerd in de Verenigde Staten tussen 1992 en 2006, die evidentie (of tegen-evidentie) gaven voor *teaching to the test* als gevolg van gestandaardiseerde toetsen (Au, 2007). Deze meta-analyse resulteerde in 34 studies waarin evidentie werd verzameld voor een vernauwing van het curriculum, maar stuitte ook op 14 studies waarin evidentie werd verzameld voor een tegengesteld effect waarbij gestandaardiseerde toetsen ertoe hadden geleid dat net een breder curriculum werd aangeboden (om het volledige spectrum van het examen te kunnen omvatten, cfr het eerder genoemde *teaching to the standard*). Sommige studies vonden evidentie van beide situaties, zoals Anagnostopoulos (2003), Luna en Turner (2001) en Segall (2003). Daarnaast vond Au (2007) niet alleen 24 studies die evidentie gaven voor het meer gefragmenteerd aanbieden van leerstof (waardoor geen 'hogere orde'-vaardigheden worden aangeleerd), maar tevens tien studies waar net een meer geïntegreerd aanbod werd opgebouwd als gevolg van gestandaardiseerde eindexamens (Au, 2007). Ten slotte bleek dat 32 studies evidentie hadden verzameld omtrent een verhoogd gebruik van leraargerichte werkvormen, terwijl in zes studies kon geconcludeerd worden dat de eindexamens aanleiding hadden gegeven tot het gebruik van meer leerlinggerichte werkvormen. Slechts twee van de 45 gereviewde studies (Gradwell, 2006; Grant, 2003) vonden een positief noch negatief effect inzake curriculum of werkvormen (Au, 2007).

In een studie in de staat Tennessee gaf 79,3% van de bevroagde leraren Engels in het hoger secundair onderwijs aan dat ze lestijd vrijhouden voor specifieke voorbereiding betreffende het high-stakes eindexamen (*English II Gateway Examination*). 46,7% van de leraren Engels spendeerde zelfs meer dan drie maanden lestijd aan deze examenvorbereiding. In tegenstelling tot de hypothese van dit onderzoek was er geen verschil inzake de frequentie van het gebruik van leraar- of leerlinggerichte werkvormen tussen de leraren die niet, matig of veel tijd vrijmaakten voor deze specifieke voorbereiding. Toch vond ruim negen op tien van deze leraren Engels dat het 'belang om leerlingen te helpen een score te behalen die hen toelaat een diploma te behalen' en het 'belang om mijn school goede eindscores te laten behalen' een grote invloed hadden op de werkvormen die ze hanteerden (respectievelijk 96,4% en 90,2%). Leraren die geen tijd hadden vrijgemaakt voor

specifieke toetsvoorbereiding bleken ook minder vaak akkoord te zijn met deze beide stellingen (respectievelijk 'slechts' 85,7% en 74,5%) (Vogler, 2006a, 2006b).

Een kwalitatieve studie waarin zes net afgestudeerde leraren werden gevolgd tijdens hun eerste jaar in het onderwijs, kwam tot de conclusie dat voor elk van de respondenten gold dat de high-stakes toetsen in hun staat (New York) sterk hun manier van lesgeven hadden beïnvloed (waarbij ze lespraktijken moesten toepassen die ze zelf van twijfelachtig allooi vonden). Bovendien leidde deze toetscultuur tot veel stress en vormde ze een bedreiging voor de professionele identiteit zoals die werd vormgegeven tijdens de lerarenopleiding (Costigan, 2002).

In een andere kwalitatieve studie die eerder besproken werd, stelde men vast dat *state tests* in verschillende scholen fungeren als startpunt voor verder onderzoek en planning, bijvoorbeeld door een onderzoek van een 'data team' in de school of door tussentijdse toetsen om al op eerdere momenten een overzicht te krijgen van de prestaties van leerlingen. Daarbij werden de toetscondities zo identiek mogelijk gekopieerd (inzake terminologie, itemtypes en berekening van individuele scores). Daarnaast kunnen de *state tests* gebruikt worden als startpunt voor onder meer het gestandaardiseerd inplannen van toetsvoorbereiding-activiteiten, het afstellen van het gebruikte curriculum op de vereisten van de toets en het beperken van lestijd in niet-getoetste leerdomeinen (Beaver & Weinbaum, 2015).

Enkele studies uit Duitsland, waaronder de reeds eerder vermelde studie van Van Ackeren et al. (2012), kunnen hier eveneens vermeld worden. De vragenlijst van leerlingen in deze studie geeft verschillen aan inzake gebruikte werkvormen tijdens de wiskundelessen: leerlingen in de deelstaat met centrale eindexamens rapporteren minder interactieve werkvormen (groepswork, discussies) en minder invloed op de keuze van behandelde thema's. De verschillen inzake wiskundeonderwijs werden evenwel niet geconstateerd bij de lessen Duits of 'advanced' biologie, met uitzondering van een grotere hoeveelheid huiswerk. Uit de vragenlijst voor leraren in dezelfde studie wordt geconcludeerd dat leraren in de deelstaat met centrale eindexamens vaker een standaard-criterium gebruiken bij hun eigen evaluaties, terwijl in de deelstaat zonder centrale eindexamens vaker een zelf-criterium of een norm-criterium wordt gehanteerd. Hoewel er een verschil is inzake specifieke toetsvoorbereiding, vinden de onderzoekers in deze studie geen evidentie voor *teaching to the test* als vernauwing van het curriculum (Van Ackeren et al., 2012).

Brozo en Hargis (2003) wezen uit dat om leerlingen optimaal voor te bereiden op deze toetsen, leraren van de getoetste leerdomeinen in mindere mate akkoord gingen met de bewering dat ze alle curriculum-doelen hadden nagestreefd, in vergelijking met hun collega's van niet-geëxamineerde leerdomeinen. Dit verschil was statistisch significant (met sterke effectgrootte). Deze vorm van *teaching to the test* kan een 'crowding out' effect veroorzaken waarbij de intrinsieke motivatie van leerlingen in het gedrang kan komen. Een multilevel analyse toonde bovendien aan dat de school weinig invloed heeft op de mate waarin leraren overgaan tot vernauwing van het curriculum. Ook kon worden aangetoond dat *teaching to the test* minder voorkwam bij leraren met een hogere zelfeffectiviteit, maar vaker voorkwam bij leraren die eerder onzeker waren over de vereisten die tijdens de gestandaardiseerde toets worden gesteld (Jäger, Maag Merki, Oerke, & Holmeier, 2012).

Ten slotte is er evidentie voor dit onwenselijk neveneffect uit het eerder vermeld onderzoek van Collins et al. (2010). Uit deze studie bleek dat er zowel in Engeland als in Wales een sterke neiging was naar *teaching to the test*, waarbij belangrijk geachte, maar niet-getoetste vaardigheden weinig

aandacht kregen. Bovendien was er minder ruimte voor leerdomeinen die niet getoetst worden, zoals aardrijkskunde, geschiedenis “*and certainly no more art or anything remotely creative like that*” (Engelse leraar van het laatste jaar lager onderwijs, geciteerd door Collins et al., 2010, p. 280). De extra tijd geïnvesteerd in de getoetste leerdomeinen werd echter niet besteed aan effectieve lestijd, maar wel aan specifieke toetsvoorbereiding (bijvoorbeeld door het aanbieden van *practice papers* die weinig bijdragen aan de domeinspecifieke kennis of vaardigheden van de leerlingen). In Engeland werd evenwel statistisch significant meer gebruik gemaakt van specifieke voorbereidende documenten, zoals eerdere Key Stage 2 toetsen. Hoewel dit minder het geval was in Wales dan in Engeland, bleek ook in Wales (waar deze Key Stage 2 toetsen drie jaar voor het onderzoek waren afgeschaft) nog in zekere mate *teaching to the test* overeind, onder meer omdat schoolbesturen zelf vervangende toetsen hadden ingelegd. Toch rapporteert men in Wales meer tijd voor praktische en onderzoeksgerichte lessen (een deel van het curriculum dat niet direct door de Key Stage 2 toetsen werd geëvalueerd), en minder *teacher-centered* werkvormen (Collins et al., 2010).

1.2.3 Negatieve impact op motivatie van leerlingen

In een eerder vermelde studie geven leerlingen in deelstaten met centrale eindexamens minder vaak aan hun huiswerk nauwgezet in te dienen. Bovendien zijn deze leerlingen minder vaak aandachtig tijdens de lessen, beleven ze minder plezier aan de lessen en bezitten ze een minder hoog zelfbeeld (Jürges et al., 2009). Daarnaast werd in twee eerder vermelde studies wel gesteld dat de leerprestaties positief worden beïnvloed door gestandaardiseerde eindexamens, maar rapporteerde men ook dat leerlingen in de deelstaten met gestandaardiseerde eindexamens minder gemotiveerd zijn voor de getoetste leerdomeinen (Van Ackeren et al., 2012), minder plezier beleven aan wiskunde en het vaker een saai vak vinden, wat de auteurs inspireerde tot de treffende titel: “*CEEs [central exit exams] increase performance... but take the fun out of mathematics*” (Jürges & Schneider, 2010).

1.2.4 Negatieve impact op studieloopbaan

Scholen hebben er baat bij om, in het geval van een high-stakes monitoringssysteem voor scholen, de leerlingen pas aan de toets te laten deelnemen indien ze daar klaar voor worden geacht. Daardoor kunnen gestandaardiseerde toetsen leiden tot een verhoogd aantal beslissingen om leerlingen een jaar te laten overdoen (de Wolf & Janssens, 2007; Janssens et al., 2014). In een meer extreem geval, zouden leerlingen zelfs door de school kunnen worden ontmoedigd om verder te studeren, of van school kunnen worden gestuurd omdat ze een goede score voor de school zouden bedreigen (Amrein-Beardsley, Berliner, & Rideau, 2010; Horn, 2005; Ryan, 2004; Sloane, Oloff-Lewis, & Kim, 2013). Zo vond een studie in Texas dat de resultaten op de Texas Assessment of Academic Skills (TAAS) een correlatie vertoonden met drop-out rates, voornamelijk onder zwarte en Hispanic leerlingen (Haney, 2000).

1.2.5 Onwenselijk strategisch gedrag van leerlingen

Extrinsieke motivatie, zoals het behalen van goede cijfers, zou leiden tot meer oneigenlijk gedrag van leerlingen (Anderman, Griesinger, & Westerfield, 1998). Hoewel het spiekgedrag van leerlingen vooral vooral het kopiëren van antwoordgedrag van andere leerlingen omvat (Skorupski & Egan, 2012), bereiden leerlingen zich meer strategisch voor indien de leerdoelen centraal geëxamineerd worden door bijvoorbeeld toetsen van voorgaande jaren te maken. Dit laatste effect zou zich

voornamelijk voordoen in de eerste jaren na de implementatie van gecentraliseerde examens, zoals blijkt uit de hoge score op dit item in een deelstaat waar net centrale examens waren ingevoerd (Van Ackeren et al., 2012).

1.2.6 Werkt ongelijkheid in de hand

Zoals eerder vermeld werd in sectie 1.1.1, geven verschillende auteurs aan dat bepaalde groepen van leerlingen door gestandaardiseerde toetsen worden benadeeld. Dit argument staat lijnrecht tegenover de bewering dat gestandaardiseerde toetsen een instrument kunnen zijn om bias en ongelijkheid in evaluatie te vermijden. Meer bepaald wordt verondersteld dat de vorm van de meeste gestandaardiseerde toetsen (met name: vaak schriftelijk, vaste antwoordcategorieën) bepaalde studenten, bijvoorbeeld met een bepaalde culturele of etnische achtergrond, onterecht benadeelt omdat ze moeilijker met het toetsformat overweg kunnen (Gipps & Stobart, 2009; Klenowski, 2014). Bovendien zouden leerlingen met een minder gunstige achtergrond (zoals een lager zelfconcept) benadeeld worden, omdat de zelfeffectiviteit van leerlingen inzake academische prestaties de toetsscores onrechtmatig beïnvloedt (Rogers-Ard, Knaus, Epstein, & Mayfield, 2013).

Op basis van een overzicht van verschillende studies ter zake, trekken Neumann et al. (2011) in twijfel of gestandaardiseerde examens werkelijk tot een betere evaluatie leiden ten opzichte van de eigen evaluaties van leraren, aangezien de correlaties immers erg groot zijn in de meeste onderzochte studies. Toch geven enkele studies evidentie voor de bewering dat gestandaardiseerde eindexamens bias in evaluatie door leraren kan wegnemen. Uit een andere studie wordt afgeleid dat (éénmalige) gestandaardiseerde toetsen een minder goed beeld geven van de kennis en vaardigheden van leerlingen dan een meer informele en holistische evaluatie op basis van observaties en toetsen van leraren (Allal, 2013). Door enkel rekening te houden met de scores van gestandaardiseerde toetsen, wordt de waarde van het dagelijkse menselijke interactiepatroon tussen leraren en leerlingen in het complexe sociale klassysteem ondergewaardeerd (Haertel, 1999). Durant (2003) sluit hierbij aan en stelt dat leerlingen die geëvalueerd worden door de leraar minder kwetsbaar zijn voor storende factoren, zoals het hebben van een minder goede dag.

1.2.7 Impact op professionaliteit en motivatie van leraren

Vanuit de Zelf-Determinatie Theorie wordt geargumenteed dat een te sterke inperking van de vrijheid van leraren om zelf prioriteiten te kiezen en toe te voegen aan een curriculum, de motivatie van leraren sterk kan aantasten (Deci & Ryan, 2000) en als gevolg daarvan de prestaties van de leraren negatief kan beïnvloeden (Jäger et al., 2012). Gestandaardiseerde toetsen kunnen er mogelijk toe leiden dat de status van leraren als experts inzake leren, instructie en evaluatie wordt aangetast (Van Ackeren et al., 2012). Ook de professionele identiteit van leraren kan worden bedreigd door gestandaardiseerde toetsen, omdat deze hen in een richting sturen die mogelijk niet strookt met hun eigen opvattingen over goed onderwijs (Luna & Turner, 2001; Shepard, 1992; Smagorinsky, Lakly, & Johnson, 2002). Dit gevoel van miskenning van de professionele identiteit kan ertoe leiden dat bekwame leraren het onderwijs verlaten voor andere sectoren (Shepard, 1992).

1.3 Positieve effecten op mesoniveau

1.3.1 Betere interne kwaliteitszorg

Schoolteams krijgen vanuit een leer(winst)monitoringssysteem een belangrijke input van informatie over de kwaliteit van een aantal van hun processen. De resultaten van gestandaardiseerde toetsen kunnen een indicatie geven van ‘*areas of need*’ in een school, en zo ook aanleiding zijn om verdere data te verzamelen, om een verbetertraject op te zetten of om een discussie op te zetten over het beleid en de huidige praktijken in de school (J. A. Marsh et al., 2005; Stecher et al., 2008).

Ook scholen die niet op eigen initiatief gebruik maken van de data van een leer(winst)-monitoringssysteem kunnen door de gestandaardiseerde toetsen sterker worden aangezet om hun eigen kwaliteit te bewaken en te optimaliseren. De veronderstelde onderliggende mechanismen zijn onder meer de verhoogde monitoring van scholen door interne en externe stakeholders (en de versterkte ‘marktwerking’) die door de openbaarheid van resultaten wordt mogelijk gemaakt (Hanushek & Wössmann, 2010); of de dreiging met sancties voor scholen waarin de leerprestaties van leerlingen wijzen op onvoldoende onderwijskwaliteit, zoals in verschillende staten van de Verenigde Staten gangbaar is (Chakrabarti & Schwartz, 2013).

Wang et al. (2006) zijn evenwel niet overtuigd dat leer(winst)monitoringssystemen tot een verbetering van de kwaliteit van scholen kunnen leiden. De toetsen bieden daarvoor immers te weinig handvaten aan, aangezien de (veronderstelde) onderliggende redenen voor vele problemen in het onderwijs - zoals het *“fundamental misdesign of schools, lack of qualified teachers, and the instability of the families and communities from which students come”* (Wang et al., 2006, p. 310) - van een zodanig complexe aard is dat ze niet door de resultaten van gestandaardiseerde toetsen kunnen worden opgelost.

1.3.2 Betere inzet professionele ontwikkeling van schoolteams

Cizek (2005) verwacht dat na de implementatie van een leer(winst)monitoringssysteem scholen zich meer zullen focussen op de professionele ontwikkeling van leraren in het kader van praktijken die bijdragen tot de leerprestaties van leerlingen (Cizek, 2005). Scholen kunnen door de resultaten van hun leerlingen beter inschatten op welke domeinen men zich verder dient te ontwikkelen (J. A. Marsh et al., 2005). Bovendien kan de schooldirectie de professionele ontwikkeling meer gericht inzetten op de leraren die dit nodig blijken te hebben, en op de specifieke domeinen waarop dit meest vereist is (Loeb, 2013). Ook verwacht Cizek (2005) dat het aanbod inzake professionele ontwikkeling zal evolueren naar een aanbod dat meer gelinkt is aan het curriculum en meer gericht is op betere onderwijsvaardigheden en op een betere vakdomeinexpertise.

1.3.3 Betere inzet schoolinspecties

Indien scholen zich dienen te verantwoorden door middel van schooldoorlichtingen, is het belangrijk dat deze doorlichtingen op een geïnformeerde en betrouwbare wijze geschieden. Daarom wordt soms aangenomen dat betrouwbare informatie over de output van scholen in termen van leerprestaties, een cruciaal gegeven is in de externe kwaliteitszorg van een onderwijssysteem (Parkes & Stevens, 2003).

1.4 Onwenselijke effecten op mesoniveau

1.4.1 Onwenselijk strategisch gedrag van scholen (cheating)

Bij leer(winst)monitoringssystemen die extern worden opgelegd, bestaat de mogelijkheid dat leraren of scholen verleid worden tot onwenselijk strategisch gedrag (de Wolf & Janssens, 2007). Dit onwenselijk strategisch gedrag kan zich voordoen op zowel individueel initiatief van leraren als op een meer geïnstitutionaliseerd niveau (het schoolniveau). Dit verondersteld onwenselijk neveneffect bestrijkt bijgevolg zowel het micro- als het mesoniveau.

In een artikel over *cheating* in de Amerikaanse context, als reactie op de high-stakes toets *Arizona's Instrument to Measure Standards* (AIMS), onderscheiden Amrein-Beardsley et al. (2010) drie niveaus van '*cheating*'. '*Cheating in the first degree*' omhelst bewust en gepland vervalsen en wordt daardoor doorgaans als het meest onwenselijk beschouwd, terwijl '*cheating in the second degree*' vaak meer subtiel en toevallig is en dus niet noodzakelijk bedoeld of op voorhand gepland. De auteurs spreken van '*cheating in the third degree*' wanneer sprake is van een verstoring van de resultaten door onverschilligheid, roekeloosheid of onachtzaamheid, zonder de intentie de resultaten te vervalsen (Amrein-Beardsley et al., 2010). De verschillende voorbeelden van ongewenst strategisch gedrag worden volgens deze drie niveaus besproken. Uit het onderzoek van Amrein-Beardsley et al. (2010) bleek dat *cheating in the second degree* het meest voorkwam, gevolgd door *cheating in the third degree*. Zelden gaven leraren toe dat ze zich lieten verleiden tot *cheating in the first degree*.

Cheating in the first degree

Jacob en Levitt (2003) ontwikkelden een algoritme dat hen in staat stelde om ongewone antwoordpatronen op gestandaardiseerde toetsen in kaart te brengen. Het onderzoek wees uit dat in vier tot vijf percent van de klassen in lagere scholen gefraudeerd wordt.

Er zijn een aantal voorbeelden van bewust en gepland vervalsen. Zo is er onderzoek dat niet enkel het aanvullen of veranderen van de antwoorden van leerlingen op de gestandaardiseerde toetsen aan het licht brengt, maar dat eveneens het bestaan van fraude met examenresultaten aantoonst (Hoxby, 2012). Daarnaast vertekenen leraren bewust de resultaten door de gestandaardiseerde instructies niet op te volgen of door leerlingen ongeoorloofde hulp te bieden tijdens de toetsafname (Amrein-Beardsley et al., 2010; Baird et al., 2013). Ook het uitsluiten van bepaalde leerlingen voor de toetsen om een beter resultaat op schoolniveau te bekomen, onder meer door zwakke leerlingen te motiveren op de dag van toetsafname thuis te blijven, te schorsen of zelfs naar een andere school te verwijzen, blijkt een gangbare praktijk te zijn (Haney, 2000; Horn, 2005; Sloane et al., 2013). Dit verschijnsel wordt in de Angelsaksische literatuur ook '*reshaping the test pool*' genoemd (Janssens et al., 2014). Een onwenselijk gevolg van *reshaping the test pool* is dat relatief zwakke leerlingen worden gedemotiveerd om hun schoolloopbaan door te zetten. Dit wordt bevestigd door een analyse van de *drop-out* cijfers van leerlingen tijdens het hoger secundair onderwijs, waarin men vaststelde dat de tien staten in de Verenigde Staten waar de *drop-out* cijfers het hoogst lagen ten tijde van het onderzoek, allen gebruik maakten van high-stakes eindexamens om na te gaan of een leerling een diploma krijgt of kan doorstromen naar een bepaalde vervolgstudie. In tegenstelling vond men bij vijf van de tien staten met de laagste *drop-out* cijfers dat er geen verplichte gestandaardiseerde toetsen waren, terwijl de gestandaardiseerde toetsen in de andere vijf staten enkel low-stakes waren - er moet hierbij evenwel worden opgemerkt dat deze correlatie niet noodzakelijk op een oorzakelijk

verband duidt. Volgens de auteurs kunnen minder presterende leerlingen reeds vroeg gedemotiveerd worden door de school om deel te nemen aan gestandaardiseerde toetsen waarvan de resultaten ook op schoolniveau high-stakes zijn (Clarke, Haney, & Madaus, 2000). Een gelijkaardige analyse bevestigt dat bij de 20% laagst presterende leerlingen ('*bottom quintile*') de kans op drop-out 25% hoger ligt in staten met eindexamens in vergelijking met staten zonder deze high-stakes eindexamens (Jacob, 2001).

Tot slot kunnen scholen er bewust voor kiezen om bij metingen van toegevoegde waarde, waarbij uitgegaan wordt van een begin- en opvolgmeting, leerlingen op de beginmeting relatief slechte scores te laten behalen. Op die manier wordt de leerwinst van leerlingen en bijgevolg de toegevoegde waarde van de school artificieel groter gemaakt dan ze in werkelijkheid is (Castellano & Ho, 2013).

Cheating in the second degree

Leerlingen motiveren om bepaalde oefeningen nog eens opnieuw te proberen (Amrein-Beardsley et al., 2010), is een voorbeeld dat aansluit bij 'tweede-orde vervalsingen'. Ook het geven van extra hulp aan leerlingen, zoals het verklaren van moeilijke woordenschat, het toestaan van woordenboeken en het gebruiken van didactische posters die de leerlingen kunnen helpen bij de toetsafname bewust laten hangen, vallen onder deze categorie. Leraren kunnen leerlingen tevens herinneren aan processen die tijdens de lessen werden aangeleerd, waardoor leerlingen op het juiste pad worden gezet bij één of meer toetsitems. Het voorlezen van toetsitems of het uitklaren van wat precies verwacht wordt door de ontwikkelaars van de toets helpt de leerlingen ook in bepaalde mate vooruit tijdens de toetsafname (Amrein-Beardsley et al., 2010). Tot slot blijkt dat in sommige scholen, desondanks de instructies, bepaalde leerlingen meer tijd krijgen voor de toetsafname (Amrein-Beardsley et al., 2010; Baird et al., 2013);

Hoewel de resultaten uit het onderzoek van Amrein-Beardsley et al. (2010) door de lage responsgraad moeilijk te veralgemenen zijn, gaven ruim 500 *educators* in deze studie in Arizona (Verenigde Staten) aan dat ze leerlingen hadden aangemoedigd om op bepaalde testitems een ander antwoord te geven, terwijl ruim 400 *educators* toegaven dat ze leerlingen meer tijd geven voor de toetsafname dan was toegestaan door de formele tijdslimiet. Andere regelmatig voorkomende voorbeelden in dit onderzoek zijn het begeleiden ('*coaching*') van leerlingen tijdens de toetsafname en het niet verwijderen van ongeautoriseerde hulpmiddelen (Amrein-Beardsley et al., 2010).

Cheating in the third degree

Soms worden deze 'derde-orde vervalsingen' ook omschreven als onvrijwillige vervalsingen, omdat ze veroorzaakt worden zonder de geplande intentie om hierdoor de resultaten te vertekenen. Voorbeelden hiervan zijn onder meer '*teaching to the test*' (zie 1.2.2), het gebruiken van vroegere toetsen om leerlingen voor te bereiden op de volgende toetsen en het aanbieden van items waarvan men verwacht dat ze ook tijdens de gestandaardiseerde toets zullen worden gevraagd (Amrein-Beardsley et al., 2010).

1.4.2 Rankings

Het lijkt bijna een automatisch gevolg van openbaarheid van toetsresultaten op schoolniveau dat, indien de overheid het zelf niet doet, media rankings zullen opmaken van scholen op basis van de prestatie van leerlingen (de Wolf & Janssens, 2007; Elstad, 2009; Sloane et al., 2013). Daarbij gaat

aandacht naar de 'beste scholen', maar heel vaak ook naar de 'slechtste scholen'. De rankings en daaraan gekoppelde '*naming-and-blaming*-cultuur' (Elstad, 2009) waarbij scholen onderaan de ranglijst door de media door het slijk worden gehaald, zoals in vele Angelsaksische onderwijssystemen het geval is, zijn onwenselijke neveneffecten van gestandaardiseerde toetsen. Volgens tegenstanders zijn rankings demoraliserend voor scholen en leraren (Elstad, 2009) en werken ze sociale ongelijkheid en competitie tussen scholen in de hand (Jürges, Büchel, et al., 2005; Sloane et al., 2013). Deze verhoogde competitiviteit vormt bovendien een bedreiging voor de samenwerking tussen scholen in scholengemeenschappen of andere samenwerkingsstructuren (Hadfield & Chapman, 2009).

Toch hoeft deze verhoogde competitiedrang tussen scholen niet negatief te zijn. Vanuit een economische benadering zien bepaalde auteurs voordelen in een competitie tussen scholen, aangezien die scholen tot het uiterste zou drijven om de beste kwaliteit van onderwijs te voorzien (Black & Wiliam, 2005; Thomas, Peng, & Gray, 2007).

1.5 Positieve effecten op macroniveau

1.5.1 Beter zicht op onderwijskwaliteit

Verschillende auteurs geven aan dat het wenselijk is dat het onderwijsbeleid de vinger aan de pols houdt wat de onderwijskwaliteit van scholen betreft door middel van het regelmatig nagaan van de leerprestaties van leerlingen. Een leer(winst)monitoringssysteem met uiteenlopende gestandaardiseerde toetsen is daartoe volgens Baird et al. (2013) het meest aangewezen middel.

1.5.2 Bijstellen van het curriculum

Wanneer een leer(winst)monitoringssysteem gelinkt is aan de eindtermen of leerplandoelen, kan het geheel van prestaties van leerlingen een overheid (of een onderwijskoepel) informeren over de haalbaarheid van bepaalde gestelde doelen binnen de leerdomeinen. Het initiatief van deze toetsing kan uitgaan van de verantwoordelijken voor het curriculum, maar kan ook onder druk van andere stakeholders ontstaan. Zo kan een leer(winst)monitoringssysteem volgens Cizek (2005) tot een grotere betrokkenheid van leraren op de curriculumdoelen leiden, aangezien leraren deze in vraag kunnen stellen. Op die manier kunnen leraren ertoe bijdragen dat deze doelen verfijnd en geoptimaliseerd worden.

1.5.3 Onderzoeksdoeleinden

De beschikbaarheid van de resultaten van gestandaardiseerde data kan een belangrijke voedingsbodem zijn voor academisch onderzoek. Er zijn zeer veel voorbeelden van studies die op basis van zulke resultaten hebben geleid tot een indringend inzicht inzake onder meer:

- de impact op leerprestaties van factoren die scholen niet (geheel) onder controle hebben, zoals het leereffect van leerlingen tijdens de vakantieperiodes (Lin, Hsieh, & Chen, 2015), de mobiliteit van leerlingen tussen scholen (Strand & Demie, 2007), ouderbetrokkenheid (Fan, 2001), (veranderende) leerlingenpopulatie (Driessen, 2002; Luyten & de Wolf, 2011) of schoolgrootte (Gershenson & Langbein, 2015; V. E. Lee & Loeb, 2000);
- de effectiviteit van bestaande (macro-)onderwijskundige praktijken, zoals zittenblijven (Goos, Van Damme, Onghena, Petry, & de Bilde, 2013);

- de impact op leerprestaties van factoren op mesoniveau, zoals professionele leergemeenschappen (Siguroardottir, 2010) of verschillende schooltypes (Sins & Van der Zee, 2015);
- de impact op leerprestaties van specifieke onderwijskundige vernieuwingen (microniveau), bijvoorbeeld in lees-onderwijs (Van Keer & Verhaeghe, 2005) chemie-onderwijs (Pentecost & Barbera, 2013), voor leerlingen met dyslexie (Gijsel & Bosman, 2010), of de impact van fysieke activiteiten tussen lessen (Kall, Malmgren, Olsson, Linden, & Nilsson, 2015). In Nederland kunnen de eigen projecten van scholen geëvalueerd worden door (onder meer) gebruik te maken van de resultaten van gestandaardiseerde toetsen (Heyma et al., 2015);
- de effectiviteit van leraargebonden factoren, bijvoorbeeld de impact van het gebruik van 'authentieke' werkvormen (Saye, 2013), de ervaring, het ras en het opleidingsniveau van leraren (Muñoz & Chang, 2007), en van de houding van leraren ten aanzien van dyslexie (Hornstra, Denessen, Bakker, van den Bergh, & Voeten, 2010) op de studieresultaten van leerlingen.

1.5.4 Publieke tevredenheid

Verschillende studies in de Verenigde Staten hebben aangegeven dat een meerderheid van het 'publiek' (ouders, bedrijfsleiders en politici) zich voorstander van gestandaardiseerde toetsen noemt (Phelps, 2005a; Driesler, 2001).

1.6 Onwenselijke effecten op macroniveau

1.6.1 Grotere segregatie

Sommige auteurs gaan ervan uit dat wanneer de resultaten van een leer(winst)monitoringssysteem openbaar beschikbaar zijn, dit zal leiden tot een verhoogde segregatie tussen sterke en zwakke scholen. Scholen die sterke leerprestaties kunnen voorleggen, zouden hierdoor sterke leerlingen kunnen aantrekken. Bovendien zouden sterk gekwalificeerde leraren ook de voorkeur kunnen geven aan scholen waar goede leerprestaties worden behaald (Ladd & Walsh, 2002). Bovendien wordt gesteld dat de ouders van leerlingen uit bevoorrechte milieus hun scholen sneller zullen aanspreken op de beschikbare resultaten en met een groter 'sociaal en onderwijskundig kapitaal' hun rol als actieve partner zullen kunnen vervullen. Scholen met een minder bevoorrechte instroom zouden zo nog meer achterstand kunnen oplopen (Reay, 2008; Redden & Low, 2012).

1.6.2 Financiële impact

Tegenstanders van gestandaardiseerde toetsen wijzen vaak op de zware financiële kost voor de ontwikkeling, afname en correctie van gestandaardiseerde toetsen in een leer(winst)-monitoringssysteem (White, 2013).

Ook Phelps (2000) stelt de financiële kost ruimer voor dan enkel de aankoop van testmaterialen en bijhorende diensten (correctieservice, ...) en brengt ook andere aspecten in rekening. Een eerste aspect waar rekening met dient gehouden te worden, is de tijdscomponent. Zo neemt een toetsafname tijd in beslag voor zowel de school, leraar als leerling. Leraren kunnen immers worden ingezet bij zowel het afnemen van de toetsen als bij het voorzien van feedback aan de leerlingen en de ouders. Ook wordt er personeel voorzien dat instaat voor het verzenden van de resultaten per e-

mail of post. In sommige gevallen wordt ook de tijd die leerlingen spenderen aan de voorbereiding en het afleggen de toets mee in rekening gebracht. De administratieve verwerking is eveneens een meerkost in het proces van toetsontwikkeling en -afname. Ook is het mogelijk dat de zoektocht naar het juiste materiaal veel tijd kost.

Niet alleen het informeren van de ouders inzake procedures, data en resultaten kost dus tijd en geld, maar ook de aankoop van computersoftware om de resultaten te analyseren, is een grote meerkost. Tot slot telt Phelps (2000) ook het gebruik en onderhoud van de gebouwen mee.

2 BOUWSTENEN VAN LEER(WINST)MONITORINGSSYSTEMEN

De mate waarin een leer(winst)monitoringssysteem leidt tot bepaalde effecten en neveneffecten, hangt sterk samen met enkele voorwaarden. Het enkel implementeren van gestandaardiseerde toetsen is niet voldoende: *“let us not forget that assessment alone cannot be expected to do the magic, much like taking the temperature of a sick child is not going to heal the pain”* (Wang et al., 2006, p. 323). In dit deel worden de condities besproken die vanuit het literatuuronderzoek naar de voorgrond komen om de positieve effecten van een leer(winst)monitoringssysteem te maximaliseren en de negatieve effecten zo veel mogelijk te vermijden.

Uit de vorige sectie bleek duidelijk dat de reacties op (de resultaten van) een leer(winst)-monitoringssysteem erg uiteenlopend kunnen zijn. Onder meer de vraag of de gestandaardiseerde toetsen aanleiding kunnen geven tot het in kaart brengen van de leerwinst van individuele leerlingen en/of de toegevoegde waarde van scholen, kan daarin een belangrijke beïnvloedende factor zijn. De literatuur in verband met deze vraag wordt behandeld in de eerstvolgende sectie. Daarna verzamelen we de literatuur in verband met drie andere groepen van beïnvloedende factoren die worden onderscheiden in het model van Schildkamp et al. (2012): kenmerken van de gestandaardiseerde toetsen, individuele kenmerken van de leraren in de scholen en kenmerken van de scholen zelf. Ten slotte bespreken we literatuur inzake de invloed van kenmerken op macroniveau (Black & Wiliam, 2005).

2.1 Mogelijkheid tot berekening van leerwinst en/of toegevoegde waarde

Zoals eerder toegelicht wordt in deze studie over leerwinst gesproken wanneer het individuele leerlingen betreft (microniveau). De term toegevoegde waarde wordt gereserveerd voor de bijdrage van scholen aan het leren of aan de leerwinst van leerlingen.

2.1.1 Mogelijkheid tot leerwinstmeting

Verschillende auteurs menen dat de waarde van de resultaten van een monitoringssysteem sterk wordt beïnvloed door het al dan niet creëren van de mogelijkheid om leerwinst te berekenen. Een monitoringssysteem wint enorm aan belang voor gebruik op microniveau wanneer de gestandaardiseerde toetsen niet enkel een eenmalige meting van de leerprestaties van leerlingen weergeven, maar er tevens in slagen om de leerwinst aan te duiden die individuele leerlingen hebben doorgemaakt. Het meten van leerwinst ondersteunt onder meer de diagnostische functie van gestandaardiseerde toetsen (Ver Eecke, 2004). Bovendien motiveert het scholen om voor elke leerling in te zetten op een maximale ontplooiing van capaciteiten, en niet enkel op het bereiken van

de minimumdoelen (Popham, 2005; Ver Eecke, 2004). Leerwinstmetingen helpen de school tevens bij het beantwoorden van de vraag of een bepaalde leerprestatie voor een leerling nu voldoende kan geacht worden (Janssens et al., 2014).

Bovendien is er door het berekenen van leerwinst, waarbij de voortgang tussen twee momenten wordt bekeken, minder aanzet voor zowel leerlingen als leraren om over te gaan tot vormen van onwenselijke strategisch gedrag (Elstad & Turmo, 2011). De empirische evidentie inzake de invloed van leerwinstmonitoring op het gebruik van gestandaardiseerde toetsen is evenwel zeer beperkt.

Daarnaast is het ook belangrijk om leerwinst in kaart te brengen voor het bepalen van de toegevoegde waarde van scholen en klassen. Men gaat er immers van uit dat modellen die uitgaan van leerwinst in plaats van leerprestaties een meer rechtvaardig beeld van de school schetsen (Goldschmidt, Choi, & Beaudoin, 2012) (zie 3.2.2 voor een meer uitgebreide omschrijving van toegevoegde waarde-modellen op basis van leerwinst).

2.1.2 Mogelijkheid tot toegevoegde waarde-meting

Het berekenen van de toegevoegde waarde van de school brengt in kaart in welke mate de school bijdraagt tot de leerprestaties en/of leerwinst van de leerlingen. Ook hierbij is de empirische kennisbasis erg beperkt. Het wordt verondersteld dat toegevoegde-waardeberekeningen een positief effect kunnen hebben op de interne kwaliteitszorg van scholen, en op die manier op het streven naar een maximale ontplooiing van de capaciteiten van elke leerling. Dit gebeurt onder meer omdat een toegevoegde waarde-berekening de minder effectieve scholen (in het bereiken van leerwinst van leerlingen) kan aanzetten tot een verhoogde effectiviteit (Ver Eecke, 2004). Loeb (2013) omschrijft drie manieren waarop informatie inzake de toegevoegde waarde de school of het schoolbestuur kan helpen bij de verdere schoolontwikkeling. Allereerst kan deze informatie ertoe leiden dat het eigen onderwijs en specifieke innovaties kunnen worden geëvalueerd. Ook zorgt de informatie ervoor dat de professionele ontwikkeling meer gericht wordt ingezet op leraren die dat binnen bepaalde domeinen het meest nodig hebben (Loeb, 2013).

Naast bovenstaand ontwikkelingsgericht doel, hebben toegevoegde waarde-metingen vaak een verantwoordingsgericht perspectief: ze geven informatie aan een schoolbestuur, aan ouders of aan een overheid over de mate waarin de school bijdraagt tot het leren van de leerlingen (Keeves, Hungi, & Afrassa, 2005; Loeb, 2013; Saunders, 1999). Tot slot kunnen op basis van deze metingen bonussen voor goede leraren worden toegekend (Loeb, 2013).

Een probleem met toegevoegde waarde-metingen is de complexiteit van het begrip, waardoor het bijgevolg een moeilijke uitdaging is om dit concept aan scholen en leraren helder uit te leggen (Scheerens, 2014), zeker in het geval van multilevel analyse (Sloane et al., 2013; van de Grift, 2009). Bovendien is het onduidelijk welke factoren voor welke leerdomeinen op welke manier in rekening moeten worden gebracht, zodat enkel assumpties kunnen gemaakt worden over de causaliteit tussen schooleffectiviteit en de leerprestaties of de leerwinst die uit de gestandaardiseerde toetsen blijkt (Raudenbush (2004) in Janssens et al., 2014).

De beperkte empirische basis ondersteunt evenwel niet volledig de bewering dat de berekening van toegevoegde waarde sterk bijdraagt tot het ontwikkelingsgerichte doel. Uit empirisch onderzoek in Engeland bleek de impact van toegevoegde-waarde informatie opgeleverd door het ALIS-systeem in een eerste fase relatief beperkt te zijn (Williamson, 1990). Toch stelde men in een opvolgstudie

(Williamson, Tymms, & Haddow, 1992) vast dat meer scholen na verloop van tijd actie hadden ondernomen. Tevens gaf men in deze opvolgstudie aan dat de informatie enkel bij *school managers* een extra bewustzijn had gecreëerd (en noch bij de leidinggevenden, noch bij leraren van het departement Engels) en dat de statistische data een groot struikelblok vormden. Scholen beschikten op dat ogenblik (nog) niet over adequate structuren om op basis van data een beleid te voeren (Williamson et al., 1992).

Uit een kwalitatieve studie van de National Foundation of Educational Research (NFER) in negen scholen in Engeland bleek dat het gebruik van een instrument voor het in kaart brengen van toegevoegde waarde van scholen eerder beperkt was: in zes van de negen scholen was men niet verder gekomen dan het proberen begrijpen (op niveau van het *senior management*) wat de data betekenden (hoewel NFER van mening was dat het een zeer duidelijk rapport gecreëerd had). Men concludeerde dat heel wat leraren het moeilijk hadden om betekenis te geven aan de term '*added value*' waardoor de resultaten van de gestandaardiseerde toetsen op een verkeerde wijze werden geïnterpreteerd en de data oneigenlijk werden gebruikt (door bijvoorbeeld voorspellingen te maken op basis van data die daartoe niet geschikt zijn). Daarnaast lijken de volgende factoren bij te dragen tot het gebruik van de toegevoegde waarde-data: (1) een actieve ondersteuning en begeleiding door de ontwikkelaars van het meetinstrument, (2) '*championing*' van de data door een personeelslid (meestal iemand van het directieteam die de data analyseert en die iedereen voortdurend aanmoedigt om de data om te zetten in acties) en (3) minimaal drie opeenvolgende jaren ontvangst van deze data. Ook (4) als school goed scoren op toegevoegde waarde wanneer de school het op traditionele rankings eerder slecht doet, blijkt sterk tot gebruik van aan te zetten (Saunders, 2000).

2.2 Opzet van de gestandaardiseerde toetsen

In een eerder vermelde meta-analyse van kwalitatieve evidentie inzake de effecten van toetsen op de breedte van het curriculum werden zowel voorbeelden van positieve effecten als van onwenselijke neveneffecten verzameld. De analyse leidt de auteur tot de conclusie dat de effecten sterk afhankelijk zijn van het opzet van de toetsen zelf (Au, 2007). In de literatuur worden uiteenlopende aspecten van het opzet van gestandaardiseerde toetsen in een leer(winst)-monitoringssysteem bediscussieerd. Deze literatuur wordt in onderstaande secties besproken.

2.2.1 Sterke psychometrische onderbouwing

Door verschillende auteurs wordt beklemtoond dat toetsen moeten voldoen aan de vereiste condities van betrouwbaarheid en validiteit (Black & Wiliam, 2005; Sireci, 2005). Bovendien blijkt dat ook scholen een hoge mate van betrouwbaarheid en validiteit verwachten van gestandaardiseerde toetsen (Janssens et al., 2014).

Betrouwbaarheid

De betrouwbaarheid van gestandaardiseerde toetsen is de mate waarin toetsscores consistent zijn doorheen verschillende afnames. De meeste toetsontwikkelaars gebruiken een maat van interne consistentie om de betrouwbaarheid van een instrument na te gaan. De interne consistentie kan mathematisch worden nagegaan door te berekenen in welke mate de vragen aan elkaar gerelateerd zijn (Buckendahl & Hunt, 2005). Om de betrouwbaarheid te verhogen, doet men er goed aan om tijdens het proces van toetsontwikkeling expliciet op zoek te gaan naar *biased items*, items die door

een bepaalde groep van leerlingen beter kunnen begrepen worden omwille van bijvoorbeeld hun culturele achtergrond (Goodman & Hambleton, 2005). Betrouwbaarheid houdt daarnaast de mate in waarin een toets erin slaagt om zoveel als mogelijk 'valse positieve' (leerlingen die onterecht worden bestempeld als 'geslaagd', of als iemand die een bepaald competentieniveau beheerst), alsook 'valse negatieve' (leerlingen die onterecht worden bestempeld als 'niet geslaagd', of als iemand die een bepaald competentieniveau niet zou beheersen) diagnoses te vermijden (Nyström, 2005). Bovendien vereisen gestandaardiseerde toetsen, zeker in het geval ze leiden tot berekeningen van leerwinst en/of toegevoegde waarde, dat er een betrouwbare (consistente) schaal wordt opgemaakt die doorheen verschillende toetsafnames kan worden gehanteerd (Goldschmidt et al., 2012).

De betrouwbaarheid van toetsen is evenwel niet volledig in handen van de toetsontwikkelaar. De afname van de toets moet volgens een gestandaardiseerde procedure gebeuren en de omstandigheden voor elke leerling zouden identiek moeten zijn. Zoals eerder gesteld, kunnen scholen echter onder bepaalde omstandigheden geneigd zijn om de procedures niet strikt toe te passen. Zelfs indien aan deze condities zou voldaan worden, zijn toetsen nog steeds niet volledig betrouwbaar: er zijn ook factoren op leerlingniveau die de betrouwbaarheid beïnvloeden. Een leerling kan immers net een 'goede' of een 'slechte' dag hebben, waardoor toetsen bijgevolg altijd af te rekenen hebben met een zekere meetfout (Ladd & Walsh, 2002).

Tot slot wordt het algemeen als fair beschouwd om leerlingen met een beperking (zoals een visuele beperking, dyslexie, ...) compenserende maatregelen aan te bieden. Er bestaat evenwel geen standaard die kan bepalen hoe de toetsresultaten voor deze leerlingen op dezelfde manier als voor andere leerlingen betrouwbaar kunnen worden ingeschaald (Randall, Cheong, & Engelhard, 2011).

Validiteit

De validiteit van gestandaardiseerde toetsen houdt in dat een hoge score voor een leerling of een school moet overeenkomen met respectievelijk een goede *mastery* van de getoetste leerinhoud of met een hoge onderwijskwaliteit (William, 2010), aangezien dat de zaken zijn die gestandaardiseerde toetsen (meestal) in kaart willen brengen. Volgens Buckendahl en Hunt (2005) veronderstelt validiteit dat de toets curriculum-gerelateerd is, omdat ze moet toetsen wat de leerling heeft geleerd - dit geldt evenwel enkel voor high-stakes toetsen die uitspraken willen doen over leerlingen, leraren of scholen, want bij andere gestandaardiseerde toetsen is het mogelijk dat men net een ander aspect van de leerling (zoals niet-curriculum-specifieke rekenvaardigheden) in kaart wil brengen.

Twee grote bedreigingen voor toetsvaliditeit zijn *construct underrepresentation* en *construct-irrelevant variance* (William, 2010). *Construct underrepresentation* betekent dat verschillen in scores tussen leerlingen geen adequate reflectie zijn van effectieve verschillen tussen leerlingen voor het construct dat men wil meten (bijvoorbeeld rekenvaardigheid) - deze meting is dus eigenlijk 'te smal'. Bij *construct-irrelevant variance* is de meting net 'te breed', waardoor verschillen in scores niet (enkel) gerelateerd zijn aan het construct dat men wil meten: het toetsinstrument meet ook andere constructen, naast het bedoelde construct (William, 2010).

Construct underrepresentation kan worden vermeden door er zo nauwgezet mogelijk over te waken dat de toets beoordeelt wat scholen behoren te onderwijzen (of wat leraren willen weten over de leerlingen). Zo wordt vaak de kritiek geformuleerd dat gestandaardiseerde toetsen slechts ten dele

valide zijn omdat ze meestal gericht zijn op reproductie en zich slechts beperkt focussen op 'hogere orde'-vaardigheden (White, 2013).

Er zijn in de literatuur vele voorbeelden van *construct-irrelevant variance* beschreven. Pollitt, Ahmed, Baird, Tognolini, en Davidson (2008) stellen dat de toegankelijkheid van de toetsen een belangrijke conditie is, aangezien verschillende gestandaardiseerde toetsen, naast het bedoelde construct, ook de taalvaardigheid van leerlingen evalueren. Als leerlingen de toetsopgave anders interpreteren dan bedoeld is, lopen zij ten onrechte punten mis (Pollitt et al., 2008). Daarnaast meten toetsen, naast het bedoelde construct, ook steeds de toetsvaardigheid van de leerling. Zo kan de beste *test taker* een hogere score behalen op een toets, ondanks dat hij of zij niet per se meer kennis heeft van het getoetste onderwerp dan een andere student die minder toetsvaardig is (Ladd & Walsh, 2002). Uit een analyse uitgevoerd op data van de digitaal aangeboden nationale toetsen Engels van het lager onderwijs in Noorwegen, blijkt dat er een significante samenhang is tussen de zelfgerapporteerde ICT-vaardigheden van leerlingen en de resultaten op de toetsen. Dit suggereert dat ook digitale vaardigheden worden meegenomen in de nationale toetsen Engels. De - eveneens digitale - nationale toetsen Engels in het secundair onderwijs vertoonden dan weer geen samenhang met de ICT-vaardigheden, maar bleken wel samen te hangen met de score van leerlingen op een non-verbale redeneervaardigheidstest (Manger, Vold, & Eikeland, 2009).

Een ander aspect van toetsvaliditeit is het correct bepalen van *cut-off scores* of *passing standards* (Buckendahl & Hunt, 2005). Een valide benadering daarbij is dat er niet wordt gewerkt vanuit een groepsnorm-criterium (bijvoorbeeld vanuit een Gauss-curve), maar dat door een groep van experts wordt bepaald welk niveau vereist is om te 'slagen' voor een bepaalde toets of om een bepaalde score toegekend te krijgen (met andere woorden: een standaard-criterium) (Sireci, 2005).

Ten slotte is de predictieve validiteit in dit kader van belang. Deze gaat na in welke mate de toetsresultaten een goede voorspeller van verder studiesucces zijn (Babo, Tienken, & Gencarelli, 2014; Kobrin, Kim, & Sackett, 2012). In het geval van bijvoorbeeld gestandaardiseerde eindtoetsen aan het eind van een onderwijsfase moeten de gebruikers erop kunnen vertrouwen dat wie slaagt voor de eindtoets een grotere kans heeft om de volgende fase in het onderwijs succesvol af te leggen (in vergelijking met wie niet slaagt). Dit vraagstuk werd reeds in de jaren '70 en '80 van de vorige eeuw uitgebreid onderzocht door verschillende studies in Jordanië op basis van het toenmalige GSECE-eindexamen, waarbij bleek dat deze eindexamens geen goede voorspeller waren voor het academische succes van de leerlingen (verschillende studies besproken in Jaradat, 1986). Intussen zijn evenwel ook al studies (op andere gestandaardiseerde toetsen) uitgevoerd die wel wijzen op een hoge predictieve validiteit (een voorbeeld uit het hoger onderwijs is Barton, Willson, Langford, & Schreiner, 2014). Bovendien kan de voorspellende waarde van een gestandaardiseerde toets verschillend zijn voor verschillende leeftijdsgroepen of voor jongens en meisjes, zoals werd aangetoond op een analyse van predictieve validiteit op eindexamens in Kenya (Othuon & Kishor, 1994).

Wanneer zijn psychometrische kwaliteitskenmerken belangrijk?

Het belang van betrouwbaarheid en validiteit wordt hoger ingeschat naarmate het leer(winst)-monitoringssysteem meer kadert in een verantwoordingsgericht perspectief en naarmate de toetsen high-stakes worden voor leerlingen of leraren (Elstad, 2009; Wiliam, 2010). Omdat het ook bij high-stakes toetsen niet altijd mogelijk is dat alle leerlingen op hetzelfde ogenblik een toets afleggen - en

omwille van het high-stakes karakter van de toetsen - kan men geen zelfde toetsitems op twee verschillende tijdstippen gebruiken. Een goed onderbouwde IRT-analyse is dan zeker nodig om de vergelijkbaarheid van twee verschillende toetsen te garanderen (Taylor & Lee, 2009).

Perceptie

Voor scholen is het niet makkelijk om de betrouwbaarheid en de validiteit van gestandaardiseerde toetsen in te schatten. Nochtans hangt de mate waarin positieve effecten zich voordoen en waarin onwenselijke neveneffecten vermeden worden eerder af van de perceptie van betrouwbaarheid en validiteit, dan van de feitelijke realisatie van deze psychometrische kenmerken.

Het is evenwel onduidelijk wanneer gestandaardiseerde toetsen in de perceptie van scholen voldoende psychometrisch onderbouwd zijn om als bruikbare input te worden beschouwd. Een eerder besproken studie in Duitsland vond dat, ondanks de sterke kwaliteitszorg bij de correctie van high-stakes eindexamens, de huidige procedures niet volstonden. Zo wenste men dat de examens door meer dan twee correctoren zouden worden nagekeken om een nog grotere objectiviteit te bekomen (Van Ackeren et al., 2012).

Geen enkele toets is echter 100 procent betrouwbaar en valide. Hoewel het een streven moet zijn om de hoogst mogelijke betrouwbaarheid en validiteit na te streven, is het tevens belangrijk om te erkennen dat toetsen per definitie imperfect en onvolledig zijn (Loeb, 2013). Zo zijn klachten over de betrouwbaarheid en/of validiteit van high-stakes toetsen bijna onvermijdelijk.

Buckendahl en Hunt (2005) formuleren enkele belangrijke adviezen om klachten zoveel mogelijk te vermijden. Zo draagt, naast het opvolgen van de psychometrische aanbevelingen, de ontwikkeling en het onderhoud van een technisch draaiboek bij tot een verhoogde betrouwbaarheid en validiteit. Tot slot wordt het consulteren van experts in psychometrische vraagstukken aangeraden. Instanties voor toetsontwikkeling zouden zich niet alleen moeten laten adviseren door (interne of externe) experts inzake betrouwbaarheid en validiteit, maar ze zouden hun proces ook moeten laten evalueren door een externe expert ter zake (Buckendahl & Hunt, 2005).

2.2.2 Inzet van de toetsen

Analyses uit de Verenigde Staten zijn erg positief over de impact van gestandaardiseerde toetsen op de leerprestaties van leerlingen. Een vergelijking tussen verschillende staten van Carnoy en Loeb (2002) wijst uit dat er een sterke correlatie bestaat tussen de '*accountability index*' (de mate waarin gestandaardiseerde toetsen worden ingezet voor verantwoordingsgerichte doelen) en de leerprestaties van leerlingen in het achtste leerjaar. Bovendien is het verschil nog groter bij enkele groepen van leerlingen die traditioneel zwakker scoren (zoals Afro-Amerikanen en Hispanics) dan bij blanke leerlingen (Carnoy & Loeb, 2002).

Amrein en Berliner (2002) onderzochten het verschil in leerprestaties tussen studenten in staten met, en staten zonder high-stakes toetsen. De resultaten wijzen uit dat er geen causaal verband bestaat tussen de aanwezigheid van high-stakes toetsen en een toename van de leerprestaties van studenten. Ook Rosenshine (2003) onderzocht de resultaten van leerlingen in dergelijke onderwijscontexten. Zij concludeerde echter dat de NAEP-scores van leerlingen in staten met een high-stakes toetsensysteem wel sterker toegenomen waren dan in staten met low-stakes toetsen.

Vaak wordt verondersteld dat low-stakes gestandaardiseerde toetsen minder neveneffecten hebben (Wang et al., 2006), terwijl de meningen over positieve effecten verdeeld zijn: sommigen

veronderstellen dat een hoge inzet nodig is om hogere motivatie en betere focus van leraren en een betere prestatie van leerlingen te bewerkstelligen (Carnoy & Loeb, 2002), terwijl anderen net menen dat toetsen zonder directe consequenties de sleutel zijn tot een verbeterde onderwijskwaliteit (Brozo & Hargis, 2003).

Onderzoek uit Florida met een beleid van '*school voucher programs*' bevestigt dat wanneer de inzet voor een school hoog is, de impact van de gestandaardiseerde toetsen op de leerprestaties wordt versterkt. In dit beleid worden waardebonnen (vouchers genoemd) verstrekt aan ouders, die kunnen worden ingeruild bij elke openbare of particuliere school die deelneemt aan het programma, zodat leerlingen naar een school van vrije keuze kunnen gaan. Deze vouchers worden in bepaalde omstandigheden aangewend als stimulans om de onderwijskwaliteit te garanderen. In Florida krijgt elke openbare school een rang gaande van A tot F, toegewezen op basis van het aantal studenten dat slaagt voor de Florida Comprehensive Assessment Test (FCAT). Leerlingen kunnen een nieuwe school kiezen (door middel van een nieuwe voucher) als hun school twee opeenvolgende negatieve evaluaties (rang F) krijgt. Greene (2001) onderzocht of de prestaties binnen deze scholen verbeteren als de dreiging van deze vouchers dichterbij komt (na één negatieve evaluatie). Uit de resultaten blijkt dat scholen met een lage FCAT-score bij een volgende afname van dit eindexamen wel degelijk betere resultaten behalen, maar dat vooral scholen met een F-grade (en de dreiging van bovenvermelde vouchers) grotere winsten genereren. Onderzoek van Chakrabarti (2008a) treedt deze resultaten bij. Hoewel aanvankelijk gedacht werd dat de prestaties van de scholen verbeterden omwille van de vrees van een blijvend stigma en niet voor de dreiging van de vouchers, blijkt uit onderzoek van Greene, Winters en Forster (2003) dat scholen een minder grote leerwinst boeken als de dreiging van de vouchers wegvalt.

Bovenstaande onderzoeken tonen aan dat de inzet van toetsen een rol speelt in de manier waarop scholen met de resultaten omgaan. Maar ook het gedrag van leerlingen wordt beïnvloed door de inzet van de toetsen. Zo zijn leerlingen minder geneigd om vals te spelen op toetsen wanneer deze niet gebonden zijn aan leerlingenrankings, zittenblijven, financiële stimuli of het uitreiken van diploma's (Skorupski & Egan, 2012). Daarnaast gaf eerder vermeld onderzoek aan dat al de tien staten in de Verenigde Staten waar de drop-out van leerlingen het hoogst lag, gebruik maakten van high-stakes gestandaardiseerde eindexamens. Bij de tien staten met de kleinste *drop-out* waren er geen eindexamens, of hadden ze een eerder low-stakes karakter (Clarke et al., 2000).

2.2.3 Openbaarheid, rankings, competitie

Het openbaar maken van kwaliteitsgegevens van scholen is een trend die in de meeste landen toeneemt (OECD, 2013). Toch wordt die openbaarheid van resultaten sterk gecontesteerd, omdat het - zoals eerder gesteld - rankings in de hand zou werken (zie 1.4.2). En hoewel enkele voorstanders voordelen zien in een competitie tussen scholen, blijkt uit evidentie dat de openbaarheid van de resultaten van toetsen op schoolniveau, of van toegevoegde waarde-informatie, ertoe leidt dat de bruikbaarheid van de informatie voor ontwikkelingsdoeleinden wordt bedreigd (Yang, Goldstein, Rath, & Hill, 1999).

Vanuit de literatuur kunnen dan ook enkele effecten onderscheiden worden: (1) impact op scholen, (2) impact op leraren en (3) impact op ouders.

Impact op scholen

Wanneer gegevens van scholen openbaar worden gemaakt, blijken scholen strategische acties te ondernemen (Woods, Bagley, & Glatte, 1998). Een aantal scholen zet promotiecampagnes op om de school in een positief daglicht te stellen, waarbij gretig gebruik wordt gemaakt van de resultaten van de gestandaardiseerde toetsen om op die manier nieuwe leerlingen (en dan voornamelijk *sterke* nieuwe leerlingen aan te trekken) (Woods et al., 1998).

Volgens verschillende auteurs is openbaarheid van gegevens (en de daarbij horende rankings) wel degelijk wenselijk, omdat ze scholen nog meer aanzetten om het beste uit elke leerling naar boven te halen. Toen in Wales ook de publicatie van rankings gebaseerd op de resultaten van Key Stage 4 (GCSE) in het secundair onderwijs werden afgeschaft, onderzochten Burgess, Wilson en Worth (2013) in welke mate deze beslissing een effect had op de prestaties van scholen op deze gestandaardiseerde toetsen. Deze studie wees uit dat de prestatie van scholen (berekend als het percentage leerlingen met minimaal vijf geslaagde examens op GCSE (dit is ook de maat die standaard wordt openbaar gemaakt)) significant hoger ligt voor Engelse scholen dan voor de scholen uit Wales vanaf de cohorten waarbij de toetsresultaten niet meer worden gepubliceerd. Bovendien wordt het verschil tussen Engeland en Wales sinds de afschaffing van de schoolresultaten jaar na jaar groter. Ook wanneer scholen worden vergeleken met 'vergelijkbare' scholen qua uiteenlopende factoren (onder meer de prestaties van leerlingen op Key Stage 3 toetsen) blijft hetzelfde beeld overeind (Burgess et al., 2013). Een opvallende bevinding van deze studie is dat er bij de scholen die op basis van de Key Stage 3 toetsen in het hoogste kwartiel werden ingedeeld geen verschil is tussen Wales en Engeland. Het verschil is wel significant voor het tweede, derde en vierde kwartiel en is het grootst voor scholen in deze laatste categorie. Burgess et al. (2013) gingen ook na of de verschillen in scores niet het gevolg konden zijn van onwenselijke neveneffecten, zoals *teaching to the test* of onwenselijk strategisch gedrag. Door aan te tonen dat de verschillen tussen Engelse en Welshe scholen op de PISA-toetsen, zowel inzake wiskunde, lezen als wetenschappen, in dezelfde periode ook groter worden (in het nadeel van Wales), nemen de auteurs aan dat dit niet het geval is, aangezien er bij de PISA-toetsen voor scholen niets op het spel staat (low-stakes). Daarom besluiten Burgess et al. (2013) dat er sinds het wegvallen van de rankings in Wales werkelijk betere leereffecten zijn in Engeland. De auteurs geven echter niet aan of bovenstaande onwenselijke voorbereiding voor GCSE (zoals een specifieke voorbereiding inzake toetstechnieken, een grotere ervaring met toetsafnames, ...) niet mogelijk een doorlopend effect zouden kunnen hebben bij de afname van PISA.

Ook enkele andere studies publiceerden resultaten die het positieve effect van rankings ondersteunen. Een analyse tussen de verschillende vroegere Oost-Duitse deelstaten i leert inderdaad dat leerlingen in de ene deelstaat waarin rankings worden gepubliceerd (met name Saksen) gemiddeld betere prestaties toont op een internationaal vergelijkende studie (PISA-E) in zowel *Realschule*, als *Gymnasia*, maar niet in *Hauptschule* (Jürges, Richter, et al., 2005).

Publicatie van resultaten leidt tot een verhoogde competitie tussen scholen. Een review studie concludeerde uit een overzicht van 41 studies tussen 1972 en 2001 in de Verenigde Staten dat er een positief verband bestaat tussen een hoge mate van competitie en een hoge onderwijskwaliteit, al is de effectgrootte wel beperkt (Belfield & Levin, 2002). Figlio en Hart (2010) stellen vast dat de resultaten van leerlingen in scholen die deelnemen aan 'Florida's Corporate Tax Credit Scholarship program' sterk verbeteren als gevolg van toegenomen concurrentie. Ook Hoxby (2001) stelt dat de

prestaties in openbare scholen in Milwaukee, Michigan en Arizona aanzienlijk stijgen wanneer er sprake is van concurrentie, hoewel de financiële middelen niet werden gewijzigd.

Een interessante nuance werd gemaakt in de studie van Levačić (2004): zij stelde vast dat er geen verschil was inzake de leerprestaties van leerlingen tussen scholen in een 'competitieve markt' en scholen in een 'niet-competitieve markt' (op basis van objectieve parameters); maar dat leerlingen gemiddeld hogere leerprestaties behalen (op GCSE-examens) in scholen waarin de perceptie leeft dat men in een competitieve markt ageert. Eerder dan de objectieve mate van competitie, lijkt dus de 'gepercipieerde competitie' een effect te hebben op de leerprestaties (Levačić, 2004).

Ondanks bovenstaande voordelen stellen Karsten, Visscher, & De Jong (2001) vast dat het openbaar maken van de schoolresultaten in rankings op groot ongenoegen van directies stuit. Directies zijn immers van mening dat deze rankings een aantal beperkingen kennen. Allereerst zouden de gepubliceerde data te ruw zijn en worden de indicatoren van schoolkwaliteit eenzijdig gedefinieerd. Zo richt men zich hoofdzakelijk op de academische prestaties, terwijl andere kwaliteitsaspecten, zoals extra-muros-activiteiten, buiten beeld blijven. Daarnaast bevatten deze data geen informatie over de onderwijsleerprocessen binnen de school en worden de resultaten van alle leerlingen in kaart gebracht, ongeacht hun deelname aan de eindexamens. Bovendien zou er sprake zijn van een oneerlijke vergelijking tussen de scholen (bijvoorbeeld tussen selectieve scholen en niet-selectieve scholen). Directies geven ook toe dat het verantwoordingsgericht karakter van de rankings één van de hoofdzorgen van de school is, waardoor minder geïnvesteerd wordt in minder capabele leerlingen. Tot slot blijkt het gebruik van deze rankings in de media een doorn in het oog van vele directies te zijn omdat de media niet geneigd zijn om ook de context en achtergrondkenmerken van de school mee in rekening te brengen: "*their main goal is a sensational story*" (Karsten et al., 2001, p. 238).

Het onderzoek van Collins et al. (2010) sluit hierbij aan en stelt dat het afschaffen van de Key Stage 2 en de daarop gebaseerde *School and College Assessment and Attainment Tables* (SCAATs) in Wales door leraren en directies sterk verwelkomd werd, onder meer omdat per definitie slechts een beperkt beeld van de taken van de school in kaart wordt gebracht.

Een studie van Hoxby (2003) onderzoekt of de concurrentie tussen scholen, die ontstaat door vrije schoolkeuze, openbare scholen aanzet om op een constructieve manier te reageren, door bijvoorbeeld de eigen productiviteit te verhogen.

Ondanks de voordelen moeten beleidsmakers bedacht zijn op één van de meest belangrijke strategieën die scholen hanteren om betere resultaten te verkrijgen. Scholen trachten in een verantwoordingsgericht perspectief vaak de instroom van hun leerlingen onder controle te houden, zodat na selectie alleen 'goede' leerlingen overblijven ('cream skimming'). "*it is not so much what the school can do for its students, but what the students can do for their school*" (Ball, 1999, p. 93, in Karsten et al., 2001). Dit verschijnsel duidt op het feit dat de kwalitatief betere onderwijsinstellingen geneigd zijn om leerlingen toe te laten die het minst ondersteuning nodig hebben om goede resultaten te boeken, waardoor leerlingen uit achterstandsgroepen naar minderbegoede scholen gestuurd worden en de ongelijkheid verder toeneemt. De ongelijkheid tussen leerlingen van uiteenlopende sociaaleconomische achtergrond neemt op die manier toe (Horn, 2005). Uit een studie van Gillborn (1996) blijkt dat het ontzeggen van schooltoegang voor storende leerlingen toeneemt in het eerste jaar na het implementeren van rankings. Andere acties die scholen

ondernemen, zijn: het uitsluiten van deelname aan examens of aan bepaalde vakken, of 'makkelijkere' vakken aanbieden. In onderzoek van Burgess et al. (2013) kon dit verondersteld neveneffect van gestandaardiseerde toetsen echter weerlegd worden, aangezien men geen evidentie vond dat de segregatie tussen scholen inzake leerlingeninstroom was verlaagd in Wales nadat de rankings niet meer werden gepubliceerd (Burgess et al., 2013). Ook Hoxby (2003) stelt het bestaan van '*cream-skimming*' in vraag. Uit dit onderzoek blijkt dat de prestaties van leerlingen in voucher- of charterscholen verbeteren, hoewel er van '*cream-skimming*' geen sprake is. Opvallend is dat deze scholen net leerlingen aantrekken die minder goede resultaten behaalden in publieke scholen, maar doordat deze cijfers sterk wisselen doorheen de tijd kunnen de resultaten niet worden doorgetrokken. Hoxby (2003) besluit dat de selectie van leerlingen eerder afhankelijk is van de schoolorganisatie dan van het statuut van voucher- of charterschool.

Impact op leraren

Onderzoek van Perryman, Ball, Maguire, & Braun (2011) in vier secundaire scholen in Engeland toont aan dat leraren Engels en wiskunde zich onder druk voelen gezet naar aanleiding van het jaarlijks publiceren van rankings. Deze leraren ervaren de druk om goede resultaten te behalen niet enkel voor zichzelf, maar voelen zich ook verantwoordelijk voor zowel de leerlingen als de gehele school. Daarnaast rapporteren ze een toename van stress, een verzwakte creativiteit en voelen zij zich gefrustreerd omdat ze zich moeten houden aan een extern opgelegde agenda, waardoor deze leraren zich voelen aangetast in hun professionaliteit (Perryman et al., 2011). Dit onderzoek bevestigde de resultaten van Gillborn (1996) die in het verleden reeds de impact van rankings op leraren onderzocht.

Daarnaast kan het publiceren van rankings ook leraren demotiveren. De *naming-and-blaming*-cultuur zorgt voor een negatieve impact op de samenwerking tussen leraren (Barber, 1998; Elstad, 2009). Leraren frustreren zich vooral in de beperkingen van de publicaties, waarbij achtergrondkenmerken van leerlingen niet in rekening worden gebracht.

Impact op ouders

Onderzoek heeft evidentie aangeleverd voor het zogenaamde '*little-fish-big-pond-effect*' (H. W. Marsh, 1987), waarbij een sterke leerling in een gemiddelde groep zichzelf hoger inschat dan wanneer hij in een sterke klasgroep zou zitten. Dit heeft tot gevolg dat het zelfconcept en de motivatie van sterke leerlingen in zeer sterke klasgroepen vaak erg laag is. Indien deze leerlingen hun kennis en vaardigheden kunnen vergelijken met een volledige populatie (in plaats van enkel met hun eigen klasgroep), kan hun zelfconcept en motivatie echter sterk toenemen (Trautwein, Gerlach, & Lüdtke, 2008; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006).

Geïnteresseerde ouders vragen zich vaak af hoe hun kind op school presteert in vergelijking met klasgenoten en in vergelijking met leeftijdsgenoten overheen de verschillende scholen. Daarnaast kunnen ouders zich eveneens de vraag stellen welke vooruitgang hun kind heeft gemaakt in vergelijking met een voorgaande evaluatie (Wiggins, 1990). Vanuit het verantwoordingsgericht perspectief kunnen gestandaardiseerde toetsen of examens vergelijkbare informatie opleveren over de onderwijskwaliteit van diverse scholen en de leerprestaties van leerlingen. Ouders en leerlingen kunnen deze data ook aanwenden om de juiste schoolkeuze te maken (Wössmann, 2007). Niet alle ouders zijn echter in staat om de resultaten van rankings correct te interpreteren (Karsten et al., 2001; Wössmann, 2007). Daarnaast zijn niet alle onderzoekers van mening dat deze rankings de

juiste informatie bevatten om naar ouders toe te rapporteren (Goldstein & Spiegelhalter, 1996; Goldstein & Leckie, 2008). Volgens Karsten et al. (2001) kunnen ouders drie acties ondernemen wanneer uit de publicatie van de leerresultaten blijkt dat de school minder goede resultaten behaalt. Allereerst is er de *exit option*, waarbij ouders op zoek gaan naar een alternatieve school met betere kwaliteit. Daarnaast kunnen ouders feedback geven aan de onderwijsinstelling met het oog op kwaliteitsverbetering (*voice option*). Wiggins en Tymms (2000) concluderen echter dat ouders uit Engelse en Schotse basisscholen niet geneigd zijn om de resultaten van centrale toetsen te bespreken met de school. De laatste optie die ouders hebben is de *loyalty option*. Hier zijn de ouders de school genegen omwille van bijvoorbeeld religieuze redenen of zijn ouders, ondanks de tegenvallende cijfers, toch tevreden met het geboden onderwijs.

2.2.4 Gelinkt aan het curriculum

Verschillende auteurs geven aan dat het voor de scholen belangrijk is dat het leer(winst)-monitoringssysteem gelinkt is aan het curriculum om ten volle van de toetsen gebruik te kunnen maken op meso- en microniveau (Goodman & Hambleton, 2005; Popham, 2002; Shepard, 1992; Wang et al., 2006; Wiliam, 2010). Sommige auteurs geven aan dat bij een goede link tussen de toetsinhoud en het curriculum '*teaching to the test*' eigenlijk een positief effect is, met name '*teaching to the standard*' (Cizek, 2005; Ver Eecke, 2004). De expertengroep waarover Wiliam (2010) rapporteert, kwam wel tot de conclusie dat veel van de curriculumdoelen in de Verenigde Staten hiervoor moesten worden geherformuleerd. Ze werden breder geformuleerd, zodat per leerdomein ongeveer twaalf doelen overbleven. Ook het taalgebruik werd aangepast zodat het doel voor leraren zeer duidelijk is.

2.2.5 Onderwerp van de toetsen

In de meeste leer(winst)monitoringssystemen zijn gestandaardiseerde toetsen beperkt tot het meten van kennis binnen een aantal basis-leerdomeinen, zoals rekenen/wiskunde en moedertaalonderwijs. In het kader van het schooleffectiviteitsonderzoek wordt evenwel al geruime tijd geargumenteed dat ook andere vaardigheden binnen een ruimer aantal leerdomeinen moeten worden getoetst, omdat niet kan worden verondersteld dat de cognitieve resultaten binnen rekenen en taal kunnen veralgemeend worden naar andere vormen van output (vaardigheden, attitudes, competenties) en andere leerdomeinen (Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988; Opdenakker & Van Damme, 2000; Purkey & Smith, 1983). Wanneer er gestandaardiseerde toetsen worden ontwikkeld voor meerdere leerdomeinen, vermindert de kans op een vernauwing van het curriculum (zie 2.2.2). Leraren en leerlingen worden dan gestimuleerd om alle leerinhouden te verwerken. Ondanks het pleidooi voor 'brede toetsing', kan het belang van de meer frequente onderwerpen van gestandaardiseerde toetsen niet worden ontkend. Zo bracht een analyse aan het licht dat er een causaal verband bestaat tussen enerzijds de cognitieve wiskunde- en wetenschapsprestaties van leerlingen, en de economische groei van een bepaald land, waarbij dit verband niet door andere factoren kon worden verklaard (Hanushek & Kimko, 2000)

2.2.6 Verplichte of vrijwillige deelname

De meeste literatuur die de effecten van leer(winst)monitoringssystemen beschrijft, is gericht op verplichte examens. Toch hoeven leer(winst)monitoringssystemen niet altijd een verplichte status te hebben. Sommige auteurs menen dat het beter is om deze systemen enkel ter beschikking te stellen

voor scholen die er zelf voor kiezen om ze voor alle of voor specifieke leerlingen af te nemen (Barton, 1999).

De keuzevrijheid kan niet enkel op schoolniveau, maar ook op het individuele leerlingenniveau worden gelegd. Leerlingen kunnen dan kiezen in welke leerdomeinen ze toetsen afleggen en welke moeilijkheidsgraad van toets ze verkiezen. Er zijn verschillen tussen onderwijssystemen in de mate waarin deze keuzes gebonden zijn aan het gekozen studietraject van de leerling: in sommige landen leggen leerlingen voor opgelegde vakken een examen af op de moeilijkheidsgraad dat ze gestudeerd hebben (bijvoorbeeld een *basic* of *advanced* course), terwijl in andere landen er geen automatische koppeling is tussen gevolgde lessenspakket en de examens (Klein & Van Ackeren, 2012).

Klein en Van Ackeren (2012) stellen daarnaast vast dat er ook nog differentiatie bestaat inzake de vrijheid voor leerlingen om de leerdomeinen te selecteren waarbij men een gestandaardiseerde toets (in dit geval voornamelijk een eindexamen) afneemt. In de meeste regio's is zowel het afleggen van toetsen in de moedertaal als het afleggen van enkele hoofdvakken verplicht. Toch bestaan er grote verschillen tussen de verschillende regio's: terwijl in Engeland een zeer grote vrijheid in leerdomeinen bestaat (voor bijvoorbeeld het GCSE), is de keuzevrijheid in Finland eerder beperkt. Leerlingen leggen naast het examen in de moedertaal ook vier examens af en mogen hierbij kiezen uit vijf mogelijkheden. Het grote voordeel van een gesloten keuze is het verschaffen van meer duidelijkheid aan scholen en leerlingen, maar zou ook verschuivingen in het curriculum (*teaching to the test*) in de hand kunnen werken. Een open keuze heeft dan weer als mogelijk nadeel dat leerlingen worden aangemoedigd om leerdomeinen te studeren waar men de grootste opportuniteit (op goede punten) ziet (Klein & Van Ackeren, 2012).

2.2.7 Toetsvorm

Gestandaardiseerde toetsen nemen meestal de vorm aan van een geheel van korte, makkelijk te verbeteren items (Baird et al., 2013; Kühn, 2011). Vaak worden meerkeuze-items ingezet omdat de betrouwbaarheid van de resultaten van meerkeuzevragen meestal hoger ligt: om een toets te maken met open antwoorden die even betrouwbaar is, zijn veel meer items nodig, waardoor zowel de afname als de verbetering ervan heel arbeidsintensief worden (Cunningham, 2005). De validiteit van meerkeuzetoetsen is dan weer vaak lager dan bij open antwoorden, omdat zowel leraren als leerlingen vaak minder het gevoel hebben dat er effectief getoetst wordt wat er werd aangeleerd, en omdat het moeilijker is om door middel van meerkeuze-items 'hogere orde'-vaardigheden te meten (Haladyna, Downing, & Rodriguez, 2002).

Verschillende auteurs menen dat leer(winst)monitoringssystemen ook meer 'authentieke' taken zouden moeten bevatten. Enkel op die manier kan het resultaat van een gestandaardiseerde toets immers een valide weergave te zijn van de kennis en vaardigheden van de leerlingen (Banta, Griffin, Flateby, & Kahn, 2009; Gandal & McGiffert, 2003; Shepard, 1992; Wiggins, 1990). Bovendien is het nodig om 'hogere orde'-vaardigheden te meten in gestandaardiseerde toetsen om leraren net aan te zetten om deze vaardigheden tijdens het lesgeven als doel van het leerproces te beschouwen - en dus om het beoogde effect dat gestandaardiseerde effecten zullen leiden tot beter lesgeven - te bereiken (zie 1.1.3) (Longo, 2010). Ook voor een ander beoogd effect, met name dat de eigen evaluatie van leraren door gestandaardiseerde toetsen naar een hogere kwaliteit zal worden getild (zie 1.1.5) is de inclusie van authentieke taken nodig: *"Ideally, the end of course examination should*

be complex, demanding and stimulating, so that the backwash from the examination encourages teachers to develop complex and varied teaching material to support students' critical thinking, participation and motivation" (Baird et al., 2013, p. 21).

Authentieke taken zijn echter moeilijker te standaardiseren: de verhoogde validiteit van de taken gaat dan ten koste van de betrouwbaarheid. Bovendien is de afname ervan veel intensiever (Wang et al., 2006). In het werk van de eerder genoemde expertengroep gerapporteerd door Wiliam (2010) wordt besloten dat authentieke taken nodig zijn, maar spaarzaam moeten worden ingezet. Hathcoat en Penn (2012) pleiten er daarentegen voor om bij het evalueren en vergelijken van authentieke taken van leerlingen een groot aantal taken per student te verzamelen. Zij stellen dat er een aantal contextgebonden factoren zijn die de kwaliteit van de taak sterk kunnen beïnvloeden. Zo kan afhankelijk van de inhoud van de toets de motivatie van een leerling verschillend zijn, wat de prestaties op de toets beïnvloedt (Hathcoat & Penn, 2012). Ook andere toetsomstandigheden kunnen de kwaliteit van zulke taak beïnvloeden: denk maar aan een leerling met beperkte ICT-vaardigheden, waardoor een authentieke taak niet enkel de schrijf-, maar ook de typvaardigheden van een leerling mee in rekening zal brengen (Russell & Plati, 2000) - zie ook de sectie 2.2.1 rond validiteit van toetsing.

2.2.8 Frequentie van toetsafname

De frequentie waarmee leerlingen een gestandaardiseerde toets wordt aangeboden, hangt idealitiek samen met de beoogde doelstelling. Om data te verzamelen die op microniveau bruikbaar zijn, is een hogere frequentie nodig dan wanneer op meso- of macroniveau doelstellingen worden nagestreefd. Shepard (1992) geeft aan dat ook verantwoordingsgerichte toetsen op microniveau (waarbij leerlingen hun kennis moeten aantonen in de context van een high-stakes eindexamen) over verschillende toetsen zouden moeten worden gespreid. Zij citeert een expertengroep die fungeerde binnen het Amerikaanse National Education Goals Panel: *"Consideration should be given to permitting students to accumulate examination credits over several years rather than sitting for a single pass-or-fail test. This method could accommodate different rates of learning, yet hold all students to a high standard of achievements"* (NEGP Resource Group (1991), in Shepard, 1992, p. 9).

In dezelfde lijn wordt gesteld dat het evalueren van leraren op basis van de leerprestaties van leerlingen rekening moet houden met verschillende metingen op verschillende tijdstippen, omdat éénmalige resultaten weinig betrouwbare indicatoren van de kwaliteit van het lesgeven van individuele leraren zijn (Loeb, 2013).

Er is evenwel weinig recent onderzoek bekend naar de frequentie waarmee gestandaardiseerde toetsen moeten worden afgenomen om een optimaal effect te hebben. Een reviewstudie bracht de resultaten bijeen van 40 studies in de Verenigde Staten tussen 1928 en 1989, en stelde vast dat leerlingen op een eindexamen betere studieresultaten behalen als ze regelmatig op een gestandaardiseerde wijze getoetst worden. Men stelde vooral een groot verschil vast tussen leerlingen in de groep die in het semester voor het eindexamen één of twee gestandaardiseerde toetsen hadden gekregen, en de groep leerlingen die voor het eindexamen nog geen enkele gestandaardiseerde toets had afgelegd. Het aanbieden van meer dan twee voorbereidende toetsen bleek nog altijd te leiden tot betere leerprestaties, maar hier was het verschil minder uitgesproken (Bangert-Drowns, Kulik, & Kulik, 1991). Het blijft evenwel onduidelijk of het in deze (oudere) studies

effectief gaat om betere leerprestaties (bijvoorbeeld door aangepaste instructie op basis van de feedback van eerdere toetsen), of louter om een effect van toetsvoorbereiding op een high-stakes examen.

Om zinvolle informatie op te leveren met betrekking tot leerwinst en toegevoegde waarde, wordt in het pilootproject 'Leerwinstmonitoring' aangeraden om de tijd tussen verschillende metingen beperkt te houden. Zo menen zij dat het moeilijk is betekenis toe te kennen aan een toegevoegde waarde-meting gebaseerd op cijfers aan het begin en aan het eind van het lager onderwijs (Janssens et al., 2014). Om ontwikkelingsgerichte doelen op microniveau na te streven is het dan weer nodig om niet aan het eind van een schooljaar of cyclus te toetsen, maar eerder bij aanvang en halverwege het leerproces: zo levert de gestandaardiseerde toets informatie op die gebruikt kan worden om tijdig bij te sturen (Black & Wiliam, 2005; Janssens et al., 2014).

2.2.9 Wie ontwikkelt de gestandaardiseerde toetsen?

In bijna alle onderwijssystemen waar de overheid investeert in een leer(winst)monitoringssysteem, is er een centrale autoriteit die de verantwoordelijkheid draagt voor de ontwikkeling, de afname en de verbetering van de gestandaardiseerde toetsen. Deze centrale autoriteit is een non-profit organisatie die vaak gelieerd of ondergeschikt is aan het Ministerie van Onderwijs. Er zijn evenwel enkele uitzonderingen, zoals het Verenigd Koninkrijk - toetsontwikkeling is daar overgelaten aan private organisaties - en het CITO dat in Nederland instaat voor de toetsontwikkeling. In beide gevallen is er evenwel een autoriteit die toezicht houdt op het opgeleverde werk van deze organisaties (Klein & Van Ackeren, 2012).

Verschillende auteurs bepleiten de betrokkenheid van leraren in het ontwikkelingsproces. Dit kan immers de kwaliteit van de toetsen versterken, alsook de aanvaarding ervan door het onderwijsveld positief beïnvloeden (Upton & Supovitz, 1996). Toch zijn er ook nadelen verbonden aan deze betrokkenheid van leraren. Het ontwikkelingsproces vraagt immers niet enkel meer tijd, maar wordt ook duurder. Door het vergroten van het aantal betrokkenen kan ook de moeilijkheidsgraad van de toetsen dalen: *"the more people who come to the table to arrive at consensus over the content of new tests, the more likely it is that ambitious, challenging content will be negotiated out"* (Shepard, 1992, p. 16).

2.2.10 Wie verbetert de afgenomen gestandaardiseerde toetsen?

De verbetering van afgenomen gestandaardiseerde toetsen is uiteraard afhankelijk van de gekozen toetsvorm: digitale toetsen met eenvoudige antwoordtypes kunnen meestal automatisch worden verbeterd. In andere gevallen is de werkwijze voor de verbetering van toetsen vaak afhankelijk van het doel van de toetsen. Wanneer het doel eerder verantwoordingsgericht is, zal met meer zorg gewaakt worden over de kwaliteit van het proces van verbetering (soms ten koste van de efficiëntie). In de meeste Angelsaksische landen wordt slechts één evaluator ingezet, terwijl gestandaardiseerde toetsen in andere landen vaak door twee, van elkaar onafhankelijke, evaluatoren wordt verbeterd - met, zoals in Denemarken, de mogelijkheid tot een derde evaluator wanneer de resultaten inconsistent zijn (Klein & Van Ackeren, 2012). De evaluatoren zijn slechts in uitzonderlijke gevallen leraren van de school zelf: meestal gaat het om leraren van andere scholen, of leden van een specifiek hiertoe opgerichte commissie. Interne evaluatoren hebben het voordeel dat eventuele verrassingen in de resultaten niet kunnen toegeschreven worden aan een onfaire

correctie (Klein & Van Ackeren, 2012). Daar staat evenwel tegenover dat - zeker bij high-stakes toetsen die een invloed hebben op de verdere studieloopbaan van leerlingen - interne evaluatoren ertoe neigen om minder strikt om te gaan met de vooropgestelde criteria en '*soft criteria*' als inzet en gedrag in de klas het oordeel over kennis en vaardigheden mee laten beïnvloeden (Neumann et al., 2011). Bij externe evaluatoren wordt ervan uitgegaan dat het geheel meer gestandaardiseerd verloopt en dat het scholen nog meer aanzet tot kwaliteitsverbetering (maar tegelijk mogelijk ook tot onwenselijk strategisch gedrag) (Klein & Van Ackeren, 2012).

2.2.11 Triangulatie van resultaten met andere bronnen van informatie

Indien de resultaten van een gestandaardiseerde toets effect hebben op de verdere studieloopbaan van leerlingen, geven verschillende auteurs aan dat de resultaten niet de enige bron mogen zijn waarop eventuele beslissingen zijn gebaseerd. Triangulatie met andere bronnen van informatie kan bovendien de aanvaarding door scholen verhogen (Goodman & Hambleton, 2005; Lamprianou & Christie, 2009). Gestandaardiseerde toetsen zouden dan idealiter één element vormen binnen een geheel van formatieve toetsen op klasniveau (door de leraar/school ontwikkeld), waarbij deze formatieve toetsen en de gestandaardiseerde toetsen als complementair worden beschouwd om een summatieve evaluatie te maken van elke leerling (Wang et al., 2006). Door andere auteurs wordt dit principe evenwel anders ingevuld: zo meent Cizek (2005) dat slagen voor een gestandaardiseerde toets één van de voorwaarden kan zijn voor een leerling om een diploma toegekend te krijgen (naast andere voorwaarden, als toetsen binnen de school), maar dat niet-slagen voor een gestandaardiseerde toets automatisch leidt tot het weigeren van het diploma (zelfs indien de toetsen binnen de school aantonen dat de leerling over de vereiste kennis of vaardigheden beschikt).

Hetzelfde idee wordt geopperd wanneer het gaat over verantwoording op het niveau van individuele leraren (bijvoorbeeld beslissingen inzake benoemingen of promotie): indien de gestandaardiseerde toetsen hiertoe bijdragen, mogen ze niet de enige bron van informatie zijn (Goodman & Hambleton, 2005). Zelfs indien maximaal wordt gecorrigeerd voor kenmerken van leerlingen in een klas, blijft een meting van prestaties van leraren op basis van leerprestaties van leerlingen weinig accuraat (Loeb, 2013). Ook hier geldt dat de betrouwbaarheid en validiteit van de meting door triangulatie kan worden verhoogd, maar ook dat de bruikbaarheid van de resultaten (in het kader van kwaliteitsverbetering van leraren) erdoor wordt versterkt (Harris, 2011, 2012).

2.2.12 Andere kenmerken in verband met het toetsopzet

Naast bovengenoemde kenmerken van leer(winst)monitoringssystemen, wordt er door één of meerdere auteurs ook aan enkele bijkomende elementen aandacht besteed.

Inzichtelijkheid van de resultaten

Gestandaardiseerde toetsen kunnen maar tot de beoogde effecten leiden in die mate waarin de resultaten door de betrokken actoren ook correct worden begrepen. Dit is evenwel geenszins altijd het geval: Goodman en Hambleton (2005) beweren dat de resultaten van gestandaardiseerde toetsen nog al te vaak niet voldoende kunnen geïnterpreteerd worden, noch door ouders, noch door onderwijsprofessionals in scholen en evenmin door beleidsmakers, onder meer omdat de vereiste kennis van statistiek onvoldoende aanwezig is. Zij suggereren daarom om per doelgroep

(bijvoorbeeld school, overheid, ouders) een apart specifiek rapport van de school te maken (Goodman & Hambleton, 2005).

Bruikbaarheid van de resultaten

In een kwalitatieve studie naar uiteenlopende reacties van leraren inzake een nieuwe vorm van gestandaardiseerde toetsen in Pennsylvania, werden sommige strategieën van leraren gekenmerkt als 'diepgaand' (*pervasive*), andere gebruikte strategieën als eerder oppervlakkig (*non-pervasive*). De mate waarin het gebruik diepgaand was, lijkt volgens deze onderzoekers samen te hangen met de inschatting van leraren inzake de bruikbaarheid van de resultaten. Er werden drie primaire obstakels geïdentificeerd inzake deze bruikbaarheid, met name: (1) twijfels of een éénmalige toets voldoende de variatie in leerprestaties kan meten en dus adequaat het leerniveau van leerlingen kan capteren; (2) de bekommernis dat leerlingen doorheen hun studieloopbaan de gestandaardiseerde toetsen (die voor hen low-stakes zijn) steeds minder ernstig nemen en (3) de vaststelling dat de state tests weinig nieuwe informatie opleveren voor de leraren (Beaver & Weinbaum, 2015). Om een ontwikkelingsgericht effect te hebben op microniveau is het noodzakelijk om een gestandaardiseerde toets te beperken tot een erg fijnmazige meting van een klein aspect van een bepaald leerdomein. Door die grote mate van detail wordt diezelfde toets evenwel onbruikbaar voor verantwoordingsgerichte aspecten (zeker op meso- of macroniveau), die meer baat hebben bij een meting van een volledig leerdomein (Cizek, 2005).

Hoewel de literatuur op dit vlak nog niet erg uitgebreid is, wordt hier en daar gerapporteerd over pogingen om gedifferentieerde informatie te verschaffen over de verschillende vaardigheden die leerlingen nodig hebben om één bepaald item correct op te lossen. Dit gaat via *Diagnostic Classification Models*, al duidt voorlopig onderzoek aan dat deze modellen niet altijd betere informatie geven (Kunina-Habenicht, Rupp, & Wilhelm, 2009).

Eenduidig doel

Verschillende auteurs waarschuwen voor het vooropstellen van verschillende doelen aan een gestandaardiseerde toets. Daarbij wordt vooral verwezen naar de nefaste gevolgen van een toets in ontwikkelingsgericht perspectief, wanneer aan dezelfde toets ook een verantwoordingsgerichte component is verbonden (Wang et al., 2006). Dit was ook de conclusie van een studie waarin de formatieve en summatieve functies van evaluatie werden nagegaan in vier verschillende onderwijssystemen (met name Engeland, Duitsland, Frankrijk en de Verenigde Staten): verantwoordingsgerichte aspecten maken dat het ontwikkelingsgerichte gebruik gemarginaliseerd wordt (Black & Wiliam, 2005). Bovendien kunnen de mogelijke nadelen verbonden aan verantwoordingsgerichte toetsen (bijvoorbeeld teaching to the test) maken dat de resultaten onbruikbaar worden voor verdere kwaliteitsontwikkeling (Wang et al., 2006).

Daarnaast wordt het ook niet aangeraden om met één gestandaardiseerde toets zowel doelen op microniveau als op meso- of macroniveau na te streven. Op verschillende niveaus is immers een andere vorm van informatie (en een andere mate van detail) nodig, en deze moet zich vertalen in de inhoud van de gestandaardiseerde toets (Skedsmo, 2011).

Adaptieve toetsing

Bij een adaptieve toets wordt het niveau van de aangeboden toetsitems maximaal aangepast aan het kennis- of vaardigheidsniveau van de leerling. Op basis van een beoordeling van de correctheid

van het antwoord op alle vorige items, wordt daaropvolgend een moeilijker respectievelijk makkelijker item aangeboden. Hierdoor wordt de toets meer uitdagend voor elke leerling en wordt de inschatting van de kennis of vaardigheid van de leerling telkens meer verfijnd (Brossman & Guille, 2014; Magyar, 2015; Wang & Kolen, 2001). Een voorwaarde hiervoor is dat de toets digitaal wordt afgenomen, en dat er meer items worden aangemaakt die elk op hun moeilijkheidsniveau worden ingeschaald (via Item Response theory) (Frey & Seitz, 2009; Magyar, 2015; Wang & Kolen, 2001). Recent onderzoek wees uit dat traditionele papieren toetsen kunnen dienen als basis voor digitale adaptieve toetsen, en dat deze adaptieve toetsen met een grotere precisie de vaardigheid van de leerling kunnen inschatten (Brossman & Guille, 2014; Frey & Seitz, 2009; Magyar, 2015; Molnár & Magyar, 2015), of dat minder items nodig zijn om dezelfde precisie te bereiken, wat leidt tot een grotere toetsefficiëntie (Frey & Seitz, 2009). 'Getrapt adaptieve toetsen' (ook 'multi-stage testing') - waarbij met sets van items worden gewerkt waarna wordt overgegaan naar een makkelijker of moeilijker set van items - zijn makkelijker om op te stellen dan volledig adaptieve toetsen, terwijl slechts in geringe mate informatie of precisie verloren gaat (Brossman & Guille, 2014; Molnár & Magyar, 2015).

2.3 Individuele kenmerken van leraren

Reacties op (de resultaten van) een leer(winst)monitoringssysteem verschillen niet alleen tussen scholen, maar ook tussen verschillende leraren binnen dezelfde school (Beaver & Weinbaum, 2015). De houding van leraren ten opzichte van de bruikbaarheid gestandaardiseerde toetsen bepaalt sterk in welke mate de beoogde effecten op het microniveau kunnen worden bereikt (Rozenwajn, 2014). De impact van uiteenlopende attitudes ten opzichte van de resultaten op het gebruik van de gestandaardiseerde toetsen, werd inderdaad vastgesteld door een studie van de NFER: waar negatieve attitudes werden vastgesteld, bleken deze vaak gerelateerd aan *ignorance* (het niet begrijpen van de data), aan een gebrek aan vertrouwen in de data, of aan de overtuiging dat een overheid geen gestandaardiseerde toetsen aan scholen mocht opleggen (Saunders, 2000). Die laatste bevinding duidt op de impact van de mate waarin de gestandaardiseerde toetsen als een bedreiging worden ervaren (Saunders, 2000). Ook uit een studie van Wikeley (1998) bleek dat leraren in de onderzochte secundaire scholen een wantrouwen hadden ten aanzien van kwantitatieve data, en meer vertrouwden op hun eigen 'professioneel oordeel' dan op de statische analyses.

Dat niet alle leraren overtuigd zijn van de meerwaarde van de resultaten op gestandaardiseerde toetsen, blijkt ook uit een Amerikaans onderzoeksrapport. Een groot aantal leraren beschouwen de resultaten die gestandaardiseerde toetsen opleveren als weinig representatief, aangezien er geen gegevens van leerwinstmonitoring in de resultaten worden opgenomen. Toch ziet een minderheid van de leraren wel degelijk een meerwaarde in de resultaten, hoewel ze ook vragende partij zijn naar gegevens op school- en klasniveau (Scholastic & Bill & Melinda Gates Foundation, 2012). Resultaten uit andere studies (Kerr, Marsh, Ikemoto, Darilek, & Barney, 2006; Supovitz & Klein, 2003) sluiten hier bij aan. Kerr et al. (2006) stellen eveneens vast dat een groot aantal leraren deze data zelfs volledig aan de kant legt omdat ze vinden dat resultaten van gestandaardiseerde toetsen niet tijdig worden aangeleverd of niet relevant bevonden worden.

Onderzoek naar het gebruik van de resultaten van gestandaardiseerde toetsen in scholen en districten in Pennsylvania (Dembosky, Pane, Barney, & Christina, 2005) toont echter aan dat het

merendeel van de leraren en directies deze data wel waardevol vindt en ook actief op zoek gaat naar manieren om deze data aan te wenden voor procesverbeteringen. Toch geven leraren in dit onderzoek aan dat de toegang tot bestaande data vaak belemmerd wordt door een gebrek aan datageletterdheid en weinig toegankelijke datasystemen. In een aantal studies wordt vastgesteld dat leraren, hoewel ze niet steeds het nut inzien van deze data, vaak verplicht worden door de school of overheid om er alsnog met aan de slag te gaan (Dembosky et al., 2005; Gallagher, Means, & Padilla, 2008; Hamilton, Berends, & Stecher, 2005).

2.4 Kenmerken van scholen

De mate waarin leer(winst)monitoringssystemen leiden tot positieve effecten of onwenselijke neveneffecten vertoont tevens samenhang met hoe de (resultaten van) deze systemen binnen de school door leraren, directies en schoolbesturen worden gebruikt (Firestone et al., 2002).

2.4.1 Datageletterdheid

De resultaten van leer(winst)monitoringssystemen leiden niet automatisch tot kwaliteitsverbetering in scholen. Scholen hebben een zekere vorm van datageletterdheid nodig om de data juist te interpreteren en om ze te gebruiken als leidraad voor relevante kwaliteitsverbeteringen (Beaver & Weinbaum, 2013; Tully, 2008).

Vanuit de literatuur kan volgende opdeling worden gemaakt inzake het verzamelen en gebruiken van data die verzameld werden vanuit gestandaardiseerde toetsen (Beaver & Weinbaum, 2013): (1) het gebruik van de reeds beschikbare gegevens als stimulans om verder onderzoek uit te voeren met behulp van datagerichte teams en tussentijdse evaluaties, waarbij mogelijk nieuwe gegevens worden verzameld; (2) het gebruik van data om acties te ondernemen die gericht zijn op het verbeteren van de onderwijskwaliteit van de gehele school (zoals het afstemmen van het curriculum op de eindtermen of het professionaliseren van leraren) en (3) het gebruik van data om middels remediëring en differentiatie leerlinggericht te werken. In datzelfde onderzoek stellen Beaver en Weinbaum (2013) vast dat de meeste strategieën van scholen gericht zijn op het jaarlijks verbeteren van de resultaten op de toetsen.

2.4.2 Het beleidsvoerend vermogen van de school

Effectief leiderschap heeft een impact op zowel de interne kwaliteitszorg (Wallace, 2002) als op de leerprestaties van leerlingen (Leithwood & Riehl, 2003; Scheerens, 2012). Ook schoolinspecties beoordelen een school als succesvol wanneer die in staat blijkt te zijn om haar eigen beleid vorm te geven (Perryman, 2006). Het beleidsvoerend vermogen – ofwel het vermogen van scholen om beleid te voeren – wordt als volgt gedefinieerd door Vanhoof, Deneire en Van Petegem (2011, p.15):

“De mate waarin een school de haar beschikbare beleidsruimte succesvol aanwendt om te komen tot een voortdurend proces van behouden of veranderen van haar functioneren met als doel het verbeteren van haar onderwijskwaliteit en het bereiken van de haar opgelegde en eigen doelen.”

Op basis van empirische studies worden volgende acht kenmerken van beleidsvoerend vermogen onderscheiden: doeltreffende communicatie, ondersteuning en samenwerking, gedeeld

leiderschap, gezamenlijke doelgerichtheid, responsief vermogen, innovatief vermogen, geïntegreerd beleid en reflectief vermogen (Vanhoof et al., 2011).

Uit onderzoek van Vanhoof, Verhaeghe, Van Petegem en Valcke (2012) kan worden geconcludeerd dat de meerderheid van de basisscholen in Vlaanderen schoolfeedback-data niet benut om de werking van de school in vraag te stellen en dat de mate waarin schoolresultaten worden gebruikt sterk verschilt tussen de scholen onderling. In scholen met een sterk gedeeld leiderschap worden deze data vaker ingezet dan in andere scholen (Vanhoof et al., 2012; Lange, Range, & Welsh, 2012). Cosner (2011) stelt dat schoolleiders niet enkel belangrijke 'betekenis-gevers' zijn om met data aan de slag te gaan, maar dat zij ook de manier waarop leraren deze data gebruiken sterk beïnvloeden. Wikeley (1998) besluit uit onderzoek dat data over de toegevoegde waarde van de school maar gebruik werden in die afdelingen van scholen die over sterke zelfevaluatie-mechanismen beschikten. Schildkamp en Kuiper, (2010) rapporteren dat slechts een beperkt aantal schoolleiders de leraren waarvan de leerlingen onderpresteren een verplichte verdere professionalisering oplegt. Ramsteck, Muslic, Graf, Maier en Kuper (2015) hebben onderzoek gedaan naar de interpretatie en het gebruik van feedback van de 'VergleichsArbeiten' (VerA), een low-stakes variant van de CITO-toets in Duitsland, door schoolleiders en schoolbesturen. Uit de resultaten blijkt dat het opzet van beleidsmakers, namelijk het verhogen van het beleidsvoerend vermogen binnen scholen, niet makkelijk te realiseren is. Hoewel schoolleiders de meerwaarde van het verplicht evaluatiesysteem erkennen, blijkt dat een minderheid van de schoolleiders inspanningen levert om het gebruik en de reflectie ervan in te bedden in de dagelijkse werking van de school. De bespreking van de resultaten beperkt zich veeleer tot een moment tijdens personeelsvergaderingen of tot gesprekken op leraarniveau. Toch werd er een positieve impact op datagebruik door schoolleiders vastgesteld (Ramsteck et al., 2015).

2.4.3 Aanwezigheid van een draagvlak binnen de school

Zelfs indien aan bovenstaande voorwaarden wordt voldaan, is de kans klein dat men tot een effectief gebruik van de resultaten van een leer(winst)monitoringssysteem zal overgaan indien in scholen de overtuiging leeft dat dit systeem hen door een overheid of een koepelorganisatie wordt opgelegd (Fielding, 1999). Onder meer het onderzoek van Saunders (2000) wees er inderdaad op dat niet-aanvaarding van gestandaardiseerde toetsen op zich een voldoende reden kan zijn voor scholen of individuele leraren om enkel oppervlakkig van de resultaten gebruik te maken. Zo moet de beslissing om aan gestandaardiseerde toetsen deel te nemen, gedragen zijn door het schoolteam. Ramsteck et al. (2015) concluderen dat een dergelijk evaluatiesysteem door leraren wordt aanzien als een bedreiging van hun professionele identiteit. Wanneer leraren aanvoelen dat data enkel 'van bovenaf' worden opgelegd, is dit een ernstige bedreiging voor het ontwikkelingsgericht gebruik (Wikeley, 1998).

Dit draagvlak hangt sterk samen met de gepercipieerde kwaliteit van het leer(winst)-monitoringssysteem. De mate waarin een school deze data zal inzetten, is sterk afhankelijk van de opinie van de school inzake de bijdrage van de gestandaardiseerde toetsen tot het stimuleren van interne kwaliteitsprocessen. Daarom moeten deze data een representatieve evaluatie zijn van doelen die belangrijk en haalbaar geacht worden alsook moet het gebruik ervan de doelen en de missie van een school ten goede komen (Massell, 2001).

2.5 Kenmerken van het onderwijssysteem

2.5.1 Aanwezigheid van een draagvlak binnen de overheid

Onderzoek naar datagebruik op macroniveau toont aan dat beleidsmakers onderling vaak verschillende percepties hebben over de waarde en het doel van de door een leer(winst)monitoringssysteem aangeleverde data (Coburn & Talbert, 2006). Daar waar een beleidsmaker de voorkeur kan geven aan het evalueren van de kwaliteit van een toets, zijn andere personen dan weer meer te vinden voor het beoordelen van academische doelen of onderwijskwaliteiten van leraren. Coburn en Talbert (2006) benadrukken dan ook dat uiteenlopende doelen zowel het gebruik van data als het gewicht dat aan de data wordt toegekend sterk kunnen beïnvloeden. Kerr et al. (2006) identificeren twee factoren die het datagebruik op macroniveau sturen: de mate van het verantwoordingsgericht karakter van het toetssysteem en de toegankelijkheid en de actualiteit van de aangeleverde data.

2.5.2 Investerings in professionele ontwikkeling en begeleiding van scholen

Een sleutelkenmerk in het succes van gestandaardiseerde toetsen, zoals bepaald door de eerder besproken expertengroep in onderzoek van Wiliam (2010), is dat er sterk geïnvesteerd wordt in de professionele ontwikkeling. In dit kader wordt professionele ontwikkeling en diepgaande begeleiding bepleit inzake datageletterdheid van scholen (zie 2.4.1). Daarnaast moet ook worden geïnvesteerd in het bevorderen van een datacultuur in school om deze data in te zetten in het proces van interne kwaliteitszorg en in een verhoogde mate van 'opbrengstgericht werken' (Janssens et al., 2014; Ladd & Walsh, 2002). Ook moet worden ingezet op de capaciteit van leraren om zowel hun eigen evaluatiemechanismen te versterken (Boyd, 2008) als om gericht te differentiëren in de klas (Roeber, 1995). Uit een empirisch onderzoek in Engeland bleek immers dat de houding van leraren ten aanzien van het gebruik van de data positiever was in een groep van scholen waarbij begeleiding werd opgezet, in vergelijking met leraren in een scholen die geen begeleiding kregen bij dezelfde data (Tymms, 1997).

2.5.3 Duidelijk kader voor de onderwijsinspectie

Hoewel het eigenlijk opzet van de centrale examens in Nederland een vergelijking tussen de diverse scholen is, maakt de schoolinspectie ook gebruik van de resultaten om een kwaliteitsoordeel uit te spreken over scholen. Enkel wanneer deze data onregelmatigheden aan het licht brengen, gaat de inspectie de schoolprocessen verder onderzoeken. Wanneer scholen er dan niet in slagen om de resultaten te verbeteren overheen een afgesproken termijn, worden er sancties getroffen. De inspectierapporten worden, samen met de examenresultaten, publiekelijk beschikbaar gemaakt op de website van de inspectie.

Wilcox en Gray (1996) stellen dat de inspectie kan worden beschouwd als een deel van een audit waarin het gebruik van kwantitatieve data kan worden verbonden met het evalueren van de effectiviteit, kwaliteit en prestaties. Recent onderzoek dat het gebruik van centrale toets- of examenresultaten door onderwijsinspecties in kaart brengt, ontbreekt voorlopig.

3 CONCEPTUALISERING VAN LEERWINST EN TOEGEVOEGDE WAARDE

In het eerste deel van dit boek werd een eerste afbakening van de begrippen 'leerwinst' en 'toegevoegde waarde' gegeven. Leerwinst heeft steeds betrekking op het leren van individuele leerlingen (microniveau). De term 'toegevoegde waarde' wordt gereserveerd voor de bijdrage van scholen aan het leren of aan de leerwinst van leerlingen. Toegevoegde waarde situeert zich dus op het mesoniveau. Zoals uit onderstaande bespreking van de literatuur zal blijken, zijn er evenwel nog zeer uiteenlopende conceptualiseringen van beide begrippen gangbaar.

3.1 Leerwinst

Met leerwinst wordt de voortgang van leerlingen in een bepaalde vaardigheid bedoeld (De Maeyer & Rymenans, 2004). Leerwinst is een maat voor de groei van individuele leerlingen, in het algemeen of binnen een specifiek leerdomein (Janssens et al., 2014). Zelfs wanneer is afgelijnd dat de term leerwinst zich enkel richt op het microniveau van leerlingen, kan de term nog steeds verschillende ladingen dekken. Verschillende auteurs bedoelen erg uiteenlopende zaken wanneer ze schrijven over leerwinst. Ook in de praktijk worden verschillende conceptualiseringen door elkaar gebruikt. Ver Eecke (2004, p.150) spreekt in dit kader van een "babylonische spraakverwarring". Er bestaat dan ook een grote verscheidenheid in de bestaande modellen van leerwinst.

In een veel gebruikte definitie wordt leerwinst geconceptualiseerd als het verschil tussen een bepaald prestatieniveau en een verwacht prestatieniveau. Doordat er een vergelijking wordt gemaakt met een verwacht prestatieniveau, is het bij deze conceptualisering mogelijk om een inschatting te maken van 'voldoende' of 'onvoldoende' leerwinst (Castellano & Ho, 2013).

Het verwachte prestatieniveau kan theoretisch op twee manieren worden bepaald: (a) op basis van leerlingkenmerken of (b) op basis van het eerdere prestatieniveau van de leerling. De eerste manier (a) is de enige definiëring van leerwinst waarbij één meting van de leerprestaties volstaat om leerwinst te kunnen bepalen. Leerwinst is dan het verschil tussen een actuele leerprestatie en de verwachte prestatie op basis van de achtergrondkenmerken van de betreffende leerling. Deze conceptualisering wordt evenwel minder frequent gehanteerd, onder meer omdat ze de impliciete boodschap geeft aan leerlingen uit een bepaalde (minoriteits-)groep dat van hen minder wordt verwacht dan van hun meer bevoorrechte klasgenoten. Deze vorm van leerwinst wordt voornamelijk gebruikt als intermedium om in een volgende fase toegevoegde waarde volgens een 'status'-model te meten (Ladd & Walsh, 2002) (zie 3.2.1 voor informatie over status-modellen inzake toegevoegde waarde).

In beide gevallen wordt uitgegaan van een normgroep: men verwacht dat een leerling een bepaald leerniveau heeft bereikt, gebaseerd op het gemiddeld leerniveau van leerlingen met hetzelfde profiel (op basis van kenmerken, of van eerder prestatieniveau). Leerwinst verwijst hier naar een 'meer dan normale' vooruitgang van leerlingen. De leerwinst van leerlingen kan ook negatief zijn, als een leerprestatie onder de verwachte leerprestatie blijft.

Beide benaderingen kunnen ook worden gecombineerd, zoals bij Ladd en Walsh (2002). Zij stellen voor om leerwinst te berekenen volgens de formule

$$A_{it} = \lambda A_{it-1} + \alpha_v S_t + \beta_t F_{it} + e_{it}$$

waarbij A_{it} de leerprestatie van een leerling i is op het meetmoment t , en A_{it-1} de leerprestatie van diezelfde leerling op een eerder meetmoment. Andere factoren die in deze berekening worden genomen zijn kenmerken van de school (S_i), achtergrondkenmerken van de leerlingen (F_{it}) en een errorterm (e_{it}). Daarbij is opvallend dat de errorterm eigenlijk gelijk wordt gesteld aan de leerwinst van een bepaalde leerling (met bepaalde kenmerken) in een bepaalde school. In een andere toepassing wordt naast de voorgaande prestatie ook gecorrigeerd voor het geslacht, de etniciteit, een maat voor armoede, speciale onderwijsbehoeften en beheersing van de moedertaal (Ray, 2006). Er zijn evenwel vele mogelijke correctiefactoren, en welke precies al dan niet in rekening worden genomen, is onderwerp van discussie (Wang et al., 2006). Bovendien moet daarna nog gekozen worden hoe deze factoren zullen worden geoperationaliseerd: de 'socio-economische status van leerlingen' kan bijvoorbeeld gemeten worden aan de hand van onder meer het opleidingsniveau van (één van) de ouders, en/of van het beroep van (één van) de ouders. Deze keuzes kunnen een grote invloed hebben op de uiteindelijke berekening van leerwinst (De Maeyer & Rymenans, 2004).

Hieronder wordt dieper ingegaan op de verschillende conceptualisering van leerwinst op basis van het eerdere prestatieniveau (b). Het meest omvattend overzicht van verschillende leerwinstmodellen op basis van eerdere prestatiemomenten kan worden teruggevonden bij Castellano en Ho (2013). Zij onderscheiden zes vormen van leerwinstmodellen die verschillen in de mate waarin ze antwoord kunnen geven op de volgende twee vragen:

- In welke mate is er (absolute of relatieve) groei? (beschrijven van de groei)
- Welke leerprestatie mag van leerlingen verwacht worden op basis van eerdere en huidige leerprestaties? (voorspellen van de groei)

De onderverdeling die hieronder gehanteerd wordt, is evenwel ook op een nog ander onderscheid gebaseerd. Geënt op de statistische onderbouw wordt elk van de zes modellen ondergebracht bij ofwel de *gain-based models* ($n=3$), ofwel de *conditional status models* ($n=3$).

Bij *gain-based models* wordt uitgegaan van de groei van leerlingen op basis van twee of meerdere toetsmomenten. Specifiek aan deze modellen is dat er wordt betrokken van een gemeenschappelijke verticale schaal, waarop alle toetsen kunnen worden gescoord. Daardoor kunnen de resultaten van verschillende toetsen met elkaar worden vergeleken. Deze verticale schalen zijn evenwel moeilijk te ontwikkelen terwijl ze niet steeds noodzakelijk zijn om uitspraken te doen over leerwinst of toegevoegde waarde.

De *conditional status models* beoordelen de prestatie van een leerling aan de hand van een vergelijking met een vooraf bepaalde verwachting. Deze verwachting kan op voorhand voor elke leerling vastgelegd worden op basis van de eerdere prestaties van deze en andere leerlingen (en indien dat wenselijk wordt geacht ook op basis van andere factoren, bijvoorbeeld relevante factoren die buiten de invloedssfeer van de school liggen). Deze modellen vereisen geen gemeenschappelijke verticale schaal.

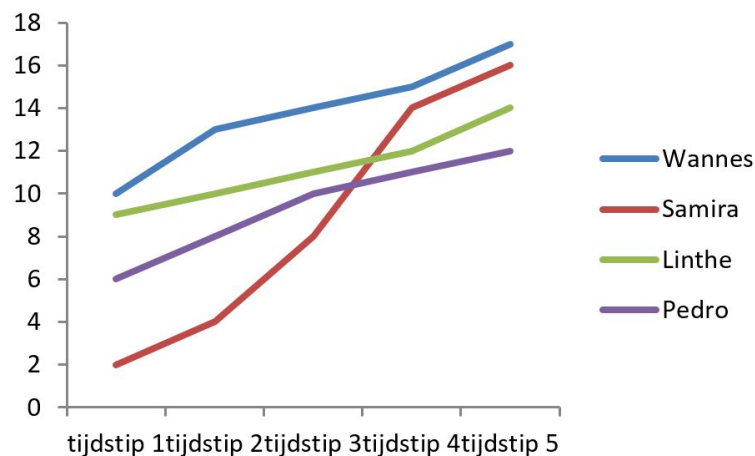
Tabel 7 geeft een overzicht van de zes modellen volgens zowel de statistische onderbouw als de vraag die er primair wordt beantwoord (Castellano & Ho, 2013). Uit deze tabel blijkt duidelijk een sterke samenhang tussen de gestelde richtvraag enerzijds, en de statistische onderbouw van de modellen anderzijds.

Tabel 7 - Manieren om leerwinst te berekenen naar statistische onderbouwing en richtvraag

	Beschrijving van de groei	Voorspellen van de groei
Gain-based models	Gain score model Categorical model	Trajectory model
Conditional status models		Residual gain model Projection model Student growth percentile model

3.1.1 Gain-based models

Bij een *gain score model* wordt de leerwinst van alle leerlingen berekend op basis van de verschilscore tussen twee toetsen die op eenzelfde schaal zijn ingeschaald. Leerwinst wordt gebruikt als term om te refereren naar de voortgang van leerlingen. Voorbeelden zijn de afnames van verschillende toetsen met eenzelfde schaal op verschillende tijdstippen, of het evalueren van het aantal competenties dat een bepaalde leerling op verschillende momenten in diens studieloopbaan verworven heeft.

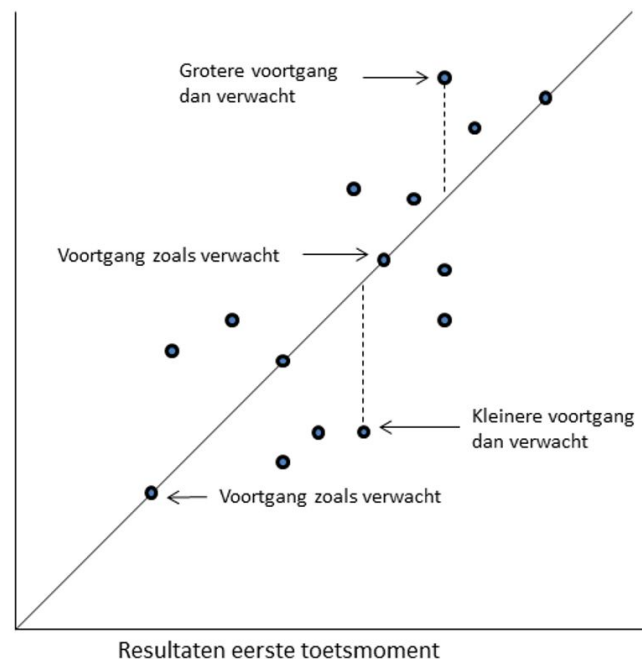


Figuur 4 - Gain score model (op basis van terugkerende evaluatie van aantal bereikte competenties)

Van de grafiek in Figuur 4 kan bijvoorbeeld worden afgeleid dat Samira 14 nieuwe competenties heeft aangeleerd (van 2 naar 16), terwijl Pedro in dezelfde tijdspanne maar zes nieuwe competenties heeft aangeleerd (van 6 naar 12). Wannes beschikt aan het eind over het meest verworven competenties, maar zijn 'leerwinst' bestaat slechts uit zeven nieuwe competenties (van 10 naar 17).

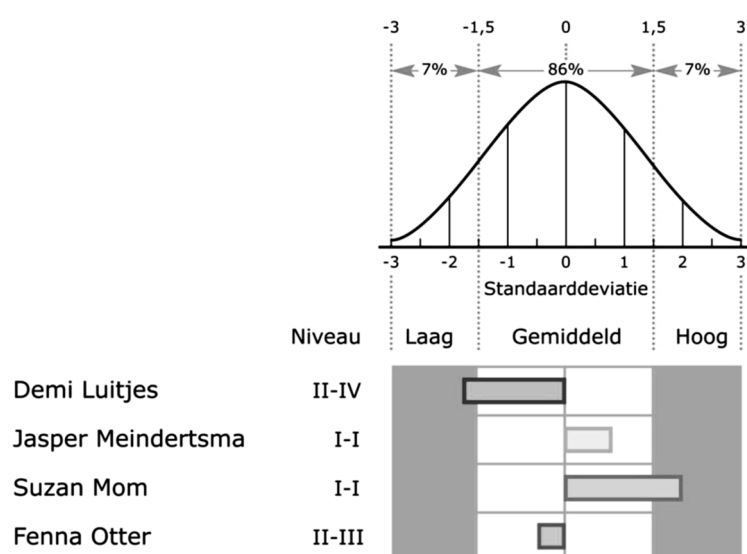
Bovenstaande conceptualisering geeft echter geen kader om de leerwinst te interpreteren: er is geen maat voor 'voldoende' of 'onvoldoende' leerwinst, omdat er geen verwachte eindscore wordt vooropgesteld (Castellano & Ho, 2013; Janssens et al., 2014) - hierdoor voldoet bovenstaande conceptualisering dan eigenlijk ook niet aan de omschrijving van leerwinst als het verschil tussen een bepaald prestatieniveau en een verwacht prestatieniveau. In de meeste toepassingen van een *gain score model* wordt daarom de verwachte leerprestatie bepaald op basis van één voorgaande meting van het leerniveau (Castellano & Ho, 2013; Goldschmidt et al., 2012). Figuur 5 stelt leerwinst voor volgens deze benadering. Leerwinst is hier de verticale afstand tussen de prestatie van een leerling en de lijn die de gemiddelde verwachting aanduidt. Die lijn wordt opgemaakt op basis van de gehele populatie (Castellano & Ho, 2013). Een andere alternatieve voorstelling van deze

berekening gaat uit van percentielscores, waarbij aan elke leerling wordt aangegeven hoe zijn/haar leerwinst zich verhoudt ten opzichte van de leerwinst van een referentiegroep (Janssens et al., 2014).



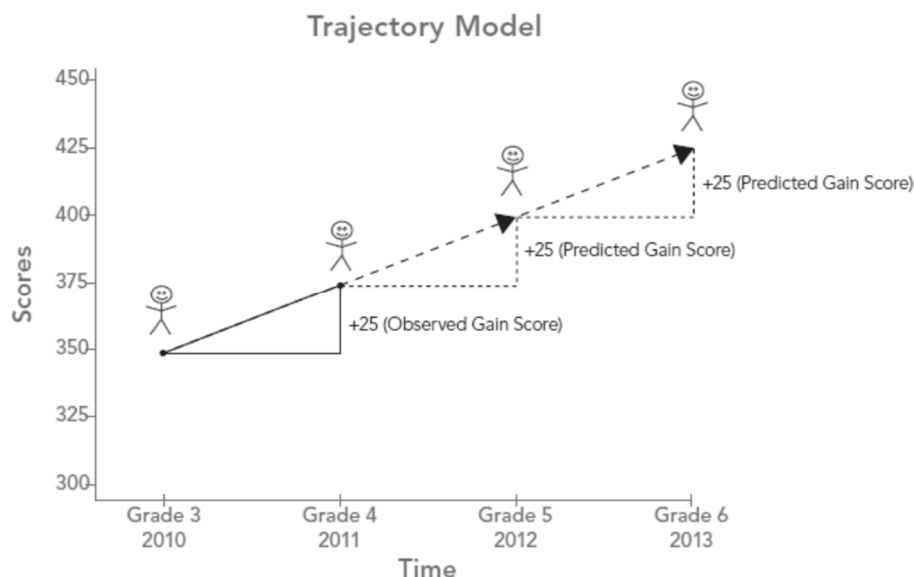
Figuur 5 - Gain score model (leerwinst als verschil met verwachte score op basis van eerder toetsresultaat)

Daarnaast kan binnen dit *gain score model* ook een standaarddeviatie worden berekend: in welke mate wijkt de groep leerlingen, met een gelijke beginscore, af van het voor hen verwachte punt op de regressielijn? Een voorbeeld hiervan is de rapportering ontwikkeld in het pilootproject Leerwinst en Toegevoegde Waarde in Nederland (zie figuur 6) (Janssens et al., 2014).



Figuur 6 - Voorbeeldrapportage pilot Leerwinst en Toegevoegde Waarde (Janssens et al., 2014)

In het *trajectory model* wordt voor elke leerling apart een ‘verwachte voortgang’ opgemaakt op basis van minimaal twee voorgaande toetsen (zie Figuur 7). Dit model vertrekt vanuit de assumptie dat de groei van leerlingen volgens een lineaire lijn verloopt: van een leerling die de voorbije twee trimesters telkens drie punten gestegen is op een verticale schaal, wordt verwacht dat hij het volgende trimester ook drie punten zal stijgen (Castellano & Ho, 2013).



Figuur 7 - Voorbeeld van verwachte leerwinst volgens het trajectory model

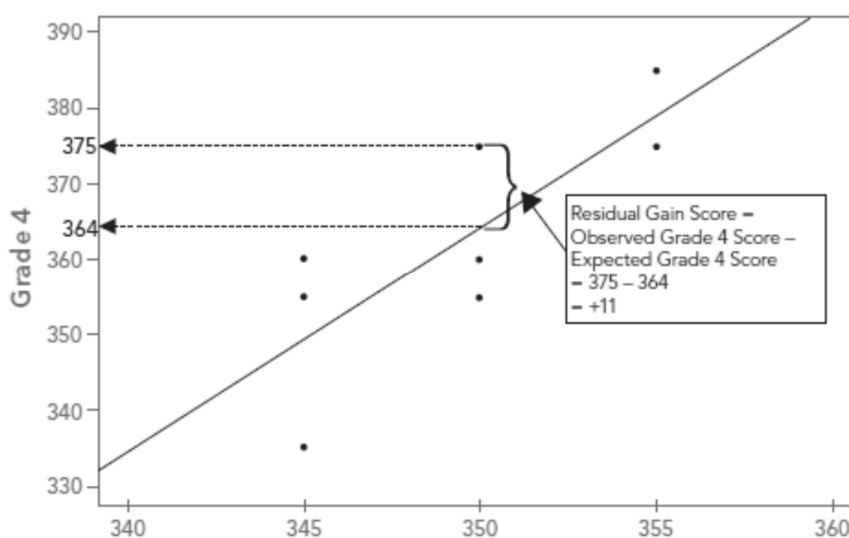
Het *categorical model* is zeer gelijkend met het *gain score model*, enkel wordt hier met categorieën van prestaties gewerkt (bijvoorbeeld in categorieën van 1 tot 5, waarbij 1 de 20% hoogste presteerders aangeeft, 2 de leerlingen tussen 20% en 40%, enz.). Deze categorisering draagt de impliciete verwachting dat leerlingen binnen een bepaalde categorie zullen blijven, en maakt het daardoor makkelijker om de groei van leerlingen niet enkel te beschrijven, maar ook om ze te voorspellen (waarbij opnieuw een actuele categorie met een verwachte categorie kan worden vergeleken) (Castellano & Ho, 2013).

Elk van deze *gain-based* modellen gebruikt een ‘verticale schaal’, via *Item Response Theory (IRT)*. In de context van gestandaardiseerde toetsen is de meest gebruikte vorm van IRT de inschaling van items volgens een Rasch-analyse. Rasch is een techniek die toelaat om de resultaten van twee toetsen met elkaar te vergelijken, omdat ze op een identieke schaal worden ingeschaald, zelfs wanneer ze verschillende items bevatten (Taylor & Lee, 2009). Vaak wordt IRT ingezet om de resultaten van toetsen afgestemd op de leerstof van verschillende leerjaren toch met elkaar te kunnen vergelijken. De ontwikkeling van zulke verticale schaal is erg arbeidsintensief. Bovendien is het niet altijd mogelijk om een gemeenschappelijke schaal te vinden, bijvoorbeeld als twee leerstofdomeinen te ver uit elkaar liggen (bijvoorbeeld twee ‘takken’ van wiskunde) (Castellano & Ho, 2013). IRT is daarom makkelijker in te zetten bij niet-curriculum gerelateerde toetsen die een algemene ontwikkeling in kaart willen brengen.

3.1.2 Conditional status-models

Conditional status-models onderscheiden zich in de eerste plaats van de *gain-based models*, door de afwezigheid van een verticale schaal waarop verschillende toetsen zijn ingeschaald.

Bij het *residual gain model* wordt een voorspelling gemaakt van toekomstige scores op basis van een lineaire regressielijn. Deze regressielijn wordt bijvoorbeeld berekend op basis van de prestaties van leerlingen uit dezelfde cohorte, die op hetzelfde ogenblik de initiële toets en vervolgtoets afleggen. Van deze regressielijn kan op basis van het eerdere toetsresultaat van elke leerling (op de initiële toets) een voorspelling worden gemaakt voor de prestatie op de vervolgtoets. In tegenstelling tot bij de *gain-based models* worden de items van beide toetsen niet op dezelfde schaal ingeschaald. Door middel van lineaire regressie op alle resultaten wordt de lijn bepaald die het best de relatie weergeeft tussen de voorspellende toets en de toets die de leerwinst van de leerling moet meten. Ook hier bestaat de leerwinst uiteindelijk uit het verschil tussen de behaalde toetsscore en de verwachte toetsscore (Castellano & Ho, 2013). In Figuur 8 wordt een voorbeeld gegeven van een regressielijn die de gemiddelde scores in een vierde leerjaar voorspelt op basis van prestaties in het derde leerjaar.



Figuur 8 - Voorbeeld van leerwinst volgens het residual gain model (Castellano & Ho, 2013)

In het *projection model* wordt eveneens een score voorspeld door een lineaire regressielijn. In tegenstelling tot bij het *residual gain model* gebeurt dit op basis van de toetsprestaties van leerlingen (uit één van de voorgaande cohorten) die eerder zowel de voorspellende toets als de outcome-toets hebben afgelegd (Castellano & Ho, 2013).

In het *student growth percentile model*, ten slotte, wordt leerwinst geconceptualiseerd als de percentielrank van een leerling wanneer zijn leerprestatie wordt vergeleken met deze van leerlingen met vergelijkbare eerdere toetsresultaten. Een leerling die bijvoorbeeld een percentiel 75 behaalt, scoort beter op een toets dan 75% van de andere leerlingen met een soortgelijk initieel profiel, en behaalt dus een leerwinst ruim boven de mediaan. Idealiter gezien bestaat de referentiegroep enkel uit leerlingen die in het verleden identieke toetsresultaten behaalden, maar dit levert vaak een relatief kleine en dus weinig betrouwbare vergelijkingsbasis op (zeker wanneer meerdere voorgaande toetsen in rekening worden genomen) (Castellano & Ho, 2013; Goldschmidt et al., 2012).

3.1.3 Meer complexe modellen

Daarnaast zijn er nog meer complexe berekeningen, waarbij leerwinst anders dan als een lineaire voortgang wordt voorgesteld. Er werd immers aangetoond dat het leren van leerlingen niet lineair, maar wel curvi-lineair verloopt (William, 2010). Er zijn heel uiteenlopende modellen die de leerwinst op deze manier trachten in kaart te brengen. In hun studie overwegen Guldemonnd en Bosker (2009) tien verschillende modellen om polynomiale (niet-lineaire) groeitrajecten te berekenen, naast een lineair model. Elk van die modellen bleek relatief gelijk in de mate waarin ze variantie in leerlingresultaten kunnen verklaren (R^2). De verklaarde variantie varieert van 67,1% tot 71,0% (Guldemonnd & Bosker, 2009).

3.2 Toegevoegde waarde

Met het concept 'toegevoegde waarde' richten we ons in dit boek op de 'prestatie' van de school: het is een maat voor de bijdrage van de school aan het leren (of de leerwinst) van leerlingen (Janssens et al., 2014). Of een dergelijke maat kan bepaald worden, is al voorwerp van discussie op zich. Zo wordt geargumenteed dat verschillende scholen uiteenlopende doelstellingen hebben en eigen accenten leggen die niet altijd door een gestandaardiseerde toets in kaart kunnen worden gebracht (Ladd & Walsh, 2002).

Een rechtvaardige berekening van toegevoegde waarde betekent bovendien dat men corrigerende berekeningen uitvoert voor factoren die buiten de invloedssfeer van de scholen liggen, bijvoorbeeld de etniciteit, de socio-economische status of de intelligentie van leerlingen (Janssens et al., 2014; Ladd & Walsh, 2002). In dit verband wordt ook wel gesproken van een 'bruto' prestatieniveau en een 'netto' prestatieniveau, waarbij in het tweede rekening is gehouden met kenmerken van de leerlinginstroom (Scheerens & Bosker, 1997). Er bestaat evenwel momenteel nog geen empirisch ondersteund overzicht van de factoren die enerzijds buiten de controle van scholen vallen, maar anderzijds wel een wezenlijke invloed hebben op de leerprestaties (Wang et al., 2006). Bovendien kunnen bepaalde niet-schoolse factoren een differentiële invloed hebben op zowel verschillende leerlingen als op verschillende leerdomeinen (bijvoorbeeld op de verwerving van taalvaardigheden versus de verwerving van rekenvaardigheden). Ook bestaat er over sommige factoren discussie of ze nu wel of niet door de school kunnen beïnvloed worden (bijvoorbeeld ondersteuning door ouders) (Janssens et al., 2014).

Daarnaast rijst de vraag of 'vergelijkbaarheid' überhaupt mogelijk is, aangezien de verschillende factoren buiten de invloedssfeer van de school op een verschillende manier interageren met elkaar en met specifieke factoren die net wel onder de controle van de school liggen, waardoor elke vergelijking per definitie een sterke vereenvoudiging van een complexe realiteit is (Standaert, 2014). Een laatste punt van kritiek is dat de meeste toegevoegde waarde modellen een corrigerende factor voor elke school op dezelfde manier, en even zwaar, laten doorwegen, terwijl dit niet noodzakelijk strookt met de realiteit (Janssens et al., 2014). Zo bleek uit een eerder onderzoek in Vlaanderen dat het IQ van leerlingen een grote impact heeft op wiskundeprestaties van leerlingen, maar dat het effect niet voor alle groepen leerlingen hetzelfde is (De Maeyer & Rymenans, 2004). Elke berekening van toegevoegde waarde van scholen is daarom niet meer dan een benadering van de feitelijke toegevoegde waarde. In sommige modellen wordt ook deze onzekerheid mee gekwantificeerd (Janssens et al., 2014).

In academisch (effectiviteits)onderzoek wordt overigens vaker uitgegaan van verschillende deelgroepen in scholen: men gaat dan niet corrigeren voor de verschillende achtergrondvariabelen, maar bekijkt wat de toegevoegde waarde is van de school voor de leerlingen met een verschillend geslacht, socio-economische status en etniciteit. In dat verband wordt gesproken van 'differentiële effecten' van de school (De Maeyer & Rymenans, 2004; Scheerens & Bosker, 1997). Een verwante vraag wordt gesteld wanneer op zoek wordt gegaan naar de mate waarin de school erin slaagt om de kansengelijkheid tussen leerlingen weg te werken (Bosker, Creemers, & Stringfield, 1999; De Maeyer & Rymenans, 2004).

Daarnaast wordt het meetvraagstuk van de 'toegevoegde waarde van scholen' (*hoe* wordt de toegevoegde waarde berekend?) op verschillende manieren benaderd. De discussie over hoe deze bijdrage van de school moet bepaald worden, is lang en complex (Sloane et al., 2013). In grote lijnen zijn twee stromingen te onderscheiden. Enerzijds zijn er de berekeningen van toegevoegde waarde op basis van de leerprestaties van leerlingen, terwijl anderzijds berekeningen op basis van de leerwinst bestaan. De eerste stroming worden de 'statusmodellen' of '*snapshot* modellen' genoemd (Castellano & Ho, 2013; Harris, 2011; Ladd & Lauen, 2010). Hierbij wordt, op basis van een éénmalige toetsafname bij één of meerdere cohorten leerlingen in een school, een uitspraak gedaan over de toegevoegde waarde van de school, eventueel na correctie voor enkele van bovengenoemde relevante factoren die buiten de invloedssfeer van de school liggen. Berekeningen volgens de tweede stroming, op basis van leerwinst, noemt men 'groeimodellen' of '*value added* modellen' (Castellano & Ho, 2013; Ladd & Lauen, 2010). Wanneer ook de leerwinst gecorrigeerd wordt voor relevante factoren buiten de invloedssfeer van de school, wordt gesproken van '*contextual value added* modellen' (Janssens et al., 2014).

3.2.1 Statusmodel: toegevoegde waarde op basis van leerprestaties

Statusmodellen houden vaak een vergelijking in van de leerprestaties van leerlingen in een school met de leerprestaties van leerlingen in andere scholen. Het gaat hierbij om een 'benchmark' waarbij scholen op basis van één toetsmoment een vergelijking krijgen aangeboden met andere scholen. In de meeste statusmodellen wordt daarbij enkel vergeleken met 'gelijkaardige scholen' (wat betreft context en/of leerlinginstroom) of wordt er een correctie toegepast voor factoren die buiten de invloedssfeer van scholen liggen. Een voorbeeld hiervan zijn de feedbackrapporten die Vlaamse scholen momenteel ontvangen bij deelname aan de peilingen.

Daarnaast bestaan er ook vormen van conceptualisering van toegevoegde waarde waarbij de leerprestaties met school-interne gegevens worden vergeleken. Deze vormen van conceptualisering kunnen dan ook vooral zicht geven op de toegevoegde waarde op klasniveau of op longitudinale tendensen binnen de school. Deze conceptualisering van toegevoegde waarde kunnen bestaan uit (a) toetsen van verschillende cohorten van verschillende leeftijden op eenzelfde ogenblik (bijvoorbeeld tweede, vierde en zesde leerjaar in hetzelfde schooljaar) of (b) uit toetsen van verschillende cohorten van dezelfde leeftijd op verschillende ogenblikken (bijvoorbeeld tweede leerjaar in drie opeenvolgende schooljaren). Bij (a) kan de gemiddelde score van verschillende leerjaren informatie geven over de voortgang van leerjaar tot leerjaar; bij (b) - dat ook wel het '*cohort-to-cohort*-perspectief' wordt genoemd - kunnen de resultaten de voortgang van de school in het bereiken van betere resultaten binnen een bepaald leerjaar in kaart brengen (Castellano & Ho, 2013). In geen van beide gevallen wordt echter gewerkt met leerwinst op individueel leerlingniveau.

3.2.2 Groeimodel: toegevoegde waarde op basis van leerwinst

Bij een conceptualisering van toegevoegde waarde in een 'groeimodel' gebeurt de berekening van toegevoegde waarde op basis van meer dan één meting van een bepaald cohort binnen de school. Hierbij wordt de leerwinst van leerlingen in een school vergeleken met de leerwinst van leerlingen in andere scholen. Toegevoegde waarde-berekeningen op basis van leerwinst verdienen volgens vele auteurs de voorkeur boven oordelen over de scholen op basis van leerprestaties van leerlingen. Berekeningen van toegevoegde waarde op basis van leerwinst zijn immers meer precies en meer correct dan berekeningen op basis van leerprestaties, omdat ze minder afhankelijk zijn van factoren die niet onder de controle van de school vallen (Elstad & Turmo, 2011; Goldschmidt et al., 2012). Uit vele studies blijkt immers dat scholen met een meer bevoorrechte leerlingenstroom meestal hoger scoren bij statusmodellen, dan scholen met een leerlingeninstroom met minder bevoorrechte kenmerken (bijvoorbeeld inzake socio-economische status en thuistaal) (Koretz, 2008; Ladd & Walsh, 2002; Popham, 1999; Sloane et al., 2013). Toegevoegde waarde op basis van leerwinst is daarom zinvol omdat hierdoor expliciet wordt erkend dat leerlingen die in een school komen met een achterstand ten aanzien van het gemiddelde, de school ook kunnen verlaten met een leerachterstand, zelfs wanneer de school onderwijs van maximale kwaliteit heeft voorzien (Ladd & Lauen, 2010; Ladd & Walsh, 2002). Hierdoor verschuift de nadruk in het beschrijven van de kwaliteit van de school van een beoordeling van de 'uitkomsten' naar een beoordeling van het proces (Wang et al., 2006).

Bij deze modellen wordt meestal enkel vergeleken met vergelijkbare scholen inzake leerling-input en context (de zogenaamde *contextual value added* modellen), al bestaat er een grote differentiatie in de factoren die in rekening worden genomen om te bepalen welke scholen 'vergelijkbaar' zijn (Wang et al., 2006). Het is evenwel nog onduidelijk of in deze modellen met zulke corrigerende factoren rekening dient te worden gehouden. Aangezien deze corrigerende factoren ook al doorwegen in de eerdere toetsprestaties menen sommige auteurs dat een correctie overbodig of zelfs foutief zou zijn aangezien de resultaten dan onterecht twee maal worden gecorrigeerd. Andere auteurs stellen dan weer dat bepaalde factoren buiten de invloedssfeer van de school niet enkel het beginniveau van de leerlingen, maar ook de daaropvolgende groei beïnvloeden (Janssens et al., 2014).

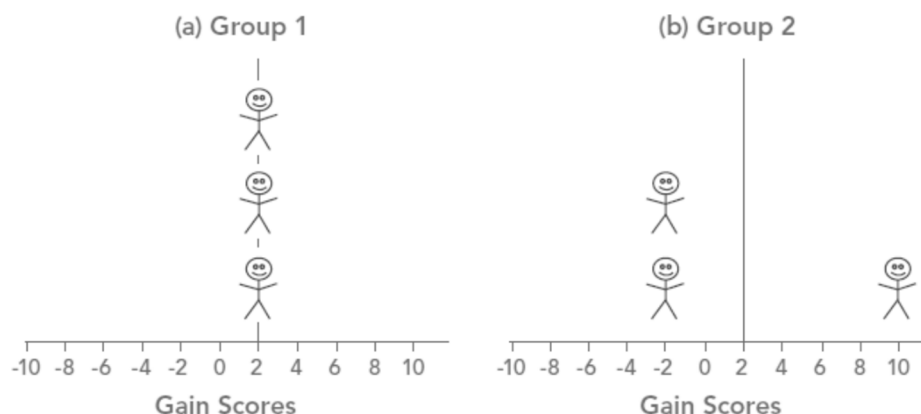
Uiteraard wordt de conceptualisering van 'toegevoegde waarde' in dit geval sterk beïnvloed door de manier waarop de leerwinst werd gedefinieerd. Daarom worden ook hier 'gain-based' modellen en 'conditional status modellen' onderscheiden (Castellano & Ho, 2013).

Gain-based models

Wanneer het '*gain score*'-model wordt gebruikt voor de berekening van leerwinst, vormt de gemiddelde leerwinst een maat voor de toegevoegde waarde van de klas of de school (Castellano & Ho, 2013; Goldschmidt et al., 2012). Een probleem met dit model is dat dit gemiddelde de spreiding van de resultaten binnen de groep kan verhullen. Zoals blijkt uit Figuur 9 kunnen twee groepen met beide een gemiddelde leerwinst van +2 toch sterk van elkaar verschillen.

Indien leerwinst wordt geconceptualiseerd volgens het *trajectory*-model of het *category* model, wordt de toegevoegde waarde aangeduid volgens een percentage. Bij het *trajectory model* gaat het om het percentage van alle leerlingen in een klas of school die '*on track*' zijn (dus die een leerprestatie op of boven hun persoonlijke regressielijn hebben getoond); bij het *categorical model* wordt

berekend welk percentage van de leerlingen in een klas of in een school 'on track' is (leerlingen die in dezelfde categorie blijven) of 'above track' is (leerlingen die een categorie hoger presteren dan voorgaande prestaties). Men kan betekenis geven aan de toegevoegde waarde van een school door dit percentage als een absoluut gegeven te beschouwen, of door het te vergelijken met andere scholen (Castellano & Ho, 2013).



Figuur 9 - Verschillende verdelingen van 'gain scores' met een zelfde gemiddelde score

Deze *gain-based* modellen worden vaak toegepast, bijvoorbeeld door Ladd en Walsh (2002) en Isenberg en Hock (2011). Een belangrijk nadeel van deze *gain-based* modellen is evenwel dat er geen rekening wordt gehouden met de geneste structuur van de data (met name dat leerlingen behoren tot klasgroepen) wat tot een vertekende rapportering kan leiden (Janssens et al., 2014). Er bestaat ook een alternatief *gain-based* model waarbij de variantie wel gesplitst wordt in variantie op leerlingniveau en variantie op schoolniveau, toegepast door bijvoorbeeld Ray (2006). Deze alternatieve modellen worden ook *random intercept* modellen genoemd en zijn vanuit methodologisch standpunt verkiesbaar boven de standaard *gain-based* modellen omdat ze (naast rekening houden met de geneste structuur) ook de corrigerende covariaten op het juiste niveau kunnen plaatsen, met name leerlingkenmerken op leerlingniveau, en schoolkenmerken op schoolniveau (Janssens et al., 2014). De wiskundige redenering onder deze *random intercept* modellen is nog relatief eenvoudig, met name:

$$Y_{ij} = \beta_0 + u_{0j} + e_{0ij}$$

waarbij Y_{ij} staat voor de ruwe leerwinst (of groei-)score, β_0 ageert als het intercept en staat voor de gemiddelde leerwinst van alle leerlingen, u_{0j} is de afwijking van school j ten opzichte van het gemiddelde, en e_{0ij} de afwijking van leerling i in school j ten opzichte van het gemiddelde. Hieruit vloeit voort dat de 'error' in deze modellen wordt beschouwd als de individuele leerwinst van leerlingen. Aan dit wiskundig model kunnen nog corrigerende covariaten worden toegevoegd (Janssens et al., 2014).

Er kan ook gedifferentieerd worden in de toegevoegde waarde van een school. Zo kan de bijdrage van de school tot de leerprestaties/leerwinst van leerlingen opgesplitst worden voor verschillende groepen (bijvoorbeeld volgens SES, volgens intelligentie of volgens eerdere leerprestaties). Een alternatieve vorm van differentiatie zijn de zogenaamde *quantile regression analysis modellen* (Haile & Nguyen, 2008), waarbij niet van het gemiddelde van de groep leerlingen van de school wordt uitgegaan. In plaats daarvan wordt een aparte vergelijking gemaakt van verschillende percentielen (bijvoorbeeld P10, P25, P50, P75, P90), waardoor een beter zicht wordt verkregen op de

toegevoegde waarde van de school voor de verschillende groepen (sterk/medium/zwak presterende) leerlingen in de school (Janssens et al., 2014).

Conditional status models

Indien leerwinst wordt geconceptualiseerd volgens het *residual gain model* wordt de toegevoegde waarde van een klas of school bepaald door de gemiddelde leerwinst van de verschillende leerlingen in de klas of school. Die toegevoegde waarde van de klas of school kan dan vergeleken worden met de toegevoegde waarde van andere klassen of scholen. Bij een andere mogelijkheid - het zogenaamde *covariate adjusted model* - wordt het 'lidmaatschap' van een leerling tot een bepaalde klas of school in het model opgenomen. Op die manier resulteert het model in een aparte regressielijn voor elke klas/school waarbij aangegeven wordt wat de leerwinst van een cohorte leerlingen in een school tussen twee meetmomenten (Goldschmidt et al., 2012). Bij een conceptualisering van leerwinst volgens het *projection model* kan toegevoegde waarde berekend worden als het gemiddelde van de verschillende leerlingen in de klas of school, of als het percentage van de leerlingen dat een toetsresultaat haalt boven de voorspelde waarde (Castellano & Ho, 2013). Wanneer leerwinst wordt berekend volgens het *student growth percentile model*, is de meest gebruikelijke wijze om ofwel de mediaan ofwel het gemiddelde van een groep leerlingen (klas of school) als toegevoegde waarde te beschouwen. Ook in elk van deze laatste modellen kan de toegevoegde waarde van de school daarna vergeleken worden met de toegevoegde waarde van andere scholen (Castellano & Ho, 2013; Goldschmidt et al., 2012).

3.2.3 Multivariate models

Terwijl de groeimodellen gericht zijn op leerwinst, en dus uitgaan van de vragen naar een beschrijving of voorspelling van de groei, gaan 'multivariate models' in de eerste plaats op zoek naar een beschrijving van welk deel van de groei aan de school (of aan leraren) kan worden toegeschreven. Of in andere woorden: waar maakt de school werkelijk het verschil?

Multivariate modellen zijn statistisch meer complexe modellen die rekening kunnen houden met een veelheid van factoren, zoals eerdere resultaten van leerlingen, met hun klas en school (en eventueel studierichting), maar ook bijvoorbeeld met de leraren waarvan de leerlingen tevoren les kregen. Op basis van deze vele data wordt een verwachte score voor elke leerling berekend voor een bepaalde toets, maar wordt vooral een directe inschatting gemaakt van de bijdrage van elke leraar en elke school tot het leren van de leerlingen (Janssens et al., 2014). In tegenstelling tot eerdere modellen houden deze modellen rekening met de geneste structuur van de data. De multivariate modellen kunnen bovendien voor veel meer relevante factoren controleren dan de eerder besproken minder complexe modellen (Castellano & Ho, 2013; Janssens et al., 2014). Indien bijvoorbeeld een toets wordt afgenomen in vier opeenvolgende leerjaren, wordt gesteld dat de score van de leerling in een klas in het eerste leerjaar (Y_{ij1}) wordt gezien als:

$$Y_{ij1} = \beta_1 + u_{j1} + e_{ij1}$$

De score wordt immers samengesteld door de gemiddelde score van het eerste leerjaar (β_1), plus een schatting van de toegevoegde waarde van de leraar van het eerste leerjaar (u_{j1}), plus een individuele error op leerlingniveau in het eerste leerjaar (e_{ij1}). In het vierde leerjaar worden ook de schattingen van de vorige leraren van de leerlingen in rekening genomen. Daarom wordt het de score van een leerling in een klas in het vierde leerjaar als volgt gezien:

$$Y_{ij4} = \beta_4 + u_{j1} + u_{j2} + u_{j3} + u_{j4} + e_{ij4}$$

waarbij de score van de leerling wordt samengesteld door de gemiddelde score van het vierde leerjaar (β_4), plus een schatting van de toegevoegde waarde van elk van de leraren van het eerste tot en met vierde leerjaar ($u_{j1} + u_{j2} + u_{j3} + u_{j4}$), plus een error op leerlingniveau in het vierde leerjaar (e_{ij4}) (Castellano & Ho, 2013). Deze schattingen worden meestal als causale verbanden begrepen, al is dat zeker niet altijd terecht (Goldschmidt et al., 2012).

De toegevoegde waarde van de school wordt 'genormeerd' door middel van het bepalen van hoeveel standaarddeviaties een school verschilt van een gemiddelde. Vanuit een verantwoordingsperspectief kunnen daarbij normen worden opgesteld, waarbij bijvoorbeeld wordt bepaald dat de scholen onder een bepaald percentiel, of onder een bepaald aantal standaarddeviaties onder de baseline, verdere verantwoording dienen af te leggen (Castellano & Ho, 2013).

Voorbeelden van toepassingen van toegevoegde waarde metingen volgens multivariate modellen zijn Choi en Seltzer (2005), Fan (2001) en Guldemonnd en Bosker (2009). Nog andere vormen van multivariate modellen zijn de 'cross-classificatie random-effect groeicurve modellen', die ook rekening houden met zittenblijvers (die daarom een toets meermaals afnemen) en in- en uitstromende leerlingen in een school (Janssens et al., 2014); en de 'gelaagde modellen'. Die laatste zijn multivariate modellen die data verwerken van verschillende toetsmomenten in verschillende leerdomeinen, en waarbij rekening gehouden wordt met effecten van vroegere leraren en voorgaande scholen (Goldschmidt et al., 2012).

Doordat meer factoren in de berekening worden opgenomen, discrimineren multivariate modellen vaak sterker tussen scholen: de verschillen tussen scholen zijn groter dan kan worden aangetoond met status- of groeimodellen (Guldemonnd & Bosker, 2009). Bovendien leveren de multivariate modellen door de veelheid aan factoren meer robuuste en consistente resultaten op (Goldschmidt et al., 2012). Een nadeel van elk van deze multivariate modellen is evenwel dat het moeilijker is om betekenis te geven aan de resultaten ervan (Castellano & Ho, 2013). Het wordt daarom gesteld dat de keuze voor een model van toegevoegde waarde een *trade off* is tussen technische accuraatheid en complexiteit enerzijds, en het kunnen afleiden van begrijpelijke en praktisch hanteerbare interferenties anderzijds (Goldschmidt et al., 2012).

4 CONCLUSIES UIT DE LITERATUURSTUDIE

Het is niet evident om eenduidige conclusies te trekken uit de literatuurstudie. Zo zijn opinie en empirie vaak sterk verweven en moeilijk van elkaar te onderscheiden. Bovendien is de empirische kennisbasis sterk gericht op één bepaalde invulling van leer(winst)monitoringssystemen, met name de implementatie van high-stakes eindexamens die beoordelen of een leerling een bepaald diploma krijgt, dan wel of een school aan bepaalde kwaliteitsvereisten voldoet. Bovendien is de kennisbasis bedreigd door de moeilijkheden die gepaard gaan met het meten van de effecten van leer(winst)monitoringssystemen.

De resultaten van de uitgebreide literatuurstudie zijn ingedeeld in drie thema's: de impact van gestandaardiseerde toetsen als basis van leer(winst)monitoringssystemen, de bouwstenen van

effectieve leer(winst)monitoringssystemen, en de conceptualisering van de begrippen 'leerwinst' en 'toegevoegde waarde'.

In het luik omtrent de impact van gestandaardiseerde toetsen worden (veronderstelde) effecten, alsook (veronderstelde) onwenselijke neveneffecten op micro-, meso- en macroniveau onderscheiden. Op microniveau werd op basis van empirische evidentie aangetoond dat leraren via gestandaardiseerde toetsen een beter beeld kunnen krijgen van de leerprestaties van leerlingen, en dat gestandaardiseerde toetsen op die manier een grotere garantie bieden dat alle leerlingen overheen de verschillende scholen op een gelijkmatige manier en volgens een gelijke maatstaf worden beoordeeld. Toch blijkt hier (net als op verschillende andere plaatsen) dat de empirische kennisbasis weinig eenduidig is. Zo wordt er door verschillende studies inderdaad een verschil vastgesteld tussen de prestaties van leerlingen op toetsen die door leraren zelf worden opgesteld, en de prestaties van diezelfde leerlingen op gestandaardiseerde toetsen. Er is evenwel geen eenduidige betekenis te geven aan dit verschil: zowel voor- als tegenstanders van gestandaardiseerde toetsen gebruiken het om de meerwaarde van de eigen meting te propageren. Enerzijds gaan stemmen op dat gestandaardiseerde toetsen leiden tot een 'meer gelijke lat voor alle leerlingen', anderzijds staat hier tegenover dat de gestandaardiseerde toets per definitie een beperking inhoudt ten opzichte van de vele elementen waarmee een leraar in zijn beoordeling rekening kan houden. Binnen dit laatste standpunt past de evidentie waarin wordt aangetoond dat de implementatie van gestandaardiseerde toetsen (door het gebruikte format) net ongelijkheid tussen leerlingen in de hand werkt.

Een ander voorbeeld van incoherentie in de empirische kennisbasis is gerelateerd aan het veronderstelde effect dat de implementatie van gestandaardiseerde toetsen (in het onderzoek veelal geoperationaliseerd als gestandaardiseerde eindexamens) door uiteenlopende mechanismen zou leiden tot betere leerprestaties. Hoewel er inderdaad verschillende studies deze thesis bevestigen, blijkt uit andere studies dat dit mogelijk kan verklaard worden door onwenselijke neveneffecten als 'teaching to the test'. Daarbij gaat de gerichtheid van het onderwijs op wat men verwacht dat getoetst zal worden ten koste gaat van andere leerstofonderdelen of niet-getoetste kennis, vaardigheden of attitudes. Andere studies geven aan dat de verhoogde leeruitkomsten gepaard gaan met een negatieve impact op de autonome motivatie van leerlingen.

Andere positieve effecten op microniveau waarvoor empirische evidentie werd verzameld, zijn het positieve effect op het lesgeven van leraren (al werden ook verschillende vormen van 'teaching to the test' empirisch gedocumenteerd), een positief effect op de doelmatigheidsbeleving van leerlingen (alhoewel we hierboven reeds tegenstrijdige evidentie aanhaalden) en een betere evaluatie door leraren zelf. Andere veronderstelde positieve resultaten (waarvoor geen empirische evidentie kon worden aangehaald) zijn de betere onderbouwing van de schoolkeuze en van de studiekeuze van leerlingen. Het meest in het oog springende negatieve effect op microniveau is de zware emotionele impact die gestandaardiseerde toetsen kunnen hebben op leerlingen. Ook hierbij dient te worden gesteld dat de empirische evidentie voornamelijk gericht is op (high-stakes) eindexamens. Tevens werd empirische resultaten besproken met betrekking tot de negatieve impact van gestandaardiseerde toetsen op de studieloopbaan van leerlingen, en onwenselijk strategisch gedrag van leerlingen. Ten slotte werd de (veronderstelde) negatieve impact op de professionaliteit en de motivatie van leraren besproken.

Hoewel ook veel is geschreven over de (neven)effecten van leer(winst)monitoringssystemen op het mesoniveau en macroniveau, blijkt dat de empirische kennisbasis hieromtrent erg beperkt is. De resultatensectie in verband met deze niveaus bevat dan ook voornamelijk beleidsmatige literatuur of literatuur van opiniërende aard. Op mesoniveau wordt in de literatuur veel aandacht besteed aan het veronderstelde positieve effect op de interne kwaliteitszorg van scholen, en de opportuniteit om zowel professionele ontwikkeling van schoolteams als schoolinspecties meer effectief en efficiënt in te zetten. Daar staat echter tegenover dat er veel evidentie is omtrent twee onwenselijke neveneffecten van gestandaardiseerde toetsen, met name onwenselijk strategisch gedrag van scholen, en het ontstaan van rankings (al beschreven we ook de visie van auteurs die het ontstaan van rankings als een opportuniteit, eerder dan een bedreiging beschouwen). Beide onwenselijk geachte neveneffecten zijn vooral gedocumenteerd in een high-stakes toetsomgeving met openbaarheid van resultaten.

Op macroniveau worden vier positieve effecten onderscheiden, met name: het onderwijsbeleid (en de verschillende stakeholders) die geholpen worden door een beter beeld van de kwaliteit van het onderwijssysteem, de geïnformeerde basis om het curriculum te evalueren, de beschikbaarheid van data voor onderzoeksdoeleinden, en de publieke tevredenheid. Daar staat tegenover dat verschillende auteurs aangeven dat door de implementatie van gestandaardiseerde toetsen een onderwijssysteem kan vervallen in een overmatige toetscultuur en dat een grotere segregatie tussen sterke en zwakke scholen zou kunnen ontstaan. Ten slotte wordt gewezen op de financiële impact van de ontwikkeling en implementatie van een leer(winst)monitoringssysteem.

Een rode draad doorheen heel deze beschrijving van effecten en neveneffecten is dat er sterke verschillen zijn tussen verschillende scholen inzake de effectiviteit van een leer(winst)-monitoringssysteem, alsook inzake de globale effectiviteit van verschillende operationaliseringsmodi van een leer(winst)monitoringssysteem. Het tweede luik van de literatuurstudie onderscheidt daarom verschillende bouwstenen die in de literatuur worden aangegeven als (mogelijke) relevante factor die kan verklaren in welke mate bepaalde (neven)effecten zich voordoen. Deze bouwstenen omvatten de mogelijkheid tot de berekening van leerwinst en van toegevoegde waarde, het opzet van de gestandaardiseerde toetsen in een leer(winst)monitoringssysteem (waarbij de literatuur voornamelijk maar niet uitsluitend aandacht heeft voor het belang van psychometrische kenmerken van de toetsen, de inzet van de toetsen, en het openbare karakter van de resultaten van de toetsen), kenmerken van leraren, kenmerken van scholen, en kenmerken van het onderwijssysteem.

In het derde luik van de literatuurstudie worden de concepten 'leerwinst' en 'toegevoegde waarde' verder ontrafeld, en verschillende mogelijkheden tot operationalisering worden besproken. Hoewel de literatuur verschillende leerwinst-modellen onderscheidt waarbij leerwinst op basis van slechts één meting wordt geoperationaliseerd, blijkt in de praktijk dat in zowat alle modellen twee of meer metingen worden gehanteerd. Bij *gain-based models* wordt uitgegaan van de groei van leerlingen op basis van twee of meerdere toetsmomenten. Specifiek aan deze modellen is dat er wordt uitgegaan van een gemeenschappelijke verticale schaal via *Item Response Theory (IRT)*. Daardoor kunnen de resultaten van verschillende toetsen met elkaar worden vergeleken. In de context van gestandaardiseerde toetsen is de meest gebruikte vorm van IRT de inschaling van items volgens een Rasch-analyse. Rasch is een techniek die toelaat om de resultaten van twee toetsen met elkaar te vergelijken, omdat ze op een identieke schaal worden ingeschaald, zelfs wanneer ze

verschillende items bevatten. Vaak wordt IRT ingezet om de resultaten van toetsen afgestemd op de leerstof van verschillende leerjaren toch met elkaar te kunnen vergelijken. De ontwikkeling van zulke verticale schaal is erg arbeidsintensief, en er werd ook aangetoond dat dit niet in alle omstandigheden mogelijk is. Binnen deze gain-based models wordt nog verder onderscheid gemaakt tussen *gain score modellen*, *trajectory modellen*, en *categorical modellen*. Naast de gain-based modellen worden in het rapport *conditional status models* besproken. Deze modellen beoordelen de prestatie van een leerling aan de hand van een vergelijking met een vooraf bepaalde verwachting. Deze verwachting kan op voorhand voor elke leerling op empirische wijze gesteld op basis van de eerdere prestaties van deze en andere leerlingen (en indien dat wenselijk wordt geacht ook op basis van andere factoren, bijvoorbeeld relevante factoren die buiten de invloedssfeer van de school liggen). Deze modellen vereisen geen gemeenschappelijke verticale schaal. Binnen de conditional status models wordt nog verder onderscheid gemaakt tussen *residual gain modellen*, *projection modellen*, en *student growth percentile modellen*.

Ondanks de vele aandacht voor de verschillende conceptualisering, geeft de literatuur geen aanwijzing omtrent welke conceptualisering in een bepaalde situatie de voorkeur geniet. Door deze lacune leidt de literatuurstudie niet tot een eenduidige conclusie omtrent het meest wenselijke model van leerwinstberekening voor de Vlaamse onderwijscontext.

De term 'toegevoegde waarde' wordt in dit boek gebruikt op mesoniveau: de bijdrage van de school aan het leren (of de leerwinst) van leerlingen. De literatuurstudie identificeert verschillende vormen van 'statusmodellen' van toegevoegde waarde, waarbij de leerprestaties van leerlingen (op een enkele meting) met elkaar worden vergeleken. Wanneer toegevoegde waarde wordt gebaseerd op de leerwinst van de leerlingen, is de berekening van de toegevoegde waarde afhankelijk van de berekeningswijze die werd gebruikt voor het bepalen van de leerwinst. Daarnaast zijn er verschillende multivariate modellen bekend waarbij meer precies de bijdrage van een leraar of een school aan het leerresultaat van een leerling kan worden bepaald. Doordat meer factoren in de berekening worden opgenomen, discrimineren multivariate modellen vaak sterker tussen scholen: de verschillen tussen scholen zijn groter dan kan worden aangetoond met status- of groeimodellen. Bovendien leveren de multivariate modellen door de veelheid aan factoren meer robuuste en consistente resultaten op. Een nadeel van elk van deze multivariate modellen is evenwel dat het moeilijker is om betekenis te geven aan de resultaten van deze modellen. Het wordt daarom gesteld dat de keuze voor een model van toegevoegde waarde een *trade off* is tussen technische accuraatheid en complexiteit enerzijds, en het kunnen afleiden van begrijpelijke en praktisch hanteerbare interferenties anderzijds. Buiten bovenstaande evenwichtsoefening kan evenwel ook inzake toegevoegde waarde op basis van de literatuurstudie geen eenduidige conclusie voor de Vlaamse onderwijscontext worden getrokken, laat staan één bepaald model naar voren worden geschoven.

DEEL 4 RESULTATEN DELPHI-STUDIE (EERSTE EN TWEEDE ONDERZOEKSRONDE)

In dit deel wordt gerapporteerd over de resultaten van de eerste en tweede onderzoeksrunde van de Delphi-studie. Aangezien de derde onderzoeksrunde een inventarisering betreft van de inschatting van het expertenpanel inzake de haalbaarheid en wenselijkheid van de werkscenario's die werden ontwikkeld, worden deze data verwerkt in Deel 7 van dit rapport. De eerste ronde van de Delphi-studie bestond uit een schriftelijke bevraging van de 24 respondenten. 22 respondenten beantwoordden deze vragenlijst. De antwoorden van de 22 respondenten werden thematisch verwerkt. De data-analyse van de eerste ronde resulteerde per thema in een tabel waarin alle gegeven antwoorden kunnen worden teruggevonden. Deze thematische tabellen werden aan alle 24 experts verzonden ter voorbereiding van de tweede onderzoeksrunde, die bestond uit een diepte-interview met elke expert.

In de tweede ronde van de Delphi-studie werden de diverse standpunten die de respondenten innemen door middel van diepte-interviews verder verkend, en werd de kans geboden om op elkaars standpunten (op basis van de tabellen uit onderzoeksrunde 1) te reageren. Van elk van de 24 respondenten werd een interview afgenomen. Aan het eind van elk diepte-interview werd aan de hand van de ingenomen standpunten door elk van de respondenten een model-scenario voorgesteld.

Deze rapportering brengt de onderzoeksresultaten van de eerste en tweede ronde van de Delphi-studie bij elkaar. Met behulp van citaten wordt een kader geschetst dat de verschillende belangen, percepties en wensen van de respondenten weergeeft. Waar dat opportuun is, wordt onderscheid gemaakt tussen vier verschillende categorieën van respondenten. Het betreft: (1) beleidsmakers (n=4; inclusief 1 inspecteur), (2) onderwijsverstrekkers (n=6), scholen (n=8; gelijk verdeeld over basisonderwijs en secundair onderwijs) en inhoudsdeskundigen (n=6).

In de onderstaande secties worden de algemene bevindingen van de respondenten beschreven. In de eerste sectie beschrijven we de algemene conclusies van de 24 respondenten inzake de wenselijkheid van een leer(winst)monitoringssysteem. Daarna worden de verwachte effecten alsook de verwachte onwenselijke neveneffecten van een leer(winst)monitoringssysteem geschetst. Vervolgens wordt stil gestaan bij de conceptualisering van de begrippen 'leerwinst' en 'toegevoegde waarde' en worden de opinies inzake de voor- en nadelen van het registreren van zowel leerwinst als van toegevoegde waarde besproken. Tot slot worden de voorwaarden weergegeven waaraan de gestandaardiseerde toetsen die deel uitmaken van het leer(winst)monitoringssysteem volgens de respondenten dienen te voldoen.

1 INSCHATTING VAN BEHOEFTE AAN EEN LEER(WINST)MONITORINGS-SYSTEEM

Er bestaat een relatief grote consensus tussen de 24 respondenten over de te verwachte positieve effecten en onwenselijke neveneffecten. Soms zijn er respondenten die de opinie van anderen in

twijfel trekken, maar dit gebeurt slechts eerder uitzonderlijk. Toch blijken de verschillende respondenten tot uiteenlopende ideaal-scenario's te komen aan het einde van de tweede onderzoeksrunde. Dat heeft dan vooral te maken met het gewicht dat aan de verschillende effecten en neveneffecten wordt toegekend: sommige respondenten menen dat de som van de positieve effecten zwaarder weegt dan de som van de onwenselijke neveneffecten, terwijl andere respondenten menen dat de balans in de andere richting overhelt. Daarnaast zijn er verschillen inzake het gewicht toegekend aan effecten en neveneffecten op micro-, meso-, en macroniveau, waardoor sommige respondenten menen dat een leer(winst)monitoringssysteem zich in de Vlaamse context eerder op het micro-, of eerder op het meso- of macroniveau dient te richten.

21 van de 24 respondenten in de Delphi-studie vinden na twee onderzoeksrondes dat het huidige Vlaamse onderwijssysteem baat zou hebben bij de invoering van een leer(winst)monitoringssysteem. In het perspectief van bijna elke respondent werd meer dan één doel gekoppeld aan het leer(winst)monitoringssysteem. In 20 van deze 21 scenario's die door respondenten werden uitgetekend in het kader van leer(winst)monitoring staat de interne kwaliteitszorg van de scholen - een ontwikkelingsgericht doel op mesoniveau - als één van de prominente doelen centraal. Daarnaast wordt in 18 van deze 20 scenario's (alsook in het enige andere scenario) gesteld dat het leer(winst)monitoringssysteem bijkomende doelen heeft die zijn gericht op het microniveau van individuele leerlingen en leraren, tevens in een ontwikkelingsgericht perspectief.

"Ik zie het dan vooral in kader van de interne kwaliteitszorg van scholen en de begeleiding van leerlingen. Die twee aspecten. Dus die interne kwaliteitszorg van scholen, ik denk dat het goed is dat een school op verschillende manieren zicht krijgt op haar eigen kwaliteit. En gestandaardiseerde toetsen kunnen daar een manier voor zijn op voorwaarde dat het benchmarks zijn, dat ze zich kunnen spiegelen aan bepaalde benchmarks die dan ook niet te grof zijn hé. Ik bedoel, je gaat geen BSO-school gaan vergelijken met een ASO-school en de ene nog in het platteland en de andere nog in de stad. Daarnaast ook naar de leerlingen toe kan het een ondersteunend element zijn in het ja, het begeleidingstraject dat leerlingen krijgen om bepaalde einddoelen te behalen. In die zin staan wij open voor gestandaardiseerde toetsen" (onderwijsverstrekkers, respondent 1)

Naast bovenstaande doelstellingen, hebben acht respondenten ook doelstellingen op macroniveau toegevoegd (bijvoorbeeld het nagaan van de validiteit van de eindtermen of ontwikkelingsdoelen). In geen enkel scenario staat het macroniveau evenwel centraal.

Door elf respondenten worden ook één of meerdere doelen in een verantwoordingsgericht perspectief toegevoegd. Meestal is dat op microniveau (een impact op beslissingen over het al dan niet slagen van leerlingen), soms ook op mesoniveau (de externe beoordeling van scholen). Op één uitzondering na, staan deze verantwoordingsgerichte doelen echter niet centraal in de door respondenten ontwikkelde scenario's.

Er zijn duidelijke verschillen tussen de categorieën van respondenten. Bij de 'beleidsmakers' kennen twee van de vier respondenten enkel ontwikkelingsgerichte doelstellingen toe aan het leer(winst)monitoringssysteem; bij een derde en vierde respondent staat ook verantwoording op meso- en macroniveau centraal. Slechts twee van de respondenten bij de 'onderwijsverstrekkers'

zien heil in een leer(winst)monitoringssysteem met een verantwoordingsgericht perspectief (respectievelijk op mesoniveau van scholen en op microniveau van leerlingen), maar dan enkel als secundair doel. In de eerste plaats gaat het scenario van elk van de onderwijsverstrekkers uit van een ontwikkelingsgericht perspectief, en dan voornamelijk op mesoniveau (interne kwaliteitszorg van de scholen).

Ook elk van de acht respondenten in de categorie 'scholen' stelt dat een bijdrage tot de interne kwaliteitszorg van scholen (ontwikkelingsgericht, mesoniveau) een doel moet zijn van een leer(winst)monitoringssysteem. Vijf van deze acht respondenten benoemen aanvullend een doel op microniveau in ontwikkelingsgericht perspectief (bijvoorbeeld een betere ondersteuning van het leren van de leerlingen). Vier van deze acht respondenten in de categorie 'scholen' menen dat een leer(winst)monitoringssysteem ook gericht moet zijn op de externe evaluatie van scholen (verantwoordingsgericht op mesoniveau), alsook vier respondenten die menen dat de resultaten van een leer(winst)monitoringssysteem een element kunnen zijn in de evaluatie van leraren (waarbij het opvallend is dat geen enkele respondent uit een andere categorie deze opinie deelt). Bij vijf van deze respondenten kan een leer(winst)monitoringssysteem ook op een verantwoordingsgerichte manier op het niveau van de individuele leerlingen worden ingezet (bij beslissingen over slagen of niet slagen) en bij één respondent op het macroniveau.

Bij de inhoudsdeskundigen, ten slotte, zijn vier respondenten ervan overtuigd dat een leer(winst)monitoringssysteem in de eerste plaats ontwikkelingsgericht op het schoolniveau moet opgebouwd worden. Eén respondent meent dan weer dat een leer(winst)monitoringssysteem enkel tot ontwikkelingsgerichte doeleinden op microniveau kan leiden. Eén respondent voegt er aan toe dat de resultaten van een leer(winst)monitoringssysteem tevens verantwoordingsgericht op microniveau kunnen worden ingezet. In het scenario van één respondent uit de categorie 'inhoudsdeskundigen' wordt in de eerste plaats de focus gelegd op het verantwoordingsgericht perspectief, met name op het niveau van de school en daarnaast op het niveau van de leerlingen.

Drie respondenten zijn van mening dat aan het huidige systeem van leer(winst)monitoring in de Vlaamse onderwijscontext - bestaande uit de peilingen en paralleltoetsen, koepelgebonden toetsen (interdiocesane toetsen en OVSG-toetsen), leerlingvolgsystemen van private aanbieders en de andere gestandaardiseerde toetsen beschreven in Deel 1 - geen extra gestandaardiseerde toetsen moeten worden toegevoegd. Zij ervaren geen behoefte in het onderwijsveld om een bijkomend leer(winst)monitoringssysteem te ontwikkelen en te gebruiken aangezien zij menen dat de onwenselijke neveneffecten zwaarder zullen doorwegen dan de positieve impact op het onderwijs. Wel geven zij aan dat de finaliteit van elk van de huidige toetsen duidelijker moet worden gemaakt, en dat er sterk moet worden ingezet op de ondersteuning van scholen en leraren om de effectiviteit van het huidige systeem te versterken. Opmerkelijk is dat elk van deze drie respondenten uit de categorie 'onderwijsverstrekkers' komt.

Daarnaast is het evenwel belangrijk te vermelden dat de scenario's van verschillende respondenten erg voorwaardelijk zijn. Zo meent men dat aan een bepaalde nood in het onderwijssysteem tegemoet moet worden gekomen opdat een leer(winst)monitoringssysteem effectief de gestelde doelen zou kunnen bereiken. Deze noden zijn binnen de beschrijving van de verwachte effecten van een leer(winst)monitoringssysteem neergeschreven onder de 'beïnvloedende factoren' (zie sectie 2). Bovendien benadrukken een aantal respondenten die wel voorstander zijn van de ontwikkeling van een leer(winst)monitoringssysteem ook wel de hoge kostprijs die het invoeren van een

leer(winst)monitoringssysteem met zich meebrengt. Van daaruit volgt de vraag of het momenteel wel een geschikt moment is om tot de implementatie ervan over te gaan. Zo zijn er respondenten die zich afvragen of de kosten opwegen tegen de baten en die andere prioriteiten in het huidige onderwijssysteem vooropstellen waarin geïnvesteerd moet worden.

“En dat komt wel, ja, en als we dan de kost die daartegenover staat gaan afwegen tegenover de noden die zich stellen, dan weet ik echt niet of dat vandaag is waar we moeten op inzetten.” (onderwijsverstrekkers, respondent 2)

2 IMPACT VAN EEN LEER(WINST)MONITORINGSSYSTEEM

In deze sectie worden zowel de verwachte effecten als de verwachte onwenselijke neveneffecten van het monitoren van de leerprestaties van leerlingen besproken. Er wordt daarbij onderscheid gemaakt tussen positieve effecten en neveneffecten op micro-, meso- en macroniveau.

2.1 Microniveau

2.1.1 Verwachte effecten op microniveau

Ontwikkelingsgerichte effecten

Er bestaat een consensus in het expertenpanel dat de implementatie van een leer(winst)-monitoringssysteem kan leiden tot positieve effecten op microniveau. In elke categorie van respondenten wordt door meerdere respondenten aangegeven dat er een ontwikkelingsgericht effect kan verwacht worden op zowel de individuele leerlingen als op de leraren. Op het niveau van de leerlingen kan een leer(winst)monitoringssysteem bijdragen tot een betere oriëntering van leerlingen doorheen hun schoolloopbaan: leerlingen kunnen door middel van de gestandaardiseerde toetsen een beter beeld krijgen van hun eigen capaciteiten bij de overstap naar een volgende onderwijsvorm of bij heroriëntering (vooral in secundair onderwijs). Ook kunnen leerlingen een beter inzicht krijgen in hun eigen ontwikkeling ten opzichte van de prestaties van andere leerlingen en ten opzichte van de eindtermen, ontwikkelingsdoelen of leerplandoelen, wat motiverend kan werken. De leer(winst)monitoring kan leerlingen eveneens aanzetten tot het bereiken van betere leerprestaties en leerlingen kunnen een beter inzicht krijgen in de doelen die van hen worden verwacht.

“Dat lijkt me ook heel interessant om dat als leerling te weten: hoe doe ik het nu eigenlijk in vergelijking met andere leerlingen? Dat kan ook wel motiverend zijn” (onderwijsverstrekkers, respondent 4)

Ook leraren worden geacht gemotiveerd te worden door de resultaten van een regelmatige leer(winst)-monitoring:

“Je hebt dus hele goede leraren, die gaan zich gewaardeerd voelen en die gaan zien dat ze jaar na jaar goed scoren en die gaan zien van ik ben een goede leraar dus die motivatie kan verhogen. Leraren die minder goed zijn, die gaan beter gaan zien waar het fout loopt en die gaan betere begeleiding krijgen dus die gaan ook opwaartse druk. Dus je krijgt verbetering”. (inhoudsdeskundigen, respondent 2)

Daarnaast wordt ook in elke categorie van respondenten bevestigd dat een leer(winst)-monitoringssysteem een ontwikkelingsgerichte bijdrage kan hebben op het niveau van het lesgeven in de klas. Leraren krijgen inzicht in de eigen onderwijskwaliteit aan de hand van de resultaten van de leerlingen op de gestandaardiseerde toetsen en kunnen op basis daarvan hun didactisch handelen bijsturen en optimaliseren; of de resultaten kunnen net de bevestiging geven dat men als leraar de gewenste impact heeft op de leerlingen.

“Als je gestandaardiseerde toetsen hebt en je gaat dat bekijken wat het resultaat is voor u en je hebt benchmarks, dan kun je gaan zeggen van: hoe komt dat dat ik jaar na jaar onder die benchmark zit?” (onderwijsverstrekkers, respondent 1)

Leraren kunnen ook een beter inzicht krijgen in de ontwikkeling van elke leerling (zowel de reeds aanwezige leerlingen als neveninstromers) en krijgen een betrouwbare basis om beter te differentiëren en leerlingen individueel te begeleiden. Met behulp van de resultaten kan een leraar ook de leerlingen aanzetten om zichzelf te evalueren.

“Empowerment zowel als school, in teamverband, maar ook voor de leraar is het een vorm, die zelfevaluatie kan losmaken, die zorgt voor een betere werking van de eigen klassenpraktijk. Een goede leraar gaat zijn leerlingen ook op dezelfde manier benaderen om hen aan zelfevaluatie te laten doen.” (scholen, respondent 3)

Net als bij de leerlingen geldt dat leraren door het leer(winst)monitoringssysteem een beter beeld kunnen krijgen van de doelen die ze met de leerlingen moeten bereiken; en dat ze (hierdoor) het onderwijsaanbod sterker kunnen richten op deze verwachte doelen. Het gebruik van een leer(winst)monitoringssysteem kan de leraren meer zekerheid verschaffen omtrent de verwachtingen die de overheid koestert ten aanzien van evaluatie-instrumenten.

“Waarom gebruiken leraren liever methodes die door de uitgeverijen op de markt worden aangeboden dan dat ze zelf vanuit hun leerplannen lessen gaan ontwikkelen? (...) voor een stuk heeft dat ook te maken met veiligheid en ik denk dat in diezelfde lijn een dergelijke manier van gestandaardiseerde toetsen voor hen ook veiligheid kan bieden.” (onderwijsverstrekkers, respondent 5)

Een leer(winst)monitoringssysteem kan daarnaast zorgen voor betere en meer kwaliteitsvolle evaluatie van de leerlingen. Bovendien zijn twee respondenten van mening dat een leer(winst)-monitoringssysteem kan leiden tot een verbeterde relatie tussen leerling en leraar (als een soort ‘partners’ die samen een gemeenschappelijke externe uitdaging aangaan) en voor een grotere beschikbaarheid van tijd om lessen voor te bereiden (aangezien de evaluatie extern gebeurt).

“Leraren spenderen één derde van de tijd aan evaluaties dus het opstellen van evaluaties, corrigeren en bespreken. Als je dat gewoon al wegneemt, ga je veel efficiënter naar je leerproces gaan.” (inhoudsdeskundigen, respondent 3)

De veronderstelling dat dit laatste effect werkelijk opgaat, wordt evenwel slechts door een minderheid van de respondenten ondersteund: meer respondenten menen dat de implementatie van een leer(winst)monitoringssysteem net zal leiden tot minder beschikbare onderwijstijd (zie 1.2.1).

Verantwoordingsgerichte effecten

Daarnaast kan een leer(winst)monitoringssysteem ook op een verantwoordingsgerichte wijze op leerlingniveau worden gebruikt. Leerlingen moeten dan via gestandaardiseerde toetsen aantonen over een bepaald niveau van kennis of vaardigheid te beschikken als voorwaarde om bijvoorbeeld een diploma te behalen. Op deze manier draagt een leer(winst)monitoringssysteem bij tot een gelijke beoordeling voor alle leerlingen in elke school. Dat idee wordt ook expliciet gesteund door zeven van acht respondenten in de categorie 'scholen', door één respondent uit de categorie 'onderwijsverstrekkers' en door twee van zes respondenten in de categorie 'inhoudsdeskundigen'.

"Ik denk dat het toch moet mogelijk zijn om voor de basis, hé het allermimum wat iedereen zou moeten kennen van, die bijvoorbeeld twee uur wetenschappen per week heeft in een vierde jaar, moet het toch mogelijk zijn om echt een minimum te zoeken, waar dan iedereen toch moet voor geslaagd zijn om te kunnen naar een derde graad gaan. Met hetzelfde aantal of meer uren wetenschappen. Er is op dit moment ook veel te veel verschil tussen verschillende scholen. Iemand kan gerust in één school 80% halen en dezelfde leerling in een andere school 60%. Het ligt echt ver uit elkaar." (scholen, respondent 7)

Toch moeten deze resultaten volgens de respondenten altijd in samenhang bekeken worden met andere bronnen van informatie over de kennis, vaardigheden en competenties van de leerlingen.

"Het moet zeker naast die zijn hele schoolcarrière staan, naast de klasprestaties ook nog omdat ja zo'n toets, toetsen zijn altijd momentopnames. Dus stel dat ik zie dat een heel schooljaar dat kind het wel goed gedaan heeft en op die toetsing niet, dan... Dus ja, een grote waarde mag er aan gehecht worden maar geen unieke bepalende waarde" (scholen, respondent 7)

Een respondent uit de categorie 'scholen' meent dat het belangrijkste doel van een leer(winst)monitoringssysteem moet zijn dat het leerlingen met enige verplichting naar een bepaalde studierichting kan leiden - een gegeven dat momenteel ontbreekt in het Vlaamse onderwijs.

"Maar er zijn nu eenmaal, ja, leerlingen die echt nog niet eum... juist georiënteerd zitten en dan moet je dat als leraar ook niet trachten toch te overbruggen. Het is dan beter dat je eerlijk gewoon zegt van: "Het is niet uw richting. Je hebt andere talenten. Kies een weg die bij u past." (scholen, respondent 4)

Beleidsmakers noch onderwijsverstrekkers ondersteunen het idee dat een leer(winst)-monitoringssysteem moet bijdragen tot de summatieve beoordeling van leerlingen. In elke categorie van respondenten wordt wel ondersteund dat de resultaten van een leer(winst)monitoringssysteem ertoe kunnen bijdragen dat alle leerlingen (indirect) op een meer gelijkaardige maatstaf zullen worden beoordeeld bij de toekenning van diploma's, getuigschriften of certificaten.

Een andere vorm van verantwoordingsgericht gebruik op microniveau is eerder een symbolisch gebruik van de resultaten van het leer(winst)monitoringssysteem. Een respondent geeft aan dat het voor leraren makkelijker is om in de contacten met ouders te kunnen verwijzen naar 'betrouwbare' gegevens, zelfs wanneer dezelfde conclusie reeds op basis van de eigen toetsen of eigen observaties kon getrokken worden.

“Wij evalueren onze kinderen gewoon op basis wat wij alle dagen in de klas zien. En heel veel ouders hebben daar moeite mee want (...) zij hebben het gevoel dat dan allemaal erg subjectief is (...). Heel vaak gebruik ik testen niet zozeer voor mijzelf, maar eerder om iets te bewijzen aan ouders. Om aan ouders aan te tonen van: ‘Kijk, uw kind heeft eigenlijk wel niet de basisleerstof bereikt die die nodig heeft om in een vierde leerjaar in te stappen” (scholen, respondent 6)

2.1.2 Verwachte onwenselijke neveneffecten op microniveau

In elke categorie van respondenten worden punten van zorg geuit omtrent onwenselijke neveneffecten op het microniveau. Deze bezorgdheden kunnen samengebracht worden onder de noemers ‘teaching to the test’ en ‘emotionele impact’.

Teaching to the test

De meerderheid in elk van de vier categorieën van respondenten is bedacht op het fenomeen van ‘teaching to the test’. Vier van de verschillende vormen die in de literatuurstudie werden beschreven, worden ook genoemd door respondenten in de Delphi-studie. Het gaat in de eerste plaats om een vernauwing van het curriculum tot die leerstofdomeinen die zullen getoetst worden, al blijkt dat niet volgens alle respondenten per definitie een probleem te zijn. Wanneer de vernauwing betekent dat leraren hun onderwijs meer richten op de leerplandoelen, kan dit zelfs een wenselijk effect zijn. Dit effect werd in de literatuurstudie ‘teaching to the standard’ genoemd. Toch geeft een ruime meerderheid van de respondenten aan dat de eventuele implementatie van een leer(winst)-monitoringssysteem een bedreiging kan vormen voor belangrijke geachte aspecten van het onderwijs (bijvoorbeeld holistische persoonsontwikkeling, ontwikkeling van sociale vaardigheden, vakoverschrijdende eindtermen en duurzaam leren).

“Ik ben ervan overtuigd dat dat een enorme vershraling van het onderwijs met zich meebrengt. Ik zeg het, dan hebben wij geen tijd meer om hier een week met die gasten te werken rond een project in Bangladesh en hebben we geen tijd voor een flashmob en hebben we geen tijd voor een liedje te zingen of een liedje te maken en hier in de gang allemaal staan samen te zingen. Dan krijg je een vershraling van onderwijs en daar ben ik, dat is mijn grootste vrees.” (scholen, respondent 1)

Ook een andere vorm van vernauwing van het curriculum - met name een vernauwing van het lesgeven tot lagere orde vaardigheden die makkelijk toetsbaar zijn - wordt door respondenten in elke categorie als een te vermijden neveneffect genoemd. Daarnaast vrezen respondenten voor nog twee andere invullingen van *teaching to the test*: een verenging van de manieren waarop leerstof wordt aangeboden, waardoor er minder ruimte is voor creatieve lesmethoden; en een training op de evaluatievormen die in de gestandaardiseerde toetsen worden gebruikt:

“En dus dat is ook een vorm van teaching to the test en niet alleen inhoudelijk, maar ook op formats, op de evaluatievorm gaat werken. Hé, bijvoorbeeld meerkeuzevragen is daar denk ik een mooi voorbeeld van.” (inhoudsdeskundigen respondent 6)

Verondersteld wordt ook dat deze training op toetsvaardigheden ten koste zal gaan van de onderwijstijd.

Een vijfde en zesde vorm van teaching to the test die in de literatuurstudie werd onderscheiden (leerlingen die in de focus worden geplaatst ten koste van andere leerlingen en leraren die de

toetsitems zelf gaan voorschotelen aan leerlingen tijdens de voorbereiding), worden door geen enkele respondent vernoemd.

Emotionele impact

Verder wordt in elke categorie van respondenten aangegeven dat - afhankelijk van het gebruik - gestandaardiseerde toetsen die deel uitmaken van een leer(winst)monitoringssysteem kunnen leiden tot een verhoging van stress en toetsangst bij leerlingen, onder meer als gevolg van een hogere competitiviteit tussen leerlingen en tussen klassen.

“Ik vraag mij gewoon af of dat niet allemaal wat te veel gaat worden. Te veel, de competitiviteit misschien tussen leerlingen en klassen bevorderen. Stress, voor leerlingen met faalangst is dat denk ik niet evident om in zo’n context te moeten functioneren. Tenzij dat je, ja, misschien dat het niet altijd zo duidelijk moet zijn voor de leerlingen van: Dit is nu een toetsmoment. Maar ik weet niet of dat kan.”
(beleidsmakers, respondent 1)

Verhoogde stress en toetsangst bij leerlingen worden voornamelijk verwacht bij high-stakes toetsen waarbij de resultaten van het leer(winst)monitoringssysteem een impact hebben op de studieloopbaan van de leerlingen. Toch wordt ook gesteld dat het voor een betrouwbare meting van leerprestaties van leerlingen net ook nuttig kan zijn dat er voor de leerlingen iets op het spel staat: indien leerlingen niet gemotiveerd zijn om hun best te doen tijdens de toetsafname, krijgt de gebruiker geen realistisch beeld van de leerprestaties of leerwinst van de leerlingen.

“Want bijvoorbeeld voor peilingstoetsen zien we of denken we dat de motivatie soms laag is omdat de stakes laag zijn.” (beleidsmakers, respondent 2)

Daarnaast meent een respondent dat de emotionele impact van een leer(winst)monitoringssysteem niet noodzakelijk hoeft te verschillen van de huidige situatie, zelfs bij de implementatie van high-stakes toetsen.

“Voor de leerlingen maakt het niet uit of ze nu een test krijgen van de leraar of van een centrale toets. Dat is nu ook. Nu moet een leerling ook goed scoren om een diploma te halen. Als die toets nu komt van een centraal instituut waar een heleboel mensen achter zitten of van een individuele leraar, voor die leerling maakt dat niet veel uit.” (inhoudsdeskundigen, respondent 2)

Ook meent een beperkt aantal respondenten dat de emotionele impact op leerlingen niet moet worden overschat, en dat het leren omgaan met minder goede toetsresultaten leerlingen in staat stelt om zich tijdig te herpakken of te heroriënteren.

2.1.3 Beïnvloedende factoren

Brede evaluatie blijven garanderen

Opdat een leer(winst)monitoringssysteem effectief leraren en scholen nieuwe inzichten kan aanreiken over de individuele leerlingen, meent een meerderheid in elk van de vier categorieën van respondenten dat het van belang is dat de gestandaardiseerde toetsen die erin worden aangereikt zo breed mogelijk zijn. Dat wil zeggen dat volgens deze respondenten naast kennis ook vaardigheden, attitudes en (beroeps)competenties moeten geëvalueerd worden. Daarnaast is een

beperkt aantal respondenten van mening dat er aandacht moet geschonken worden aan het toetsen van vakoverschrijdende eindtermen.

“Er moet breder geëvalueerd worden, je moet je niet beperken tot wiskunde en taal. Wat doe je met leren leren, wat doe je met ICT-vaardigheden, wat doe je met sociale vaardigheden? Dit kun je natuurlijk heel heel moeilijk in gestandaardiseerde toetsen gaan gieten.” (scholen, respondent 3)

Bovendien bevestigen elf respondenten dat het toetsaanbod in het leer(winst)monitoringssysteem niet beperkt mag worden tot de leerdomeinen die meest frequent aan bod komen in andere leer(winst)monitoringssystemen, zoals rekenen en taal. Zeven respondenten - waarvan vier uit de categorie ‘scholen’ - opperen het idee van een toetsenbank, waarin uiteenlopende gestandaardiseerde toetsen in een breed gamma aan leerdomeinen kunnen ontleend worden.

“Maar voor ons is het belangrijk dat we, binnen evaluatie, dat je het geheel meehebt van de ontwikkeling die we bij leerlingen willen nagaan. Dat het niet alleen taal en wiskunde is bijvoorbeeld. Maar evenzeer muzische vorming, evenzeer wereldoriëntatie, evenzeer lichamelijke opvoeding, evenzeer de vaardigheden en attitudes. Alle dingen die we verwachten van leerlingen. Dat je de persoon in zijn totaliteit, in al zijn facetten, een stuk bekijkt.” (onderwijsverstrekkers, respondent 5)

Tegelijkertijd erkennen de meeste respondenten, zoals in het eerste citaat, dat uiteenlopende vaardigheden, attitudes en competenties zeer moeilijk via gestandaardiseerde toetsen te evalueren zijn. Nochtans is dat aldus de respondenten niet enkel van belang in het licht van het bereiken van de ontwikkelingsgerichte doeleinden. Het heeft ook een impact op de relevantie van de verantwoordingsgerichte doelen: de voortgang in de studieloopbaan van leerlingen mag niet enkel afhankelijk zijn van kennistoetsen in de meest prominente leerdomeinen. Ten slotte is ‘brede toetsing’ ook van belang in het licht van het neveneffect *teaching to the test*: als het onderwijsaanbod van leraren zich inderdaad voornamelijk zou gaan richten op die zaken die door het leer(winst)monitoringssysteem gemeten worden, is het cruciaal dat de meting zo breed als mogelijk wordt gehouden.

Aanvullend menen drie verschillende respondenten (categorie scholen en inhoudsdeskundigen) dat een leer(winst)monitoringssysteem ook mogelijk moet maken dat meer socio-emotionele zaken, zoals het welbevinden van leerlingen, op een betrouwbare en valide manier in kaart worden gebracht. Dit is dan weer minder het geval bij respondenten die menen dat de focus van het leer(winst)monitoringssysteem voornamelijk op mesoniveau moet liggen:

“Als je kijkt naar variantie op schoolniveau, daar zit weinig variantie op dat soort zaken als welbevinden. Eigenlijk hebben scholen daar minder een impact op en dat varieert ook veel meer. Hoe dat je in je je in uw vel voelt, heeft misschien te maken met de scheiding van je ouders thuis of een ruzie die je 's morgens hebt gehad...” (onderwijsverstrekkers, respondent 4)

Leer(winst)monitoringssysteem gebaseerd op de eindtermen of leerplandoelen

Alle respondenten zijn het erover eens dat, indien er geïnvesteerd wordt in de ontwikkeling van een leer(winst)monitoringssysteem, de gestandaardiseerde toetsen die hiervan deel uitmaken gebaseerd moeten zijn op een vooropgesteld curriculum. Volgens drie van de respondenten moeten

deze gestandaardiseerde toetsen gericht zijn op de leerplannen van de onderwijskoepels. Drie andere respondenten menen dat zowel de leerplannen als de eindtermen in rekening moeten worden genomen bij de toetsontwikkeling. De andere negen respondenten die hun opinie hierover hebben geuit, menen dat het leer(winst)monitoringssysteem moet geënt zijn op de eindtermen en ontwikkelingsdoelen, met bijkomend (volgens zes van deze negen respondenten) uitbreidingsdoelen zodat niet enkel het minimale verwachte niveau wordt getoetst.

Leer(winst)monitoringssysteem levert snelle en duidelijke feedback op

Daarnaast is het volgens de respondenten belangrijk om de resultaten zo snel mogelijk ter beschikking te stellen, waarbij individuele, digitale schoolfeedbackrapporten de voorkeur krijgen. Twee respondenten uit de categorie 'scholen' opperen de mogelijkheid om bij specifieke profielen van leerlingen meteen mogelijke actiepunten te laten suggereren door het leer(winst)monitoringssysteem. Dezelfde respondenten menen dat het vereist is voor het bereiken van een ontwikkelingsgericht effect op microniveau dat de verkregen feedback zeer gedetailleerd kan worden gegeven en dat de analyse zowel per leerling als per item mogelijk wordt gemaakt.

Cultuuromslag in vele scholen

Tot slot stellen verschillende respondenten dat onafhankelijk van de kracht van een leer(winst)-monitoringssysteem, de effectiviteit ervan op het microniveau beperkt zal blijven indien niet wordt geïnvesteerd in een cultuuromslag in vele scholen, waarbij sterker wordt ingezet op differentiatie tussen leerlingen, en op het toekennen van verantwoordelijkheid in het leerproces aan de leerlingen zelf.

2.2 Mesoniveau

2.2.1 Verwachte effecten op mesoniveau

Ontwikkelingsgerichte effecten

Er is een consensus tussen de 21 respondenten die menen dat, indien Vlaanderen in een leer(winst)monitoringssysteem zou investeren, dit systeem een positieve impact kan hebben op de interne kwaliteitszorg van scholen. Input over de leerprestaties en/of leerwinst van de leerlingen kan een betrouwbare basis zijn om het schoolteam te informeren over de sterktes en zwaktes van de school:

"Het geeft een beeld van wat leraren, wat misschien de blinde vlek kan zijn van een leraar. En kan je daar, weer binnen uw school, de interne kwaliteitszorg versterken door een vorm van coaching op basis van dunne plekken die aan het licht komen."
(scholen, respondent 3)

Verschillende respondenten in de categorie 'onderwijsverstrekkers' geven evenwel aan dat dit slechts een beperkt aspect van de school betreft, aangezien de onderwijskwaliteit om meer gaat dan de leerprestaties en/of leerwinst van leerlingen in de meetbare aspecten van een beperkt aantal leerdomeinen (zie ook 2.2.3 rond 'brede evaluatie').

Opdat het leer(winst)monitoringssysteem werkelijk een impact kan hebben op de interne kwaliteitszorg van de school is het belangrijk dat de scholen in staat zijn om betekenis te geven aan de resultaten. Dit kan onder meer door deze resultaten in de tijd te vergelijken met de resultaten van

eerdere cohorten van leerlingen in de eigen school, of door deze resultaten te vergelijken met de resultaten andere scholen. Vooral deze laatste piste krijgt veel bijval vanuit de verschillende categorieën van respondenten, al is er ook hier tegenkanting. Zo menen vier respondenten (categorieën beleidsmakers, scholen en inhoudsdeskundigen) dat de school zich in haar interne kwaliteitszorg door haar eigen proces moet laten leiden, en niet door de vergelijking met andere scholen.

Net zoals de tegengestelde verwachtingen inzake de impact van een leer(winst)monitoringssysteem op leerlingen (microniveau), kunnen er ook op mesoniveau tegengestelde verwachtingen worden teruggevonden inzake de impact op de motivatie van schoolteams. Enerzijds kunnen de resultaten van een leer(winst)monitoringssysteem een schoolteam bevestigen of positief aanmoedigen om aan bepaalde aspecten van de school te werken, terwijl anderzijds het bestaan van een leer(winst)monitoringssysteem ook de professionele identiteit van een schoolteam kan aantasten.

Verantwoordingsgerichte effecten

Een leer(winst)monitoringssysteem zou ook een bijdrage kunnen leveren aan de verantwoording die scholen dienen af te leggen ten aanzien van de overheid. Er bestaat geen consensus over de vraag hoe de implementatie van een leer(winst)monitoringssysteem de werking van de onderwijsinspectie kan versterken. Negen respondenten (waarvan vijf uit de categorie 'scholen' maar geen enkele uit de categorie 'onderwijsverstrekkers') menen dat de onderwijsinspectie de resultaten van de leer(winst)monitoring mag gebruiken binnen de beoordeling van de school, of binnen een voorafgaandelijke risicoanalyse. Zeven van deze negen respondenten kunnen zich evenwel ook vinden in een alternatief dat een grotere gedragenheid van de 16 andere respondenten geniet. In dit alternatief zal de inspectie het bestaan van een leer(winst)monitoringssysteem gebruiken als een instrument om na te gaan hoe de school de resultaten ervan heeft ingezet binnen de eigen werking en binnen de interne kwaliteitszorg.

"Als wij dan in gesprek gaan met onderwijsinspectie, dan lezen wij liever niet in doorlichtingsverslagen: "van deze school is een goede school want ze scoren boven het gemiddelde van Vlaanderen". We lezen liever van "deze school gaat aan de slag met de resultaten". Ze hebben er een analyse van gemaakt en een actieplan is er uit voort gekomen en op die manier gaan ze opvolgen of dat effectief werkt."
(inhoudsdeskundigen, respondent 5)

Daarbij wordt door de onderwijsinspectie niet rechtstreeks naar de output gekeken, maar worden de resultaten gebruikt als vertrekpunt van een analyse van procesfactoren in de school. Op die manier wordt onwenselijk strategisch gebruik van het leer(winst)monitoringssysteem door scholen vermeden.

In ieder geval moet erover worden gewaakt dat inspectie van scholen niet vervalt tot een technocratisch gegeven waarbij het oordeel van de inspecteurs gelijkloopt met de geaggregeerde resultaten opgeleverd door het leer(winst)monitoringssysteem. De meeste respondenten zijn niet gewonnen voor een verantwoordingssysteem dat zich eenzijdig focust op deze outputgegevens. Eén respondent geeft aan dat de onderwijsinspectie bij het bestaan van een leer(winst)monitoringssysteem zelfs voornamelijk zou moeten focussen op de inspanningen van de school om ook de niet-meetbare aspecten van onderwijskwaliteit te verwezenlijken.

Daarnaast kunnen de resultaten op schoolniveau ook gebruikt worden in de verantwoording van de school ten aanzien van het eigen schoolbestuur. Verder menen voornamelijk respondenten in de categorie 'scholen' dat de resultaten van een leer(winst)monitoringssysteem ook verantwoordingsgericht kunnen worden gebruikt op het niveau van de beoordeling van individuele leraren. In geen enkele andere categorie van respondenten vonden we evenwel steun voor dat idee. Integendeel, men meent net dat het een mogelijk onwenselijk neveneffect is dat schoolleiders de resultaten van een leer(winst)monitoringssysteem op oneigenlijke wijze zullen gebruiken binnen de evaluatie van leraren. Toch wordt aangegeven dat de informatie van de leerlingenprestaties schooldirecties wel kunnen helpen om een inschatting te maken van de kwaliteiten van verschillende leraren - maar dan enkel in een ontwikkelingsgericht perspectief:

"Stel dat je dat verhaal [leerwinst van leerlingen] heel concreet in kaart brengt, zeg je: 'Ho, wacht eens eventjes, maar bij mij meester Jan doen die dat heel goed en bij juf Sarah, hohoho, die bakt er precies niets van.' Het is altijd goed voor de directeur dat die die informatie krijgt. Maar het is vooral wat die met die informatie doet. Als dat gebeurt, als dat gebruikt wordt om juf Sarah te kijk te stellen of, dan is dat natuurlijk een ander paar mouwen!" (inhoudsdeskundigen, respondent 6).

2.2.2 Verwachte onwenselijke neveneffecten op mesoniveau

Hoewel de respondenten menen dat de resultaten van een leer(winst)monitoringssysteem zinvol kunnen zijn in het kader van de interne kwaliteitszorg van de school, menen enkele respondenten dat hier ook een gevaar om de hoek schuilt. Zo wordt aangegeven dat indien de interne kwaliteitszorg te sterk wordt geïnspireerd door wat meetbaar kan gemaakt worden in het leer(winst)-monitoringssysteem, de aandacht voor andere belangrijke aspecten van onderwijskwaliteit mogelijk wordt bedreigd:

"De vakoverschrijdende thema's, die zitten nu al wat in het verdomhoekje en dat zijn dan al typisch dingen die je heel moeilijk kunt gaan meten. Die je heel moeilijk kunt evalueren. (...). Die gaan dan nog meer naar de achtergrond worden gedrukt. Want dan zal de roep, als we resultaten op die gestandaardiseerde toetsen hebben, de roep zal dan eerder zijn van: 'Ah, daar zitten we toch niet zo goed, dus we moeten daar meer op gaan inzetten'." (beleidsmakers, respondent 1)

Daarnaast uiten verschillende respondenten hun vrees dat de implementatie van een leer(winst)monitoringssysteem - afhankelijk van het gebruik ervan - zal leiden tot onwenselijke strategisch gedrag van scholen, waarbij de resultaten kunnen worden gemanipuleerd en (sommige) scholen zich zullen inspannen om een zo gunstig mogelijke leerlingenpopulatie aan te trekken, waarbij minder sterke leerlingen van de school worden verwezen. In dezelfde scholen zou men ook makkelijker kunnen overgaan tot een beslissing om een leerling een jaar te laten overdoen, of te verwijzen naar een andere studierichting.

2.2.3 Beïnvloedende factoren

Openbaarheid van de resultaten

De mate waarin onwenselijke neveneffecten op mesoniveau zich voordoen hangt sterk samen met de beslissing of de resultaten van het leer(winst)monitoringssysteem al dan niet openbaar zullen

worden gemaakt. Deze openbaarheid blijkt een sterk overheersende factor in de discussie tussen de 24 respondenten.

Vier respondenten in het expertenpanel, verdeeld over de categorieën 'scholen' en 'inhoudsdeskundigen', zijn van mening dat de resultaten van een leer(winst)monitoringssysteem, geaggregeerd op schoolniveau, openbaar mogen worden gemaakt. Bijkomend meent één respondent dat de resultaten niet globaal mogen worden gepubliceerd, maar dat ouders (bijvoorbeeld via de oudergeleding in de schoolraad) wel recht hebben op deze informatie. Eén respondent wil hierin geen standpunt innemen, terwijl de achttien andere respondenten uitgesproken tegenstander zijn van openbaarheid van resultaten. De grootste vrees van de tegenstanders van openbaarheid is dat rankings zullen worden opgemaakt die op basis van beperkte gegevens een onterecht als betrouwbaar beschouwde (want op een beperkt deel van het hele takenpakket van de school) uitspraak zullen doen over de kwaliteit van de school. De gebruikers van deze rankings (bijvoorbeeld lezers van media die de ranking opmaken en publiceren) beschikken immers niet over de achtergrondkennis om op een genuanceerde manier met deze gegevens om te gaan. Hierdoor krijgen de rankings op basis van toetsen in een leer(winst)monitoringssysteem onterecht een 'waarheidswaarde' waarin de verschillen tussen scholen onterecht worden uitvergroot.

Bij een eventuele ranking zou daarom in principe moeten aangegeven worden tussen welke scholen de resultaten statistisch significant (met zekere effectgrootte) van elkaar verschillen, maar deze nuances zullen bij publicatie in de media niet de nodige aandacht krijgen. Daarnaast wordt onder meer de unidimensionaliteit van de rankings door verschillende respondenten bekritiseerd: een ranking suggereert dat er één bepaalde maat bestaat voor onderwijskwaliteit, waarop alle scholen kunnen worden ingeschaald.

"Er is geen 'beste school'. Er is een 'beste school voor die of die leerling'. Ik heb 3 kinderen gehad. Die zijn alle 3 naar een andere school geweest in het secundair. Waarom? Je moet een match maken tussen soort van kind en soort van school."
(Onderwijsverstrekkers, respondent 3).

Een ander nadeel van openbaarheid van gegevens en onvermijdelijk geachte rankings is dat er een sociale segregatie zal ontstaan tussen scholen, waarbij sterke scholen voornamelijk sterke leerlingen (met een gunstige socio-economische status) zullen aantrekken - zowel door een actief beleid van scholen om de 'beste' leerlingen aan te trekken, als door (geïnformeerde) ouders die hun kinderen naar de scholen willen brengen die bovenaan de ranking staan. De verschillen tussen het leerlingenpubliek van verschillende scholen zouden daardoor groter worden, aldus deze respondenten.

Verschillende van de in secties 2.2 en 2.4 beschreven veronderstelde onwenselijke neveneffecten worden gestimuleerd door de openbaarheid van de resultaten - terwijl de meer wenselijke effecten zoals de impact op de interne kwaliteitszorg van scholen, verstoord worden door de openbaarheid van de resultaten.

Hoewel de vier voorstanders van openbaarheid verschillende van bovenstaande argumenten accepteren, nemen ze op basis van een aantal tegen argumenten een andere positie in. Het publiceren van de resultaten van gestandaardiseerde toetsen zou bijdragen tot het borgen en zelfs bevorderen van de onderwijskwaliteit in scholen door het toevoegen van een externe stimulans.

Daarnaast wordt door drie respondenten nog een bijkomend voordeel van de openbaarheid van resultaten gegeven, namelijk dat deze openbaarheid zou kunnen leiden tot een beter geïnformeerde schoolkeuze van ouders voor hun kinderen – al loert daar het gevaar van een grotere sociale segregatie weer om de hoek, zoals blijkt uit volgend citaat van één van deze voorstanders:

“Het probleem is dat meestal de moeder van Fatimah die vraag zich niet stelt, maar de vader van Jan zich die vraag wel stelt. Maar ja, ik heb dan het liefst,... mijn grootste bezorgdheid is dan natuurlijk dat die moeder van Fatimah dat dan ook op de best mogelijke manier doet.” (inhoudsdeskundigen, respondent 6)

Beleidsvoerend vermogen van de school

Het aanwenden van data als een instrument om het beleid van scholen vorm te geven, is volgens de respondenten geen vanzelfsprekendheid. De mate waarin een leer(winst)monitoringssysteem een impact kan hebben op de interne kwaliteitszorg van de scholen, alsook de mate waarin scholen zich laten verleiden tot ongewenst strategisch gedrag, is volgens respondenten uit verschillende categorieën sterk gelinkt aan het beleidsvoerend vermogen van scholen. De veronderstelling luidt dat hoe sterker dit beleidsvoerend vermogen is ontwikkeld, hoe groter de positieve impact zal zijn.

“En dat zij [gestandaardiseerde toetsen] de school een spiegel kunnen voorhouden, maar is dat dan voldoende? Ik weet dat niet. Dat hangt dan weer af, denk ik, van, ja, wat dat ze dan beleidsvoerend vermogen van de school noemen.” (beleidsmakers, respondent 1)

Bovendien wordt verondersteld dat scholen met een sterk ontwikkeld beleidsvoerend vermogen minder onwenselijke neveneffecten zullen vertonen, omdat de aanvaarding van een leer(winst)monitoringssysteem wellicht groter is, alsook omdat er in sterkere mate een cultuur heerst van vertrouwen dat de resultaten binnen een ontwikkelingsgericht perspectief zullen worden gebruikt en kunnen leiden tot kwaliteitsverbetering ten bate van de leerlingen. Scholen met een beperkte datageletterdheid kunnen verkeerde conclusies trekken en daardoor ondoordachte acties op het getouw zetten.

“Als scholen zich sterk genoeg voelen en een sterk eigen beleid kunnen voeren op basis van instrumenten die ze krijgen ook om leerwinst in kaart te brengen, om daardoor hun eigen beleid binnen de school, binnen de klassen, met medeweten van leerlingen, te voeren... Dat je mooi, sterk in je schoenen staat, dan heeft geen enkele school problemen met de blik van derden.” (scholen, respondent 3)

De meest genoemde relevante dragers van beleidsvoerend vermogen in dit kader zijn het ‘reflectief vermogen’ van scholen, alsook het ‘innovatief vermogen’. De mate waarin scholen over de capaciteit en bereidheid beschikken om de eigen werking in vraag te stellen alsook de mate waarin ze kunnen en willen experimenteren met nieuwe invalshoeken die geïntegreerd worden in de dagelijkse werking, zal de openheid van het schoolteam over de resultaten van het leer(winst)monitoringssysteem vergroten. Het betreft evenwel een wisselwerking volgens de respondenten, want de implementatie van een leer(winst)monitoringssysteem kan ook het reflectief vermogen en het innovatief vermogen van scholen positief beïnvloeden. Dat kan onder meer omdat de gegevens van het leer(winst)monitoringssysteem kunnen gebruikt worden om de eigen innovaties te evalueren en als gevolg daarvan bij te sturen:

“Ik herinner mij het voorbeeld van een school die een bepaalde leesinterventie deed. Die hebben dus eigenlijk de leerwinst vergeleken van leerlingen die in dat programma zaten, met een aantal leerlingen die dan niet in dat programma zaten. En dan hebben ze zo het verschil in leerwinst bekeken. Dat vind ik wel een heel mooi voorbeeld.”
(onderwijsverstrekkers, respondent 4)

Een andere drager van beleidsvoerend vermogen die in dit kader wordt genoemd, wordt gevormd door de ‘professionele ondersteunende relaties tussen collega’s’: het is nodig dat de verschillende onderwijsprofessionals constructief durven en kunnen samenwerken en zich als een hecht team kunnen profileren opdat de interne kwaliteitszorg van de school effectief de vruchten kan plukken van de resultaten van het leer(winst)monitoringssysteem. Tegelijk zagen we evenwel - net als hierboven bij het innovatief en reflectief vermogen - dat bepaalde respondenten menen dat net de resultaten van het leer(winst)monitoringssysteem een impuls kunnen geven tot de professionele ondersteunende relaties in de scholen. Door respondenten uit de categorie ‘scholen’ wordt aangegeven dat concrete informatie over leerprestaties en/of leerwinst leraren kan stimuleren om deze resultaten met collega’s te bespreken, waardoor de dialoog tussen leraren wordt gestimuleerd en tegelijk meer doelgericht zou kunnen worden.

Een meerderheid van de respondenten in elke categorie is er niet van overtuigd dat het beleidsvoerend vermogen van de meeste scholen momenteel al in voldoende mate is ontwikkeld om op een adequate manier met een leer(winst)monitoringssysteem om te gaan. Negen respondenten, verdeeld over de vier categorieën, geven dan ook aan dat bijkomende professionalisering en ondersteuning vanuit de overheid, vanuit de toetsontwikkelaars of vanuit de onderwijsverstrekkers een cruciale voorwaarde is om de beoogde effecten van een leer(winst)monitoringssysteem op mesoniveau te bewerkstelligen en om onwenselijke neveneffecten (en onwenselijk strategisch gedrag) te vermijden als verkeerd gebruik door de scholen. De investeringen in het versterken van de datageletterdheid hoeven zich overigens niet te beperken tot leraren en scholen, maar ook de datageletterdheid van schoolbesturen en ouders is van belang om succesvol met data van een leer(winst)monitoringssysteem aan de slag te gaan.

2.3 Macroniveau

2.3.1 Verwachte effecten op macroniveau

Respondenten in elke categorie geven daarnaast aan dat de resultaten van een leer(winst)-monitoringssysteem het onderwijsbeleid van zinvolle data kunnen voorzien om het Vlaamse onderwijssysteem verder vorm te geven en te optimaliseren, alsook om aan het publiek en aan relevante instellingen aan te geven wat de waarde van dit onderwijssysteem is. Er is dus ook op dit niveau sprake van potentieel gebruik in zowel een ontwikkelingsgericht als een verantwoordingsgericht perspectief.

Ontwikkelingsgericht perspectief

Volgens vele respondenten stelt een leer(winst)monitoringssysteem de overheid in staat om een beter beeld te krijgen van de algemene onderwijskwaliteit in Vlaanderen, en van de onderwijskwaliteit voor specifieke subgroepen van leerlingen. Respondenten geven aan dat een leer(winst)-monitoringssysteem aanleiding kan geven tot het trekken van lessen op beleidsniveau

om op die manier de onderwijskwaliteit op scholen te bewaken en te verbeteren. Eén respondent stelt dat uit de resultaten van de peilingen duidelijk naar voor komt dat er een *“mismatch bestaat tussen het basisonderwijs en de eerste graad secundair onderwijs”* (beleidsmakers, respondent 2). Middels een leer(winst)monitoringssysteem kan de overgang tussen beide leerjaren beter opgevolgd worden. Het onderzoeken, evalueren en verder ontwikkelen van het huidig onderwijssysteem is in dat opzicht een item dat door meerdere respondenten aan bod komt tijdens de interviews.

“Als de overheid niet met gegevens begint te werken, moet je het dan vragen aan duizend individuele scholen?” (inhoudsdeskundigen, respondent 3)

Bovendien wordt een databank aan gegevens verzameld die toepassingsgericht, beleidsgericht en fundamenteel onderzoek mogelijk zou maken. Daarnaast geven verschillende respondenten - uit de categorieën beleidsmakers, onderwijsverstrekkers en inhoudsdeskundigen - aan dat een leer(winst)monitoringssysteem ook gebruikt zou moeten kunnen worden om de eindtermen te evalueren.

“Wellicht ook, dat is, ruimer bekeken, voor de overheid ook interessant van die eindtermen die geformuleerd zijn, in welke mate worden die gehaald en om daar conclusies uit te trekken? Die conclusies moeten niet altijd zijn van, ja, de scholen zullen een tandje moeten bijsteken. Het zou ook kunnen dat die eindterm slecht geformuleerd is. Dus dat is een, allez, ruimer om te bekijken, maar naar het beleid toe is dat een interessant instrument.” (onderwijsverstrekkers, respondent 1)

Toch is niet elke respondent ervan overtuigd dat er werkelijk een impact zal zijn op het onderwijsbeleid. Dit voorbehoud wordt ingegeven door de gepercipieerde beperkte impact die de peilingen hebben op het onderwijsbeleid in Vlaanderen.

“Het is nu twee keer vastgesteld met een peiling van wiskunde basisonderwijs dat herleidingen dat die ondermaats scoren. Want er gebeurt niks. Dus je zou denken van ja dan gaan we toch eens kritisch kijken. Niet alleen van hoe komt dat nu dat dat niet gaat, maar is onze eindterm wel op niveau basisonderwijs?” (inhoudsdeskundigen, respondent 5)

Bovendien kan het onderwijssysteem op macroniveau de vruchten plukken van een leer(winst)monitoringssysteem doordat de externe kwaliteitsbewaking van scholen beter kan worden georganiseerd. Zoals reeds bleek, zijn niet veel respondenten voorstander van een directe link tussen de (geaggregeerde) resultaten van leerlingen in een school en een kwaliteitsoordeel over de school, maar alle respondenten menen wel dat de externe kwaliteitsbewaking door middel van de doorlichtingen van de onderwijsinspectie erdoor kunnen worden versterkt.

Ten slotte menen respondenten uit de categorie ‘beleidsmakers’ dat er een effect van een leer(winst)monitoringssysteem kan zijn op macroniveau doordat onderwijsverstrekkers de begeleiding aan scholen meer gericht kunnen inzetten en beter trajecten op maat kunnen voorzien. Dit effect kan uiteraard enkel bereikt worden indien de onderwijsverstrekkers zelf over de resultaten ervan beschikken.

Ook kunnen scholen die sterke resultaten behalen op de toetsen in het leer(winst)monitoringssysteem gevraagd worden om hun expertise te delen zodat *good practices* verspreid worden.

“Kijk in heel die discussie heb je natuurlijk ook vaak heel erg aan het begin van het secundair onderwijs of aan beginpunten van: Ja, maar zie eens wat, onze leerlingen die kunnen toch niets meer. Maar als de mensen, als die jongeren en die kinderen de eindtermen hebben gehaald, dan kunnen die wél alles, en dan moet het secundair onderwijs niet mekkeren. Ze moeten daar op voortbouwen. Ze moeten vooral hun eigen lat daar niet ten opzichte, tegenover leggen.” (inhoudsdeskundigen, respondent 6)

Verantwoordingsgerichte effecten

Verschillende respondenten geven aan dat het onderwijssysteem zich door middel van de resultaten van een leer(winst)monitoringssysteem op macroniveau zou kunnen verantwoorden voor haar organisatie ten aanzien van de verschillende stakeholders. Binnen het verantwoordingsperspectief worden verder nog genoemd dat de impact van eventuele innovaties door middel van het leer(winst)monitoringssysteem zou kunnen worden geëvalueerd en verantwoord, en dat de allocatie van middelen (bijvoorbeeld toedeling van SES-middelen) door deze resultaten zou kunnen worden gelegitimeerd.

Een ander voordeel op macroniveau is dat een onderwijsniveau zich kan verantwoorden ten opzichte van het vervolgonderwijs: als immers een objectief geacht leer(winst)monitoringssysteem heeft aangetoond dat leerlingen de eindtermen hebben bereikt, kan dit door het vervolgonderwijs niet als excuus worden gebruikt om eventuele problemen in de studievoortgang van leerlingen (bijvoorbeeld door zittenblijven) af te wentelen op het beginniveau van leerlingen:

2.3.2 Verwachte onwenselijke neveneffecten op macroniveau

In totaal maken negen respondenten - waarvan de meerderheid uit de groep onderwijsverstrekkers - zich zorgen over de autonomie en de pedagogische vrijheid van scholen. Door de implementatie van een leer(winst)monitoringssysteem worden scholen immers in een bepaalde richting geduwd waardoor hun autonomie zou worden bedreigd. Daar staat evenwel tegenover dat twee respondenten menen dat scholen net meer autonomie zouden kunnen krijgen wanneer een leer(winst)monitoringssysteem beschikbaar is: enerzijds zouden scholen waaruit blijkt dat ze goede resultaten (in termen van leerprestaties en/of leerwinst) kunnen voorleggen, een stimulans kunnen krijgen om hun autonomie net ruimer in te vullen, anderzijds zou aan dezelfde scholen door het onderwijsbeleid een nog grotere autonomie kunnen worden toegekend.

Daarnaast zijn er verschillende respondenten die menen dat het debat over de kwaliteit van het onderwijs mogelijk wordt verengd omdat men vooral oog zal hebben voor de resultaten uit het leer(winst)monitoringssysteem. Aangezien verondersteld wordt dat enkel makkelijk meetbare eindtermen geëvalueerd zullen worden, kan dit systeem aanleiding geven tot een ongebalanceerde kijk op onderwijskwaliteit.

Ten slotte uit een aantal respondenten een bezorgdheid over het ontstaan van een toetscultuur.

“Obsessieve cijfertjescultuur bestaat in Vlaanderen nog niet. Goh ja, ik zou dat ook niet gunstig vinden.” (inhoudsdeskundigen, respondent 4)

2.3.3 Beïnvloedende factoren

Cultuuromslag in het onderwijsveld

Elf respondenten - waarvan zes in de categorie 'scholen' - geven bovendien aan dat het effectief gebruik van een leer(winst)monitoringssysteem een cultuuromslag in het onderwijsveld vereist waarin 'vertrouwen' centraal staat opdat scholen effectief van de resultaten van een leer(winst)monitoringssysteem gebruik kunnen maken. Scholen moeten aanvoelen dat ze in vrijheid kunnen werken en dat de overheid hen vertrouwt dat ze op een zinvolle manier met het leer(winst)monitoringssysteem aan de slag gaan.

"Onze doorlichtingen werken nog altijd binnen het concept van theorie X: 'eigenlijk zijn scholen en leraars en directeurs niet te vertrouwen als ze zeggen dat ze goed werken. Ik wil cijfers zien die dat aantonen dat ze goed werken'. Ik zeg: Nee, vertrek eens van theorie Y: 'we gaan ervan uit dat je goed bezig bent tot we duidelijk het tegenoverstelde komen te zien natuurlijk'. Dat is wat anders en dan kom je in de situatie waarin dat je zegt van: 'Ik vertrouw erop dat je goed bezig bent. Ik kom eens checken hoe dat je met dat proces bezig bent'." (scholen, respondent 2)

De gevolgtrekking uit deze vraag naar vertrouwen is evenwel niet gelijklopend: voor sommige respondenten betekent dit een argument om geen verantwoordingsgericht perspectief toe te kennen aan de resultaten van een leer(winst)monitoringssysteem. Andere respondenten (allen uit de categorie 'scholen') menen integendeel dat scholen zich door middel van de resultaten van een leer(winst)-monitoringssysteem kunnen verantwoorden voor het gegeven vertrouwen.

Gerelateerd aan de nood voor de verdere professionalisering in scholen en de ontwikkeling van het beleidsvoerend vermogen van scholen, wordt verder gesteld dat verschillende stakeholders op grote schaal moeten voorbereid worden op een onderwijssysteem dat zich meer door gestandaardiseerde data wil laten informeren. Het Vlaamse onderwijssysteem moet verder doorgroeien naar een 'kwaliteitscultuur' waarin zaken in vraag worden gesteld, en waarin leerprestaties en leerwinst in de mate van het mogelijke gemeten worden. Verwacht wordt dat dit een werk van lange adem is, omdat in de Vlaamse onderwijscontext slechts in beperkte mate in output-termen wordt gedacht:

"Als je zo'n systeem invoert, ik denk dat je dan als beleid minimaal voorafgaand, zeer zeker flankerend, moet inzetten op precies een aantal van die belangrijke dingen als scholen leren omgaan met een meetcultuur. Leraren leren omgaan met een meetcultuur. Ouders leren omgaan met een meetcultuur. Maar ik denk dat je daar heel veel, een hele andere mindset voor nodig hebt." (inhoudsdeskundigen, respondent 6)

Actief werken aan een draagvlak

Tien respondenten, verspreid over elk van de vier categorieën, wijzen op het belang van de aanvaarding van het leer(winst)monitoringssysteem door uiteenlopende stakeholders in het Vlaamse onderwijsveld. De overheid wordt op haar taak gewezen om bij de beslissing tot invoering van een leer(winst)monitoringssysteem expliciet een campagne op te zetten met onder meer publicaties en vormingsdagen die de meerwaarde hiervan aantoonen voor de verschillende betrokkenen. Daarnaast moeten in deze communicatie mogelijke gevolgen (onder meer het ontstaan van rankings) stevig en onderbouwd worden ontmoedigd.

Uiteenlopende respondenten schatten dat er makkelijker een draagvlak zal gevonden worden indien louter op ontwikkelingsgerichte perspectieven wordt ingezet waarbij de scholen een soort 'eigenaarschap' ontwikkelen van het leer(winst)monitoringssysteem. Het ontwikkelingsgericht effect zou, volgens bepaalde respondenten, versterkt worden door adviseurs ter beschikking te stellen die scholen ondersteunen in het gebruik van de resultaten van het leer(winst)monitoringssysteem.

Daarnaast geven een aantal respondenten aan dat het interessant kan zijn om te vertrekken vanuit de bestaande koepelgebonden toetsen en eerst deze te valideren. Zo hoeft het bestaand en aanwezig materiaal niet te worden weggegooid. Men meent ook dat hierdoor het draagvlak tegenover een leer(winst)monitoringssysteem vanuit de koepels vergroot kan worden. Opmerkelijk is dat dit idee gesteund wordt door vier respondenten uit de categorie 'inhoudsdeskundigen', maar door andere respondenten niet wordt vermeld (ook niet tegengesproken evenwel).

Proeffase

De aanvaarding kan ook vergroten door bij de implementatie van een leer(winst)monitoringssysteem te starten met een kleine steekproef van scholen die vrijwillig deelnemen, waarbij grondig wordt nagegaan wat de effectiviteit van het leer(winst)monitoringssysteem is, tot welke onwenselijke neveneffecten de implementatie leidt, en aan welke voorwaarden voldaan moet zijn om de effectiviteit te garanderen. Vijf respondenten benadrukken het belang van een proeffase.

Geleidelijke ontwikkeling

Daarnaast zijn er een aantal respondenten die aangeven dat het relevant kan zijn om van start te gaan met een beperkte set toetsen, gericht op hoofdvakken en op de ontwikkeling van scholen, om daarna de lijn verder door te trekken naar andere vakken en om eventueel ook verantwoordingsgerichte doelen te integreren.

"Ik denk dat je eigenlijk hier stapsgewijs kunt gaan. Ik ben realistisch genoeg om te beseffen dat het onmogelijk is om op relatief korte termijn een gestandaardiseerde set te gaan hebben die de volledige breedte van het secundaire onderwijs gaat nagaan. Dat beseft ik heel goed maar dat moet daar stapsgewijs bekeken worden. Ik zou kunnen leven met een eerste fase waarin we zeggen van oké laat het ons nu de gemakkelijkste dingen, laat ons die eerst nemen. Maar er moeten fasen op volgen. En dat dat niet met alle studierichtingen tegelijk is, oké, maar laat ons er dan keuzes in maken voor welke studierichtingen we verder willen gaan en in de toekomst nog verder willen opbouwen [...]. Laat ons starten met de gemakkelijke zaken maar met de ambitie om verder te gaan." (scholen, respondent 1)

3 LEERWINST EN TOEGEVOEGDE WAARDE

In onderstaand luik worden de opinies van de 24 respondenten besproken inzake de vraag of bij de implementatie van een leer(winst)monitoringssysteem ook ingezet moet worden op de mogelijkheid om de leerwinst van individuele leerlingen en de toegevoegde waarde van scholen in kaart te brengen.

3.1 Leerwinst

In deze sectie wordt gerapporteerd hoe leerwinst volgens de 24 respondenten moet worden geconceptualiseerd. Daarna bespreken we de mate waarin leerwinstberekening als onderdeel van een monitoringssysteem wenselijk wordt geacht.

3.1.1 Conceptualisering

De meeste respondenten zijn het erover eens dat leerwinst best berekend wordt op basis van twee metingen - een beginmeting en een eindmeting - van een bepaald kennisdomein of een vaardigheid. Deze respondenten noemen het verschil tussen beide metingen daarbij de leerwinst, op twee respondenten na (beide uit de categorie 'inhoudsdeskundigen') die leerwinst zien als het verschil tussen een verwachte score (op basis van de beginmeting) en een geobserveerde score (op de eindmeting).

Drie respondenten (twee inhoudsdeskundigen, één respondent uit de categorie 'scholen') menen dat, waar mogelijk, leerwinst rekening moet houden met meer dan twee meetmomenten. Doorheen de resultaten van verschillende opeenvolgende toetsen wordt dan per leerling een regressielijn getekend. Vier respondenten geven het belang aan van gekalibreerde toetsen (via Item-Response Theory (IRT)) wanneer twee verschillende toetsen worden gebruikt om de leerwinst in kaart te brengen. Toch is er ook bij de inhoudsdeskundigen een respondent die meent dat het arbeidsintensieve werk van IRT-inschaling niet per definitie nodig en mogelijk is:

"Dat is serieus werk het IRT en het is een beetje omstreden. Er zijn mensen die zeggen: "Nee, dat op één dimensie brengen over langere jaren, dat gaat niet. De wiskunde in een hoger leerjaar is iets anders dan de wiskunde in het eerste leerjaar. Een kritische IRT-analist zegt: "Nee, er zitten meerdere dimensies in. (...) Dus ik zie toch regelmatig mensen die zeggen: "Nee, percentielscores van representatieve groepen, dat is even goed enzovoort." Dus ik zeg u gewoon: Je moet daar toch over nadenken." (inhoudsdeskundigen, respondent 3)

Hoewel er dus wel consensus is dat leerwinstberekening minimaal twee opeenvolgende toetsen vereist, vinden de meeste respondenten het vrij moeilijk om een inschatting te geven van de noodzakelijke frequentie van toetsing om leerwinstmonitoring mogelijk te maken. Het meest gegeven antwoord is dat toetsing nodig is één maal per schooljaar, maar de antwoorden variëren van drie maal per jaar tot éénmaal per twee jaar.

3.1.2 Wenselijkheid

Van de 21 respondenten, die menen dat Vlaanderen moet investeren in de ontwikkeling van een monitoringssysteem, vinden 16 respondenten dat dit systeem de gebruikers in staat moet stellen om de leerwinst van individuele leerlingen in kaart te brengen. Er worden zowel op micro-, meso- als op macroniveau verschillende voordelen genoemd van leerwinstmonitoring, waardoor deze 16 respondenten het registeren van leerwinst verkiezen boven het louter meten van de leerprestaties van leerlingen. Dit impliceert dat een groot aantal respondenten aangeven voorstander te zijn van het gebruik van een zelfgericht criterium waarbij de evolutie van de leerling ten aanzien van zijn/haar eerdere prestaties van belang is.

“Ik heb daar niets aan als mijn leerlingen vergeleken worden met de rest van Vlaanderen of misschien zelfs Nederland of met het katholiek onderwijs (...). Ik heb daar niets aan om dat te weten. Ik heb liever dat ik weet: Zeg, die leerling heeft nogal een evolutie gemaakt.” (scholen, respondent 6)

Het belangrijkste argument is dat leerwinstgegevens een eerlijker en vollediger beeld verschaffen van elke leerling. Terwijl leerprestaties een éénmalig beeld geven dat door allerlei omstandigheden kan vertekend zijn, geeft leerwinstmeting op basis van verschillende toetsen een meer genuanceerd beeld: daarom meent men dat leerwinstmonitoring betrouwbaarder kan zijn dan éénmalige toetsen. Op die manier geven leerwinstgegevens een betere basis om geïndividualiseerde trajecten van leerlingen te ontwikkelen of te valoriseren.

Daarnaast uiten respondenten de opinie dat leerwinstgegevens het verantwoordelijkheidsgevoel en de motivatie bij leraren bevorderen. Leraren kunnen immers makkelijker opvolgen tot welk resultaat hun inspanningen leiden, ook bij leerlingen waarbij het moeilijker is om de prestaties op het niveau van de eindtermen te brengen. Schooldirecties krijgen door de leerwinstgegevens dan weer zicht op de prestaties van leraren.

“Ik denk eigenlijk dat je als leraar wilt weten hoe goed je aan het doen bent, dan is het uiteindelijk... Het gaat uiteindelijk over: wat breng ik mijn leerlingen bij, eerder dan: wat kunnen ze al?” (onderwijsverstrekkers, respondent 4)

Een ander argument waarom leraren meer baat hebben bij leerwinstmonitoring is omdat dit hen meer nieuwe informatie oplevert dan het louter meten van prestaties. Dat laatste kunnen leraren tot op zekere hoogte zelf in kaart brengen, terwijl leerwinstmonitoring meestal niet door leraren zelf kan worden uitgevoerd.

“Het is ook een knelpunt hé. Dat wordt altijd gezegd van: ‘Je moet, de leerwinst is het eigenlijke doel en je moet de leerwinst meten.’ Er is voor scholen en voor directies, voor leraren niets zo moeilijk dan leerwinst meten” (scholen, respondent 2)

Bovenstaande geldt evenwel voornamelijk voor het secundair onderwijs, want respondenten in het lager onderwijs geven aan dat er al leerlingvolgsystemen bestaan waarmee leraren ook zicht krijgen op de leerwinst van leerlingen doorheen hun schoolloopbaan.

Ook leerlingen zelf zijn, aldus deze respondenten, meer gebaat met leerwinstmonitoring. Deze resultaten kunnen ze immers aanwenden om zicht te krijgen op de eigen vorderingen. Leerlingen die hun eigen groei kunnen zien, zijn volgens de respondenten niet enkel beter gemotiveerd, maar worden meer betrokken in het eigen leerproces.

Andere respondenten stellen dat gegevens over leerwinst bij leerlingen ook op macroniveau een bijkomende input kunnen zijn voor onderwijskundig onderzoek. Een ander wenselijk effect dat op macroniveau zou worden versterkt, is dat het debat over de onderwijskwaliteit sterker gericht zou zijn op leerwinst:

“Als je vandaag ziet, de impact van bijvoorbeeld de PISA-resultaten op het maatschappelijk debat; als we dan toch naar gestandaardiseerde toetsing gaan die beleidsmatig zou inzetten, dan zou ik het wel een meerwaarde vinden wanneer het dan zou gaan over leerwinst in plaats van over prestatie.” (onderwijsverstrekkers, respondent 5)

Naast een groter positief effect, menen de respondenten die voorstander zijn van het meten van leerwinst dat een aantal onwenselijke neveneffecten worden ingeperkt wanneer een monitoringssysteem in dienst staat van het meten van leerwinst in plaats van leerprestaties. Zo wordt verwacht dat de emotionele impact op leerlingen en leraren minder groot is, en dat leraren minder verleid zullen worden tot allerlei vormen van teaching to the test:

“Een aantal van mijn zorgen of opmerkingen die ik heb met betrekking tot het meten van prestaties, namelijk die teach to the test en zo van die dingen. Een aantal daarvan vallen volgens mij wel weg wanneer het gaat over het meten van leerwinst. Dus alle, ik vind, er is nogal wat voor te zeggen om leerwinst te meten.” (beleidsmakers, respondent 3)

Een aantal respondenten vreest echter dat de verschillende toetsmomenten in het kader van leerwinstmonitoring een toetscultuur met zich zullen meebrengen. De vijf respondenten, die menen dat de overheid niet moet investeren in leerwinstmonitoring, zijn meestal voorstander van een monitoringssysteem dat zich richt op het mesoniveau, waarbij men meent dat gegevens over de mate waarin leerlingen de eindtermen/leerplandoelen hebben bereikt, volstaan. Andere argumenten tegen leerwinstmonitoring zijn de complexiteit ervan, alsook op gepercipieerde noodzaak om tussendoelen te bepalen op het eind of zelfs het midden van elk leerjaar (terwijl eindtermen momenteel veel algemener zijn geformuleerd, bijvoorbeeld enkel einde basisonderwijs).

3.2 Toegevoegde waarde

Gelijklopend met de vorige sectie bespreken we eerst hoe de 24 respondenten invulling geven aan het begrip toegevoegde waarde. Daarna bespreken we de mate waarin een berekening van de toegevoegde waarde van de school als onderdeel van een monitoringssysteem wenselijk wordt geacht.

3.2.1 Conceptualisering

De respondenten geven aan dat toegevoegde waarde van de school in de Vlaamse onderwijscontext moet bepaald worden door een vergelijking van de leerprestaties of leerwinst van de leerlingen in de betreffende school ten opzichte van de leerprestaties/leerwinst van leerlingen in vergelijkbare scholen. Voor elk van deze respondenten moeten inputkenmerken van de leerlingen (bijvoorbeeld thuistaal, opleidingsniveau ouders en socio-economische achtergrond) in de scholen in rekening worden genomen bij het identificeren van ‘vergelijkbare scholen’, en twee respondenten geven bijkomend aan dat contextkenmerken - met name de ligging van de school - hieraan moeten worden toegevoegd. Toch waarschuwen twee respondenten voor een gevaar dat schuilt in het gebruik van input- en contextkenmerken: men voelt aan dat hierdoor de boodschap wordt gegeven dat van leerlingen met bepaalde achtergrondkenmerken per definitie minder wordt verwacht.

“Als je gaat controleren voor anderstaligheid, dan ga je eigenlijk op leerlingniveau beschouwen als: die zit in een ander groep, dat is een anderstalige, dus we steken die daar. Ik denk dat dat echt wel hokjesdenken is op leerlingniveau. Op schoolniveau heeft dat dan wel wat anders, maar op leerlingniveau... Dan heeft die leerling, een anderstalige, die zit in dat groepje... Ik denk dat je daar een naar gevoel bij hebt.” (onderwijsverstrekkers, respondent 3)

Een aantal respondenten argumenteert bovendien dat het quasi onmogelijk is om die achtergrondkenmerken volledig correct in rekening te brengen.

Een aantal respondenten (uit de categorie 'inhoudsdeskundigen') gaat een stap verder in de conceptualisering van toegevoegde waarde. In plaats van enkel een vergelijking tussen te scholen te maken, opperen zij de optie om per school een verwachte score te berekenen, en de effectieve gemiddelde score van leerlingen daarmee te vergelijken. Toegevoegde waarde wordt dan gedefinieerd als het verschil tussen de effectieve leerprestaties/leerwinst van de leerlingen in vergelijking tot de leerprestaties/leerwinst die op basis van de input-kenmerken van de leerlingen in de school verwacht zou mogen worden. Het voorbehoud in verband met de uitdaging om deze achtergrondkenmerken correct in rekening te brengen, gaat uiteraard ook in deze conceptualisering van toegevoegde waarde op.

Bovendien verdient het volgens verschillende respondenten in elke categorie aanbeveling om bij de bovenstaand beschreven conceptualisering van de toegevoegde waarde van scholen niet (enkel) de geaggregeerde leerprestaties van leerlingen per school te vergelijken, maar ook de geaggregeerde leerwinst te vergelijken: dat geeft immers een eerlijker beeld, aldus deze respondenten. Hierbij wordt de toegevoegde waarde van scholen berekend als het verschil tussen de geaggregeerde leerwinst van leerlingen en de te verwachten leerwinst op basis van het beginniveau van de leerlingen. Bovendien wordt beweerd dat verschillen op scholen groter zijn wanneer met leerwinst rekening wordt gehouden, dan wanneer louter met leerprestaties wordt gewerkt. Hierdoor komen eventuele verschillen in de resultaten van scholen dus duidelijker naar voren:

De meeste respondenten onthouden zich van een antwoord bij de vraag of er bij toegevoegde waarde op basis van leerwinst ook rekening moet gehouden worden met input- en contextkenmerken van scholen. Onder degenen die hierover wel een uitspraak doen, zijn de meningen verdeeld. Ten slotte wordt opgemerkt door twee van de inhoudsdeskundigen dat bij berekening van toegevoegde waarde op basis van leerwinst een nulmeting op zeer jonge leeftijd (niet later dan het tweede leerjaar, één respondent pleit zelfs voor een meting aan het eind van het kleuteronderwijs) moet worden georganiseerd, omdat anders al te grote verschillen tussen scholen worden opgenomen in de nulmeting.

3.2.2 Wenselijkheid

Het in kaart brengen van de toegevoegde waarde van scholen door een vergelijking te voorzien met de output in andere scholen, kan een krachtig instrument zijn in de interne kwaliteitszorg van scholen. Dat standpunt wordt ingenomen door 17 van de 21 respondenten die menen dat Vlaanderen een leer(winst)monitoringssysteem moet implementeren. Daarbij wordt het ontwikkelingsgericht perspectief op mesoniveau sterk benadrukt, waarbij gegevens over de toegevoegde waarde van een school kunnen worden gebruikt binnen een vorm van zelfevaluatie om kernprocessen op school bij te sturen.

“Omdat ik vind dat dat een onderdeel uitmaakt van onze job om regelmatig te reflecteren (...) over wat gebruiken we, waar zijn we mee bezig, tot waar brengen we onze leerlingen? Daar een soort zelfevaluatie in zien. Dat vind ik heel belangrijk. Zonder dat die ja, zonder dat die moet leiden tot vingerwijzing of weet ik wat, nee gewoon met zijn allen: ‘Ah tiens, onze kinderen bereiken nu telkens eind tweede

leerjaar dat niveau op lezen. Is dat omdat wij dat op die manier doen of op die manier? Zouden we dat nog kunnen verbeteren? Waar zitten die hiaten? Of waar doen we het goed, wat loopt goed?’ Dat vind ik heel belangrijk en dat mis ik toch wel.” (scholen, respondent 7)

Naast bovenstaand beschreven ontwikkelingsgericht perspectief, maakt het in kaart brengen van toegevoegde waarde van scholen ook het verantwoordingsgericht gebruik van de resultaten van een leer(winst)monitoringssysteem op mesoniveau makkelijker. Slechts een kleine minderheid van de respondenten is evenwel gewonnen voor een beoordeling van scholen louter op basis van output-data uit dit leer(winst)monitoringssysteem.

Daarnaast pleiten een aantal respondenten ervoor om ook data in verband met de toegevoegde waarde op klas- en leraarniveau beschikbaar te maken. Leraren kunnen deze gegevens dan gebruiken om de effectiviteit van hun didactische aanpak te evalueren. Toch zijn respondenten ook bedacht op oneigenlijk gebruik van deze data in een verantwoordingsgericht perspectief. Er gaan dan ook een groot aantal stemmen op om deze resultaten niet te gebruiken bij het evalueren van leraren.

Een ander argument luidt dat toegevoegde waarde-berekeningen overbodig zijn omdat scholen zich niet moeten vergelijken met andere scholen. Deze respondenten argumenteren dat het belangrijk is dat scholen binnen het kader van interne kwaliteitszorg rekening houden met het zelfcriterium (hoe zijn de leerprestaties en leerwinst van onze leerlingen nu in verhouding tot vroegere cohorten van leerlingen) en standaardcriterium (in welke mate bereiken onze leerlingen de leerplandoelen of eindtermen en ontwikkelingsdoelen). Slechts in mindere mate dienen scholen zich laten leiden door de prestaties van leerlingen in andere scholen. Dit standpunt wordt ook in zekere mate gedragen door andere respondenten, al meent de meerderheid dat dit niet betekent dat er dan helemaal geen toegevoegde waarde-informatie ter beschikking moet worden gesteld:

“Eigenlijk die norm, die standaard, die je jezelf opstelt: uw visie, uw doel, moet primeren boven: vergelijk u nu eens met andere scholen. Los daarvan kan het wel perspectief bieden van: Hoe doen andere scholen dat? Oh, daar zitten we precies wel boven. Hoe komt dat nu zo? Want anders heb je daar niet echt een zicht op.” (onderwijsverstrekkers, respondent 4)

4 ANDERE AANDACHTSPUNTEN UIT DE DELPHI-STUDIE

Hieronder bespreken we nog enkele punten die prominent aan bod kwamen tijdens de eerste en/of tweede ronde van de Delphi-studie.

4.1 Psychometrisch sterk onderbouwde toetsen

De criteria van betrouwbaarheid en validiteit van de toetsen worden in de eerste plaats door de respondenten uit de categorie ‘inhoudsdeskundigen’ genoemd, maar ook andere respondenten gaan akkoord dat een sterke psychometrische kwaliteit van de toetsen van belang is voor zowel de aanvaarding van het leer(winst)monitoringssysteem door de verschillende stakeholders als voor de effectiviteit ervan. Daarom moet aandacht uitgaan naar een degelijke expertise van de toetsontwikkelaars op verschillende vlakken. Daarom geeft men aan dat naast experts in

toetsontwikkeling, ook experts binnen het getoetste leerdomein, alsook ervaren leraren (met vakexpertise) moeten worden ingezet in het proces van toetsontwikkeling. De meeste respondenten geven er voorkeur aan dat dit gebeurt binnen een onafhankelijk toetsinstituut. Inhoudsdeskundigen waarschuwen er wel voor dat het ontwikkelen van gestandaardiseerde toetsen met een hoge betrouwbaarheid en validiteit een arbeidsintensief proces is, dat dus niet zonder een behoorlijke financiële inspanning kan gerealiseerd worden.

“Wel... mijn enige grote zorg bij dit alles is met dat financiële. Is dat uitgerekend, als we het toelaten dat dat absoluut, dat we het alsjeblift goed doen? En laten we niet een Vlaamse mentaliteit hebben [...] denken dat we dat allemaal wel eventjes tussen de soep en de patatten kunnen doen.” (inhoudsdeskundigen, respondent 6)

Enkele respondenten stellen zich de vraag of het wel mogelijk is om een betrouwbare en valide toets (of leerwinst- en/of toegevoegde waarde-meting) te garanderen, aangezien toetsresultaten mogelijks vertekend worden door de beleving van de leerling (verhoogde stress), door een manier van bevragen die verschilt van wat leerlingen gewoon zijn en door de ‘taligheid’ van de toets (bijvoorbeeld ook om wiskundige taken op te lossen, moet men voldoende taalvaardig zijn om de onderdelen van de vraag correct te begrijpen).

4.2 Toetsformat

Respondenten stellen dat een schriftelijke of digitale toets, ondanks zijn beperkingen (het in kaart brengen van vooral cognitieve kennis in plaats van hogere orde vaardigheden of competenties), het meest geschikte instrument is om op grote schaal toe te passen. Een aantal respondenten brengt ook het gebruik van observaties ter sprake, al zouden deze weinig tastbare informatie opleveren met het oog op reflectie.

“De meeste praktijkvoorbeelden zijn dat, dus ik denk dat er wel mogelijkheden zijn om het anders te maken, maar de vaststelling is toch dat het schriftelijke medium zich het meeste leent tot standaardisering. Dat is gewoon een objectief gegeven denk ik. Het is niet dat er geen andere mogelijkheden zijn maar het is minder evident.” (scholen, respondent 2)

Het gedigitaliseerd aanbieden van gestandaardiseerde toetsen zou een meerwaarde kunnen zijn in de verwerking en rapportage van de resultaten. Toch zijn de respondenten behoed voor een toename in de infrastructurele kosten (aankoop computers). Vier respondenten (twee uit de categorie ‘scholen’, en twee uit de categorie ‘inhoudsdeskundigen’) opperen het idee van een adaptief leer(winst)monitoringssysteem waarbij de moeilijkheidsgraad van de gestandaardiseerde toetsen wordt aangepast aan het niveau van de leerling.

“En als je dat dan niet meer op papier doet, waarom zou je dan nog iedereen dezelfde toets geven? Waarom gaan we dan niet gedifferentieerde, individuele aangepaste toetsen geven?” (inhoudsdeskundigen, respondent 4)

4.3 Vrijwillig of verplicht karakter

De 21 respondenten die menen dat een vorm van leer(winst)monitoring geïmplementeerd kan worden in de Vlaamse onderwijscontext, zijn erg verdeeld over de vraag of de afname van de

gestandaardiseerde toetsen die hiervan deel uitmaken al dan niet verplicht moet worden door de overheid. Het standpunt van de respondenten is sterk gekoppeld aan het vooropgestelde doel: respondenten die een leer(winst)monitoringssysteem eerder in een verantwoordingsgericht perspectief plaatsen, zijn geneigd om aan te geven dat de gestandaardiseerde toetsen verplicht moeten worden. Bij een ontwikkelingsgericht perspectief is dat minder het geval.

In totaal menen acht respondenten dat de gestandaardiseerde toetsen voor alle scholen verplicht moet zijn (zij het niet altijd voor alle leerlingen in elke school, en niet altijd voor alle toetsen die deel uitmaken van het leer(winst)monitoringssysteem). Bijkomend stelt één respondent voor om een toetsenbank te ontwikkelen, waarvan het gebruik verplicht is, maar waarbij de school kan kiezen welke toetsen eruit worden gekozen. Nog een andere respondent stelt voor om het gebruik van het leer(winst)monitoringssysteem niet te verplichten, maar er wel over te waken dat het binnen 15 à 20 jaar in die mate is ingeburgerd dat alle scholen er gebruik van maken.

12 respondenten menen dat het gebruik van een leer(winst)monitoringssysteem facultatief moet blijven. Volgens deze respondenten is het belangrijk dat scholen zelf beslissen of en in welke mate ze gebruik maken van het leer(winst)monitoringssysteem. Verschillende respondenten menen wel dat de overheid er sterk op moet inzetten om het gebruik te stimuleren (mogelijk door er een financiële incentive voor te voorzien), alsook om goede praktijkvoorbeelden te tonen:

“Als ze misschien in eerste instantie niet geneigd zijn, maar als ze merken van: Oh, heel veel scholen maken daar wel gebruik van en die school daar heeft dat gedaan en dat is precies toch op die en die vlakken wel wat bijgestuurd en verbeterd, dat ze dan op die manier sneller geneigd gaan zijn vanuit een betere motivatie, vanuit echt interne motivatie, eerder dan een verplichting die extern wordt opgelegd.”
(beleidsmakers, respondent 1)

4.4 Andere voorwaarden

Zeven respondenten, waarvan vier uit de categorie ‘scholen’, menen dat er bij de implementatie van een leer(winst)monitoringssysteem vanuit de verschillende stakeholders expliciet over moet worden gewaakt dat het Vlaamse onderwijssysteem niet vervalt in een soort van toetscultuur, waarbij toetsen de norm worden.

Eveneens zeven respondenten, waarvan vier uit de categorie ‘onderwijsverstrekkers’, wijzen erop dat de investering van scholen beperkt moet blijven. Het gaat zowel om het idee dat de implementatie van het leer(winst)monitoringssysteem niet tot extra planlast mag leiden, als om het idee dat het leer(winst)monitoringssysteem geen financiële inspanning van scholen mag vragen. Hierbij aansluitend geven twee respondenten dat het de taak is van de overheid om regelmatig te evalueren of het leer(winst)monitoringssysteem de druk op leraren niet nodeloos heeft verhoogd.

Er is geen consensus tussen respondenten over de wijze waarop de resultaten van gestandaardiseerde toetsen aan de leden van de onderwijsinspectie dienen te worden bezorgd. Wel vinden de respondenten het in grote mate belangrijk dat de resultaten worden gekaderd in het licht van interne kwaliteitszorg eerder dan in het kader van de controle van leerprestaties als dusdanig.

DEEL 5 RESULTATEN INTERNATIONALE STUDIE

Zoals besproken in Deel 2 werden vier internationale casestudies uitgevoerd. De selectie van cases in dit onderzoek omvat 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). Uit elk van deze casestudies werden uiteenlopende aandachtspunten gehaald voor de eventuele implementatie en het ontwerp van een leer(winst)monitoringssysteem in Vlaanderen. Een overzicht van deze aandachtspunten wordt besproken in dit deel.

Binnen het bestek van dit boek is er geen ruimte om elk van de verschillende cases uit te werken in de diepgang die ze verdienen. Voor de lezer die geïnteresseerd is in meer informatie over elk van de cases worden de vier casestudies ter beschikking gesteld op de website van de onderzoeksgroep Edubron (www.edubron.be).

1 BESLISSING OVER IMPLEMENTATIE

1.1 Bepalen van een doel en rationale

Uit de case Noorwegen komt de boodschap dat een duidelijke doelgerichtheid een belangrijke voorwaarde is die de ontwikkeling van gestandaardiseerde toetsen of een toetsensysteem moet voorafgaan. In deze rationale dient helder gesteld te worden 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. Zonder deze rationale bestaat het gevaar dat verschillende toetsen onvoldoende op elkaar zijn afgestemd 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 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 casestudie dat men, achteraf gezien, misschien te snel van start is gegaan bij de eerste nationale toetsen. Hierdoor heeft men pas sinds 2014 de psychometrische expertise 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 toegekend. 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 waardoor men niet steeds inzielt 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 casestudies - 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). Doordat toetsen dienen afgestemd te zijn op een duidelijk onderwijskundig probleem of dienen om een mogelijk probleem in de toekomst te vermijden, moet de toetsontwikkeling 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 hierin te geven. 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, is de impact van het evaluatiesysteem er momenteel 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.

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 casestudies verschillende voorbeelden waarin verantwoordingsgerichte doelen toegevoegd worden aan toetsen die werden vormgegeven 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)monitoringssystemen 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 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 casestudies ook concluderen dat men zich bij de ontwikkeling van gestandaardiseerde toetsen best richt 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. Toch is het een moeilijk evenwicht voor toetsontwikkelaars om gestandaardiseerde toetsen te vervaardigen die informatie vergaren die zinvol is op zowel het mesoniveau als op het microniveau. De toetsen die leraren nieuwe informatie kunnen aanleveren 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.

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 een hogere (gepercipieerde) inzet van een toets tot meer onwenselijke neveneffecten kan leiden. 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 casestudies 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 Engeland, 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 casestudie 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 bevindingen 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.

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.

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 alle 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)

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 zowel scholen als ouders te wijzen op het voordeel van betrouwbare informatie inzake de leeruitkomsten van leerlingen. Daarbij is het tevens noodzakelijk om sterk in te zetten op een dialoog in scholen (of met scholen) op basis van de resultaten van de gestandaardiseerde toetsen.

Australische respondenten zien, zeker bij aanvang, weinig heil in verplichte toetsen: 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 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 leer(winst)monitoringssystemen 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. In de aanvangsfase 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 langzaamaan 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 is het aangewezen 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.

2 EIGENSCHAPPEN VAN HET LEER(WINST)MONITORINGSSYSTEEM

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 het moment van toetsafname en de 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 5.2.2) en dat in de rapportering automatische linken kunnen worden voorzien naar delen van het curriculum (en eventueel naar concrete adviesmogelijkheden) (zie 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)

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 aanleveren, 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.

2.3 Al dan niet curriculum-gerelateerd toetsen

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 aan over meer generieke capaciteiten van leerlingen 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. De ervaring bij CEM leert echter 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 curriculumwijziging moeten worden nagekeken 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 naar curriculum-gerelateerde toetsen wordt sinds enkele jaren echter ook door CEM meer geïnvesteerd in de ontwikkeling hiervan. 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).

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 leesvaardigheden. 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 wordt bij CEM een 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 Engeland (bij CEM).

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 duidelijkere richtlijnen opstellen bij de toetsontwikkeling opdat de verschillende toetsen meer op elkaar zijn afgestemd. Op die manier kan men ook hier evolueren naar een echte toetsenbatterij.

2.6 Sterke psychometrische onderbouwing

In elke casestudie wordt er uitvoerig gedocumenteerd welke procedures er worden gevolgd om de betrouwbaarheid van de toetsen te garanderen. Zo 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 casestudies.

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 definitief worden vastgelegd nadat de toetsresultaten zijn geanalyseerd.

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.

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 beschreven procedures die worden gehanteerd bij toetsen met een hoge inzet worden ervaren als een cruciaal element binnen de wijdverspreide 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 beschreven. 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.

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 casestudie dat bij zowat elke toets een samenwerking tot stand wordt gebracht tussen inhoudelijke experts, toetsexperts 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 zowel CEM in Engeland als 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 Inspira Assessment.

2.9 Gebruik maken van bestaande (internationale) toetsen

Verskillende 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.

3 LEERWINST

Verskillende systemen stellen zich voor als leerwinstmonitoringssystemen. 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 casestudies van Australië, Engeland en Nederland.

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, hetgeen door verschillende respondenten als een 'gemiste kans' wordt omschreven. Zoals eerder gesteld, is 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. Indien de toetsen ook betrouwbare informatie genereren over het leerproces en de leerwinst van leerlingen, zou de meerwaarde van dergelijke toetsen veel groter zijn. 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.

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 alsook van de gebruiksvriendelijkheid van de rapportage.

Leerwinst zou voornamelijk zinvolle informatie voor scholen opleveren 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.

4 TOEGEVOEGDE WAARDE

Geen enkele casestudie resulteerde in een voorbeeld 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 vergelijkingen gemaakt tussen de behaalde leerwinst in scholen. Dit zorgt voor een meer genuanceerde voorstelling 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 het meten van toegevoegde waarde van scholen 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 casestudies) 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 RAPPORTERING

5.1 Gedetailleerde en interactieve rapportering

Een brede en gedetailleerde rapportering is belangrijk. De rapportering moet toelaten om informatie te geven over zowel de prestatie op schoolniveau als de prestaties van individuele leerlingen en klassen in vele subdomeinen van de toets. Daarnaast is het belangrijk dat de gegevens snel ter beschikking staan 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 eveneens concluderen dat dit principe niet makkelijk te realiseren is. Dit blijkt onder meer uit de eigen evaluatie van het pilootproject in Nederland, dat - ondanks vele bijstellingen op basis van eerdere feedback - nog steeds niet heeft geleid tot een overtuigend resultaat.

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 dit af. Wel is iedereen in de Australische casestudie 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).

6 ONDERSTEUNING VAN SCHOLEN

Een ondersteuningsaanbod 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 casestudies als een belangrijke succesvoorwaarde beschouwd. Dat aanbod gebeurt best op maat van de school, zoals uit de vermelde ‘kritische kantelmomenten’ in de Nederlandse casestudie 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 verspreid door 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)

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.

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 tevens 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 voorzien beide bezochte scholen op elke woensdagnamiddag '*professional learning time*', begeleid door interne '*professional leading teams*' die bestaan 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 bevraagde 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 genoemd 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 managementteam 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 casestudie 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 casestudie 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 geschreven 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)

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 boek 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.

DEEL 6 BOUWSTENEN VAN DE SCENARIO'S

In dit deel worden acht werkscenario's voorgesteld voor het Vlaamse onderwijsbeleid. Deze werkscenario's zijn voornamelijk opgebouwd op de resultaten van de eerste en tweede ronde van de Delphi-studie (zie Deel 4, maar houden ook rekening met de resultaten uit Deel 3 (de literatuurstudie) en Deel 5 (de internationale studie). Deze acht werkscenario's zijn geen definitieve outputs van dit rapport. De acht werkscenario's worden immers nadien - in de derde ronde van de Delphi-studie - aan de 24 respondenten voorgelegd met de vraag elk van de werkscenario's te beoordelen inzake wenselijkheid en haalbaarheid. Over de resultaten van deze derde ronde van de Delphi-studie wordt gerapporteerd in Deel 7.

In dit deel geven we eerst een overzicht van het ontwikkelingsproces van de werkscenario's en worden keuzes die hierin werden genomen toegelicht. Daarna wordt dieper ingegaan op de sleutelfactoren die werden gebruikt bij de ontwikkeling van de werkscenario's. Daarna geven we een integraal overzicht van de acht werkscenario's die werden gebruikt tijdens de derde ronde van de Delphi-studie.

1 ONTWIKKELING VAN DE ACHT SCENARIO'S AAN DE HAND VAN BOUWSTENEN

Uit de eerste en tweede ronde van de Delphi-studie blijkt enerzijds de relatief grote consensus tussen de 24 respondenten over de te verwachte positieve effecten en onwenselijke neveneffecten, alsook over de bouwstenen die de effectiviteit van een leer(winst)monitoringssysteem in gunstige of ongunstige zin kunnen beïnvloeden; anderzijds blijkt dat toch zeer uiteenlopende scenario's worden uitgewerkt door de respondenten door verschillen in het gewicht dat aan de verschillende effecten en neveneffecten wordt toegekend. Deze verscheidenheid aan invalshoeken betekent dat, ondanks (of net omwille van) de rijkheid aan data die binnen de uitklaring van de verschillende standpunten werd bereikt, moeilijk eenduidige conclusies kunnen worden getrokken. Dit impliceert dat bij het opmaken van werkscenario's ten behoeve van de derde ronde van de Delphi-studie een breed perspectief zal moeten worden gehanteerd waarbij uiteenlopende doelen aan bod komen.

Globaal kan worden gesteld dat een ruime meerderheid van de respondenten in de eerste twee rondes van de Delphi-studie (met name 21 van de 24 respondenten) menen dat het huidige Vlaamse onderwijssysteem baat zou hebben bij de invoering van een leer(winst)monitoringssysteem. De aandacht gaat daarbij voornamelijk uit naar ontwikkelingsgerichte doelen op micro- en mesoniveau, maar ook verantwoordingsgerichte doelen, alsook doelen op macroniveau worden door verschillende respondenten genoemd.

Verder kunnen op basis van de resultaten in Deel 3, Deel 4 en Deel 5 een groot aantal bouwstenen worden geïdentificeerd. De werkscenario's dienen op differentiële wijze variatie te voorzien in de invulling van deze bouwstenen: bij sommige bouwstenen is er op basis van de voorgaande deelstudies een duidelijke voorkeur voor één bepaalde invulling (bijvoorbeeld de relatie met het curriculum), terwijl dezelfde deelstudies in verband met andere bouwstenen weinig eenduidige

resultaten hebben opgeleverd. Elk van deze bouwstenen, alsook de variatie die hierin werd aangebracht doorheen de scenario's wordt in de volgende sectie besproken.

2 OVERZICHT VAN DE BOUWSTENEN

In deze sectie bespreken we eerst de 'primaire bouwstenen': dit zijn de bouwstenen die meer bepalend zijn dan andere bouwstenen, omdat de invulling ervan ook consequenties heeft voor de invulling van de andere (secundaire) bouwstenen. Uit het internationale onderzoeksluik werd geconcludeerd dat het van groot belang is om eerst het doel van het leer(winst)monitoringssysteem te bepalen. Dit doel (enerzijds gerichtheid op micro-, meso-, of macroniveau; anderzijds het ontwikkelingsgericht of verantwoordingsgericht perspectief) wordt dan ook als een primaire bouwsteen beschouwd.

Daarnaast blijkt ook het verplichtend (dan wel facultatief) karakter van het leer(winst)-monitoringssysteem een factor die cruciaal is bij de invulling van andere bouwstenen. Daarom wordt ook de vraag of de toetsen binnen een bepaald scenario door alle leerlingen (van een bepaalde cohorte) in elke school moeten worden afgelegd, als een primaire bouwsteen beschouwd.

2.1 Primaire bouwstenen

Uit de resultaten van de Delphi-studie blijkt dat respondenten over het algemeen vragende partij zijn voor toetsresultaten op zowel microniveau (leerling- en leraarniveau) als mesoniveau (schoolniveau). Dit is overeenkomstig met de resultaten van de internationale studie, waar respondenten ook voornamelijk heil zien in instrumenten die een rechtstreekse impact hebben op leerlingen, leraren en scholen. Ook achten enkele respondenten het wenselijk dat een leer(winst)monitoringssysteem informatie genereert die kan worden aangewend op macroniveau. Onder meer uit de internationale studie kon geconcludeerd worden dat doelen op micro-, meso- en macroniveau moeilijk verenigbaar zijn in hetzelfde leer(winst)monitoringssysteem. Om tegemoet te komen aan deze bevindingen wordt dan ook geopteerd om de doelstellingen in de scenario's zo zuiver mogelijk te richten op ofwel het micro-, ofwel het meso-, ofwel het macroniveau - met de nadruk op de eerste twee niveaus.

Een tweede spanningsveld tussen de opinies van de verschillende respondenten in de eerste twee ronden van de Delphi-studie is dat tussen ontwikkelings- en verantwoordingsgerichte evaluatie. Hoewel de primaire focus bij de meeste respondenten op het ontwikkelingsgerichte aspect lag, werden door verschillende respondenten ook micro- en op mesoniveau concrete verantwoordingsgerichte perspectieven toegevoegd. Zoals blijkt uit de bespreking in Deel 3, is een aantal respondenten van mening dat een leer(winst)-monitoringssysteem moet worden ingezet om de kwaliteit van het geboden onderwijs te onderzoeken en ze dus een eerder controlerende functie toegewezen krijgen waarbij de school zich tegenover diverse stakeholders (overheid, maatschappij, ouders en leerlingen) dient te verantwoorden over de geleverde onderwijskwaliteit. Verantwoordingsgericht gebruik op microniveau heeft dan weer tot doel dat leerlingen zich door middel van gestandaardiseerde toetsen moeten 'verantwoorden' voor het verkrijgen van een getuigschrift, diploma of toegang tot een hoger leerjaar of vorm van vervolgonderwijs. De meerderheid van de respondenten legt evenwel de nadruk op ontwikkelingsgerichte doeleinden. Zij

zien het leer(winst)monitoringssysteem als een middel om de interne kwaliteitszorgprocessen in de scholen te stimuleren en vinden het absoluut niet gewenst om de resultaten publiek te maken. Op microniveau wordt dan de focus gelegd op onder meer het versterken van de leerprestaties en de individuele begeleiding van leerlingen. Ook in het internationale luik en de literatuurstudie blijkt het continuüm tussen ontwikkeling en verantwoording in sterke mate de toetsinzet en het toetsformat van het leer(winst)monitoringssysteem te bepalen. Aangezien uit beide onderzoeksluiken ook het advies kan worden afgeleid om bij het ontwerp van een leer(winst)monitoringssysteem van een éénduidig doel uit te gaan, werd bij het scenario-ontwerp in de mate van het mogelijke vermeden om ontwikkelingsgerichte en verantwoordingsgerichte doeleinden toe te kennen aan de toetsen binnen één scenario. Enkel bij scenario 5 wordt een in origine ontwikkelingsgericht leer(winst)monitoringssysteem tevens in een verantwoordingsgericht perspectief geplaatst.

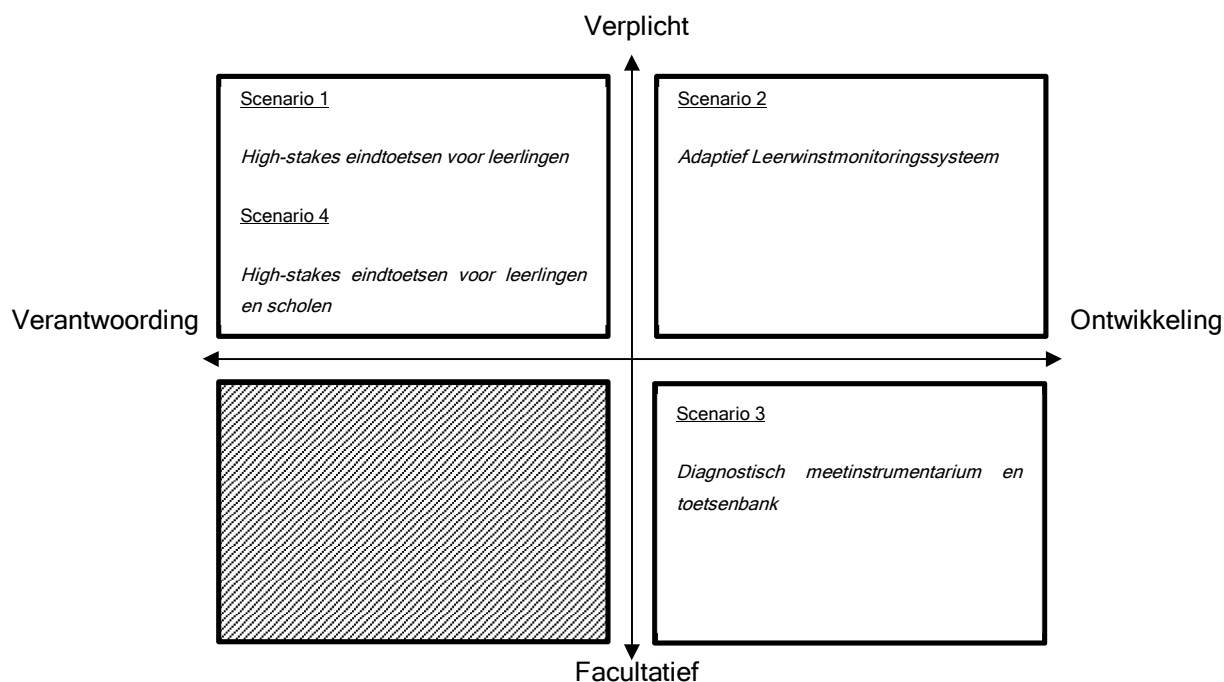
Tot slot blijkt ook de mate van vrijblijvendheid in deelname aan de toetsen een twistpunt te zijn tussen de diverse respondenten binnen dit onderzoek. Een aantal respondenten is van mening dat een verplichte deelname aan gestandaardiseerde toetsen de autonomie van scholen aantast en het principe van *teaching to the test* in de hand werkt, terwijl andere respondenten positief staan tegenover een centraal systeem waaraan elke school verplicht deelneemt. Het internationale luik wijst bovendien uit dat beide systemen ook naast elkaar kunnen bestaan.

Wanneer het doel van de toetsen (verantwoording versus ontwikkeling) op een horizontale as gezet wordt en de mate van vrijblijvendheid (verplicht versus facultatief) op een verticale as, dan ontstaat er per niveau (micro-, meso- en macroniveau) een figuur met vier kwadranten. Elk van die kwadranten geeft dan een mogelijke combinatie weer tussen het doel van de toetsen en de mate van vrijblijvendheid. In sommige gevallen wijst de praktijk uit dat de ontstane situatie geen wenselijke situatie is en zal het kwadrant leeg zijn. Dit is concreet het geval bij verantwoordingsgerichte leer(winst)monitoringsystemen met een vrijblijvend karakter: wanneer toetsen een summatief of controlerend doel hebben, dienen ze door leerlingen in alle scholen verplicht te worden afgelegd. In totaal werden acht afzonderlijke scenario's overheen de verschillende niveaus ontwikkeld.

2.1.1 Microniveau

Zoals blijkt uit Figuur 10 werden er op microniveau in totaal vier werkscenario's ontwikkeld.

In het eerste scenario 'high-stakes eindtoetsen voor leerlingen' worden gestandaardiseerde toetsen ontwikkeld die als eindtoets worden ingezet bij elk van de onderwijsniveaus in het leerplichtonderwijs. De resultaten van deze verplichte eindtoetsen beïnvloeden in progressief opgebouwde mate of de leerling een getuigschrift basisonderwijs, een oriënteringsattest van de eerste of tweede graad secundair onderwijs of een diploma secundair onderwijs behaalt. Het verantwoordingsgerichte perspectief wordt dus expliciet op het microniveau van leerlingen gericht: het zijn leerlingen die door middel van de gestandaardiseerde toetsen dienen te bewijzen dat ze een getuigschrift, oriënteringsattest of diploma verdienen. De opeenvolging van deze eindtoetsen vormt tevens de basis voor een leerwinstmonitoringsysteem.



Figuur 10 - Overzicht van de verschillende scenario's op microniveau

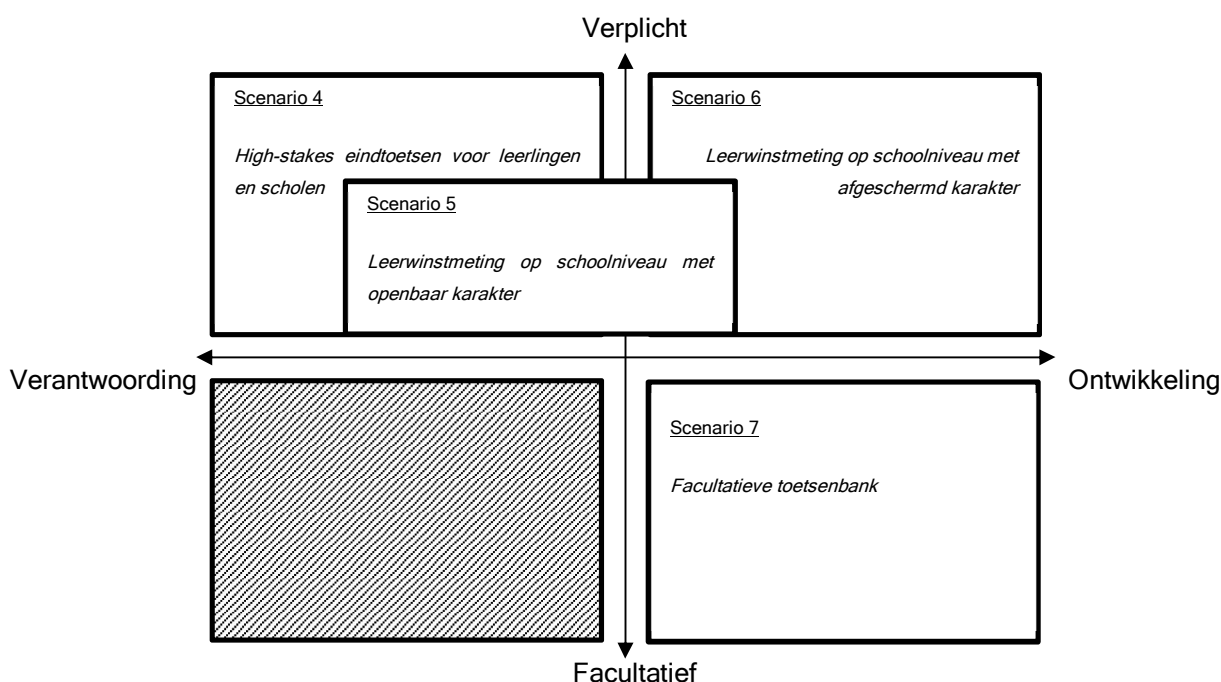
Het tweede scenario met de titel 'adaptief leerwinstmonitoringssysteem' bevindt zich in het kwadrant ontwikkelingsgericht/verplicht. Binnen dit leerwinstmonitoringssysteem worden een reeks op elkaar volgende digitale adaptieve toetsen aangeboden. Het hoofddoel ervan is de leraar inzicht te geven in de ontwikkeling en in de leerwinst van de leerlingen inzake enkele essentieel geachte vaardigheden. Hiertoe moeten alle leerlingen minimaal één maal per jaar een toets per vaardigheid afleggen. De resultaten van deze toetsen zijn in geen enkel geval bindend voor attestering of diplomering, maar kunnen gebruikt worden als informatiebron bij bijvoorbeeld differentiatie of studieloopbaanbegeleiding.

Het derde scenario wordt een 'diagnostisch meetinstrumentarium en toetsenbank' genoemd. In dit scenario in het kwadrant 'ontwikkelingsgericht' en 'facultatief' wordt een leer(winst)-monitoringssysteem als een toetsenbank met gestandaardiseerde toetsen ontwikkeld. Scholen en leraren kunnen uit deze toetsenbank gestandaardiseerde toetsen gebruiken om na te gaan welke aspecten van een bepaald leerdomein door de leerling reeds beheerst worden en waar die leerling nog remediëring kan gebruiken. De toetsen zijn gericht op zeer specifieke leerstofonderdelen.

Het vierde en laatste scenario op microniveau ('high-stakes eindtoetsen voor leerlingen en scholen') situeert zich net als het eerste scenario in het verantwoordingsgericht en verplicht kwadrant. Het is het enige scenario dat zich over twee niveaus verspreidt, met name het micro- en mesoniveau. Binnen dit scenario worden gestandaardiseerde toetsen ingezet met een tweeledig hoofddoel: enerzijds het in kaart brengen van de geaggregeerde leerprestaties van leerlingen in scholen; anderzijds het objectiever maken van de beslissing over het uitreiken van getuigschriften en diploma's aan de leerlingen. Het eerste doel is verantwoordingsgericht op mesoniveau, het tweede doel is eveneens verantwoordingsgericht maar dan op microniveau.

2.1.2 Mesoniveau

Ook op mesoniveau kunnen vier werkscenario's onderscheiden worden. Zoals eerder gesteld (zie 2.1), bevindt het vierde scenario 'High-stakes eindtoetsen voor leerlingen en scholen' zich ook hier. Figuur 11 geeft een visueel overzicht van de scenario's 4 tot en met 7.



Figuur 11 - Overzicht van de verschillende scenario's op mesoniveau

Het vijfde scenario 'leerwinstmeting op schoolniveau met openbaar karakter' situeert zich tussen twee kwadranten. In dit scenario dragen de resultaten van de ontwikkelde gestandaardiseerde toetsen in het leerwinstmonitoringssysteem bij tot de kwaliteitszorg binnen scholen, waardoor dit scenario haar plaats binnen het ontwikkelingsgericht kwadrant verdient. Hoewel deze verplichte en schriftelijke toetsen een formatief karakter hebben (en als low-stakes gezien worden voor leerlingen), worden ze door de openbaarheid van gegevens inzake toegevoegde waarde op schoolniveau als high-stakes beschouwd, waardoor dit scenario ook thuishoort in het verantwoordingsgericht kwadrant.

In het zesde scenario 'leerwinstmeting op schoolniveau met afgeschermd karakter' dragen de resultaten van het leer(winst)monitoringssysteem bij tot de kwaliteitszorg binnen scholen. De digitale toetsen in dit monitoringssysteem zijn verplicht voor alle scholen, maar hebben een formatief karakter en de resultaten worden niet openbaar gemaakt.

Het zevende scenario bevindt zich in het kwadrant 'facultatief' en 'ontwikkelingsgericht'. In dit scenario wordt leer(winst)monitoringssysteem als een toetsenbank ontwikkeld waaraan scholen toetsen kunnen ontlelen die hen helpen bij het in kaart brengen van de mate waarin het onderwijs in de school de leerlingen leidt tot het behalen van de eindtermen, aangevuld met andere vragenlijsten. Het gebruik van deze toetsen past binnen de interne kwaliteitszorg op schoolniveau en is vrijblijvend. De resultaten worden enkel aan de school ter beschikking gesteld. Het verschilt van scenario 3 door de gerichtheid van de toetsen op het mesoniveau: ze zijn in de eerste plaats gericht op het genereren van informatie op schoolniveau.

2.1.3 Macroniveau

Uit de eerste en tweede ronde van de Delphi-studie bleek bij de bevroagde respondenten weinig expliciete vraag naar de ontwikkeling van een leer(winst)monitoringssysteem dat gericht is op het genereren van informatie ten behoeve van het macroniveau. Aangezien in de huidige situatie de initiatieven van de overheid inzake ontwikkeling en afname van gestandaardiseerde toetsen voornamelijk gericht zijn op het macroniveau (met name door de peilingsproeven), werd de vraag van respondenten om zoveel mogelijk binnen het huidig kader te blijven gecatalogiseerd als een scenario op het macroniveau. Dit achtste - en tevens laatste - scenario met de naam 'de huidige trend wordt verdergezet' wil evenwel niet zeggen dat niets hoeft te veranderen. In dit scenario wordt immers sterker ingezet op het optimaliseren van de randvoorwaarden van het bestaande monitoringssysteem, zoals het valideren van bepaalde toetsen, om zo de effectiviteit van deze toetsen in het licht van onderwijskwaliteit te versterken. Daarnaast besteedt het scenario aandacht aan de ondersteuning en verdere professionalisering van scholen en lerarenteams in het kader van datageletterdheid.

2.2 Secundaire bouwstenen

Een scenario omvat uiteraard meer dan een omschrijving van het doel van het beoogde leer(winst)monitoringssysteem en de bepaling of de gestandaardiseerde toetsen al dan niet een verplicht karakter hebben. Ook inzake verschillende andere bouwstenen diende een variatie te worden beoogd.

Zo is er de bouwsteen 'de publieke beschikbaarheid van data'. De invulling van deze bouwsteen wordt al in zekere mate bepaald door het doel van het werkscenario (met name het ontwikkelings- dan wel verantwoordingsgericht karakter). Geen enkele van de drie voorgaande deelstudies gaf een eenduidig advies hieromtrent, al halen we voornamelijk uit de internationale studie de aanbeveling om openbaarheid te vermijden. Wel kan duidelijk geconcludeerd worden dat eventuele openbaarheid enkel op het schoolniveau kan. Concreet wordt publieke beschikbaarheid van data in twee van de vier werkscenario's op mesoniveau ingebracht.

De verschillende deelstudies pleiten elk op een eigen manier voor de inclusie van leerwinst bij de ontwikkeling van een monitoringssysteem, maar de internationale studie en de literatuurstudie wijzen uit dat dit niet in alle omstandigheden evident is. In vier werkscenario's wordt de leerwinst van leerlingen berekend waarbij verschillende conceptualiseringen van het begrip 'leerwinst' worden gehanteerd. In zes scenario's wordt de toegevoegde waarde van scholen in kaart gebracht, al beperken we ons in de meeste gevallen (op aangeven van de internationale studie en de eerste ronde van de Delphi-studie) bijna steeds tot een relatief eenvoudige vergelijking tussen scholen inzake leerresultaten of leerwinst.

Zowel de literatuurstudie als de eerste twee ronden van de Delphi-studie pleiten voor gestandaardiseerde toetsen die gebaseerd zijn op het curriculum. Enkel in de internationale studie zien we ook andere voorbeelden. We kiezen er daarom voor om in één scenario af te wijken van de eindtermen en ontwikkelingsdoelen en baseren ons daarbij op het niet-curriculumgerelateerde model dat we zagen in Noorwegen. We volgen ook in bijna elk scenario de aanbeveling om zo breed mogelijk te evalueren, zeker waar het high-stakes toetsen betreft.

Tabel 8 - Overzicht van de verschillende werkscenario's

Scenario's	1 - High-stakes eindtoetsen voor leerlingen	2 - Adaptief leerwinst- monitoringsysteem	3 - Diagnostisch meetinstrumentarium & toetsenbank	4 - High-stakes eindtoetsen voor leerlingen en scholen	5 - Leerwinstmeting op schoolniveau met openbaar karakter	6 - Leerwinstmeting op schoolniveau met afgeschermd karakter	7 - Facultatieve toetsenbank
Verplichting	Verplicht	Verplicht	Facultatief	Verplicht	Verplicht	Verplicht	Facultatief
Karakter	Verantwoordingsgericht	Ontwikkelingsgericht	Ontwikkelingsgericht	Verantwoordingsgericht	Ontwikkelingsgericht	Ontwikkelingsgericht	Ontwikkelingsgericht
Beschikbaarheid resultaten	Niet openbaar	Niet openbaar	Niet openbaar	Openbaar op schoolniveau	Openbaar op schoolniveau	Niet openbaar	Niet openbaar
Leerwinst	Leerwinst residual gain	Leerwinst als groei	Geen leerwinst	Geen leerwinst	Leerwinst Categorical model	Leerwinst Residual gain	Geen leerwinst
Toegevoegde waarde	Toegevoegde waarde residual gain model	Toegevoegde waarde Statusmodel facultatief	Geen toegevoegde waarde	Toegevoegde waarde Residual gain model	Toegevoegde waarde Gain score model	Toegevoegde waarde Statusmodel	Toegevoegde waarde Statusmodel
Doel microniveau	Invoed op oriëntering en attestering	Leerwinstmonitoring	Diagnostische informatie	Invoed op oriëntering en attestering	Beperkt - Leerwinstmonitoring	Beperkt - Leerwinstmonitoring	Beperkt
Doel mesoniveau	Verantwoording (inspectie en openbaarheid) Interne kwaliteitszorg	Interne kwaliteitszorg	Beperkt (Informatiebron CLB)	Verantwoording (inspectie en openbaarheid) Interne kwaliteitszorg	Interne kwaliteitszorg Verantwoording door openbaarheid	Interne kwaliteitszorg	Interne kwaliteitszorg
Doel macroniveau	Evaluatie van het onderwijssysteem	Evaluatie van het onderwijssysteem	/	Evaluatie van het onderwijssysteem	Evaluatie van het onderwijssysteem	/	/
Toetsvorm	Schriftelijk	Digitaal - adaptief - IRT	Digitaal	Schriftelijk	Schriftelijk - IRT	Digitaal - IRT	Digitaal + observatie
Curriculum	Eindtermen	'Cruciale' vaardigheden	Leerplandoelen, eindtermen Competenties	Eindtermen	Eindtermen	Eindtermen	Eindtermen en andere aspecten
Inhoud	Verschillende leerdomeinen	Rekenen, lezen, schrijven, wetenschappen	Verschillende (incl. socio-emotionele) leerdomeinen	Verschillende leerdomeinen	Verschillende leerdomeinen	Verschillende (incl. socio-emotionele) leerdomeinen	Verschillende (incl. socio-emotionele) leerdomeinen
Ontwikkeling	Jaarlijks hernieuwd	Jaarlijkse controle van items en gedeeltelijke vernieuwing	Jaarlijkse controle van items en gedeeltelijke vernieuwing	Jaarlijks hernieuwd	Jaarlijks hernieuwd	Jaarlijkse controle van items en gedeeltelijke vernieuwing	Jaarlijkse controle van items en gedeeltelijke vernieuwing
Correctie	Extern	Automatische correctie	Automatische correctie	Extern	Extern (automatisch)	Automatische correctie	Automatische correctie

De bouwstenen 'correctie' en 'ontwikkeling' van de gestandaardiseerde toetsen hangen sterk samen met het vooropgestelde doel: uit de internationale studie wordt geconcludeerd dat verantwoordingsgerichte toetsen een jaarlijkse hernieuwing vergen en dat ze tevens gecorrigeerd dienen te worden door externe evaluatoren.

3 SCENARIO'S

Om het inzicht in de verschillende scenario's te vergroten, beschrijft deze sectie de verschillende bouwstenen binnen elk van de acht ontwikkelde scenario's. De scenario's zijn hierin opgenomen op dezelfde manier als de respondenten van de Delphi-studie ze te zien kregen ter voorbereiding van de derde ronde van de Delphi-studie. De beschrijving van de scenario's geeft het opzet van de toetsen in zowel het basisonderwijs als het secundair onderwijs weer. Vervolgens worden de diverse gebruiksmogelijkheden op micro-, meso- en/of macroniveau beschreven. Er gaat eveneens aandacht uit naar de ontwikkeling van de toetsen alsook naar de correctie en rapportage van de resultaten (beschikbaarheid van de gegevens). Daarnaast wordt ook aangegeven op welke manier de leerwinstmeting en toegevoegde waarde-meting binnen het scenario zal worden uitgevoerd. Elk scenario wordt tot slot afgesloten met een inschatting van de kostprijs.

Voor informatie met betrekking tot de verwachte positieve en onwenselijke effecten wordt verwezen naar de literatuurstudie (zie Deel 3), de resultaten van de eerste twee ronden van de Delphi-studie (zie Deel 4) en van de internationale studie (zie Deel 5).

In een volgend deel van dit boek (zie Deel 7) worden de standpunten inzake de wenselijkheid en de haalbaarheid van de verschillende respondenten ten aanzien van de bovenstaande scenario's weergegeven.

3.1 Scenario 1 'High-stakes eindtoetsen voor leerlingen'

Microniveau - verplicht - verantwoordingsgericht

3.1.1 Basisonderwijs

Leerlingen in het zesde leerjaar nemen verplicht¹ deel aan een gestandaardiseerde eindtoets die zich uitstrekt over meerdere leerdomeinen² en die plaatsvindt op het einde van het schooljaar. Deze eindtoets is schriftelijk, en wordt van alle leerlingen op hetzelfde moment afgenomen. De eindtoets heeft een summatief karakter en gaat na in welke mate de leerlingen de eindtermen hebben behaald.

¹ Alternatieve piste 1: In een pilootfase kunnen scholen facultatief gebruik maken van de toetsen. Het moet evenwel een streefdoel zijn om de toetsen binnen de drie jaar verplicht in te zetten door alle scholen. Daarvoor wordt er sterk geïnvesteerd in het promoten van de visie waarin gestandaardiseerde toetsen zijn ingebed, om een ruime gedragenheid bij leraren te realiseren.

² Alternatieve piste 2: de eindtoetsen kunnen zich ook beperken tot een kleiner aantal leerdomeinen (bijvoorbeeld minimaal wiskunde en Nederlands) of nog uitgebreid worden met de leerdomeinen 'Lichamelijke Opvoeding' en 'Muzische Vorming', of de leergebied-overschrijdende eindtermen ('leren leren', 'ICT' en 'sociale vaardigheden'). Bij deze uitbreiding kunnen louter schriftelijke toetsen evenwel niet volstaan. Een inperking tot een kleiner aantal leerdomeinen betekent uiteraard dat de kostprijs wordt gedrukt. Een uitbreiding betekent een exponentiële verhoging van de kostprijs, door de veel grotere arbeidsintensiviteit van afname van niet-schriftelijke toetsen. Een mogelijke piste is dat het toetsenarsenaal gradueel wordt uitgebreid.

De resultaten van deze eindtoets worden eveneens beschouwd als het moment van de 'nulmeting' om verdere studieprogressie in kaart te brengen.

Er is een eindtoets in elk van de volgende leerdomeinen: wiskunde, Nederlands, wetenschap en techniek, Frans, mens en maatschappij. De toetsen zijn zo breed mogelijk: ze trachten naast kennis ook vaardigheden van leerlingen in kaart te brengen.

De klassenraad beslist aan het einde van het basisonderwijs, maar voor de afname van de eindtoets of de leerling het schooljaar met vrucht doorlopen heeft en recht heeft op het getuigschrift basisonderwijs. De resultaten van de gestandaardiseerde eindtoets zijn enkel bindend wanneer een leerling slaagt voor de eindtoets, terwijl het getuigschrift basisonderwijs eerder niet werd toegekend door de school. Een positieve bijstelling van de beslissing van de klassenraad is dus het enige bindende effect van de eindtoets.

3.1.2 Secundair onderwijs

In het secundair onderwijs worden oriënteringsattesten uitgereikt op het einde van elk schooljaar. In het tweede, vierde en laatste jaar van het secundair onderwijs wordt een eindtoets ingelast die grote overeenkomsten vertoont met de eindtoets in het zesde leerjaar basisonderwijs. Deze eindtoetsen in uiteenlopende leerdomeinen meten de mate waarin de leerlingen de eindtermen van de eerste, tweede en derde graad secundair onderwijs hebben behaald. Idealiter vervangen deze eindtoetsen op het einde van de bovengenoemde schooljaren de examenperiode in juni.

In het secundair onderwijs wordt de bindende kracht van de resultaten op de eindtoetsen incrementeel opgebouwd. Op basis van een transparante formule wordt aan elke leerling een eindcijfer toebedeeld. De toetsresultaten van de gestandaardiseerde eindtoets tellen voor 15% van het jaargemiddelde binnen een bepaald leerdomein mee in het tweede jaar secundair onderwijs en 30% in het vierde jaar secundair onderwijs. De overblijvende 85%/70% van het eindcijfer wordt ingevuld door de school op basis van door de school verzamelde gegevens. In deze school-evaluatie kan bijvoorbeeld sterker worden ingezet op die vaardigheden en competenties die moeilijker door schriftelijke gestandaardiseerde toetsen kunnen worden gemeten. De eindcijfers in het secundair onderwijs hebben een bindend effect: het al dan niet slagen van de leerlingen, alsook oriënteringsattesten, worden rechtstreeks gekoppeld aan de eindcijfers van de leerlingen in de verschillende leerdomeinen.

In het laatste jaar tellen de resultaten van de eindtoets mee voor 50% van het eindcijfer per leerdomein³. Deze eindtoetsen worden ingelegd op het einde van het laatste jaar secundair onderwijs (bij voorkeur mei) en vervangen de examenperiode in juni. Of een leerling het diploma van secundair onderwijs krijgt, wordt rechtstreeks gekoppeld aan de eindcijfers van de leerlingen in de verschillende leerdomeinen. Er wordt evenwel een correctie toegepast op het cijfer dat leerlingen van de eigen school krijgen om de vergelijkbaarheid overheen scholen te garanderen. Dit kan door een factor toe te voegen aan het individuele cijfer van elke leerling op de school-evaluatie, op basis

³ Alternatieve piste 3: het percentage dat aan de eindtoets wordt gekoppeld, kan uiteraard variëren. Er wordt evenwel aangeraden dit percentage voor de eindtoetsen aan het einde van het secundair onderwijs niet groter dan 50%, en niet kleiner dan 25% te maken. Hoe kleiner het percentage, hoe geringer de kans dat enkele neveneffecten zich voordoen (bijvoorbeeld emotionele impact voor leerlingen).

van de gemiddeldes en spreiding van de toetsresultaten van leerlingen in de school op de eindtoets. Deze individuele factor wordt onafhankelijk per leerdomein berekend.

De eindtoetsen zijn in heel Vlaanderen hetzelfde voor alle leerlingen in dezelfde studierichting van hetzelfde leerjaar, met de volgende leerdomeinen die zowel in ASO, KSO als TSO aan bod komen: wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne vreemde talen (Engels en Frans). In het BSO wordt PAV (project algemene vakken) geëvalueerd. Bijkomende leerdomeinen die geëvalueerd worden, zijn afhankelijk van de gevolgde studierichting⁴. De toetsen meten in de mate van het mogelijke zowel de kennis als de vaardigheden van leerlingen.

In elk leerdomein waarin een leerling twee opeenvolgende toetsen afneemt, wordt ook de leerwinst van leerlingen in kaart gebracht. De behaalde leerwinst heeft evenwel geen invloed over de beslissingen in verband met oriënteringsattesten of diploma's.

3.1.3 Gebruik op microniveau

Het primaire doel van deze high-stakes eindtoetsen ligt op microniveau: het leren van leerlingen maximaliseren en garanderen dat alle leerlingen op een gelijke manier worden beoordeeld in het licht van hun getuigschrift/diploma. Daarnaast kunnen de leraren hun eigen functioneren evalueren aan de hand van de leerprestaties (of leerwinst in het secundair onderwijs) van hun leerlingen. De school-evaluatie en het resultaat op de eindtoets, resulteert in een eindcijfer voor elke leerling, per leerdomein.

3.1.4 Gebruik op mesoniveau

Hoewel het primaire doel zich op microniveau situeert, kunnen de scholen de resultaten gebruiken binnen hun interne kwaliteitszorg. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te analyseren en te gebruiken om de onderwijskwaliteit te optimaliseren.

Daarnaast kan ook de onderwijsinspectie de resultaten gebruiken bij hun reguliere doorlichtingen. Ze kan daarbij tijdens de doorlichting expliciet nagaan welke processen de school hanteert om te komen tot de leerresultaten/leerwinst die blijkt uit de eindtoetsen, en welke inspanningen de school doet om ook de eindtermen te behalen die minder aan bod komen bij de eindtoetsen.

3.1.5 Ontwikkeling en correctie

Omwille van de grotere garantie op betrouwbaarheid en de efficiënte werking inzake correctie, zijn de eindtoetsen schriftelijk van aard. De toetsitems worden jaarlijks volledig vernieuwd. De toets wordt opgesteld door een hiertoe opgericht onafhankelijk toetsinstituut, onder toezicht van de overheid. Bij de toetsontwikkeling is maar een relatief klein aantal mensen betrokken, om zo de ontwikkeling beheersbaar te houden en om lekken te voorkomen. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit.

Om de vergelijkbaarheid van de toetsen doorheen verschillende jaren te waarborgen, is het nodig dat een beperkt aantal 'sample items' van opeenvolgende toetsen overlappend is.

⁴ Alternatieve piste 4: Het groot aantal studierichtingen in het secundair onderwijs maakt dat zeer uiteenlopende toetsen zouden moeten ontwikkeld worden. De ontwikkelingskost kan sterk gereduceerd worden door uit elke studierichting slechts één centraal leerdomein te laten toetsen.

De toetsresultaten van de leerlingen worden bepaald per leerdomein, op de mate waarin de leerlingen de eindtermen hebben behaald. Het gaat dus om een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

De eindtoetsen van het basisonderwijs, en aan het einde van het tweede en vierde leerjaar secundair onderwijs, worden aan de hand van een correctiesleutel verbeterd door leraren. Wel blijft het belangrijk om de punten in te voeren in de web-based tool, zodat cijfers rechtstreeks gegeneerd kunnen worden en snel toegankelijk zijn voor de besproken partners. De eindtoetsen aan het eind van het secundair onderwijs worden centraal verbeterd door hiertoe opgeleide evaluatoren. De afname van deze eindtoetsen dient daarom ook reeds in de maand mei te gebeuren.

Een nieuwe web-based tool wordt speciaal hiervoor ontwikkeld, zodat zowel leraren als externe evaluatoren de (toets)resultaten rechtstreeks kunnen invullen. Op die manier kan de gelijkschakeling tussen de scholen berekend worden, maar worden de gegevens eveneens op een snelle manier toegankelijk gesteld (evaluatie op macroniveau, onderzoek (cohorten), ...)

Er worden strikte procedures opgemaakt waarmee gewaakt wordt over de kwaliteit van de toetsontwikkeling én over de kwaliteit van de correctie van de toetsen.

De eindcijfers op basis van de schooltoetsen en de gestandaardiseerde eindtoets worden rechtstreeks aan elk van de leerlingen bezorgd. Deze cijfers zijn niet publiek beschikbaar. De scholen krijgen een overzicht van de prestaties van hun leerlingen, maar mogen deze resultaten niet openbaar maken.

3.1.6 Beschikbaarheid van de gegevens

Via een te ontwikkelen web-tool beschikken scholen over de toetsresultaten, geaggregeerd op schoolniveau. Daarbij bestaat de mogelijkheid om de resultaten van de school in verhouding te plaatsen tot (anonieme) scholen met een gelijkaardige leerlinginstroom. De nadruk in deze rapportering ligt evenwel vooral in de vergelijking van scholen doorheen de tijd met zichzelf (vergelijking van verschillende cohorten met elkaar). De informatie wordt interactief aangeboden, zodat men makkelijk zelf per getoetste eindterm en per (combinatie van) sub-leerdomein(en) kan zien hoe de school zich verhoudt tot eerdere prestaties van de school. Het is eveneens mogelijk om de resultaten per doelgroep open te breken voor specifieke doelgroepen (bijvoorbeeld anderstalige leerlingen, leerlingen met leerachterstand (zittenblijvers), ...).

De resultaten van de eindtoetsen zijn niet publiek beschikbaar, maar kunnen wel door de onderwijsinspectie worden gebruikt als informatiebron over de output van de school.

Op macroniveau zijn de resultaten, geanonimiseerd, ter beschikking van het beleid. Op vraag kunnen datasets ter beschikking worden gesteld voor wetenschappelijk onderzoek.

3.1.7 Leerwinst en toegevoegde waarde

De toetsen staan op zich, en zijn in de eerste plaats bedoeld om de mate waarin de eindtermen zijn behaald, in kaart te brengen (standaard-criterium). Toch kan voor leerlingen in het secundair onderwijs ook leerwinst in kaart worden gebracht, door de resultaten van leerlingen op de eindtoetsen na het tweede jaar secundair onderwijs te vergelijken met de resultaten van dezelfde leerlingen in het basisonderwijs (voor de leerdomeinen die gelijk zijn). Aangezien het niet de bedoeling is deze toetsen op een gemeenschappelijke schaal te stellen, gaat het om een residual

gain model van leerwinst. Op dezelfde wijze kan ook voor elke leerling leerwinst worden berekend aan het eind van het vierde leerjaar secundair onderwijs en aan het eind van het secundair onderwijs.

De berekening van toegevoegde waarde van scholen in het basisonderwijs is beperkt: er wordt een vergelijking mogelijk gemaakt tussen scholen inzake de prestaties van (groepen van) leerlingen en een vergelijking van de huidige cohorte van leerlingen met eerdere cohorten in dezelfde school ("statusmodel"). In het secundair onderwijs kan bijkomend een vergelijking worden gemaakt tussen scholen inzake de gemiddelde bereikte leerwinst van leerlingen (volgens het residual gain model).

Bij de berekening van toegevoegde waarde wordt enkel rekening gehouden met leerlingen die gedurende de twee opeenvolgende meetmomenten onderwijs hebben gevolgd in dezelfde school.

3.1.8 Geraamde kostprijs

Bij de ontwikkeling van de eindtoetsen is een relatief klein aantal personen betrokken, waardoor de ontwikkelingskost gedrukt kan worden. Uiteraard moeten de items wel voldoende uitgetoetst worden. De ontwikkelingskost van één kwaliteitsvolle gestandaardiseerde eindtoets wordt op deze manier geraamd op € 100.000 (zie Tabel 9). Voornamelijk het grote aantal verschillende toetsen dat voor het secundair onderwijs moet worden ontwikkeld (cf. verschillende studierichtingen), verhoogt sterk de jaarlijks terugkerende ontwikkelingskost. In de berekening wordt uitgegaan van vijf eindtoetsen aan het eind basisonderwijs, en respectievelijk 10, 20 en 30 verschillende eindtoetsen voor het tweede, vierde, en einde van het secundair onderwijs ($65 \times € 100.000 = € 6.500.000$).

De afname van de toets is relatief duur doordat voor alle leerlingen een kopij moet worden voorzien; deze moet ter plaatse worden gebracht en moet (voor eind secundair onderwijs) terug verzonden worden naar een centrale plaats. De kost hiervan wordt geraamd op € 8 per leerling basisonderwijs en eerste en tweede graad secundair onderwijs, en € 10 per leerlingen eind secundair onderwijs (70.000 leerlingen per cohorte, totaal € 2.380.000).

Tabel 9 - Raming van de kostprijs in scenario 1

	Enmalige kost	Jaarlijks terugkerende kost
Ontwikkeling		€ 6.500.000
Afname		€ 2.380.000
Correctie		€ 2.500.000
Rapportering	€ 50.000	€ 10.000
Technische ondersteuning		
Algemene ondersteuning		€ 600.000
Totaal	€ 50.000	€ 11.990.000

Om de toetsen te corrigeren van ongeveer 70.000 laatstejaars leerlingen secundair onderwijs wordt een toets in minimaal vijf leerdomeinen afgenomen en centraal verbeterd. Deze 350.000 toetsen moeten op een korte periode handmatig worden gecorrigeerd, alsook moeten er analyses gebeuren inzake interbeoordelaarsbetrouwbaarheid op bepaalde types van vragen. Dit maakt dat veel evaluatoren moeten worden ingezet, die op voorhand ook sterk worden getraind. Geschat wordt dat 250 evaluatoren gedurende twee maanden actief zullen zijn met correctie, onder toezicht van een tiental analisten (€ 2.500.000).

Inzake rapportering, dient een gebruiksvriendelijke en interactieve website te worden opgebouwd (€ 50.000). Daarnaast dienen de gegevens jaarlijks ingebracht te worden, en dient onderhoud aan de website te gebeuren (€10.000).

Ten slotte dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten (€ 600.000).

3.2 Scenario 2 'Adaptief leerwinstmonitoringsysteem'

Microniveau - verplicht - oriënterend toetsensysteem - niet-bindend

3.2.1 Basisonderwijs

Er worden gestandaardiseerde toetsen ontwikkeld met betrekking tot drie cruciale vaardigheden die doorheen de basisschool worden aangeleerd, met name rekenvaardigheid, leesvaardigheid en schrijfvaardigheid. Deze toetsen hebben een uitgesproken formatief karakter. De afname van de toetsen één maal per jaar door alle leerlingen is verplicht, binnen een bepaalde periode (bijvoorbeeld de tweede maand van het schooljaar). Binnen deze periode kunnen scholen het precieze moment zelf vaststellen. Omwille van het adaptieve karakter van de toetsen kunnen scholen ervoor kiezen om de toetsen meerdere malen aan te bieden aan dezelfde leerling.

Het gaat om 'getrapt adaptieve toetsen' die de leerling digitaal beantwoordt (bij jonge leerlingen met ondersteuning van een leraar). Een leerling krijgt bij de eerste toetsafname een set van 10 items op zijn/haar eigen leeftijdsniveau aangeboden. Afhankelijk van het aantal correcte antwoorden, wordt vervolgens een set van 10 makkelijkere of moeilijker items aangeboden. Bij jongere leerlingen wordt de toetsafname beperkt tot 20 items, bij oudere leerlingen kan een toets gradueel uitgebreid worden tot maximaal vijf sets van 10 items, waarbij telkens een makkelijkere of moeilijker set wordt aangeboden op basis van de resultaten in de vorige set. De toetsen, die per vaardigheid worden opgesteld, zijn in heel Vlaanderen hetzelfde voor alle leerlingen, al kan het beginniveau van de eerste set van 10 items worden afgesteld op het eindniveau van de leerling bij de vorige toetsafname. Door het getrapt adaptieve karakter wordt de toets voldoende uitdagend voor elke leerling. Het getrapt adaptieve karakter vereist dat alle items op een IRT-schaal zijn ingeschaald, zodat de moeilijkheidsgraad ervan kan worden vastgesteld.

Elke leerling legt in zijn/haar studieloopbaan op de basisschool minimaal zeven keer deze gestandaardiseerde toetsen in elk van de drie domeinen af: een eerste keer in het laatste jaar kleuteronderwijs, daarna éénmaal per schooljaar. Scholen kunnen de toetsen meerdere keren aanbieden aan de leerlingen, maar worden hiertoe niet verplicht.

De items in de gestandaardiseerde toetsen zijn niet rechtstreeks gekoppeld aan eindtermen of leerplannen, maar toetsen specifieke vaardigheden die niet specifiek leerdomein-gebonden zijn.

De resultaten van de toetsen worden automatisch gekoppeld aan een tool die scholen tot hun beschikking hebben en waarin de leerprestaties van de leerlingen wordt bijgehouden. Bovendien wordt ook de leerwinst van de leerlingen berekend op basis van één of meerdere voorgaande toetsen.

3.2.2 Secundair onderwijs

In het secundair onderwijs wordt een zelfde stramien gevolgd als in het basisonderwijs: scholen worden verplicht om minimaal éénmaal per jaar, binnen een afgesproken periode (bijvoorbeeld: het eerste trimester), de algemene vaardigheid van leerlingen te toetsen inzake wiskunde geletterdheid, lees- en schrijfvaardigheden (Nederlands, Frans, Engels) en wetenschappelijke geletterdheid. Ook in het secundair onderwijs wordt leerwinst van leerlingen in kaart gebracht en zijn de toetsen formatief en niet curriculum-gerelateerd. Waar mogelijk worden de leerlingengegevens en -resultaten uit het basisonderwijs overgedragen naar het secundair onderwijs.

3.2.3 Gebruik op microniveau

In de eerste plaats is het opzet van deze vorm van standaardisatie het ontwikkelen van een leerwinstmonitoringsysteem in een aantal essentiële vaardigheden. Gegevens over leerwinst geven de school meer inzicht in de ontwikkeling die leerlingen doormaken in de gemeten onderdelen. Op basis daarvan kunnen scholen maatregelen nemen om leerlingen bij te werken waar dat nodig geacht wordt.

In het secundair onderwijs worden oriënteringsattesten uitgereikt op het einde van elk schooljaar. Niettegenstaande dit niet het doel van de toetsen is, zou de school de resultaten ook kunnen gebruiken binnen beslissingen om leerlingen te oriënteren naar een bepaalde studierichting (bijvoorbeeld op basis van verschillen in ontwikkeling van vaardigheden bij een bepaalde leerling). Er zijn evenwel geen rechtstreekse consequenties aan de toetsen verbonden waardoor de toetsen low-stakes blijven.

3.2.4 Gebruik op mesoniveau

Hoewel het primaire doel zich op microniveau situeert, kunnen de toetsresultaten gebruikt worden om aan kwaliteitszorg op schoolniveau te doen. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te gebruiken binnen hun interne kwaliteitszorg. De onderwijsinspectie bevraagt niet de resultaten van de toets, maar gaat wel na hoe ze door de school in het kader van interne kwaliteitszorg worden gebruikt.

3.2.5 Gebruik op macroniveau

Doordat de toetsafname gebeurt middels een online tool, kunnen de gegevens door beleidsmakers worden gebruikt om na te gaan of de vooropgezette essentiële vaardigheden bereikt worden in het huidige onderwijssysteem en waar de sleutelmomenten in de ontwikkeling van deze vaardigheden zich bevinden. Daarnaast kan de dataverzameling ook worden aangewend door onderzoekers om cohortonderzoek uit te voeren. De openbaarheid van de gegevens wordt gevrijwaard.

3.2.6 Ontwikkeling en correctie

De toetsafname gebeurt via een online tool en de antwoorden van de leerlingen worden onmiddellijk verbeterd. Het gaat bijgevolg om relatief eenvoudige itemtypes (meerkeuzevragen, korte open vragen, 'drag and drop'-vragen, 'hot spot'-vragen, ...). Het adaptieve karakter van de gestandaardiseerde toetsen vereist dat er een vrij ruime databank van items voorradig is in elk van de voorgestelde domeinen. Het kan daarom opportuun zijn om bij aanvang slechts toetsen in één of enkele domeinen te ontwikkelen. Bij de ontwikkeling van de toetsen kan evenwel gebruik gemaakt

worden van toetsitems in bestaande toetsen, zoals de peilingen en koepelgebonden toetsen, wat de ontwikkelingskost sterk kan drukken. De kwaliteit van de toetsitems worden regelmatig opgevolgd (in kader van bijvoorbeeld item shifting) en nieuwe items worden toegevoegd. De IRT-inschaling van nieuwe items kan vrij eenvoudig gebeuren door ze tijdens de bestaande toetsafnames extra aan te bieden aan leerlingen.

De opvolging en ontwikkeling van de toetsen en toetsitems gebeurt onder toezicht van de overheid. Bij de ontwikkeling van items worden inhoudsexperten en ervaren leraren betrokken. Veel aandacht gaat uit naar de inschaling van de items op de IRT-schaal. Er is voor scholen een hulplijn aanwezig bij de toetsontwikkelaar voor eventuele technische problemen tijdens de toetsafname.

De toetsresultaten van de leerlingen worden geëvalueerd ten opzichte van een zelfcriterium en standaardcriterium: hoe staat het met de vaardigheid van de leerling, in het licht van zijn/haar eerder aangetoonde vaardigheid en in elke mate voldoet de leerling aan de vooropgestelde verwachtingen?

3.2.7 Beschikbaarheid van de gegevens

Zowel de leerlingen als de school worden geïnformeerd over de resultaten van de gestandaardiseerde toetsen en de geboekte leerwinst. De resultaten worden niet vergeleken met de resultaten van andere leerlingen. Deze rapportering gebeurt geautomatiseerd vanuit de web-tool.

Scholen mogen de resultaten niet openbaar maken. Na het afleggen van de toetsen worden ook de ouders geïnformeerd over de prestaties en de geboekte leerwinst in de getoetste domeinen bij hun kind.

3.2.8 Leerwinst en toegevoegde waarde

De prioriteit van dit systeem is het monitoren van leerwinst. Leerwinst wordt hierbij eenvoudig geconceptualiseerd ('leerwinst als groei')⁵. Daarbij worden de resultaten van de leerlingen niet vergeleken met prestaties van anderen, maar enkel met de prestaties ten opzichte van zichzelf en het standaardcriterium.

Een maat voor toegevoegde waarde voor scholen zou kunnen worden toegevoegd, maar dan voornamelijk vanuit een statusmodel waarbij de school de leerprestaties en leerwinst van haar leerlingen doorheen de tijd (verschillende cohorten) met elkaar vergelijkt ('cohort-to-cohort-perspectief').

3.2.9 Geraamde kostprijs

Op termijn bestaan er 21 gestandaardiseerde toetsen voor het basisonderwijs (3 domeinen x 7 leerjaren) en 36 gestandaardiseerde toetsen voor het secundair onderwijs (6 domeinen x 6 leerjaren). Er kan evenwel overlap bestaan in items tussen de verschillende toetsen van een bepaald domein, zoals gebruikelijk in adaptieve toetsen. De ontwikkelingskost van kwaliteitsvolle gestandaardiseerde toetsen in elk van deze drie domeinen voor beide niveaus (basis- en secundair onderwijs) vereist dat ongeveer 2000 items zijn ontwikkeld en ingeschaald. Het ontwikkelen van een set van gestandaardiseerde toetsen voor een leerdomein wordt op deze manier geraamd op € 200.000 (zie Tabel 10). Het inschalen van elk van de items op een gemeenschappelijke schaal,

⁵ Alternatieve piste 1: bij de leerwinst kan ook rekening worden gehouden met de leerwinst in de normgroep. In dat geval komen we tot een 'gain score' model.

maakt dat de verwachte kost verdubbeld wordt. De eenmalige kost wordt bijgevolg $3 * € 400.000 = € 1.200.000$. Daarnaast dient kwaliteitscontrole en het toevoegen van nieuwe items jaarlijks te gebeuren (€ 200.000). Doordat de toetsen gradueel ontwikkeld kunnen worden, kunnen de eenmalige ontwikkelingskosten gespreid worden over meerdere jaren.

De afname van de toets is relatief goedkoop, doordat de toetsen online ter beschikking gesteld worden en automatisch gecorrigeerd worden. Er dient evenwel een gebruiksvriendelijke en interactieve webtool te worden opgebouwd (éénmalige kost, € 600.000). Daarnaast dient onderhoud aan de website te gebeuren en dienen nieuwe items te worden ingebracht. Deze web-tool zorgt ook voor correctie en rapportering, waardoor op deze posten geen extra kosten worden voorzien.

Ten slotte is een technische ondersteuning voor scholen tijdens de afname van de toetsen (€ 100.000), en dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten (€ 600.000).

Tabel 10 - Raming van de kostprijs in scenario 2

	Eenmalige kost	Jaarlijks terugkerende kost
Ontwikkeling	€ 1.200.000	€ 200.000
Afname	€ 600.000	€ 50.000
Correctie		
Rapportering		
Technische ondersteuning		€ 100.000
Algemene ondersteuning		€ 600.000
Totaal	€ 1.800.000	€ 950.000

3.3 Scenario 3 'Diagnostisch meetinstrumentarium & toetsenbank'

Microniveau - facultatief - oriënterend toetsensysteem, niet-bindend

3.3.1 Basisonderwijs

Er wordt een reeks gestandaardiseerde toetsen ontwikkeld en via een toetsenbank ter beschikking gesteld aan scholen. De resultaten van deze toetsen genereren diagnostische informatie over het leren van de leerlingen.

De ontwikkelde diagnostische meetinstrumenten richten zich op het beschrijven van het huidige leerniveau van de leerling. Daarnaast identificeert de toets ook de sterktes en zwaktes inzake de kennis en vaardigheden van de leerlingen. Op basis van de resultaten kan een individueel verbeteringstraject worden opgesteld om de tekorten weg te werken.

Om diagnostische informatie te voorzien, dient de toets erg gedetailleerde informatie op te leveren. De titels van de toetsen zijn bijgevolg zeer specifiek. Het gaat bijvoorbeeld niet om een algemene rekentoets, maar binnen leerdomeinen zijn er verschillende toetsen beschikbaar, gericht op één of een beperkt aantal eindtermen en leerplandoelen. In het leerdomein rekenen kunnen de volgende titels bijvoorbeeld worden ontwikkeld: 'breuken vierde leerjaar', 'breuken vijfde leerjaar', 'breuken zesde leerjaar', 'veelvouden van natuurlijke getallen vijfde leerjaar', ... Op geregelde tijdstippen wordt

het toetsenaanbod in samenspraak met het onderwijsveld geëvalueerd en afgestemd op de wensen van de betrokken actoren, zodat de databank blijft evolueren.

De toetsen zijn methode-onafhankelijk en hebben betrekking op de eindtermen wiskunde, Nederlands, wetenschap en techniek, Frans, mens en maatschappij. Daarnaast zijn ook observatie- en meetinstrumenten terug te vinden die betrekking hebben op competenties, vakoverschrijdende eindtermen en sociaal-emotionele aspecten van de leerling (cf. welbevinden, intelligentie, motorische vaardigheden, communicatief vermogen, samenwerkingsvaardigheden, digitale vaardigheden, mediawijsheid, kritisch denken, creativiteit, ...).

Ten slotte kunnen er ook toetsen ontwikkeld worden die gericht zijn op het detecteren van leerstoornissen, zoals dyslexie of dyscalculie. Deze toetsen zullen geen diagnostisch uitsluitsel geven, maar kunnen wel resulteren in een advies om een leerling al dan niet door een expert te laten screenen.

Waar mogelijk gaat het om digitale toetsen. Bij bepaalde competenties zal het echter eerder gaan om een (gestandaardiseerde) observatielijst die door de leraar zelf moet worden ingevuld, al dan niet in samenspraak met het CLB.

3.3.2 Secundair onderwijs

Gelijklopend met het basisonderwijs worden ook in het secundair onderwijs diagnostische evaluatie-instrumenten ter beschikking gesteld aan de scholen.

Om de diverse leerdomeinen in het secundair onderwijs te dekken, is het belangrijk om te investeren in een uitgebreid aanbod van gestandaardiseerde toetsen. In de beginfase is het wenselijk om diagnostische instrumenten voor Nederlands, wiskunde en de moderne talen (Frans en Engels) ter beschikking te stellen, hoewel op langere termijn ook andere algemene vakken (bijvoorbeeld aardrijkskunde, natuurwetenschappen, geschiedenis) in een toets kunnen worden aangeboden.

Om specifieke informatie in te winnen over de sterktes en zwaktes van de leerling, is het ook hier belangrijk om in te zetten op het evalueren van duidelijk afgebakende leerplandoelen en eindtermen. Wanneer de vergelijking met wiskunde wordt doorgetrokken, kunnen de toetsen betrekking hebben op 'natuurlijke getallen 1A', 'kommagetallen 1A', 'breuken 1A', 'gelijkvormigheid en congruentie 1A'. Er wordt dus rekening gehouden met zowel de studierichting als het leerjaar waarin de leerling zich bevindt.

Naast de cognitieve toetsen wordt ook hier het aanbod uitgebreid met observatie- en meetinstrumenten die betrekking hebben op competenties, vakoverschrijdende eindtermen en sociaal-emotionele aspecten van de leerling (cf. basisonderwijs). Voor praktijkvakken in het technisch en beroepsonderwijs worden eveneens evaluatie-instrumenten ontwikkeld die het huidige leerniveau van de leerling in kaart kunnen brengen.

Het testen op een leerstoornis kan gebeuren aan de hand van hiertoe ontwikkelde testen, eventueel in samenspraak met het CLB, al is het niet de bedoeling dat de leraar ook meteen een diagnose zal stellen. Wel kan de school een advies uitbrengen om een leerling al dan niet door een expert te laten screenen.

Net zoals in het basisonderwijs wordt ook hier de mogelijkheid ingebouwd om op geregelde tijdstippen het toetsenaanbod in samenspraak met het onderwijsveld te evalueren en zo af te stemmen op de behoeften van de betrokken actoren.

3.3.3 Gebruik op microniveau

Het voornaamste doel van deze diagnostische instrumenten is het in kaart brengen van het huidige leerniveau van de leerling. De toetsen hebben een uitgesproken formatief karakter. Niet enkel kunnen leerlingen op deze manier objectief georiënteerd worden in een bepaalde klas en graad, ook kunnen vermoedelijke leerachterstanden bij leerlingen geïdentificeerd worden. Het zorgteam op school kan dan maatregelen nemen om de leerling in kwestie te faciliteren en om gerichte leerkansen te creëren.

Scholen of individuele leraren kunnen zelf beslissen welke toetsen ze uit de toetsenbank gebruiken, en op welk moment ze deze toetsen inzetten. Men kan ervoor kiezen om een toets bij een hele cohorte leerlingen af te nemen, enkel bij bepaalde leerlingen, of enkel bij een specifieke leerling waarover men meer informatie wenst te verkrijgen. Dat laatste kan bijvoorbeeld het geval zijn bij leerlingen die in het midden van een schooljaar instromen in een nieuwe school.

De school of leraar kan er ook voor kiezen om eenzelfde gestandaardiseerde toets twee maal af te nemen (door verschillende cohorten van leerlingen), bijvoorbeeld als voor- en nameting om een didactische innovatie of een nieuwe lesmethode te evalueren.

3.3.4 Gebruik op mesoniveau

De resultaten van deze toetsen kunnen worden gebruikt door de medewerkers van het CLB. De inspectie kan de resultaten en genomen maatregelen inkijken op voorwaarde dat de school inzagerecht verleent.

3.3.5 Ontwikkeling en correctie

Omdat de school zelf toetsen uit de databank selecteert en deze afneemt bij de leerlingen op een tijdstip naar keuze, kunnen de toetsen digitaal ter beschikking worden gesteld. Scholen die dit wensen, kunnen de toetsen schriftelijk afnemen en zelfstandig corrigeren. Een nieuwe web-based tool wordt hiervoor ontwikkeld, zodat de prestaties van de leerlingen kunnen worden opgeslagen.

De ontwikkeling van de diverse evaluatie-instrumenten kan vertrekken vanuit reeds bestaande, leerplangebonden toetsen en bestaande observatie-instrumenten. Zoals eerder gesteld wordt het werkveld op regelmatige basis om advies gevraagd om de waaier aan mogelijkheden up-to-date te houden. De toetsenbank staat onder toezicht van de overheid, die indien nodig een externe instelling kan aanstellen om toetsen te ontwikkelen. Zo draagt een vijfjaarlijkse evaluatie van de validiteit en betrouwbaarheid van de toetsen bij tot het behouden van de kwaliteit. Bij de toetsontwikkeling is een relatief klein aantal mensen betrokken, om zo de ontwikkeling beheersbaar te houden. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit.

De toetsresultaten van de leerlingen worden geëvalueerd ten opzichte van een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

3.3.6 Beschikbaarheid van de gegevens

Het systeem verleent zich ertoe om een zeer duidelijk overzicht op te stellen van de sterktes en tekortkomingen van individuele leerlingen. Dit rapport wordt in geen geval openbaar gemaakt, maar kan enkel worden aangewend om een behandelingsplan op te stellen voor de problemen waar de leerling tegenaan loopt.

3.3.7 Leerwinst en toegevoegde waarde

Binnen dit scenario wordt er geen gebruik gemaakt van leerwinstmonitoring. De toegevoegde waarde wordt eveneens niet berekend. Leraren en zorgcoördinatoren richten zich op het faciliteren van het onderwijsleerproces, maar brengen het proces van de leerling hierbij niet in kaart. Wel kan geopteerd worden voor een herafname van de toets nadat de behandeling werd voltooid.

3.3.8 Geraamde kostprijs

Op termijn bestaan er meerdere gestandaardiseerde, diagnostische toetsen voor het basisonderwijs en het secundair onderwijs. De ontwikkelingskost daalt doordat IRT niet noodzakelijk is en de toetsen low-stakes zijn en slechts een klein domein wordt omvat met elke toets (geraamde kostprijs €40.000) (zie Tabel 11). Uiteraard dient wel de betrouwbaarheid en validiteit gecontroleerd te worden, maar de investering hierin is minder groot dan bij een high-stakes toets. Er wordt gerekend op een toetsenbank met 150 toetsen (totale ontwikkelingskost € 6.000.000).

Doordat de toetsenbank al kan worden opgestart met enkele basismogelijkheden, kunnen de kosten gespreid worden over meerdere jaren. Daarna dienen er om de vijf jaar kwaliteitscontroles te worden toegepast op de evaluatie-instrumenten, waardoor een jaarlijkse kost van € 200.000 bestaat.

Tabel 11 - Raming van de kostprijs in scenario 3

	Enmalige kost	Jaarlijks terugkerende kost
Ontwikkeling	€ 6.000.000	€ 200.000
Afname	€ 400.000	€ 50.000
Correctie		
Rapportering		
Technische ondersteuning		€ 100.000
Algemene ondersteuning		
Totaal	€ 6.400.000	€ 350.000

De afname van de toets is relatief goedkoop, doordat de toetsen online ter beschikking gesteld worden en automatisch gecorrigeerd worden. Er dient evenwel een gebruiksvriendelijke en interactieve website te worden opgebouwd (éénmalige kost). Daarnaast dient onderhoud aan de website te gebeuren en dienen nieuwe items te worden ingebracht. Deze web-tool zorgt ook voor correctie en rapportering, waardoor op deze posten geen extra kosten worden voorzien.

Ten slotte is een technische ondersteuning voor scholen tijdens de afname van de toetsen (€ 100.000).

3.4 Scenario 4 'High-stakes eindtoetsen voor leerlingen en scholen'

Meso- en microniveau - verplicht - openbaar - verantwoordingsgericht

3.4.1 Basisonderwijs

Alle leerlingen in het zesde leerjaar, die in aanmerking komen voor het getuigschrift basisonderwijs, nemen verplicht deel aan gestandaardiseerde eindtoetsen die zich uitstrekken over meerdere leerdomeinen en die plaatsvinden op het einde van het schooljaar (bij voorkeur in de maand mei). Deze toetsen zijn schriftelijk en worden op hetzelfde moment bij alle leerlingen afgenomen.

Er is een eindtoets in elk van de volgende leerdomeinen: wiskunde, Nederlands, wetenschap en techniek, Frans, mens en maatschappij. De toetsen zijn methode-onafhankelijk en richten zich uitsluitend op het evalueren van de mate waarin de eindtermen door de leerlingen in het zesde leerjaar bereikt worden. De toetsen zijn zo breed mogelijk: ze brengen naast kennis ook bepaalde vaardigheden van leerlingen in kaart.

Deze eindtoetsen hebben voor leerlingen een summatief karakter en gaan na in welke mate de leerlingen de eindtermen hebben behaald. De resultaten van deze gestandaardiseerde toetsen tellen mee voor 50% van het jaargemiddelde in het leerdomein. De overblijvende 50% van het eindcijfer wordt ingevuld op basis van door de school verzamelde gegevens. In deze school-evaluatie kan bijvoorbeeld sterker worden ingezet op die vaardigheden en competenties die moeilijker door schriftelijke gestandaardiseerde toetsen kunnen worden gemeten.

3.4.2 Secundair onderwijs

In het secundair onderwijs worden eindtoetsen opgelegd aan het einde van het laatste jaar. De resultaten van deze eindtoetsen leiden tot het al dan niet uitreiken van het diploma secundair onderwijs. Het betreft verplichte gestandaardiseerde eindtoetsen die de examenperiode in juni in het betreffende schooljaar kunnen vervangen.

Deze eindtoetsen zijn in heel Vlaanderen hetzelfde voor alle leerlingen in dezelfde studierichting van het laatste jaar, met de volgende leerdomeinen die zowel in ASO, KSO als TSO aan bod komen: wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne vreemde talen (Engels en Frans). In het BSO wordt PAV (project algemene vakken) geëvalueerd. Bijkomende vakken die geëvalueerd worden, zijn afhankelijk van de gevolgde studierichting. De toetsen meten de mate waarin de leerlingen de eindtermen van de derde graad secundair onderwijs hebben behaald. Waar mogelijk worden zowel kennis als vaardigheden van leerlingen geregistreerd.

De eindcijfers op de eindtoetsen in het secundair onderwijs hebben een bindend karakter: het al dan niet slagen van de leerlingen, alsook het behalen van oriënteringsattesten, worden rechtstreeks gekoppeld aan de eindcijfers van de leerlingen in de verschillende leerdomeinen⁶.

3.4.3 Gebruik op microniveau

De toetsen zijn high-stakes voor de leerlingen. Op basis van de resultaten wordt een formele beslissing genomen over het toekomstige studietraject van de leerling. Leerlingen kunnen op die manier laten zien welke capaciteiten ze hebben met betrekking tot de door de overheid gestelde eindtermen en/of ontwikkelingsdoelen, en in vergelijking met leeftijdsgenoten in heel Vlaanderen.

Ouders beschikken over vergelijkende informatie over hun kind. Door de openbaarheid van de resultaten op schoolniveau (zie onder), beschikken ouders ook over betrouwbare aanvullende informatie om een bewuste schoolkeuze te maken.

3.4.4 Gebruik op mesoniveau

Door het openbaar karakter van de toetsresultaten beschikken de scholen en het publiek over informatie die het mogelijk maakt om de kwaliteit van een bepaalde school te vergelijken met die van andere scholen en regio's. Scholen kunnen vergeleken worden met scholen met een gelijkaardige input. Hierbij wordt zowel in het basisonderwijs als het secundair onderwijs het aandeel van leerlingen binnen een bepaalde categorie van leerprestaties per leerdomein weergegeven. Er wordt onderscheid gemaakt tussen vijf categorieën.

Percentage leerlingen dat in het betreffende leerdomein...

- > 90% van de getoetste eindtermen bereikt
- > 80% van de getoetste eindtermen bereikt
- > 70% van de getoetste eindtermen bereikt
- > 60% van de getoetste eindtermen bereikt
- ≤ 60% van de getoetste eindtermen bereikt

Daarnaast kan ook de onderwijsinspectie de resultaten gebruiken bij hun reguliere doorlichtingen. Ze kan daarbij tijdens de doorlichting expliciet nagaan welke processen de school hanteert om te komen tot de leerresultaten die blijken uit de eindtoetsen, alsook welke inspanningen de school doet om ook de eindtermen te behalen die minder aan bod komen bij de eindtoetsen.

Naast de verantwoordingsgerichte doelstellingen van deze toetsen, kunnen de gestandaardiseerde toetsen in dit scenario ook ontwikkelingsgericht worden ingezet: scholen kunnen ze gebruiken

⁶ Alternatieve piste 1: in het secundair onderwijs wordt de bindende kracht van de resultaten op de eindtoetsen incrementeel opgebouwd door ook op het einde van de eerste en tweede graad eindtoetsen in te leggen. Op basis van een transparante formule wordt aan elke leerling een eindcijfer toebedeeld. De toetsresultaten van de gestandaardiseerde eindtoets tellen voor 15% mee in het tweede jaar secundair onderwijs en 30% in het vierde jaar secundair onderwijs van het jaargemiddelde binnen datzelfde domein. In het laatste jaar secundair onderwijs tellen de resultaten van het gestandaardiseerde eindexamen mee voor 50% van het jaargemiddelde in het leerdomein. De overblijvende 85%/70%/50% van het eindcijfer wordt ingevuld door de school op basis van door de school verzamelde gegevens. In deze school-evaluatie kan bijvoorbeeld sterker worden ingezet op die vaardigheden en competenties die moeilijker door schriftelijke gestandaardiseerde toetsen kunnen worden gemeten. De eindcijfers in het secundair onderwijs hebben een bindend effect: het al dan niet slagen van de leerlingen, alsook oriënteringsattesten, worden rechtstreeks gekoppeld aan de eindcijfers van de leerlingen in de verschillende leerdomeinen.

binnen hun eigen interne kwaliteitszorg. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te gebruiken binnen hun interne kwaliteitszorg.

3.4.5 Gebruik op macroniveau

De overheid kan de gegevens gebruiken om zowel de prestaties van het eigen onderwijssysteem te evalueren als de mate waarin de eindtermen behaald kunnen worden. Ook kan de gegenereerde databank gebruikt worden als basis voor verder onderzoek.

3.4.6 Ontwikkeling en correctie

Omwille van de hoge inzet van de eindtoetsen in dit scenario voor leerlingen en scholen, is de noodzaak om de betrouwbaarheid van de toetsen te garanderen bijzonder groot. Om deze reden, alsook omwille van de efficiënte werking inzake correctie, zijn de eindtoetsen schriftelijk van aard. De toetsitems worden jaarlijks volledig vernieuwd. De toetsen worden opgesteld door een hiertoe opgericht onafhankelijk toetsinstituut, onder toezicht van de overheid. Bij de toetsontwikkeling is maar een relatief klein aantal mensen betrokken, om zo de ontwikkeling beheersbaar te houden en om lekken te voorkomen. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit.

De toetsresultaten van de leerlingen worden bepaald per leerdomein, op de mate waarin de leerlingen de eindtermen hebben behaald. Het gaat dus om een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

Alle eindtoetsen worden centraal verbeterd door hiertoe opgeleide evaluatoren. De afname van deze eindtoetsen dient daarom ook reeds in de maand mei te gebeuren. Er worden strikte procedures opgemaakt waarmee gewaakt wordt over de kwaliteit van de toetsontwikkeling én over de kwaliteit van de correctie van de toetsen. Om de vergelijkbaarheid van de toetsen doorheen verschillende jaren te waarborgen, is het nodig dat een beperkt aantal 'sample items' van opeenvolgende toetsen overlappend is.

Een nieuwe web-based tool wordt ontwikkeld, waarop de resultaten van de leerlingen worden ingevuld. Deze software zorgt voor een snelle berekening van de individuele leerlingresultaten, maar wordt ook ingezet om data te genereren op hogere niveaus (school, regio, provincie, ...).

De individuele eindcijfers van de eindtoetsen worden zowel aan de leerling in kwestie als aan de schooldirectie bezorgd. Deze individuele cijfers zijn niet publiek beschikbaar.

3.4.7 Beschikbaarheid van de gegevens

Via een te ontwikkelen web-tool worden de toetsresultaten, geaggregeerd op schoolniveau, openbaar gemaakt. Daarbij bestaat de mogelijkheid om de resultaten van de school in verhouding te plaatsen met scholen met een gelijkaardige leerlinginstroom. De resultaten worden gepubliceerd per leerdomein, zodat het publiek de 'prestatie' van de school in verschillende leerdomeinen kan vergelijken. De scores van de verschillende leerdomeinen worden niet geaggregeerd.

Bovendien kunnen de resultaten per doelgroep (bijvoorbeeld anderstalige leerlingen, leerlingen met leerachterstand (zittenblijvers), ...) alsook per regio (stad, provincie, ...) worden opgebroken. De web-tool laat de gebruikers ook toe om evoluties van scholen doorheen de tijd te raadplegen. Voor

secundaire scholen kan men daarnaast raadplegen wat de gemiddelde leerwinst is van leerlingen (in totaal, en per doelgroep) ten opzichte van de resultaten op de eindtoetsen basisonderwijs.

3.4.8 Leerwinst en toegevoegde waarde

Binnen dit scenario wordt toegevoegde waarde berekend op basis van leerprestaties en een vergelijking van deze leerprestaties met andere scholen (statusmodel), weliswaar met correcties op het vlak van diverse leerlingencategorieën (moedertaal, socio-economische status, ...).

Voor secundaire scholen wordt toegevoegde waarde daarnaast ook berekend als het gemiddelde van de leerwinst van leerlingen, geconceptualiseerd als de verschilscore tussen de eindtoetsen secundair onderwijs en basisonderwijs (residual gain model). Uiteraard wordt ook hier rekening gehouden met correcties betreffende de leerlingeninstroom.

3.4.9 Geraamde kostprijs

Bij de ontwikkeling van de eindtoetsen is een relatief klein aantal personen betrokken, waardoor de ontwikkelingskost gedrukt kan worden. Uiteraard moeten de items wel voldoende getoetst worden. De ontwikkelingskost van één kwaliteitsvolle gestandaardiseerde eindtoets wordt op deze manier geraamd op €100.000 (zie Tabel 12). Voornamelijk het grote aantal verschillende toetsen dat voor het secundair onderwijs moet worden ontwikkeld (cf. verschillende studierichtingen), verhogen sterk de ontwikkelingskost. In de berekening wordt uitgegaan van vijf eindtoetsen aan het eind basisonderwijs, en respectievelijk 30 verschillende eindtoetsen voor het laatste jaar secundair onderwijs (= € 3.500.000).

De afname van de toets is relatief duur doordat voor alle leerlingen een kopij moet worden voorzien, deze moet ter plaatse worden gebracht en moet terug verzonden worden naar een centrale plaats. De kost hiervan wordt geraamd op € 10 per leerling, aan 70.000 leerlingen per cohort (= € 1.400.000).

Tabel 12 - Raming van de kostprijs in scenario 4

	Enmalige kost	Jaarlijks terugkerende kost
Ontwikkeling		€ 3.500.000
Afname		€ 1.400.000
Correctie		€ 5.000.000
Rapportering	€ 50.000	10.000
Technische ondersteuning		
Algemene ondersteuning		€ 600.000
Totaal	€ 50.000	€10.510.000

Er wordt een toets afgenomen bij ongeveer 70.000 laatstejaars leerlingen in het basisonderwijs en 70.000 leerlingen in het laatste jaar secundair onderwijs in minimaal vijf leerdomeinen, deze toets wordt centraal verbeterd. Deze 700.000 toetsen moeten op een korte periode handmatig worden gecorrigeerd, alsook moeten analyses gebeuren inzake interbeoordelaarsbetrouwbaarheid. Dit maakt dat veel evaluatoren moeten worden ingezet, die op voorhand ook sterk worden getraind. Geschat wordt dat 400 evaluatoren gedurende twee maanden actief zullen zijn met correctie, onder toezicht van een tiental analisten.

Inzake rapportering, dient een gebruiksvriendelijke en interactieve website te worden opgebouwd (éénmalige kost). Daarnaast dienen de gegevens jaarlijks ingebracht te worden, en dient onderhoud aan de website te gebeuren. Ten slotte dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten.

3.5 Scenario 5 'Leerwinstmeting op schoolniveau met openbaar karakter'

Mesoniveau - verplicht - openbaar - ontwikkelingsgericht

3.5.1 Basisonderwijs

Er worden gestandaardiseerde toetsen ontwikkeld die betrekking hebben op de eindtermen en ontwikkelingsdoelen in diverse leerdomeinen, zoals wiskunde, Nederlands, Frans, wereldoriëntatie en informatieverwerking en -verwerking. Het doel van deze toetsen is om jaarlijks van elke school na te gaan in welke mate de leerlingen de leerplandoelen en eindtermen beheersen. Door het openbare en verplichte karakter van de toetsen is het opportuun om de afname van de toetsen op hetzelfde moment voor alle leerlingen te laten doorgaan. De toetsen zijn schriftelijk van aard.

De toetsen zijn zo breed mogelijk: ze trachten naast kennis ook vaardigheden van leerlingen in kaart te brengen. De toetsen worden van alle leerlingen aan het eind van het tweede, vierde, en zesde leerjaar afgenomen. De toetsen van het tweede en vierde leerjaar zijn gebaseerd op de leerplannen, waardoor er variatie mogelijk is tussen verschillende onderwijsnetten. De toetsen van het zesde leerjaar zijn gebaseerd op de eindtermen en zijn daardoor hetzelfde voor alle leerlingen in heel Vlaanderen. Elk van deze toetsen is methode-onafhankelijk.

Op basis van de leerwinst die leerlingen boeken tussen twee toetsmomenten, wordt de toegevoegde waarde van de school berekend en via een te ontwikkelen webtool openbaar gemaakt.

3.5.2 Secundair onderwijs

In het secundair onderwijs worden eveneens gestandaardiseerde toetsen ontwikkeld die eenmaal per graad verplicht worden afgenomen bij alle leerlingen.

Op een gemeenschappelijk moment aan het einde van elke graad in het secundair onderwijs leggen leerlingen een reeks toetsen af, die als opzet het verzamelen van gegevens hebben over het aantal leerlingen dat een bepaalde eindterm onder de knie heeft.

Deze toetsen zijn gebaseerd op de eindtermen, of waar deze ontbreken, op de leerplandoelen en zijn in heel Vlaanderen hetzelfde voor alle leerlingen in dezelfde studierichting van dezelfde graad. De volgende leerdomeinen die zowel in ASO, KSO als TSO aan bod komen, worden getoetst: wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne vreemde talen (Engels en Frans). In het BSO wordt PAV (project algemene vakken) geëvalueerd. Bijkomende vakken die geëvalueerd worden, zijn afhankelijk van de gevolgde studierichting⁷. De toetsen meten, waar mogelijk, zowel kennis als vaardigheden van leerlingen.

⁷ Alternatieve piste 1: om de ontwikkelingskosten te reduceren, kan het opportuun zijn om niet elk schooljaar dezelfde leerdomeinen te evalueren. Zo kan men het ene jaar opteren om gegevens over wiskunde en PAV te verzamelen, terwijl men in het volgende schooljaar de focus legt op de taalvakken.

Ook van de secundaire scholen wordt de toegevoegde waarde op basis van de leerwinst van leerlingen (tussen twee toetsmomenten) openbaar gemaakt.

3.5.3 Gebruik op microniveau

Het primaire doel van de toetsen situeert zich op mesoniveau, de toetsen zijn low-stakes voor de leerlingen. Het is niet de bedoeling van de toetsontwikkelaars om summatieve uitspraken op leerlingniveau te doen. Wel laat het systeem toe om de geboekte leerwinst van de leerling of van een groep leerlingen in kaart te brengen, zodat de evolutie van elke leerling opgevolgd kan worden. Door de relatief lange periode tussen toetsafname en beschikbaarheid van de resultaten (door centrale verbetering van de toetsen, zie onder) zijn deze toetsen evenwel niet het meest geschikte middel om het onderwijs aan individuele leerlingen bij te sturen door bijvoorbeeld differentiatie.

Hoewel de toetsen hiertoe niet zijn opgemaakt, kan de school autonoom beslissen om een gewicht toe te kennen aan de resultaten op de toets (in het kader van attestering of oriëntering naar het volgende leerjaar).

Ook leraren krijgen inzicht in de leerprestaties van zowel individuele leerlingen als van de gehele klasgroep. Zo kan snel worden bekeken welke eindtermen of lesonderwerpen in volgende schooljaren extra aandacht verdienen of kunnen de lesmethoden in vraag worden gesteld.

3.5.4 Gebruik op mesoniveau

Het hoofddoel van dit scenario situeert zich op mesoniveau, met name het aanwenden van de resultaten van de gestandaardiseerde toetsen door de scholen in het kader van interne kwaliteitszorg op school. De resultaten op de gestandaardiseerde toetsen maken het mogelijk om betrouwbare gegevens te verzamelen op graad- en schoolniveau. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te gebruiken binnen hun interne kwaliteitszorg.

Bovendien kunnen op basis van deze resultaten scholen verplicht worden een verbeteringsplan op te stellen zodat de tekorten kunnen worden weggewerkt. De onderwijsinspectie houdt hier toezicht op.

Geaggregeerde resultaten op graad- en schoolniveau (betreffende de toegevoegde waarde op basis van geboekte leerwinst van leerlingen) worden openbaar gemaakt. Door het openbaar karakter van de toetsresultaten beschikken de scholen over informatie die het mogelijk maakt om de eigen kwaliteitszorg te vergelijken met die van andere gelijkaardige scholen. Het openbaar karakter van de resultaten van deze toetsen zorgt voor een high-stakes karakter op schoolniveau.

De onderwijsinspectie kan gebruik maken van de resultaten tijdens reguliere doorlichtingen. Enerzijds kan men nagaan welke processen de school hanteert om te komen tot de leerresultaten/leerwinst die blijkt uit de eindtoetsen. Anderzijds kan de onderwijsinspectie achterhalen op welke manier de school garanties biedt om de minder makkelijk te toetsen eindtermen te bereiken.

3.5.5 Gebruik op macroniveau

De overheid kan de gegevens enerzijds gebruiken om de prestaties van het eigen onderwijssysteem te evalueren, terwijl anderzijds informatie vergaard wordt over het bereiken van de eindtermen binnen de scholen. De verzamelde data kunnen worden aangewend voor onderzoeksdoeleinden.

3.5.6 Ontwikkeling en correctie

Om de betrouwbaarheid van de toetsen te garanderen, alsook omwille van de efficiënte werking inzake correctie, zijn de eindtoetsen schriftelijk van aard⁸.

De toetsitems worden jaarlijks volledig vernieuwd. De toetsen worden opgesteld door een hiertoe opgericht onafhankelijk toetsinstituut, onder toezicht van de overheid. Bij de toetsontwikkeling is maar een relatief klein aantal mensen betrokken, om zo de ontwikkeling beheersbaar te houden en om lekken te voorkomen. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit.

Om zowel de vergelijkbaarheid van de toetsen doorheen verschillende jaren als de leerwinstberekening te waarborgen, is het nodig dat items van opeenvolgende toetsen op één schaal worden ingeschaald (IRT)⁹.

De toetsresultaten van de leerlingen worden bepaald per leerdomein, op de mate waarin de leerlingen de eindtermen hebben behaald. Het gaat dus om een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

Omwille van het high-stakes karakter worden alle toetsen centraal verbeterd door hiertoe opgeleide evaluatoren. De afname van deze eindtoetsen dient daarom ook reeds in de maand mei te gebeuren. Er worden strikte procedures opgemaakt waarmee gewaakt wordt over de kwaliteit van de toetsontwikkeling én over de kwaliteit van de correctie van de toetsen.

Een nieuwe web-based tool wordt ontwikkeld, waarop de resultaten van de leerlingen worden bewaard. Deze software zorgt voor een snelle berekening en grafische weergave van de individuele leerlingresultaten, leerwinst en toegevoegde waarde, en wordt ook ingezet om data te genereren op hogere niveaus (regio, provincie, ...).

De individuele eindcijfers van de eindtoetsen alsook de individuele leerwinst worden zowel aan de leerling in kwestie als aan de schooldirectie bezorgd. Deze cijfers zijn niet publiek beschikbaar.

3.5.7 Beschikbaarheid van de gegevens

Via een te ontwikkelen web-tool worden de toetsresultaten, geaggregeerd op schoolniveau, openbaar gemaakt. De website maakt het percentage van leerlingen dat de vooropgestelde normen per graad per leerdomein bereikt van elke school openbaar. Er wordt enkel rekening gehouden met leerlingen die gedurende de twee opeenvolgende meetmomenten onderwijs hebben gevolgd in dezelfde school. Daarnaast wordt achtergrondinformatie verschaft over de leerlinginstroom,

⁸ Alternatieve piste 2: de schriftelijke toetsen (die handmatig moeten gecorrigeerd worden) worden vervangen door digitale toetsen. Deze hebben een grotere ontwikkelingskost, maar de correctie kan vaak automatisch gebeuren. In het basismodel van dit scenario wordt hier niet voor gekozen, omdat het dan moeilijk wordt om alle leerlingen in alle scholen de toetsen op hetzelfde ogenblik te laten afnemen (door gebrek aan ICT-infrastructuur), alsook omdat verschillen in de ICT-infrastructuur de betrouwbaarheid van de leerwinstmeting kunnen bedreigen. Beide genoemde nadelen worden meestal uitgesloten in een high-stakes scenario.

⁹ Alternatieve piste 3: Een inschaling op IRT maakt dat de leerwinstberekening volgens een gain score model kan gebeuren. Indien het voor de gebruikers/opdrachtgever volstaat om leerwinst te berekenen volgens een residual gain model, is IRT niet noodzakelijk. In dat laatste geval wordt de ontwikkelingskost van de gestandaardiseerde toetsen gereduceerd, maar verkrijgt men een minder zuiver beeld van de effectief geboekte leerwinst. De waarde van de toets overheen de jaren kan eventueel bewaakt worden door gebruik te maken van sample items die meerdere keren terugkomen.

onderwijsvisie en context van de school, alsook de meest recente inspectieverslagen. De gebruiker kan vijf scholen selecteren om hun prestaties onderling te vergelijken.

Hierbij wordt zowel in het basisonderwijs als het secundair onderwijs de gemiddelde leerwinst op de verschillende leerdomeinen weergegeven met behulp van lettercodes:

≥80% van de leerlingen behaalt de vooropgestelde norm ('sterk hoger dan verwachte leerwinst')

≥60% en <80% van de leerlingen behaalt de vooropgestelde norm ('hoger dan verwachte leerwinst')

>25% en <60% van de leerlingen behaalt de vooropgestelde norm ('verwachte leerwinst').

≤25% van de leerlingen behaalt de vooropgestelde norm ('minder dan verwachte leerwinst')

Er kan worden aangenomen dat een ruime meerderheid van de scholen voor de meeste leerdomeinen in de categorie 'verwachte leerwinst' terechtkomen.

3.5.8 Leerwinst en toegevoegde waarde

Er wordt aandacht besteed aan het berekenen van de geboekte leerwinst op basis van de verwachte resultaten voor elke leerling (categorical model).

De toegevoegde waarde wordt per graad weergegeven, op basis van het aandeel van leerlingen dat in een bepaald leerdomein een leerwinst heeft bereikt op of boven de verwachting (gain score model).

Bij de berekening van toegevoegde waarde wordt enkel rekening gehouden met leerlingen die gedurende de twee opeenvolgende meetmomenten onderwijs hebben gevolgd in dezelfde school.

3.5.9 Geraamde kostprijs

Bij de ontwikkeling van de eindtoetsen is een relatief klein aantal personen betrokken. Uiteraard moeten de items wel voldoende getoetst worden. De ontwikkelingskost van één kwaliteitsvolle gestandaardiseerde toets mét IRT wordt op deze manier geraamd op € 200.000 (zie Tabel 13). Voornamelijk het grote aantal verschillende toetsen dat voor het secundair onderwijs moet worden ontwikkeld (cf. verschillende studierichtingen), verhogen sterk de ontwikkelingskost. In de berekening wordt uitgegaan van 12 toetsen in het basisonderwijs, en respectievelijk 10, 20 en 30 verschillende eindtoetsen voor het tweede, vierde en laatste jaar van het secundair onderwijs (€ 14.400.000).

De afname van de toets is relatief duur doordat voor alle leerlingen een kopij moet worden voorzien, deze moet ter plaatse worden gebracht en moet terug verzonden worden naar een centrale plaats. De kost hiervan wordt geraamd op € 10 per leerling (6 cohorten van 70.000 leerlingen, dus € 4.200.000). Deze kost kan sterk gereduceerd worden door in te zetten op het digitaal aanbieden van de toetsen, maar daaraan zijn nadelige gevolgen verbonden in dit scenario.

Correctie: van ongeveer 420.000 leerlingen wordt een toets in minimaal vijf leerdomeinen afgenomen en centraal verbeterd. Deze 2.100.000 toetsen moeten op een korte periode handmatig worden gecorrigeerd, alsook moeten analyses gebeuren inzake interbeoordelaarsbetrouwbaarheid. Dit maakt dat veel evaluatoren moeten worden ingezet, die op voorhand ook sterk worden getraind.

Geschat wordt dat 400 evaluatoren gedurende vier maanden actief zullen zijn met correctie, onder toezicht van een tiental analisten¹⁰.

Inzake rapportering dient een gebruiksvriendelijke en interactieve website te worden opgebouwd (éénmalige kost). Daarnaast dienen de gegevens jaarlijks ingebracht te worden en dient onderhoud aan de website te gebeuren. Ten slotte dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten.

Tabel 13 - Raming van de kostprijs in scenario 5

	Eenmalige kost	Jaarlijks terugkerende kost
Ontwikkeling		€ 14.400.000
Afname		€ 4.200.000
Correctie		€ 10.000.000
Rapportering	€ 50.000	€ 10.000
Technische ondersteuning		
Algemene ondersteuning		€ 600.000
Totaal	€ 50.000	€ 29.210.000

3.6 Scenario 6 'Leerwinstmeting op schoolniveau met afgeschermd karakter'

Mesoniveau - verplicht - afgeschermd - ontwikkelingsgericht

3.6.1 Basisonderwijs

Er worden gestandaardiseerde toetsen ontwikkeld die betrekking hebben op de eindtermen en ontwikkelingsdoelen in diverse leerdomeinen, zoals wiskunde, Nederlands, Frans, wereldoriëntatie en informatieverwerking en -verwerking. Het doel van deze toetsen is om jaarlijks aan elke school betrouwbare informatie te leveren over de mate waarin de leerlingen de leerplandoelen en eindtermen beheersen. De toetsen zijn verplicht, maar de gegevens worden niet openbaar gemaakt. De toetsen zijn digitaal en worden binnen een vooraf bepaalde periode afgenomen (bijvoorbeeld de eerste twee weken van juni), waarin de school het precieze toetsmoment zelf kan bepalen¹¹.

De toetsen zijn zo breed mogelijk: ze brengen naast kennis ook vaardigheden van leerlingen in kaart. De toetsen worden van alle leerlingen aan het eind van elk leerjaar afgenomen. De toetsen van het eerste tot en met vijfde leerjaar zijn gebaseerd op de leerplannen, waardoor er variatie mogelijk is tussen verschillende onderwijsnetten. De toetsen van het zesde leerjaar zijn gebaseerd op de eindtermen en zijn daardoor hetzelfde voor alle leerlingen in heel Vlaanderen. Elk van deze toetsen is methode-onafhankelijk.

¹⁰Alternatieve piste 4: om de kosten te reduceren, is het mogelijk om de correctie van de toetsen in handen van de leraren te leggen. Het is dan noodzakelijk om te investeren in de professionalisering van leraren inzake evaluatie van toetsen. Neveneffecten als ongewenst strategisch gedrag zijn evenwel moeilijk te vermijden.

¹¹ Alternatieve piste 1: scholen met een beperkte infrastructuur kunnen een aanvraag indienen om een aantal laptops uit te lenen bij de beheerder van het systeem. De spreiding van de periode van de toetsafname draagt bij tot de haalbaarheid van deze dienst.

Op basis van de leerwinst die leerlingen boeken tussen twee toetsmomenten, wordt de toegevoegde waarde van de school berekend.

3.6.2 Secundair onderwijs

In het secundair onderwijs worden eveneens digitale gestandaardiseerde toetsen ontwikkeld. Deze worden eenmaal per graad verplicht afgenomen bij alle leerlingen. Net als in het basisonderwijs, bepalen scholen zelf het precieze moment van toetsafname binnen een vooraf bepaalde periode aan het eind van het schooljaar. Het doel van deze gestandaardiseerde toetsen is het verzamelen van gegevens over het aantal leerlingen dat de eindtermen binnen een bepaald domein onder de knie heeft.

Deze toetsen zijn gebaseerd op de eindtermen, of waar deze ontbreken, op de leerplandoelen en zijn in heel Vlaanderen hetzelfde voor alle leerlingen in dezelfde studierichting van dezelfde graad. De volgende leerdomeinen die zowel in ASO, KSO als TSO aan bod komen, worden getoetst: wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne vreemde talen (Engels en Frans). In het BSO wordt PAV (project algemene vakken) geëvalueerd. Bijkomende vakken die geëvalueerd worden, zijn afhankelijk van de gevolgde studierichting¹². De toetsen meten, waar mogelijk, zowel kennis als vaardigheden van leerlingen.

Ook van de secundaire scholen wordt de toegevoegde waarde op basis van de leerwinst van leerlingen (tussen twee toetsmomenten) berekend.

3.6.3 Gebruik op microniveau

Het primaire doel van de toetsen situeert zich op mesoniveau, de toetsen zijn low-stakes voor de leerlingen. Het is niet de bedoeling van de toetsontwikkelaars om uitspraken op leerlingniveau te doen. Wel laat het systeem toe om de geboekte leerwinst van de leerling of van een groep leerlingen in kaart te brengen, zodat de evolutie van elke leerling opgevolgd kan worden. Op basis hiervan kunnen de tekortkomingen worden weggewerkt.

Hoewel de toetsen hiertoe niet zijn opgemaakt, kan de school autonoom beslissen om een gewicht toe te kennen aan de resultaten op de toets (in het kader van attestering of oriëntering naar het volgende leerjaar).

Ook leraren krijgen inzicht in de leerprestaties van zowel individuele leerlingen als van de gehele klasgroep. Zo kan snel worden bekeken welke eindtermen of lesonderwerpen in volgende schooljaren extra aandacht verdienen of kunnen de lesmethoden in vraag worden gesteld.

3.6.4 Gebruik op mesoniveau

Het hoofddoel van dit scenario situeert zich op mesoniveau, aangezien de resultaten van de gestandaardiseerde toetsen door de scholen worden aangewend in het kader van interne kwaliteitszorg op school. Doordat de resultaten van de toets in geen geval publiekelijk beschikbaar zijn, hebben de toetsen een low-stakes-karakter op schoolniveau. De resultaten op de gestandaardiseerde toetsen verschaffen een belangrijke informatiebron voor het functioneren op

¹² Alternatieve piste 2: om de ontwikkelingskosten te reduceren, kan het opportuun zijn om niet elk schooljaar dezelfde leerdomeinen te evalueren. Zo kan men het ene jaar opteren om gegevens over wiskunde en PAV te verzamelen, terwijl men in het volgende schooljaar de focus legt op de taalvakken.

klas- en schoolniveau. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te gebruiken binnen hun interne kwaliteitszorg.

De onderwijsinspectie kan gebruik maken van de resultaten tijdens reguliere doorlichtingen, maar enkel indien de school inzagerecht verleent. Enerzijds kan men nagaan welke processen de school hanteert om te komen tot de leerresultaten/leerwinst die blijkt uit de eindtoetsen. Anderzijds kan de onderwijsinspectie achterhalen op welke manier de school garanties biedt om de minder makkelijk te toetsen eindtermen te bereiken¹³.

3.6.5 Ontwikkeling en correctie

Ondanks het verplichte karakter, zal het feit dat de toetsresultaten te allen tijde worden afgeschermd, ertoe bijdragen dat scholen minder geneigd zijn om de toetsen aan te wenden voor oneigenlijk gebruik. Hierdoor verleent het systeem zich ertoe om de periode van afname over enkele weken te spreiden en aldus om de toetsen digitaal aan te bieden. Een web-based tool wordt ontwikkeld, waarop de resultaten van de leerlingen worden bewaard en automatisch gecorrigeerd. Deze software zorgt eveneens voor een snelle berekening van de individuele leerlingenresultaten en de individuele leerwinst, maar wordt ook ingezet om data te genereren op klas- en schoolniveau.

Jaarlijks worden de toetsitems geëvalueerd (bijvoorbeeld inzake inhoudsvaliditeit, verschuivingen in leerplannen, alsook inzake item shifting) en gedeeltelijk hernieuwd. De toets wordt opgesteld door een hiertoe opgericht onafhankelijk toetsinstituut, onder toezicht van de overheid. Bij de toetsontwikkeling is een panel van ervaren leraren betrokken. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit.

Om de vergelijkbaarheid van de toetsen alsook de leerwinstberekening doorheen verschillende jaren te waarborgen, is het nodig dat items van opeenvolgende toetsen op één schaal worden ingeschaald (IRT)¹⁴.

De toetsresultaten van de leerlingen worden bepaald per leerdomein, op de mate waarin de leerlingen de eindtermen hebben behaald. Het gaat dus om een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

3.6.6 Beschikbaarheid van de gegevens

Via de te ontwikkelen web-tool worden de toetsen automatisch gecorrigeerd. De resultaten zijn enkel toegankelijk voor de school zelf op zowel schoolniveau als klasniveau. Door het ontwikkelingsgericht karakter is het belangrijk dat scholen en leraren een correct beeld van zichzelf krijgen.

Scholen beslissen zelf of men informatie over de leerprestaties zal doorgeven aan de leerling en de ouders van het kind.

Zoals eerder gesteld beschikt de onderwijsinspectie enkel over de toetsresultaten wanneer de school inzagerecht verleent. De onderwijsinspectie kan dan deze resultaten aanwenden om na te

¹³ Alternatieve piste 3: de onderwijsinspectie bevraagt niet de resultaten van de toets, maar gaat wel na hoe ze door de school in het kader van interne kwaliteitszorg worden gebruikt.

¹⁴ Alternatieve piste 4: een inschaling op IRT maakt dat de leerwinstberekening volgens een gain score model kan gebeuren indien het voor de gebruikers/opdrachtgever volstaat om een leerwinst te berekenen volgens een residual gain model is IRT niet noodzakelijk. In dat laatste geval wordt de ontwikkelingskost van de gestandaardiseerde toetsen gereduceerd, maar verkrijgt men een minder zuiver beeld van de effectief geboekte leerwinst.

gaan hoe deze worden gebruikt binnen het kader van kwaliteitsbewaking. Daarnaast is elke school vrij om ook andere informatie omtrent het gebruik van de toets en de resultaten aan te leveren. Deze informatie mag door de inspectie echter niet worden vrijgegeven.

3.6.7 Leerwinst en toegevoegde waarde

Vanuit het kwaliteitsgerichte oogpunt wordt aandacht besteed aan het berekenen van de geboekte leerwinst op basis van de verwachte resultaten voor elke leerling (residual gain model).

Bij de berekening van de toegevoegde waarde van de school wordt expliciet niet uitgegaan van de leerprestaties of leerwinst van leerlingen in anderen scholen, omdat het belangrijk is dat scholen binnen het kader van interne kwaliteitszorg rekening houden met het zelf- en standaardcriterium en zich in mindere mate laten leiden door de prestaties in andere scholen¹⁵. Een maat voor toegevoegde waarde bestaat daarom uit een statusmodel waarbij de school de leerprestaties en leerwinst van haar leerlingen doorheen de tijd (verschillende cohorten) met elkaar vergelijkt (cohort-to-cohort-perspectief).

Bij de berekening van toegevoegde waarde wordt enkel rekening gehouden met leerlingen die gedurende de twee opeenvolgende meetmomenten onderwijs hebben gevolgd in dezelfde school.

3.6.8 Geraamde kostprijs

Bij de ontwikkeling van de eindtoetsen is een relatief klein aantal personen betrokken, waardoor de ontwikkelingskost relatief kan gedrukt worden. Uiteraard moeten de items wel voldoende getoetst worden. De ontwikkelingskost van één low-stakes digitale gestandaardiseerde toets met IRT wordt op deze manier geraamd op € 120.000 (zie Tabel 14).

Tabel 14 - Raming van de kostprijs in scenario 6

	Eenmalige kost	Jaarlijks terugkerende kost
Ontwikkeling	€ 8.640.000	€ 200.000
Afname	€ 600.000	€ 50.000
Correctie		
Rapportering		
Technische ondersteuning		€ 100.000
Algemene ondersteuning		€ 600.000
Totaal	€ 9.240.000	€ 950.000

Voornamelijk het grote aantal verschillende toetsen dat voor het secundair onderwijs moet worden ontwikkeld (cf. verschillende studierichtingen), verhogen sterk de ontwikkelingskost. In de berekening wordt uitgegaan van 12 toetsen in het basisonderwijs, en respectievelijk 10, 20 en 30 verschillende eindtoetsen voor het tweede, vierde en laatste jaar van het secundair onderwijs (=€8.640.000). Omwille van jaarlijkse vernieuwing en kwaliteitscontrole wordt een jaarlijks budget van €200.000 voorzien.

¹⁵ Alternatieve piste 5: De toegevoegde waarde kan worden berekend volgens een residual gain model zodat scholen toch de mogelijkheid krijgen om zich te spiegelen tegenover scholen met gelijkaardige leerlingenkenmerken.

Er dient éénmalig een gebruiksvriendelijke en interactieve web-tool te worden opgebouwd (€ 600.000). Daarnaast dient onderhoud aan de website te gebeuren en dienen nieuwe items te worden ingebracht (€ 50.000). Deze web-tool zorgt ook voor correctie en rapportering, waardoor op deze posten geen extra kosten worden voorzien. Doordat de toets digitaal aan te bieden, valt een grote afname- en correctiekost weg.

Ten slotte is een technische ondersteuning voor scholen tijdens de afname van de toetsen (€ 100.000), en dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten (€ 600.000).

3.7 Scenario 7 'Facultatieve toetsenbank'

Mesoniveau - facultatief - toetsenbank - ontwikkelingsgericht

3.7.1 Basisonderwijs

Er wordt een toetsendatabank ontwikkeld waarin gestandaardiseerde toetsen voor leerlingen van de lagere school worden aangeboden. De school kiest zelf welke toetsen worden afgenomen bij de leerlingen. Deze toetsen zijn beschikbaar voor alle leerjaren van de lagere school. De toetsen hebben betrekking op de eindtermen en ontwikkelingsdoelen in diverse leergebieden, zoals wiskunde, Nederlands, Frans en wereldoriëntatie. Deze toetsen zijn methode-onafhankelijk en kunnen worden ingezet op het moment dat de school zich daartoe geroepen voelt.

Daarnaast zijn in de toetsendatabank ook observatie- en meetinstrumenten terug te vinden die betrekking hebben op de leergebiedoverschrijdende eindtermen (cf. ICT, leren leren en sociale vaardigheden) en kan ook worden uitgebouwd met socio-emotionele vragenlijsten (welbevinden).

Het gebruik van de toetsen-databank wordt niet verplicht. Elke school selecteert zelf de toetsen die aan bod zullen komen, en de frequentie waarmee deze toetsen worden afgenomen.

3.7.2 Secundair onderwijs

Ook in het secundair onderwijs wordt een databank ontwikkeld waarin curriculumafhankelijke en -onafhankelijke formatieve toetsen voor leerlingen worden aangeboden in de verschillende studierichtingen.

Om tegemoet te komen aan de rijkheid aan leerdomeinen in het secundair onderwijs, is het wenselijk om een uitgebreid aanbod van gestandaardiseerde toetsen te voorzien in elk van de leerdomeinen. Er worden toetsen ontwikkeld voor alle leerlingen in dezelfde studierichting van dezelfde graad, met de volgende leerdomeinen die zowel in ASO, KSO als TSO aan bod komen: wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne vreemde talen (Engels en Frans). In het BSO wordt PAV (project algemene vakken) geëvalueerd. Bijkomende vakken die geëvalueerd worden, zijn afhankelijk van de gevolgde studierichting. Naast de cognitieve toetsen wordt ook in het secundair onderwijs het aanbod uitgebreid met observatie- en meetinstrumenten die betrekking hebben op zowel vakoverschrijdende eindtermen als socio-emotionele aspecten en aanvullend beroepsmatige competenties van de leerlingen.

De school selecteert zelf de toetsen die zij wenst af te nemen bij haar leerlingen. Het gebruik van de toetsenbank is geen geval verplicht, maar zal wel door de overheid gestimuleerd worden.

3.7.3 Gebruik op microniveau

Het gebruik van de toetsenbank zorgt ervoor dat de capaciteiten van de leerlingen op een betrouwbare en valide manier in beeld worden gebracht. Het is echter niet de bedoeling om uitspraken op microniveau te doen. Omdat het doel expliciet ligt op mesoniveau, wordt sterk ontraden om de resultaten te laten meetellen binnen een summatieve beoordeling van leerlingen.

Indien de school ervoor kiest om een bepaalde toets uit de toetsenbank in een vroeg stadium van het leerproces af te nemen, kunnen de resultaten van de toetsen leraren informeren over de mate waarin de groep leerlingen bepaalde vooropgezette eindtermen bereikt. Op basis van deze resultaten kan de leraar het onderwijsleerproces bijsturen of een bepaald leerstofonderdeel opnieuw aanbieden, zodat de kwaliteit systematisch verhoogd wordt. De toetsen zijn evenwel bedoeld om de eindtermen binnen een volledig leerdomein in kaart te brengen, en zijn daarom mogelijk te weinig gedetailleerd voor dit doel op microniveau.

3.7.4 Gebruik op mesoniveau

De toetsen in de databank worden in de eerste plaats ontwikkeld in het kader van het ondersteunen van de interne kwaliteitszorg van scholen. Er wordt voor scholen algemene ondersteuning voorzien om de resultaten te gebruiken binnen hun interne kwaliteitszorg: deze hulp kan zowel nodig zijn bij de selectie van toetsen uit de databank, als bij de interpretatie van resultaten en het omzetten van deze actieplannen naar actieplannen. Doordat het gebruik van de toetsenbank niet verplicht wordt en doordat de resultaten niet openbaar worden gemaakt, kunnen scholen de toetsen als low-stakes beschouwen. Vanuit het beleid is het dan zaak om het belang van de toetsen als kwaliteitsindicator te verduidelijken. Het is immers het primaire opzet van de toetsenbank om het kwaliteitsproces op de scholen te optimaliseren vanuit objectief vastgestelde knelpunten.

De onderwijsinspectie kan enkel over de gegevens van de school beschikken (welke toetsen werden afgenomen, en welke resultaten werden behaald) op voorwaarde dat de school deze gegevens ter beschikking stelt. Ook in dat geval kan de inspectie deze uitsluitend gebruiken om na te gaan hoe de school de resultaten heeft gebruikt binnen haar proces van interne kwaliteitszorg. Daarnaast kan de school deze toetsen ook gebruiken om aan te tonen dat de school aandacht schenkt aan vakoverschrijdende eindtermen.

3.7.5 Ontwikkeling en correctie

Omdat de school zelf toetsen uit de databank selecteert en deze afneemt bij de leerlingen op een tijdstip naar keuze kunnen de toetsen digitaal ter beschikking worden gesteld aan die scholen die voldoende infrastructuur hebben om leerlingen individueel aan het werk te zetten. De toetsen worden dan ook automatisch gecorrigeerd. Scholen die dit wensen, kunnen de toetsen schriftelijk afnemen, al is het noodzakelijk dat de antwoorden gedigitaliseerd worden om het monitoringssysteem up-to-date te houden. Een nieuwe web-based tool wordt speciaal hiervoor ontwikkeld, zodat leraren de (toets)resultaten rechtstreeks kunnen invullen. Op basis van de resultaten wordt een individueel schoolfeedbackrapport opgesteld.

De toetsitems worden regelmatig vernieuwd en worden opgesteld door een hiertoe opgericht onafhankelijk toetsinstituut, onder toezicht van de overheid. Bij de toetsontwikkeling is een relatief klein aantal mensen betrokken, om zo de ontwikkeling beheersbaar te houden. Uiteraard dienen de items wel voldoende getoetst te worden op bijvoorbeeld inhoudsvaliditeit. Er worden strikte

procedures opgemaakt waarmee gewaakt wordt over de kwaliteit van de toetsontwikkeling én over de kwaliteit van de correctie van de toetsen.

De toetsresultaten van de leerlingen worden geëvalueerd ten opzichte van een standaard-criterium, waarbij de score van leerlingen niet afhankelijk is van de score van andere leerlingen.

3.7.6 Beschikbaarheid van de gegevens

De school kan beschikken over de resultaten van de leerlingen op schoolniveau (directie) en microniveau (klas- of vakleraar). Door het ontwikkelingsgericht karakter is het belangrijk dat scholen en leraren een correct beeld van zichzelf krijgen. Daarnaast worden ook gemiddelde scores en betrouwbaarheidsintervallen berekend. Scholen kunnen de resultaten ook opsplitsen per doelgroep (klas, zittenblijvers, anderstalige nieuwkomers, ...). De school beslist zelf of de individuele leerprestaties worden meegedeeld aan de leerlingen en de ouders.

De inspectie mag over de toetsresultaten beschikken om na te gaan hoe deze worden gebruikt binnen het kader van kwaliteitsbewaking, maar enkel indien de school inzagerecht verleent. Daarnaast is elke school vrij om ook andere informatie omtrent het gebruik van de toets en de resultaten aan te leveren. Deze informatie mag door de inspectie echter niet worden vrijgegeven.

3.7.7 Leerwinst en toegevoegde waarde

Standaard wordt er geen leerwinst berekend, omdat deze toetsenbank expliciet gericht is op het mesoniveau, alsook omdat scholen niet altijd verschillende toetsafnames hebben van een zelfde toets¹⁶. Ook de toegevoegde waarde van scholen wordt in dit scenario niet in vergelijking met andere scholen berekend. Wel kan een school, die een zelfde toets doorheen verschillende jaren afneemt (van verschillende cohorten van leerlingen), een schoolfeedbackrapport aanvragen om de eigen evolutie in kaart brengen (=status-model, 'cohort-to-cohort-model')¹⁷.

3.7.8 Geraamde kostprijs

Bij de ontwikkeling van de eindtoetsen is een relatief klein aantal personen betrokken, waardoor de ontwikkelingskost relatief kan gedrukt worden. Uiteraard moeten de items wel voldoende getoetst worden. De ontwikkelingskost van één low-stakes kwaliteitsvolle gestandaardiseerde toets wordt op deze manier geraamd op €60.000 (zie Tabel 15). Voornamelijk het grote aantal verschillende toetsen dat voor het secundair onderwijs moet worden ontwikkeld (cf. verschillende studierichtingen), verhogen sterk de ontwikkelingskost. In de berekening wordt uitgegaan van een gradueel opbouwend aantal toetsen in elk van de leerjaren in het basisonderwijs (totaal 40 toetsen), en respectievelijk 15, 25 en 35 verschillende toetsen voor de eerste, tweede en derde graad van het secundair onderwijs (totale kost € 6.900.000).

¹⁶ Zie alternatieve piste 1: de leerwinst wordt berekend (gain score model): de leerwinst van alle leerlingen wordt berekend op basis van de verschillscore tussen twee toetsen die op eenzelfde schaal zijn gescoord. Meestal wordt de verwachte leerprestatie bepaald op basis van één voorgaande meting van het leerniveau. Op basis van deze gegevens kan objectief worden vastgesteld of de geboekte leerwinst conform de verwachtingen is.

¹⁷ Alternatieve piste 2: er kan ook voor gekozen worden om scholen toch te vergelijken met andere scholen. Het betreft ook in dit geval evenwel nog steeds een 'statusmodel' van toegevoegde waarde, aangezien deze toegevoegde waarde op leerprestaties en niet op basis van leerwinst wordt berekend.

Tabel 15 - Raming van de kostprijs in scenario 7

	Eenmalige kost	Jaarlijks terugkerende kost
Ontwikkeling	€6.900.000	€ 200.000
Afname	€ 600.000	€ 50.000
Correctie		
Rapportering		
Technische ondersteuning		€ 100.000
Algemene ondersteuning		€ 600.000
Totaal	€ 7.500.000	€ 950.000

Doordat de toetsenbank al kan worden opgestart met enkele basismogelijkheden, kunnen de kosten gespreid worden over meerdere jaren. Daarna dienen er om de vijf jaar kwaliteitscontroles te worden toegepast op de evaluatie-instrumenten, waardoor een jaarlijkse kost van € 200.000 bestaat.

Er dient éénmalig een gebruiksvriendelijke en interactieve web-tool te worden opgebouwd (€ 600.000). Daarnaast dient onderhoud aan de website te gebeuren en dienen nieuwe items te worden ingebracht (€ 50.000). Deze web-tool zorgt ook voor correctie en rapportering, waardoor op deze posten geen extra kosten worden voorzien. Doordat de toets digitaal aan te bieden, valt een grote afname- en correctiekost weg.

Ten slotte is een technische ondersteuning voor scholen tijdens de afname van de toetsen (€ 100.000), en dienen scholen algemene ondersteuning te krijgen bij de interpretatie van de resultaten (€ 600.000).

3.8 Scenario 8 'Huidige trend wordt verdergezet'

Macroniveau - huidige situatie

3.8.1 Basisonderwijs en secundair onderwijs

Het decretale kader wordt niet uitgebreid. Elke school vult de wijze waarop de eindtermen geëvalueerd worden autonoom in. Leraren zijn zelf verantwoordelijk voor het selecteren en/of ontwikkelen van de toetsen en eventuele proeven en corrigeren deze ook. De klassenraad beslist of de leerling het schooljaar met vrucht heeft doorlopen. Er worden door de overheid geen bijkomende gestandaardiseerde toetsen voorzien.

Scholen kunnen geselecteerd worden om deel te nemen aan de peilingen¹⁸, maar kunnen ook gebruik maken van het bestaande aanbod van gestandaardiseerde toetsen, waaronder de paralleltoetsen, de zogenaamde OVSG-toetsen of interdiocesane proeven van Katholiek Onderwijs

¹⁸ Alternatieve piste 1: de steekproef van de peilingen wordt uitgebreid zodat meer scholen betrokken worden bij het genereren van data. Het deelnemen aan de peilingen wordt verplicht voor scholen. Door systematisch dezelfde cohort leerlingen te laten deelnemen, kan de toegevoegde waarde voor de school alsook de geboekte leerwinst beter in kaart worden gebracht. De kosten voor de afname en correctie zullen dan wel verhoogd worden.

Vlaanderen¹⁹. Daarnaast kunnen scholen ook beroep doen op leerlingvolgsystemen²⁰ of op het Centrum voor Schoolfeedback.

3.8.2 Inzetten op randvoorwaarden

Er wordt geïnvesteerd in bijkomende ondersteuning van scholen in het gebruikmaken van de resultaten van interne evaluatie, van gestandaardiseerde toetsen of van andere informatiebronnen in het kader van interne kwaliteitszorg. Deze ondersteuning kan intern (bijvoorbeeld via professionalisering van schoolbesturen), en/of extern via bestaande of nieuwe kanalen worden georganiseerd.

Daarnaast is er bijkomende aandacht voor de professionalisering van scholen en schoolbesturen inzake datageletterdheid. Er worden tevens initiatieven ontwikkeld die bijdragen tot het creëren van een cultuur waarin data efficiënt kunnen worden ingezet, of waarin het verzamelen en gebruiken van data wordt ingebed in de dagelijkse werking van de scholen.

3.8.3 Geraamde kostprijs

De kostprijs is sterk afhankelijk van de geïnvesteerde middelen in ondersteuning.

¹⁹ Alternatieve piste 2: om de vergelijkbaarheid van de koepelgebonden toetsen doorheen verschillende jaren te waarborgen, is het mogelijk om items van opeenvolgende toetsen op één schaal worden ingeschaald (IRT). Daarnaast is het mogelijk om een samenwerkingsverband tussen de diverse koepels op te richten zodat er jaarlijks één leerplanonafhankelijke doch eindtermgebonden eindtoets kan worden opgesteld die wordt afgenomen bij alle leerlingen in het zesde jaar basisonderwijs.

²⁰ Alternatieve piste 3: het gebruiken van een leerlingvolgsysteem kan verplicht worden op schoolniveau. Op die manier kunnen scholen op een meer systematische wijze informatie verzamelen over zowel de leerprestaties van de leerling als over de geboden onderwijskwaliteit. Dit kan dan bijdragen bij tot een efficiëntere sturing van de interne kwaliteitszorg op school.

DEEL 7 INSCHATTING VAN WENSELIJKHEID EN HAALBAARHEID VAN DE SCENARIO'S

Op basis van verschillende informatiebronnen die werden gerapporteerd in Deel 3, Deel 4 en Deel 5 (respectievelijk de literatuurstudie, de eerste en tweede ronde van de Delphi-studie, en het internationale onderzoeksluik) werden acht concrete scenario's uitgewerkt voor het Vlaamse onderwijs (zie Deel 6). In de derde ronde van de Delphi-studie werd bij elk van de 24 respondenten gepeild naar de standpunten inzake de wenselijkheid en haalbaarheid van elk van deze acht scenario's. Deze standpunten werden geregistreerd op twee manieren. Allereerst kregen respondenten de vraag om voor elk scenario aan te geven in welke mate zij dit scenario wenselijk en haalbaar achtten. Voor het geven van antwoord op de stellingen werd een Likertschaal gehanteerd die loopt van 1 = helemaal mee oneens, 2 = enigszins oneens, 3 = neutraal, 4 = enigszins eens en 5 = helemaal mee eens. Vervolgens kregen de respondenten de mogelijkheid om hun standpunt verder te verduidelijken en te motiveren in een interview.

In wat volgt, zal een overzicht gegeven worden van de standpunten van de respondenten met betrekking tot de verschillende scenario's. Eerst worden de kwantitatieve resultaten per scenario kort verduidelijkt. Daarna worden de positieve en negatieve effecten die de scenario's volgens de respondenten teweeg zouden brengen, toegelicht. Tot slot wordt de inschatting van de haalbaarheid verhelderd.

1 ALGEMENE BESPREKING

In Tabel 16 en Tabel 17 wordt een overzicht gegeven van de kwantitatieve beoordeling van respectievelijk de wenselijkheid en de haalbaarheid van de acht scenario's door de respondenten per categorie. Hoewel we benadrukken dat deze cijfers geen absoluut karakter mogen krijgen - in de interviews werden immers zeer veel nuances opgetekend die het toegekende cijfer een bepaalde betekenis toekennen - geven ze wel een eerste indicatie van de globale wenselijkheid van de acht scenario's.

1.1 Wenselijkheid

Vier scenario's kregen een gemiddelde score²¹ boven de neutrale score '3'. Het betreft de vier scenario's waarin een leer(winst)monitoringssysteem wordt ontwikkeld in een ontwikkelingsgericht perspectief. Drie scenario's kregen een gemiddelde score die duidelijk onder de neutrale score '3' ligt. Het betreft de drie scenario's waarbij een leer(winst)monitoringssysteem expliciet in een verantwoordingsgericht perspectief wordt geplaatst. We kunnen hieruit concluderen dat de

²¹ Het betreft het gemiddelde van de verschillende categorieën van respondenten. Elke categorie weegt dus even zwaar door, onafhankelijk van het aantal respondenten in de betreffende categorie.

implementatie van een leer(winst)monitoringssysteem in de Vlaamse onderwijscontext het meeste kans maakt op een breed draagvlak indien het gericht is op een ontwikkelingsgericht doel op micro- of mesoniveau. Het andere spanningsveld dat in Deel 6 als 'primaire bouwsteen' werd onderscheiden (verplichte versus facultatieve deelname van scholen) blijkt geen scherprechter te zijn in de beoordeling van de wenselijkheid door de 24 respondenten. In verband met scenario 8 is het moeilijk om een eenduidige conclusie te trekken inzake wenselijkheid: negen respondenten beoordelen dit scenario als (eerder) wenselijk, maar elf respondenten menen dat dit scenario (eerder) onwenselijk is.

Tabel 16 - Wenselijkheid van de verschillende scenario's

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde (M)
Scenario 1. High-stakes eindtoetsen voor leerlingen	2,00	1,00	3,38	2,50	2,22
Scenario 2. Adaptief leerwinstmonitoringssysteem	4,25	3,67	4,25	3,17	3,83
Scenario 3. Diagnostisch meetinstrumentarium & toetsenbank	3,25	3,33	4,00	3,67	3,56
Scenario 4. High-stakes eindtoetsen voor leerlingen en scholen	1,50	1,00	1,44	2,17	1,53
Scenario 5. Leerwinstmeting op schoolniveau met openbaar karakter	1,63	1,33	1,88	2,00	1,71
Scenario 6. Leerwinstmeting op schoolniveau met afgeschermd karakter	3,75	2,83	4,13	3,00	3,43
Scenario 7. Facultatieve toetsenbank	2,88	3,83	3,13	4,20	3,51
Scenario 8. Huidige trend wordt verdergezet	2,00	4,40	2,00	3,50	2,98

Tabel 17 - Haalbaarheid van de verschillende scenario's

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhouddeskundigen	Gemiddelde (M)
Scenario 1. High-stakes eindtoetsen voor leerlingen	2,00	2,17	2,88	2,33	2,34
Scenario 2. Adaptief leerwinstmonitoringssysteem	4,25	3,83	4,25	3,17	3,88
Scenario 3. Diagnostisch meetinstrumentarium & toetsenbank	4,75	3,67	4,25	2,67	3,83
Scenario 4. High-stakes eindtoetsen voor leerlingen en scholen	2,00	1,67	2,75	2,00	2,10
Scenario 5. Leerwinstmeting op schoolniveau met openbaar karakter	1,63	1,67	2,31	2,33	1,98
Scenario 6. Leerwinstmeting op schoolniveau met afgeschermd karakter	3,50	3,00	4,00	3,17	3,42
Scenario 7. Facultatieve toetsenbank	4,25	3,50	4,13	3,80	3,92
Scenario 8. Huidige trend wordt verdergezet	4,00	5,00	4,13	5,00	4,53

Verder valt op dat de vier scenario's die eerder positief worden beoordeeld inzake wenselijkheid, de twee scenario's omvat waarin een facultatieve toetsenbank (op microniveau ten behoeve van leerlingen, en op mesoniveau ten behoeve van scholen) wordt ontwikkeld, en de twee scenario's waarin leerwinstmonitoring wordt voorgesteld in een ontwikkelingsgericht perspectief. Belangrijke criteria die door vele respondenten worden gehanteerd bij de beoordeling van de wenselijkheid, zijn de mate waarin het leer(winst)monitoringssysteem geacht wordt een bijdrage te kunnen leveren tot het leren van leerlingen (met onderwijsprocessen op maat van elke leerling) en tot de interne kwaliteitszorg van de scholen.

1.2 Haalbaarheid

In de inschatting van haalbaarheid spelen verschillende overwegingen mee. Meest prominent komen evenwel de volgende drie criteria naar boven in de inschatting van de 24 respondenten: de mate waarin men meent dat voor het betreffende scenario een draagvlak bestaat in de Vlaamse onderwijscontext; de inschatting van de haalbaarheid van de vereiste inspanningen om het betreffende leer(winst)monitoringssysteem te ontwikkelen; en de inschatting van de vereiste

financiële investering. Voornamelijk het eerste criterium maakt dat de scores inzake haalbaarheid een sterke samenhang vertonen met de scores inzake wenselijkheid. De drie scenario's die gemiddeld als onwenselijk worden bestempeld, worden ook eerder onhaalbaar geacht, onder meer omwille van het veronderstelde beperkte draagvlak. Enkel scenario 8 vormt een uitzondering: hoewel de wenselijkheid gemiddeld als neutraal werd bestempeld, wordt de realisatie ervan als zeer haalbaar geacht. Aangezien dit scenario geen ontwikkeling van een nieuw leer(winst)monitoringssysteem voorziet, is de vereiste inspanning immers veel kleiner dan in de andere scenario's.

1.3 Verschillen tussen categorieën van respondenten

Verder zien we enkele opvallende verschillen tussen de categorieën van respondenten. Deze komen het meest tot uiting in de beoordeling van de wenselijkheid van scenario 1 en scenario 8. Scenario 1 wordt door elke respondent in de categorie 'onderwijsverstrekkers' als helemaal onwenselijk beschreven, terwijl vijf respondenten uit scholen dit scenario wel wenselijk tot zeer wenselijk achten. Dit reflecteert het verschil in opinie dat reeds in Deel 4 werd beschreven, met name dat respondenten uit de categorie 'scholen' meer dan onderwijsverstrekkers zich kunnen scharen achter het idee dat een leer(winst)monitoringssysteem deels in een verantwoordingsgericht perspectief op microniveau wordt ingezet. Scenario 8 - de huidige trend wordt verdergezet - wordt dan weer sterk gesteund door de onderwijsverstrekkers. Ook dit reflecteert de verschillen in opinies uit Deel 4, waar de drie respondenten die meenden dat Vlaanderen niet bijkomend moet investeren in een leer(winst)monitoringssysteem omwille van de verwachte neveneffecten, alle drie tot de categorie 'onderwijsverstrekkers' behoorden. In de categorie 'beleidsmakers' en 'scholen' wordt dit scenario evenwel veel minder gedragen.

2 BESPREKING PER SCENARIO

In deze tweede sectie bespreken we de resultaten van de derde ronde van de Delphi-studie meer in detail. We zoomen in op elk van de acht scenario's die werden ontwikkeld in Deel 6 van dit rapport.

2.1 High-stakes eindtoetsen voor leerlingen

Tabel 18 - Wenselijkheid en haalbaarheid van scenario 1

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	2.00	1.00	3.38	2.50	2.22
Haalbaarheid	2.00	2.17	2.88	2.33	2.34

2.1.1 Wenselijkheid

Over het algemeen wijzen de resultaten in de bovenstaande tabel uit dat dit scenario door de respondenten gemiddeld als eerder onwenselijk wordt ingeschat ($M=2.22$). Hoewel respondenten ook een groot aantal positieve effecten benoemen bij dit scenario, wegen deze gepercipieerde voordelen voor 14 van de 24 respondenten niet op tegen enkele belangrijke onwenselijke

neveneffecten. Zes respondenten beschouwen dit scenario als eerder onwenselijk, terwijl de vier overblijvende respondenten dit scenario wenselijk, noch onwenselijk achten.

“En jullie hadden zelf bij dit scenario ook een aantal nadelen en neveneffecten meegegeven waar dat ik mij zeker in kon vinden en dus in dat opzicht was mijn conclusie van ja... de voordelen, de gegevens die ge er kunt uithalen, wegen voor mij niet op tegen die nadelen.” (beleidsmakers, respondent 1)

2.1.2 Positieve effecten

Hoewel de meerderheid van de respondenten dit scenario als eerder onwenselijk inschat, gaan verschillende respondenten akkoord dat dit scenario meerdere positieve aspecten bevat. Ze doelen daarbij op het feit dat dit scenario een objectieve evaluatie van leerlingen ten goede kan komen, dat scholen min of meer ‘gedwongen’ worden om onderwijs van degelijke kwaliteit aan te bieden en dat de leerwinstmeting zowel voor de leerling als voor de school een meerwaarde kan opleveren.

Een objectieve evaluatie van de leerling

Het is opvallend dat vier van de zes respondenten die dit scenario als wenselijk beoordelen afkomstig zijn uit de categorie ‘scholen’. Volgens deze respondenten biedt dit scenario dan ook kansen om de kwaliteit van evaluaties te garanderen, zodat leerlingen op een uniforme manier geëvalueerd worden. Respondenten menen dat het scenario tegemoet komt aan het algemeen aanvaarde principe van ‘een gelijke lat voor iedereen’.

“Het voordeel daar aan is dat je alle leerlingen op een gelijke manier kunt beoordelen.” (beleidsmakers, respondent 1)

Een aantal respondenten is er dan ook sterk van overtuigd dat gevalideerde en gestandaardiseerde toetsen een degelijk instrument kunnen zijn om een objectief beeld van de leerling te schetsen. De resultaten van de toetsen kunnen dan worden gebruikt in het kader van een zelfevaluatie, waarbij de leerling zicht krijgt op zijn/haar mogelijkheden en beperkingen tegenover een bepaald referentiepunt en tegenover een referentiegroep in Vlaanderen.

Garantie op onderwijskwaliteit

Een aantal respondenten juicht het gebruik van een leer(winst)monitoringssysteem met gestandaardiseerde, verantwoordingsgerichte toetsen toe om de kwaliteit binnen scholen te borgen. In het huidige onderwijssysteem zijn er, aldus deze respondenten, te weinig instrumenten voorhanden die de onderwijskwaliteit in elke school garanderen. Dit scenario biedt daar een antwoord op, aangezien de resultaten van de toetsen door leraren en scholen kunnen worden aangewend om over de eigen onderwijskwaliteit en het onderwijsprogramma te reflecteren. Ook kan dit scenario een mogelijk antwoord bieden op de vraag van ouders en leerlingen naar meer zekerheid inzake de kwaliteit en het evaluatiebeleid van scholen, aldus de respondenten.

“Ja, dat weet ik niet goed, dat kan ik echt niet zeggen, maar ik voel gewoon dat die [ouders en leerlingen] altijd garantie en kwaliteit vragen en verantwoording en dat ze merken dat ze, als je van school wisselt, dat je dan het gevoel moet hebben dat je dan hetzelfde leert en krijgt en dezelfde kwaliteit aangeboden krijgt; en dat gevoel hebben ze nu niet.” (scholen, respondent 8)

Eveneens reageren een groot aantal respondenten positief op de verhouding tussen de resultaten op schooltoetsen en de centrale eindtoets in het secundair onderwijs zoals deze in het scenario beschreven is (incrementeel opgebouwd belang van de centrale eindtoets van 15% tot 50%) daar de eindverantwoordelijkheid van de beslissing om een oriënteringsattest of diploma toe te kennen nog steeds in grote mate in handen van de school blijft. Respondenten zijn van mening dat het vergelijken van de verschillende toetsresultaten zinvol kan zijn om als school het eigen evaluatiebeleid in vraag te stellen. Het incrementeel opbouwende karakter van de toetsresultaten wordt bij de meerderheid van de respondenten als wenselijk ervaren, zolang het aandeel van het leer(winst)monitoringssysteem in de eindbeslissing over de voortgang van leerlingen niet groter is dan 50%.

“Geen probleem, zolang als dat niet over die 50% gaat. Want dan zit je ook weer in het vaarwater van de deliberatieprocessen van de leraren die eigenlijk overruled worden door de neutrale externe toets.” (scholen, respondent 2)

Respondenten waarschuwen er echter voor dat niet alle scholen in staat zijn om de resultaten van dergelijke toetsen op een correcte manier te gebruiken in het proces van interne kwaliteitszorg. Ze achten ondersteuning dan ook noodzakelijk om tot een aantal gewenste verbeteracties te komen.

Leerwinstmeting

In het scenario staat beschreven dat voor leerlingen in het secundair onderwijs leerwinst in kaart kan worden gebracht door de resultaten van leerlingen op de eindtoetsen na het tweede jaar secundair onderwijs te vergelijken met de resultaten van dezelfde leerlingen in het basisonderwijs (voor de leerdomeinen die gelijk zijn). Op dezelfde wijze kan ook voor elke leerling leerwinst worden berekend aan het eind van het vierde leerjaar secundair onderwijs en aan het eind van het secundair onderwijs. De meerderheid van de respondenten ziet de mogelijkheid om leerwinst te registreren in dit scenario als een meerwaarde. Volgens de respondenten kan de leerwinstmonitoring zowel doelen op microniveau (progressie van de leerling), mesoniveau (interne kwaliteitszorg) als macroniveau (onderzoek naar de eindtermen en leerplandoelen) dienen.

“Ja... Het onderwijs van hieruit. Van beleid, eindtermen tot zijn het goede leerplannen, zijn het goede scholen? Ik denk op niveau van de leraar is het misschien moeilijk om dat echt na te gaan.” (beleidsmakers, respondent 2)

2.1.3 Kritiek en onwenselijke neveneffecten

Het verantwoordingsgerichte en verplichtende karakter van de centrale eindtoetsen zorgt voor heel wat weerstand bij een meerderheid van de respondenten. Ze zijn van mening dat er een groot aantal ongewenste effecten zal optreden wanneer dit scenario geïmplementeerd zou worden. Respondenten denken op mesoniveau dan in de eerste plaats aan een toename van onwenselijk strategisch gedrag door leraren en scholen, maar evengoed is een groot aantal respondenten van mening dat dit scenario ook de autonomie van scholen bedreigt en de concurrentie tussen de scholen doet toenemen. Op microniveau maakt een groot aantal respondenten zich zorgen over de impact van de toetsen op de algemene ontplooiing en emoties van de leerlingen. Tot slot stelt een aantal respondenten ook de betrouwbaarheid van de toetsen in vraag.

Onwenselijk strategisch gedrag door leraren en scholen

Zoals eerder gesteld is de belangrijkste bron van weerstand ten aanzien van het eerste scenario het verantwoordingsgericht en verplichtend karakter van de voorgestelde gestandaardiseerde toetsen. Doordat de verantwoordelijkheid van de toetsen uit handen van de leraren genomen wordt, denken respondenten dat leraren en scholen meer geneigd zullen zijn tot het stellen van onwenselijk strategisch gedrag. Een groot aantal respondenten vreest dat leraren zullen verleid worden tot *teaching to the test*. Ze zijn er dan ook van overtuigd dat het curriculum binnen scholen zodanig zal worden aangepast aan de verwachtingen van de toetsen dat andere aspecten zoals welbevinden, motivatie en zelfs het pedagogische project op de achtergrond geraken.

De meerderheid van de respondenten vindt het tevens minder wenselijk dat dit scenario zich vooral richt op het evalueren van de klassieke leerdomeinen in het basisonderwijs en minder oog heeft voor het evalueren van vakoverschrijdende eindtermen of leerdomeinen zoals muzische vorming en lichamelijke opvoeding. Omwille van deze beperkingen is een aantal respondenten ervan overtuigd dat dit scenario de totale ontplooiing van het kind niet ten goede komt en een verarming van het onderwijs met zich meebrengt (*teaching to the test* als een 'vernauwing van het curriculum').

"Dat is zo, alé ... mijn grote vrees is, dat is misschien te angstig hoor, is dat we met zo een systeem gaan uitkomen naar een prototype van mens dat we willen opvoeden of dat we gaan opvoeden. Ik vrees dat als je ... rond een beperkt aantal dingen, wiskunde, Nederlands, natuurwetenschappen, aardrijkskunde, geschiedenis, moderne talen en de rest niet, dat je dan ook naar een vorm van verschraling van uw onderwijs zou kunnen gaan." (onderwijsverstrekkers, respondent 6)

Daarnaast wijst een aantal respondenten op onwenselijke neveneffecten als het opzettelijk vervalsen van antwoorden en toetsresultaten en het verlenen van ongeoorloofde hulp aan de leerlingen tijdens de toetsafname.

Autonomie van de scholen wordt bedreigd

Wanneer de antwoorden van de verschillende categorieën met elkaar vergeleken worden, kan een groot verschil in wenselijkheid tussen onderwijsverstrekkers en scholen worden opgemerkt (de gemiddelde inschatting van wenselijkheid in Tabel 18 is respectievelijk 1.00 en 3.38). Naast het bovengenoemde onwenselijke neveneffect, is bij de meeste onderwijsverstrekkers de vrees dat de autonomie van scholen zal worden aangetast, een cruciaal element om dit scenario als onwenselijk te beschouwen.

Daarbij komt dat in dit scenario de klassenraad aan het einde van het basisonderwijs - maar voor de afname van de eindtoets - beslist of de leerling het schooljaar met vrucht doorlopen heeft en recht heeft op het getuigschrift basisonderwijs. De resultaten van de gestandaardiseerde eindtoets zijn enkel bindend wanneer een leerling slaagt voor de eindtoets, terwijl het getuigschrift basisonderwijs eerder niet werd toegekend door de school (een positieve bijstelling). Een groot aantal respondenten is hier dan ook van mening dat het advies van de klassenraad meer gefundeerd zal zijn dan de uitkomst van een gestandaardiseerde toets en vindt het daarom niet wenselijk dat het oordeel van de klassenraad kan overroepen worden door de resultaten van deze eindtoets. Ze benadrukken daarbij dat dit de professionaliteit en autonomie van leraren en scholen in grote mate bedreigt.

“Precies wegens de beperktheden en dat daarmee de verantwoordelijkheid van lerarenteams wordt afgekolfd. Je hebt tenslotte een kind negen of tien jaar in het beste geval op uw school gehad.... dan zou je toch als team in staat moeten zijn om na te gaan van: is dit kind er aan toe om naar een nastroom te gaan he?” (inhoudsdeskundigen, respondent 6)

Toenemende concurrentie tussen scholen

Hoewel in het scenario duidelijk beschreven staat dat de resultaten op schoolniveau worden afgeschermd voor publieke doeleinden, zijn respondenten toch van mening dat de concurrentie tussen scholen zal toenemen bij de implementatie van dit leer(winst)monitoringssysteem. Scholen met goede resultaten kunnen, volgens een aantal respondenten, geneigd zijn om zich te profileren en om de resultaten hoe dan ook aan te wenden in de media (bijvoorbeeld op hun website of op een informatiemoment).

“Er staat in dat eerste scenario dat de resultaten niet publiek worden gemaakt. Ik wil dat graag geloven, maar ik heb mijn twijfels of die scholen die dan goed scoren of dat die dan toch voor zich gaan houden. Volgens mij zet je scholen wel wat aan om daar wat reclame mee te maken, wat dan weer druk op andere scholen verhoogd om ook met hun resultaten buiten te komen.” (beleidsmakers, respondent 3)

Impact op de leerlingen

Zoals eerder gesteld (zie 2.1.2), zijn een aantal respondenten ervan overtuigd dat dit scenario het ‘gelijke lat’-principe zal bevorderen. Het is echter de vraag of door de implementatie van deze gestandaardiseerde toetsen, effectief een meer uniforme evaluatiemanier voor alle leerlingen wordt bereikt. Bepaalde respondenten menen dat het schriftelijke format nadelig kan zijn voor bepaalde leerlingen. Zo kunnen leerlingen met een beperking of een taalachterstand moeite hebben met het juist interpreteren van de vragen. Ook wordt de vraag gesteld of het schriftelijke format zich ertoe leent om, naast kennis, ook vaardigheden in kaart te brengen.

Tot slot maakt de meerderheid van de respondenten zich zorgen over de emotionele impact van de gestandaardiseerde toetsen in dit leer(winst)monitoringssysteem op de leerling. Niet alleen zou volgens respondenten de stress toenemen en het welbevinden van leerlingen dalen in een toetsgedreven cultuur, ook zouden leerlingen zich zorgen kunnen maken wanneer leraren lange tijd afwezig blijven of wanneer de onderwijskwaliteit niet goed blijkt te zijn. Ook het wachten op de examenresultaten zorgt voor onnodig veel stress bij zowel scholen als leerlingen.

Beperkte meerwaarde van leerwinstmeting

Een andere kritiek luidt dat in het basisonderwijs enkel een eindmeting beschikbaar is. Daarom levert dit scenario geen leerwinstmeting ten behoeve van basisscholen. Ook de meetmomenten in het secundair onderwijs (aan het einde van elke graad) zouden minder geschikt zijn om de leerwinst te meten aangezien er volgens enkele respondenten veel ruis op deze data zal zitten omdat de tussenliggende periode tamelijk lang is.

“Het is zeker ook niet op maat van de leerlingen, daar is het trouwens te laat voor hé, op het einde van de rit [...] dit is dus wat we noemen ‘de achterwaartse sturing van de onderwijskwaliteit’. Het is altijd alleen maar nuttig voor de volgende generatie, de generatie die gepasseerd is, die is weg hé.” (inhoudsdeskundigen, respondent 1)

Betrouwbaarheid

Daarnaast stellen respondenten dat de betrouwbaarheid van de toetsen in dit leer(winst)monitoringssysteem afneemt aangezien de toetsomstandigheden nooit hetzelfde kunnen zijn overheen de scholen. Alhoewel er veel aandacht wordt besteed in het scenario aan de kwaliteitszorg bij de correctie van de toetsen, twijfelen verschillende respondenten ook of deze handmatige correctie voldoende betrouwbaar kan gebeuren.

2.1.4 Haalbaarheid

De vermoedelijke afwezigheid van een gemeenschappelijk draagvlak voor dit scenario zijn de meest uitgesproken argumenten voor respondenten om dit scenario gemiddeld als eerder onhaalbaar in te schatten ($M=2.34$). In elke categorie van respondenten is de gemiddelde inschatting van de haalbaarheid van scenario 1 lager dan de neutrale score (3). Respondenten uit verschillende categorieën geven aan dat zij weten dat de onderwijsverstrekkers, onder meer omwille van een mogelijke impact op de autonomie van scholen, geen vragende partij zijn voor dit scenario.

“Bovendien weten wij toch ook dat als we het gesprek aangaan met onderwijsverstrekkers of koepels dat evaluatie van leerlingen voor hen in het hart ligt van hun pedagogische autonomie van hun vrijheid en op het moment dat je natuurlijk die evaluatie uit handen moet geven, breekt dat volgens mij dusdanig in op die pedagogische vrijheid dat het draagvlak daarvoor bij koepels ongeveer nul moet zijn.” (beleidsmakers, respondent 3)

Daarnaast wordt ook de hoge recurrente kostprijs van het leer(winst)monitoringssysteem genoemd als een reden om dit scenario als (eerder) onhaalbaar te beoordelen. Enkele respondenten menen bovendien dat dit scenario een “te grote stijlbreuk met het huidige onderwijssysteem” is (beleidsmakers, respondent 2) en dat het verplichtende karakter voor meer weerstand zal zorgen. Door bepaalde respondenten wordt evenwel gesteld dat een lange proefperiode met try-outs eventueel kan bijdragen om het draagvlak van dit scenario te vergroten.

“Ja, ik denk ook vooral omwille van die kostprijs, als dat elke keer opnieuw moet ontwikkeld worden en dat moet geactualiseerd worden, dan leek ons dat eigenlijk moeilijk haalbaar.” (onderwijsverstrekkers, respondent 1)

2.2 Adaptief leerwinstmonitoringssysteem

Tabel 19 - Wenselijkheid en haalbaarheid van scenario 2

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	4.25	3.67	4.25	3.17	3.83
Haalbaarheid	4.25	3.83	4.25	3.17	3.88

2.2.1 Wenselijkheid

De resultaten in Tabel 19 wijzen uit dat dit scenario door de respondenten over het algemeen vrij wenselijk wordt ingeschat ($M=3.83$), hoewel de inhoudsdeskundigen ($M=3.17$) iets terughoudender zijn dan de scholen en beleidsmakers (beiden $M=4.25$). Een groot aantal respondenten ($n=17$) vindt dit scenario dan ook (eerder) wenselijk, terwijl slechts drie respondenten dit scenario als eerder

onwenselijk beoordelen. De overige respondenten (n=4) schatten dit scenario noch wenselijk, noch onwenselijk in. Het ontwikkelingsgerichte karakter van het leerwinstmonitoringssysteem, waarbij echt nieuwe inzichten worden aangeleverd aan de leraren, draagt sterk bij tot de hoge score inzake wenselijkheid.

“Ik denk dat dat vooral te maken heeft met het feit dat, ja hoe moet ik dat zeggen, het is meer gekoppeld aan kwaliteitszorg en het opvolgen van de kinderen in plaats van het afleveren van een product op het einde van het zesde leerjaar.” (Scholen, respondent 4)

2.2.2 Positieve effecten

Respondenten achten het zeer realistisch dat het adaptief leerwinstmonitoringssysteem op verschillende niveaus een meerwaarde kan betekenen. Men is het erover eens dat door dit leerwinstmonitoringssysteem de ontwikkeling van de leerlingen gestimuleerd kan worden. Op mesoniveau levert het meten van vaardigheden interessante resultaten op voor de interne kwaliteitszorg- en andere beleidsprocessen.

Ontwikkeling van de leerling wordt gestimuleerd

Het getrapte adaptieve karakter van de toets wordt wenselijker onthaald door een groot aantal respondenten dan het principe van dezelfde lat voor iedereen. Respondenten geven aan dat het huidige onderwijs te weinig tegemoetkomt aan leerlingen die beter presteren dan de minimumnorm en denken dat de adaptieve toetsen mogelijk voor meer uitdaging bij de leerlingen kunnen zorgen.

“Het echt adaptief maken. Dat betekent dus op itemniveau na elke vraag, hé, dat is waar je uiteindelijk binnen dit x-jaar, ik wil er nu geen getal op plakken, maar dat is waar we naartoe gaan. Dat is ook individueel aangepaste toetsen hé. Het heeft geen zin om iedereen standaard dezelfde lat aan te bieden, dezelfde proef aan te bieden. Nee, we willen juist monitoren waar jij staat. En lukt het nu wel wat ervoor niet lukte? En kunnen we een stapje verdergaan? Terwijl een andere leerling al veel hoger zit.” (inhoudsdeskundigen, respondent 4)

Respondenten zijn eveneens van mening dat dit scenario bijdraagt tot een meer globale ontwikkeling van de leerling. Ook zou de schoolloopbaan geoptimaliseerd kunnen worden doordat leraren meer gericht gaan remediëren en differentiëren op basis van de gegevens.

“Ik vind dat heel goed omdat dat eigenlijk aansluit bij de essentie bij wat onderwijs, de vragen waarbij leraren en scholen zich moeten bezig houden, namelijk: groeien mijn leerlingen genoeg? Hoe zit het met de sterke leerlingen, kunnen we die nog voldoende uitdagingen bieden? Hoe zit het met de zwakkere leerlingen, kunnen we die meer leerwinst geven of er voor zorgen dat de kloof verkleint?” (onderwijsverstrekkers, respondent 4)

Volgens respondenten zullen leerlingen bovendien het eigen leerproces sneller in handen nemen doordat de feedback onmiddellijk voorhanden is.

Betere evaluatie van vaardigheden van leerlingen

Verscheidene respondenten zijn van mening dat in het huidige onderwijs het evalueren van vaardigheden vaak uit de boot valt, omdat een groot aantal (praktijk)leraren het moeilijk heeft om

deze vaardigheden juist te beoordelen. Het gestandaardiseerd in kaart brengen van enkele essentiële vaardigheden kan volgens respondenten dan ook een middel zijn om leraren te versterken in het evalueren van andere, meer vakgebonden vaardigheden.

Bevorderen van interne kwaliteitszorg en andere processen op school

Hoewel dit scenario gericht is op het microniveau, zien vele respondenten er ook voordelen in voor het mesoniveau. De informatie die vrijkomt uit de leerwinstmonitoring levert voor de leraar en de school interessante gegevens op in het kader van interne kwaliteitszorg. Daarnaast geven respondenten aan dat de resultaten gebruikt kunnen worden om de leerlijnen inzake deze essentiële vaardigheden beter af te stemmen op elkaar. Bovendien kan de gegenereerde informatie scholen en leraren ook helpen om het eigen evaluatiebeleid te beoordelen en bij te sturen. Respondenten benadrukken eveneens dat het gebruik van de resultaten niet alleen de interne kwaliteitszorg van de school, maar ook het zorg- en taalbeleid ten goede kunnen komen.

“Ik denk dat dit op zich wel een zeer wenselijk scenario is, ik weet niet of het dit systeem alleen zou zal maken of alleen zal volstaan, maar als een mogelijk inzetbaar instrument lijkt het mij zeker wenselijk. Waarom? Omdat volgens mij die toetsen een schat aan informatie voor scholen kunnen bevatten, dus ik denk dat die voor een eigen intern kwaliteitszorgsysteem en voor het verhogen van de kwaliteit van hun onderwijs dat daar heel veel informatie kunnen uithalen.” (beleidsmakers, respondent 3)

Een aantal respondenten vindt het echter noodzakelijk om ook de toegevoegde waarde van scholen mee in kaart te brengen, rekening houdend met achtergrondkenmerken van de school en leerling om op die manier nog betere, contextgebonden informatie van de school te verkrijgen.

2.2.3 Kritiek en onwenselijke neveneffecten

Doordat de toets een low-stakes karakter heeft, zijn respondenten ervan overtuigd dat er weinig onwenselijke effecten zullen optreden. Er is geen sprake van een toenemende druk op de leerlingen doordat de resultaten idealiter gezien niet in rekening worden gebracht bij het uitreiken van getuigschriften, oriënteringsattesten of diploma's. Tevens richt het systeem zich op de eigen groei en niet op het onderling vergelijken van de leerlingen, waardoor mogelijke competitie-effecten tussen leerlingen worden vermeden. Toch noemen enkele respondenten een aantal punten van kritiek. Net zoals in het eerste scenario is ook in dit scenario sprake van een verplichte deelname, waardoor de weerstand tegenover dit systeem kan toenemen. Hoewel het scenario gericht is op de ontwikkeling van leerlingen en scholen, kan verantwoordingsgericht gebruik de voordelen overschaduwden. Tot slot kan in de gesprekken worden vastgesteld dat heel wat respondenten sceptisch staan tegenover de mogelijkheden om vaardigheden in kaart te brengen.

Verplichte deelname

Het adaptief, ontwikkelingsgericht karakter kan op veel bijval rekenen bij de verschillende respondenten en hoewel een groot aantal respondenten geen graten ziet in het verplichtende karakter (net omwille van de ontwikkelingsgerichte focus), pleiten vooral de onderwijsverstrekkers toch voor een meer vrijblijvende deelname en geven ze aan dat de kans op weerstand groter is dan bij facultatief gebruik.

“Voor mij kan het als het eigenlijk niet verplichtend zou zijn.” (onderwijsverstrekkers, respondent 3)

De gebruikers van het adaptief leerwinstmonitoringssysteem worden verplicht om minimaal éénmaal per schooljaar gebruik te maken van dit systeem. Het is opvallend dat de meerderheid van de respondenten akkoord gaat met deze minimale afname om de leerwinstmonitoring zo accuraat mogelijk door te voeren, maar respondenten hebben toch vragen bij de haalbaarheid en de organisatie van deze 'hoge' frequentie. Ze zijn naar eigen zeggen vooral ongerust om in een overdreven toetscultuur te verzeilen.

Opletten voor verantwoordingsgericht gebruik

Respondenten gaan in grote mate akkoord met het inzagerecht door de inspectie op voorwaarde dat de focus van de inspectie ligt op hoe de school met de resultaten van het leerwinstmonitoringssysteem is omgegaan, eerder dan op de output van de toetsen. Enkele respondenten waarschuwen er wel voor dat ondanks de focus van dit scenario op de ontwikkeling van leerlingen, het inzagerecht van de inspectie in bepaalde scholen voor oneigenlijk (verantwoordingsgericht) gebruik kan zorgen.

“Dat hangt ook een beetje van dat beleidsvoerend vermogen af, als ge het dan weer te ruim laat en er is te weinig beleidsvoerend vermogen, dan kan het zijn dat ze er minder zinvol met omgaan... Als het wat minder ontwikkeld is wat dat we toch ook in sommige scholen en centra vinden. Ja dan kan het zijn dat dat misschien oneigenlijk wordt, dat het wordt ingezet. Dat zou een risico kunnen zijn.” (onderwijsverstrekkers, respondent 1)

Evalueren van niet-curriculumgebonden vaardigheden

Hoewel een beperkt aantal respondenten meent dat dit methode-onafhankelijk scenario een groter draagvlak kent doordat twistpunten over de eindtermen en leerplandoelen worden vermeden, verzet een groot aantal respondenten zich evenwel tegen de evaluatie van niet-curriculumgebonden vaardigheden.

“Ik ben bang van een derde soort doelstellingen. We hebben er al twee en het is al erg genoeg. Maar leerplandoelstellingen en eindtermen, dat we er niet nog een derde moeten bijmaken.” (inhoudsdeskundigen, respondent 3)

Naast de bovenstaande discussie, uit een groot aantal respondenten, waaronder de inhoudsdeskundigen, haar bezorgdheid tegenover het juist evalueren van vaardigheden. Zij vermoeden dat het in kaart brengen van vaardigheden moeilijk te realiseren is, waardoor ze - ondanks de hoge wenselijkheid - eerder sceptisch staan tegenover dit scenario. Zo zou een digitale toetsvorm met korte antwoord-vragen veel beperkingen met zich meebrengen en zouden observatie-instrumenten wenselijker zijn volgens een aantal respondenten.

2.2.4 Haalbaarheid

De ontwikkeling en implementatie van een adaptief leerwinstmonitoringssysteem wordt door een groot aantal respondenten als vrij haalbaar ingeschat ($M=3.88$). De meerderheid van de respondenten stelt dat dit scenario in lijn is met het huidige systeem en dat het past bij de huidige visie op onderwijs. Het draagvlak voor dit scenario is volgens de respondenten dan ook groot.

“Ja. Het is ook nieuw en aanvullend. En het vergt geen grote aanpassingen ten aanzien van wat er vandaag bestaat.” (onderwijsverstrekkers, respondent 5)

Ook de mogelijkheid om het leerwinstmonitoringssysteem geleidelijk aan op te bouwen, draagt bij tot de inschatting van een hoge haalbaarheid.

“Omdat ik ook gelezen heb dat je dus eigenlijk... van plan bent om in de loop van een aantal schooljaren toetsen nog, of items nog toe te voegen dacht ik he. Dat het niet meteen allemaal klaar en af moet zijn. Dat is natuurlijk iets dat meer haalbaar lijkt altijd.”
(inhoudsdeskundigen, respondent 5)

Een groot aantal respondenten vergelijkt dit systeem dan ook met de huidige leerlingvolgsystemen en CITO-toetsen en geeft aan dat het valideren van bestaande toetsen een grote meerwaarde kan zijn. Daarnaast juichen respondenten ook het gebruik van IRT toe. Eén respondent stelt dat een eigen Vlaams systeem toch waardevol kan zijn in het kader van eindtermen alsook verschillen inzake taalgebruik en context tussen Vlaanderen en Nederland.

“Ik erger mij er dood aan, dat zijn allemaal Nederlandse testings [...], dan zijn we ten minste al zo ver dat we in Vlaanderen een tussentijdse testing hebben die Vlaams is. Eén van de dingen zijn de CITO-toetsingen [...] Het is niet gebaseerd op onze onderwijssysteem, niet op ons leerplannen en mijn kinderen moeten begrijpend lezen doen waarbij ze lezen: “we gaan naar de markt in Leiden”. Ik krijg er grijs haar van.”
(scholen, respondent 7)

Tot slot zijn er respondenten die vinden dat dit scenario niet los kan staan van andere scenario's. De relatief lage ontwikkelingskost (1.800.000€) en jaarlijkse kostprijs (950.000€) werken dit idee in de hand.

2.3 Diagnostisch meetinstrumentarium & toetsenbank

Tabel 20 - Wenselijkheid en haalbaarheid van scenario 3

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	3.25	3.33	4.00	3.67	3.56
Haalbaarheid	4.75	3.67	4.25	2.67	3.83

2.3.1 Wenselijkheid

Tabel 20 toont aan dat de respondenten eerder positief staan tegenover een diagnostisch meetinstrumentarium en toetsenbank ($M=3.56$), hoewel de wenselijkheid minder groot is bij beleidsmakers en onderwijsverstrekkers dan bij scholen en inhoudsdeskundigen. Slechts één van de 24 respondenten vindt dit scenario onwenselijk, terwijl 13 respondenten dit scenario (eerder) wenselijk vinden. Opvallend is vooral dat tien respondenten neutraal staan tegenover scenario 3.

2.3.2 Positieve effecten

De vrijblijvendheid van deze diagnostische meetinstrumenten zorgt er volgens de respondenten voor dat gebruikers zich optimaal kunnen focussen op het ontwikkelingsgericht karakter van dit scenario. Respondenten geven dan ook aan dat er nood is aan een betere professionalisering van leraren inzake evalueren en reflecteren. Daarnaast vinden respondenten het wenselijk dat de toetsen zeer gericht kunnen worden ingezet bij leerlingen om leerproblemen te achterhalen.

Op maat van de leerling

De meerderheid van de respondenten acht het gebruik van de toetsenbank en de diagnostische meetinstrumenten wenselijk in het kader van zorgverbredingsprocessen op school. De informatie op leerlingenniveau biedt leraren de kans om op “maat van de leerlingen te werken” met behulp van het individueel verbetertraject, dat wordt opgesteld voor leerlingen die nood hebben aan aangepaste remediëring.

“Omdat ik denk dat het een zinvol instrument kan zijn binnen het geheel van aanpak van uw zorgverbreding op school. En zeker nu in het kader van de ontwikkeling met het M-decreet denk ik dat dat eigenlijk wel een nuttig instrument kan zijn om in te zetten op leerlingniveau waar je vaststelt van: kijk, hier faalt ik op een of ander ding of hier faalt de leerling op een of ander facet en dan kan ik daar eigenlijk eens kijken, doelgerichter kan ik dan zien van: waar zit het probleem?” (onderwijsverstrekkers, respondent 3)

Respondenten staan in het algemeen ook positief tegenover de observatie- en meetinstrumenten die het scenario voorziet en die betrekking hebben op competenties, vakoverschrijdende eindtermen en sociaal-emotionele aspecten van de leerling. Een aantal respondenten denkt wel dat deze instrumenten reeds in voldoende mate beschikbaar zijn. Daarnaast zijn niet alle respondenten van mening dat competenties kunnen worden geëvalueerd met behulp van een digitaal beschikbaar meetinstrument.

“En ook omdat ik hier wel gelezen heb dat je veel meer eindtermen zou proberen te dekken. Dat het niet alleen gaat over wiskunde en Nederlands. Maar ik lees hier ook wetenschap, techniek, Frans, mens en maatschappij. En dan vond ik ook heel positief dat je daarnaast ook nog observatie- en meetinstrumenten zou meegeven die dan competenties, vakoverschrijdende eindtermen en sociaal-emotionele aspecten van de leerling in beeld zou willen brengen... Of dat je daar ook nog meteen aan moet koppelen: kunnen we nog toetsen ontwikkelen om te detecteren van dyslexie en dyscalculie. Ja misschien zijn die er al wel.” (inhoudsdeskundigen, respondent 5)

Professionalisering van leraren

Hoewel een aantal respondenten van mening is dat het opstellen van toetsen tot de functie van de leraar hoort (zie 2.3.4), blijkt uit de gesprekken dat de expertise van leraren om goede toetsen te ontwikkelen in vraag wordt gesteld door een groot aantal respondenten. Bij de verschillende groepen is de vraag naar ondersteuning van leraren op vlak van evaluatie pertinent aanwezig. Respondenten denken dat deze toetsenbank “een inspiratiebron voor leraren” kan zijn en hen kan aanzetten om zelf kwalitatief goede toetsen en examens te leren ontwikkelen.

“In dit scenario acht ik dan mogelijk dat je dan echt wel een hele mooie toetsenbank krijgt met hele goede toetsvragen. Waarbij je toch ook weer aangeeft: dat zijn deze eindtermen die we daarmee dekken, maar alle andere niet he... Omdat je enerzijds nu ziet dat methodetoetsen hun kwaliteiten hebben, maar ook hun gebreken. Dat als je leraren toetsvragen laat ontwikkelen, dat die toch vaak op een heel basaal niveau zijn. Dat zij... goede toetsvragen maken is gewoon heel moeilijk (lacht). Die ook nog betekenis hebben, die zinvol zijn en zo meer.” (inhoudsdeskundigen, respondent 5)

Bovendien menen respondenten dat het gebruik van gevalideerde toetsen leraren kan aanzetten om te reflecteren over het eigen didactisch handelen. De toetsenbank kan worden

gebruikt als een soort “externe reflector” waarbij leraren middels de toetsen kunnen nagaan of de leerstof door elke leerling voldoende beheerst wordt.

Professionele communicatie tussen leraren

Ten slotte suggereren respondenten dat door de toetsen in dit leerwinstmonitoringssysteem gesprekken in het kader van leerlijnen binnen de vakgroepwerking worden gestimuleerd. Zo draagt dit scenario bij tot de professionele dialoog tussen leraren.

2.3.3 Kritiek en onwenselijke neveneffecten

Het meest onwenselijke aspect van dit scenario blijkt het achterwege blijven van een leerwinstmonitoringssysteem, mede veroorzaakt door het vrijblijvende karakter van de toets. Deze niet-verplichte deelname zorgt ook voor andere onwenselijke effecten. Respondenten vrezen dat scholen en leraren minder doelgericht te werk zullen gaan in het afnemen van deze zeer specifieke toetsen waardoor het totaalbeeld van de leerling mogelijk verloren kan gaan. Ook het interpreteren van de resultaten kan voor problemen zorgen wanneer het lerarenteam onvoldoende datageletterd blijkt te zijn.

Ten slotte vragen sommige respondenten zich af of - hoewel men het bestaan van dit soort toetsenbank wenselijk acht - het de taak is van de overheid om de ontwikkeling ervan op zich te nemen.

Geen leerwinstmeting

Een groot aantal respondenten betreurt dat in dit scenario geen sprake is van leerwinstmonitoring en dat de evolutie van de leerlingen niet wordt geregistreerd. Door het vrijblijvende karakter van de toets zijn er immers geen systematische evaluatiemomenten ingebouwd. Wanneer het tweede scenario vergeleken wordt met het derde scenario blijkt de voorkeur van de respondenten in het algemeen bij het tweede scenario te liggen, net omwille van die leerwinstmonitoring.

“Dat dat een serieuze beperking is. Dus zoals we het nu zo horen ook, denk ik dan, ook liever 2 dan 3... 3 speelt wel in op een aantal beperkingen van 2 maar de beperkingen van 3 wegen zwaarder door dan die van 2.” (onderwijsverstrekkers, respondent 2)

Doelgerichtheid daalt

Hoewel een groot aantal respondenten achter “de vrijheid van keuze” staat, denken sommigen toch dat de vrijblijvendheid van dit scenario ervoor zorgt dat leraren en scholen minder doelgericht en systematisch te werk zullen gaan. Daarnaast menen respondenten dat het facultatieve karakter ervoor zorgt dat steeds dezelfde scholen en leraren zich voelen aangesproken om met het beschikbare materiaal aan de slag te gaan.

“Informatiebron voor de interne kwaliteitszorg van scholen, maar als het facultatief is, vind ik dat te vrijblijvend om het er uit te halen.” (scholen, respondent 7)

Omdat de toetsmomenten individuele leerlinginformatie verschaffen, kan uit de gesprekken worden afgeleid dat respondenten het noodzakelijk achten om leraren te ondersteunen in het interpreteren van de resultaten. Doordat er specifieke informatie van de leerling in kaart wordt gebracht, denken respondenten dat het totaalbeeld van de leerling verloren kan gaan.

Taak van de overheid?

Uit de interviews blijkt dat in het onderwijsveld de behoefte aan gevalideerde, binnenlandse toetsen groot is, maar dat niet alle respondenten van mening zijn dat het de taak is van de overheid om een dergelijk systeem uit te bouwen.

“Ja ja. Wat het gevaar dan is, dat je een stuk naar een tweedeling gaat van je hebt het verhaal van de methodiek en de toetsjes die daar bij horen en dan daarnaast heb je de eindtermen, die ook nog eens hun eigen toetsen hebben. En als die discrepantie dan te groot is, dan krijg je al... iets heel vreemds natuurlijk. Dus ik denk dat de vraag heel legitiem is van leraren van... geef ons een kader om toetsing voor te geven die sterker georiënteerd is naar die eindtermen dan het kader waarover we nu binnen onze methode beschikken... Maar dan is de vraag of dat dit... de gepaste manier is om... om aan die behoefte tegemoet te komen. De behoefte is heel legitiem denk ik, maar... er blijven toch andere wegen.” (onderwijsverstrekkers, respondent 2)

Benaming

Tot slot is ook de benaming van de toetsen zoals ze in het scenario staan voorgesteld (bijvoorbeeld ‘breuken vierde leerjaar’) voer voor discussie volgens een aantal respondenten, omdat het de klassieke indeling van leerlingen in klasgroepen die allen hetzelfde programma moeten volgen, lijkt te bestendigen - terwijl net het tegenovergestelde beoogd werd door dit scenario.

“Ja, oké, je moet het dan, door het feit dat je daar een leerjaar bij zet, maak je jezelf eigenlijk leerplan- en methode-afhankelijk ... je zou eigenlijk een objectieve omschrijving nodig hebben van het niveau van breuken, die losstaat van de klassieke indeling in leerjaren. Want dan hangt dat samen aan leerplan. Dus, want anders stimuleer je scholen ook om te blijven denken in die klassieke termen van leerjaar en dat is dan meteen een groot verschil of een groot minpunt ten opzichte van het voorgaande scenario 2, waar je dat volledig loslaat en dat je zegt van, daar ga je dus echt naar een adaptief meten van basisvaardigheden” (inhoudsdeskundigen, respondent 1)

2.3.4 Haalbaarheid

Volgens de respondenten zou het draagvlak voor dit “niet-bedreigende” scenario best groot kunnen zijn door zowel het vrijblijvend karakter van het systeem als door het ontwikkelingsgericht perspectief op leerlingenniveau ($M=3.83$). Daarnaast is de expertise van een groot aantal instrumenten reeds aanwezig, waardoor de haalbaarheid verhoogd wordt.

Doordat het toetsenaanbod in samenwerking met het onderwijsveld wordt ontwikkeld en niet louter vanuit de overheid wordt gerealiseerd, wordt het draagvlak volgens de respondenten bijkomend vergroot. De haalbaarheid daalt echter omwille van de aanwezigheid van heel wat “verschillende meningen” die in het ontwikkelingsproces worden betrokken.

De geleidelijke opbouw, die er ook voor zorgt dat de kostprijs relatief laag uitvalt, kan eveneens op bijval rekenen. Daarnaast rijst bij een aantal respondenten ook de vraag of de overheid zal willen investeren in dit scenario omwille van het vrijblijvende karakter. Daarbij komt nog dat een aantal respondenten dit scenario liever niet als enige scenario uit de bus ziet vallen, maar zou het,

gecombineerd met een ander scenario, toch een meerwaarde kunnen betekenen in het onderwijslandschap.

2.4 High-stakes eindtoetsen voor leerlingen en scholen

Tabel 21 - Wenselijkheid en haalbaarheid van scenario 4

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	1.50	1.00	1.44	2.17	1.53
Haalbaarheid	2.00	1.67	2.75	2.00	2.10

2.4.1 Wenselijkheid

In het vierde scenario worden eindtoetsen ontwikkeld die enerzijds geaggregeerde leerprestaties van leerlingen in scholen in kaart brengen en die anderzijds objectieve beslissingen over het uitreiken van getuigschriften en diploma's moeten mogelijk maken. Wanneer de respondenten het vierde scenario beoordelen (Tabel 21), blijkt dat dit scenario voor de meerderheid van de respondenten één van de minst wenselijke scenario's is ($M=1.53$).

“Dat is het monsterscenario hé.” (onderwijsverstrekkers, respondent 2)

De meerderheid van de respondenten ($n=22$) schat de wenselijk van dit scenario (eerder) onwenselijk in. Slechts twee respondenten vinden dit scenario eerder wenselijk.

2.4.2 Positieve effecten

Hoewel de meeste respondenten negatief staan tegenover deze high-stakes eindtoetsen, erkent de meerderheid dat de positieve aspecten uit het eerste scenario ook hier aan bod zullen komen. Ze doelen daarbij niet enkel op een objectieve evaluatie van leerlingen, maar ook op het gebruik van de resultaten, die op mesoniveau interessant kunnen zijn in het kader van interne kwaliteitszorg en verder onderzoek. Respondenten denken bovendien dat de resultaten op de centrale examens het schooladvies kunnen bevestigen en dat de stijgende behoefte naar een meer transparante samenleving vervuld kan worden.

Bevestigen van schooladvies

Een aantal respondenten vindt dat de combinatie van de resultaten op schooleigen toetsen en centrale examens in het basisonderwijs voor complementariteit tussen de school en de overheid zorgt. Het oordeel van de school krijgt bijvoorbeeld meer draagkracht als de resultaten van de centrale toets aansluiten, hoewel de meeste respondenten er sterk van overtuigd zijn dat het lerarenteam over het algemeen een correct beeld van de leerling heeft.

“En eigenlijk komt het bijzonder weinig voor dat kinderen die wij het getuigschrift niet kunnen geven op basis van de gegevens die wij verzameld hebben, dat daardoor, dat dat tegengesproken wordt door die toets [toets van de onderwijsverstrekker eind zesde leerjaar]. Het komt... Ik geloof in de 23 jaar dat ik directeur ben, hebben we dat twee keer meegemaakt. Dus dat wil zeggen dat ik kan bouwen op een team dat een goed zicht heeft op de resultaten.” (scholen, respondent 3)

Transparantie in de samenleving kan zorgen voor kwaliteitstoename

Ondanks de vele nadelen van de openbaarheid van de resultaten op schoolniveau, acht een aantal respondenten het wenselijk om de resultaten te publiceren in het kader van een transparante samenleving. Deze respondenten stellen ook dat ouders vragende partij zijn hiervoor.

“Die competentietesten, die competitie tussen scholen kan voor een kwaliteitstoename zorgen tussen scholen, maar dat is te zien wat je natuurlijk verstaat onder kwaliteit hé.” (scholen, respondent 1)

2.4.3 Kritiek en onwenselijke effecten

De onwenselijke effecten die in scenario 1 werden opgesomd als gevolg van het high-stakes van het toetssysteem voor leerlingen, worden ook verwacht bij de keuze voor scenario 4. Bovendien struikelt in dit scenario een groot aantal respondenten over het openbaar, sterk verantwoordingsgericht karakter van het voorgestelde monitoringssysteem op mesoniveau. Men is van mening dat binnen dit scenario rankings onvermijdelijk zijn, waardoor de competitie tussen scholen sterk zal toenemen.

“Ja, vooral door dat openbaar. Vooral dat openbaar. Dat is voor mij het toppunt van prestatie maatschappij en competitie en streven en stress en druk en dat wil ik echt niet in een basisschool. Niet voor de leraren en niet voor de kinderen.” (scholen, respondent 4)

Doordat scholen zullen streven naar een zo goed mogelijk resultaat, is het volgens respondenten mogelijk dat scholen selectief zullen worden op het vlak van leerlingeninstroom met een toenemende sociale segregatie tot gevolg. Respondenten vermoeden dat de druk op scholen en leraren in grote mate zal toenemen, maar dat ook het welbevinden en de stress bij leerlingen negatief beïnvloed wordt omwille van het bindende karakter van de toets. Tot slot sluiten respondenten *teaching to the test* niet uit en denken ze dat aspecten zoals welbevinden en motivatie minder belangrijk zullen zijn in een prestatiegerichte school.

Rankings

De toenemende concurrentie en de profileringsdrang tussen scholen om leerlingen aan te trekken, zou volgens de meerderheid van de respondenten met zich meebrengen dat scholen meer inspanningen zullen leveren op het vlak van PR en communicatie, hetgeen ten koste kan gaan van de inspanningen inzake onderwijskwaliteit en professionaliseringsbeleid in scholen.

Respondenten denken bovendien dat het publiceren van resultaten zal resulteren in het opstellen van rankings en waarschuwen dan ook voor het ontstaan van een *‘name and shame’* cultuur. Ze vermoeden dat deze vorm van openbaarheid contraproductief werkt en dat het schoolklimaat negatief kan beïnvloed worden. Respondenten wijzen er tevens op dat door de beperkte datageletterdheid van ouders en scholen, maar ook van journalisten, resultaten verkeerd geïnterpreteerd kunnen worden waardoor het imago van scholen ernstig geschaad kan worden.

“Maar wat dan die openbare rapportage betreft, ik val ook over de openbaarheid omdat ge het bredere publiek en vooral ook de tussenlijn, de factor tussen die publicatie en dat publiek, zijnde de journalisten en de kranten, nooit of te nimmer die correcties correct gaan interpreteren.” (inhoudsdeskundigen, respondent 1)

Sociale segregatie

Respondenten menen eveneens dat het verantwoordingsgericht karakter een aanleiding kan zijn voor scholen om de leerlingeninstroom te manipuleren. Scholen zouden dan geneigd zijn om vooral leerlingen toe te laten waarvan men met een minimum aan inspanningen hoge scores verwacht.

“Daar denk ik dat we helemaal de mist in gaan, omdat dit het risico natuurlijk met zich mee brengt dat we eigenlijk in een ranking komen van onderwijs, dat die ranking met zich meebrengt dat in feite we dreigen, niet de ranking op zichzelf is fout, maar het feit dat we eigenlijk in een systeem van vrije keuze eigenlijk goede en slechte scholen gaan krijgen.” (onderwijsverstrekkers, respondent 3)

Ouders zullen zich, volgens respondenten, in hun keuzeprocess meer laten beïnvloeden door examenresultaten, eerder dan over andere gegevens over de school. Volgens de respondenten zullen vooral allochtone leerlingen en/of leerlingen uit kansarme gezinnen getroffen worden, omdat deze doelgroepen vaak minder geïnformeerd zijn over de verschillende mogelijkheden, waardoor de sociale segregatie zou toenemen. Ook het beleid in verband met zittenblijven kan volgens de respondenten worden verstrengd. Daarnaast rijst de vraag hoe leerlingen met beperkingen (handicap of leerstoornissen) zullen worden opgevangen binnen dit systeem.

“Maar ik vind het ook geen goede zaak om dat dan inderdaad, dat is sociale segregatie dat je realiseert. En je gaat meer en meer zien dat moeilijke leerlingen in achtergestelde buurten terecht komen in slecht functionerende scholen en dat die het moeten doen met, als ik ze zo mag noemen, tweederangs leraren en tweederangs directeurs. Want eersterangs leraren en eersterangs directeurs gaan kiezen voor die scholen die in de ranking heel goed staan en die het goede publiek aantrekken in de goede buurten. Daar kan je niet naast.” (scholen, respondent 2)

Emotionele impact op leerlingen en leraren

Respondenten hechten niet alleen veel belang aan het openbaar karakter van dit scenario, ook het bindend karakter van de toets (50% in het basisonderwijs en 100% in het laatste jaar van het secundair onderwijs) is voer voor discussie. Zoals eerder vermeld staat een aantal respondenten positief tegenover het gewicht dat de centrale toetsen krijgen in de summatieve beoordeling van leerlingen in het basisonderwijs, maar geen enkele respondent gaat expliciet akkoord met het absoluut karakter van de centrale toets (100%) in het secundair onderwijs. Respondenten vermoeden dat deze voorwaarde een grote emotionele impact op leerlingen teweeg zal brengen (stress, faalangst). Ook het feit dat leerlingen niet vertrouwd zullen zijn met het concept van de toets kan de stress en faalangst vergroten, waardoor de resultaten mogelijk vertekend worden, aldus de respondenten.

“Omdat we ook merken en dat geldt misschien voor alle scenario's, dat als kinderen niet vertrouwd zijn met bijvoorbeeld een digitale toets, dat je die ook echt wel moet leren omgaan. Dus in... aan de ene kant moeten ze methodeonafhankelijk zijn, daar ben ik het absoluut mee eens. Aan de andere kant... maakt dat dat je misschien toch geen echt objectief beeld van leerlingen krijgt. Niet doordat ze inhoudelijk niet kennen en kunnen wat door de toets gevraagd wordt, maar gewoon omdat ze met dat soort toets niet vertrouwd zijn.” (inhoudsdeskundigen, respondent 5)

Enkele respondenten stellen ook dat de motivatie van leraren en scholen zal afnemen wanneer het evaluatiebeleid uit handen wordt genomen en dat leraren zich bedreigd zullen voelen in hun professionaliteit.

Toch nog ongelijkheden tussen scholen inzake beoordeling

Niet alle respondenten zijn het eens met het idee dat dit scenario zal leiden tot een meer gelijke lat tussen leerlingen, of toch wat betreft het basisonderwijs. Mogelijks kan de klassenraad compenserend werken wanneer de centrale toetsresultaten een ander beeld geven.

"In basisonderwijs... vind ik net omwille van die 50% verhouding... ik vind dat zo zeer moeilijk 50... dat is arbitrair... waarom is dat nu 50%? Daar heb ik nu niet echt een goeie verklaring voor gelezen. Het wordt dan eigenlijk helemaal onduidelijk wie nu eindverantwoordelijkheid draagt voor de evaluatie voor die leerling... als je weet dat de punten voor 50% op basis van die toets zijn, gaan klassenraden daar op een of andere manier zich laten beïnvloeden door die 50% die ze niet in de hand hebben? Gaan die compenseren beginnend denken?" (beleidsmakers, respondent 3)

Teaching to the test

Nog meer dan in het eerste scenario denken respondenten dat scholen en leraren binnen dit verantwoordingsgericht scenario geneigd zullen zijn tot het stellen van onwenselijk strategisch gedrag, zoals *teaching to the test*. Dat komt onder meer omdat de toets zich teveel richt op een beperkt aantal leergebieden, terwijl respondenten het tevens noodzakelijk achten om vakoverschrijdende aspecten (zoals houding, motivatie en doorzettingsvermogen) mee in rekening te brengen bij een uiteindelijke beoordeling.

2.4.4 Haalbaarheid

Uit de interviews blijkt tevens dat bijna alle respondenten vermoeden dat voor dit scenario geen maatschappelijk draagvlak bestaat omdat de nadelen voor zowel leerlingen, leraren en scholen te groot zijn ($M=2.10$).

"Ik denk Vlaanderen daar absoluut niet klaar voor is, ik denk dat er maar een paar mensen zijn die dat echt als positief gaan inschatten, dat is toch mijn inschatting, ik denk dat ge dan toch een hele revolutie teweeg zal brengen als ze dat gaan afvuren op de mensen." (onderwijsverstrekkers, respondent 4)

Ook de hoge recurrente kost zal volgens respondenten de maatschappelijke weerstand verhogen. Ondanks alle nadelen, schat een aantal respondenten de haalbaarheid van het scenario op technisch en organisationeel vlak enigszins hoog in, hoewel de tewerkstelling van de 400 externe evaluatoren toch vragen oproept.

"Naar, naar aantal mensen niet, maar ik denk... ja is dat haalbaar om 400 bekwame mensen te vinden die dan in die cruciale periode, zeg maar, dat werk willen doen? Misschien moet je dan eerst eens denken aan net gepensioneerden of zo (lacht). Het is ook een heel moeilijke en lastige periode voor scholen. Dus ik denk niet dat die daarop staan te springen dat dan leraren uit de school zouden gehaald worden bijvoorbeeld." (inhoudsdeskundigen, respondent 5)

Andere respondenten menen dan weer dat de haalbaarheid van dit scenario nog meer afneemt omwille van een aantal discussiepunten tussen de diverse actoren in het onderwijsveld over de cesuurbepaling en de gerichtheid van de toets op de eindtermen (in plaats van op leerplandoelen). Enkele respondenten menen weliswaar dat die discussie beperkt kan blijven omdat de resultaten over het al dan niet bereiken van minimumdoelen (eindtermen) niet zulke grote gevolgen met zich mee zouden brengen, aangezien elke school in staat moet zijn om deze te bereiken bij haar leerlingenpubliek.

Tot slot denkt een aantal respondenten dat een geleidelijke opbouw van de openbaarheid de haalbaarheid doet toenemen. Toch geven ook de twee respondenten die dit scenario wel wenselijk achten aan dat er momenteel geen draagvlak bestaat voor dit scenario:

“Ja, ik denk gewoon dat dat een brug te ver is voor, voor Vlaanderen gewoon. Ik denk niet dat daar heel veel animo voor kan zijn.” (inhoudsdeskundigen, respondent 2)

2.5 Leerwinstmeting op schoolniveau met openbaar karakter

Tabel 22 - Wenselijkheid en haalbaarheid van scenario 5

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	1.63	1.33	1.88	2.00	1.71
Haalbaarheid	1.63	1.67	2.31	2.33	1.98

2.5.1 Wenselijkheid

Net zoals in het vierde scenario, blijkt ook nu uit de resultaten (Tabel 22) dat de respondenten het eerder onwenselijk achten dat de resultaten van een leer(winst)monitoringssysteem openbaar worden gemaakt op schoolniveau ($M=1.71$). Het is dan ook niet onverwacht dat de meerderheid van de respondenten ($n=20$) dit scenario (eerder) onwenselijk vindt. Twee respondenten achten dit scenario echter wel wenselijk, terwijl de twee overige respondenten een neutraal standpunt innemen.

2.5.2 Positieve effecten

De meeste respondenten geven bij een berekening van toegevoegde waarde er de voorkeur aan dat dit gebeurt op basis van leerwinstmonitoring, eerder dan op basis van leerprestaties. Deze data over leerwinst kunnen tevens worden ingezet bij interne kwaliteitszorgprocessen.

Toegevoegde waarde-meting

De grote meerderheid van de respondenten ziet een grote meerwaarde in het monitoren van de toegevoegde waarde op basis van leerwinst omdat het rendement van de school er beter door zichtbaar wordt gemaakt. Daarnaast kunnen deze resultaten gebruikt worden om op macroniveau uitspraken te doen over verbanden tussen schoolkenmerken in het kader van beleidsbeslissingen.

Verschillende respondenten hameren er op dat de gegevens moeten worden geplaatst binnen de context van de school en dat er rekening moet worden gehouden met sociaal-economische achtergrondgegevens van de leerlingen.

“Ja, die leerwinst, als het over het cognitieve gaat, die gaat automatisch een stuk groter zijn in rustige scholen met een goed leer- en klasklimaat dan in andere, en dat is een sociaaleconomisch gegeven. Op dat vlak... in welke buurt staat die school en met welke leerlingen moet die werken?” (scholen, respondent 2)

Een aantal respondenten is echter van mening dat het aantal factoren, waarmee men rekening dient te houden, zo verschillend is, dat het onmogelijk is om alles te achterhalen. Andere respondenten menen dan weer dat in deze vorm van toegevoegde waarde-berekening scholen met een minder gunstige leerlinginstroom bevoordeeld worden:

“En daar geldt ook dat ik zeg van: je mag leerwinst niet op deze manier alleen feedbacken. Je moet leerwinst altijd... contextualiseren. Als je vertrekt van een zwak publiek ga je hogere leerwinst boeken.” (inhoudsdeskundigen, respondent 1)

Interne kwaliteitszorg

Respondenten vinden dat scholen en pedagogische begeleidingsdiensten de resultaten van de leerwinstmeting kunnen aanwenden om het beleid van de school op een kritische wijze te benaderen. De outputanalyse van de school kan dan ook een meerwaarde zijn in het kader van interne kwaliteitszorg.

“Wat mij hier dus goed lijkt, is dat hier de leerwinst eigenlijk iets zegt op schoolniveau, niet op leerlingniveau.” (onderwijsverstrekkers, respondent 3)

Een groot aantal respondenten vindt het tevens een voordeel dat scholen de kans krijgen om hun eigen interne kwaliteitszorg af te toetsen bij vergelijkbare scholen. Ze geven aan dat het noodzakelijk is om de onderwijskwaliteit binnen de school tijdig te evalueren en dat leraren aan de hand van de resultaten (op de brede toetsen) hun klaswerking beter kunnen bijstellen. Door de resultaten te vergelijken met andere scholen kan een representatief kader van de eigen schoolwerking tot stand komen.

De lange periode tussen toetsafname en feedback zal volgens sommige respondenten evenwel maken dat het nut van de resultaten voor de interne kwaliteitszorg verkleind wordt, waardoor de opgeleverde feedback zijn doel kan missen.

2.5.3 Kritiek en onwenselijke effecten

Ook in dit scenario wijzen respondenten op de negatieve gevolgen die de openbaarheid van de resultaten met zich meebrengt. Door de rankings en profileringsdrang van scholen, wordt vermoed dat, net zoals in het vierde scenario, de sociale segregatie zal toenemen. Daarnaast worden ook andere onwenselijke neveneffecten van de openbaarheid bij dit scenario vermeld, zoals een emotionele impact op leraren en *teaching to the test*. Aangezien dezelfde argumenten hier opgaan als in scenario 4, worden deze hier niet hernomen.

Daarnaast geven respondenten aan dat de verantwoording van scholen te sterk in dit scenario is vervlochten, waardoor het doel om de kwaliteitsontwikkeling te bevorderen, wordt bedreigd. Tot slot wijzen respondenten op een aantal voorwaarden binnen het scenario die de autonomie van scholen mogelijks ondermijnen, en is er kritiek op het gebruik van de voorgestelde categorieën waarin scholen worden ingedeeld.

Verantwoordingsgericht gebruik van scholen

Omwille van bovengenoemde risico's (concurrentie, rankings en de name en blame-cultuur) vermoeden enkele respondenten dat scholen meer geneigd zullen zijn om de resultaten te bevorderen omwille van verantwoordingsgerichte motieven, zoals het hoog houden van de reputatie van de school. Dit zou leiden tot onwenselijk strategisch gedrag van de school, wat dan weer het beoogde effect van een bijdrage aan de kwaliteitszorg van de school bedreigt: de resultaten zullen volgens verschillende respondenten immers geen betrouwbare weergave zijn van de realiteit.

Inbreuk op de autonomie van scholen

Doordat scholen verplicht kunnen worden om een verbeteringsplan te ontwikkelen, stellen sommige respondenten de zelfverantwoordelijkheid van de scholen in vraag. De controlerende werking van de onderwijsinspectie lokt gelijkaardige reacties uit. Ook de volgorde van de toetsen zou volgens sommige respondenten een inbreuk zijn op de autonomie van de school, aangezien op die manier de organisatie in een vast jaarklassensysteem zou worden bestendigd.

De autonomie van de scholen om zelf een gewicht toe te kennen aan de resultaten op leerlingenniveau lijkt dan weer een twistpunt te zijn voor verschillende respondenten. Ze menen dat dit de vergelijkbaarheid tussen de scholen negatief beïnvloedt, aangezien er hogere cijfers verwacht worden bij scholen die de resultaten wel laten meetellen dan bij scholen die de toetsresultaten enkel aanwenden in het kader van kwaliteitszorg.

Gebruik van categorieën

Scholen worden ingedeeld in categorieën per leerdomein volgens het percentage leerlingen dat de verwachte leerwinst heeft bereikt. Dit biedt volgens sommige respondenten evenwel geen meerwaarde en maakt de interpretatie van de resultaten nodeloos ingewikkeld met een verhoogd risico op stigmatisering van scholen die onderaan in de ranking scoren.

“En dan denk ik ook ... wat is dan precies die waarde van die categorieën? Is dat dan niet een versterking en stigmatisering?” (inhoudsdeskundigen, respondent 6)

2.5.4 Haalbaarheid

Naast het openbaar karakter van de toets zorgt ook de bijzonder hoge kostprijs voor de inschatting dat dit scenario onhaalbaar is ($M=1.98$). De hoge maatschappelijke kosten en “de eenzijdige focus op mesoniveau” wegen volgens de respondenten niet op tegenover de voordelen die het scenario biedt, waardoor de haalbaarheid sterk daalt. Respondenten menen dan ook dat er geen draagvlak aanwezig is bij de verschillende onderwijsactoren om met dit scenario effectief aan de slag te gaan.

“En ja, die jaarlijkse kostprijs van dertig miljoen euro. Dat lijkt mij compleet te gek voor woorden. Ik denk dat we daarmee toch 600, 700 mensen te werk zouden kunnen stellen in het basisonderwijs. Dat lijkt mij op zich toch wel een veel interessanter idee dan 30 miljoen per jaar in een toets te gaan stoppen.” (beleidsmakers, respondent 3)

Een geleidelijke aanpak kan volgens respondenten het draagvlak mogelijks vergroten, al menen sommige respondenten dat het ook relevant kan zijn om de bestaande peilingen uit te breiden naar dit scenario waarbij op regelmatige basis een representatieve steekproef van scholen geëvalueerd wordt.

2.6 Leerwinstmeting op schoolniveau met afgeschermd karakter'

Tabel 23 - Wenselijkheid en haalbaarheid van scenario 6

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	3.75	2.83	4.13	3.00	3.43
Haalbaarheid	3.50	3.00	4.00	3.17	3.42

2.6.1 Wenselijkheid

Hoewel dit scenario veel gelijkenissen vertoont met scenario 5, wordt de wenselijkheid ervan gemiddeld veel hoger ingeschat ($M=3.43$). Dit kan worden verklaard door het afschermen van de resultaten, waardoor er geen gegevens publiek bekend worden gemaakt. Hierdoor worden heel wat van de veronderstelde neveneffecten van het vorige scenario vermeden. 14 van de 24 respondenten geven aan dat ze dit scenario (eerder) wenselijk inschatten, terwijl slechts drie respondenten dit scenario beoordelen als (eerder) onwenselijk. Eveneens interessant is het relatief groot aantal respondenten ($n=7$) dat geen uitgesproken standpunt inneemt en het scenario wenselijk noch onwenselijk inschat.

"Ik heb mij afgevraagd, voor mezelf: ben ik als directeur hier rapper dan geneigd om er hier mee te sjoemelen? Want teaching to the test noem ik sjoemelen hé. Pas op, dat is verantwoord sjoemelen want iedereen doet het. Maar ben ik daartoe geneigd hier? Nee!" (scholen, respondent 1)

2.6.2 Positieve effecten

De respondenten menen dat dit scenario heel wat kansen biedt aan scholen en leraren op vlak van interne kwaliteitszorg, autonomie en verdere professionalisering.

Leerwinstmeting in het kader van interne kwaliteitszorg

De respondenten vinden het een voordeel dat de resultaten van de leerwinstmeting op mesoniveau bruikbaar zijn in het kader van interne kwaliteitszorg. Om de focus op ontwikkeling en kwaliteitszorg te houden, vindt een aantal respondenten (net als bij scenario 5) dat het onwenselijk is om scholen zelf te laten bepalen of ze al dan niet een gewicht toekennen aan de resultaten van de leerlingen. Het toevoegen van een verantwoordingsgericht perspectief op microniveau door sommige scholen, zou immers voor een scheeftrekking van de resultaten kunnen zorgen.

Verder zou de validiteit van de toetsen een meerwaarde kunnen betekenen voor een meer objectief en kwalitatief evaluatiebeleid op de scholen. Ook de digitale, automatische verbetering zorgt voor een neutrale beoordeling en werkt bovendien tijdbesparend, waardoor respondenten denken dat scholen meer ruimte krijgen om andere noden op de agenda te plaatsen.

Autonomie van de school wordt gerespecteerd

Door scholen een zekere vorm van vrijheid te bieden in het moment van de toetsafname en in het verlenen van inzage recht voor de leden van de onderwijsinspectie, vinden respondenten dat de autonomie van de school in grote mate behouden blijft. Dit scenario behoudt volgens een aantal respondenten het best evenwicht tussen de belangen van de overheid en de belangen van scholen.

“Ja, ik ben eigenlijk wel gecharmeerd door dit scenario omdat ik vind dat het best of het meest het midden houdt tussen enerzijds een krachtig instrument met betrekking tot kwaliteitsvorm en anderzijds voldoende verantwoordelijkheid laten voor scholen, voldoende autonomie en voldoende pedagogische vrijheid geven. Mij lijkt dit een scenario waar zowel, om het nu een beetje karakterieel te zeggen, waar zowel de belastingbetalers ‘ik financier hier zoveel miljard onderwijs dus mag ik er vanuit gaan dat dat kwaliteitsvol onderwijs is’ als het principe dat meer vast hangt aan die pedagogische vrijheid en autonomie, volgens mij worden die hier toch wel het meest met elkaar verzoend. Beide zullen een beetje ongelukkig zijn, maar ik zou hier persoonlijk durven voor pleiten.” (beleidsmakers, respondent 3)

Beter lesgeven door leraren, betere opvolging van leerlingen

De leerwinstmonitoring van individuele leerlingen kan voor leraren een handig instrument zijn om de opvolging van leerlingen te optimaliseren, maar respondenten geven ook aan dat dit systeem een meerwaarde kan zijn om leraren te stimuleren om meer vanuit doelen te werken dan vanuit de methoden.

“Ik denk dat dat zelf een verbetering zou kunnen zijn, omdat methodes zijn iets te heilig momenteel in het basisonderwijs. En het zou de leraren duidelijk ertoe aanzetten om hun methode kritischer te bekijken en ook anders te gaan denken omdat die toets, die OVSG toets, zal ik maar zeggen, als die doorgetrokken wordt in alle leerjaren, die is toch altijd wel net iets anders. Ik zou dat een verbetering vinden, ja.” (scholen, respondent 4)

De feedback kan er bovendien voor zorgen dat leraren een correct beeld van zichzelf en hun didactische aanpak krijgen.

2.6.3 Kritiek en onwenselijke effecten

Door het afschermen van de resultaten van de leerwinstmeting wordt volgens de respondenten een groot aantal neveneffecten vermeden. Zo zou tussenschoolse concurrentie vermoedelijk niet toenemen omdat er geen vergelijking tussen scholen mogelijk is. Uit de gesprekken blijkt verder dat de respondenten weinig tegenargumenten tegenover dit scenario op tafel leggen. Wel menen enkele respondenten dat het leerwinstmonitoringssysteem de scholen slechts in beperkte mate aanzet tot een verbetertraject, en dat het verantwoordingsgebruik binnen de schoolmuren een risico kan zijn. Tot slot vinden respondenten dat de rol van de onderwijsinspectie niet eenduidig is.

Beperkte stimulus tot verbetering

Het afgeschermd karakter kan er volgens een groot aantal respondenten toe leiden dat scholen en leraren minder gemotiveerd zullen zijn (in vergelijking met meer high-stakes scenario's) om inspanningen te leveren die de onderwijskwaliteit op school zullen bevorderen. Een aantal respondenten meent dat enkel de “goede” leraren aan de slag zullen gaan met de verkregen feedback, terwijl leraren die weinig veranderingsgezind zijn meer weerstand zullen bieden.

“En als het afgeschermd is ja, dan krijg je iets zoals vroegtijdig schoolverlaten cijfers, die gewoon op schoolniveau zitten. En niemand weet van de andere scholen in de buurt waar het wel zit. Dus ja... dat bestuur weet het niet, zelfs de stad weet het niet. Je

krijgt gewoon eigenlijk geen prikkel om het anders en beter te doen.”
(inhoudsdeskundigen, respondent 2)

Verantwoordingsgericht gebruik binnen de school

Een ander onwenselijk gevolg waaraan respondenten denken, is het verantwoordingsgericht gebruik van de resultaten binnen de schoolmuren, waarbij leraren zich bijvoorbeeld moeten verantwoorden indien er onvoldoende leerwinst geboekt wordt. Hoewel het scenario uitgaat van een low-stakes leerwinstmeting waarbij de schoolontwikkeling centraal staat, kan de toets snel een high-stakes karakter voor zowel leraren als scholen krijgen.

“Moest ik directeur zijn zou ik hier heel blij mee zijn eigenlijk, dan zou ik denken van, ik kan mijn leraren motiveren, ik kan werken aan mijn interne kwaliteitszorg, ik kan leraren vergelijken met elkaar eigenlijk van wie, wie geeft er het beste les, dat mag niet uitgesproken worden, maar ik vind dat dat wel moet kunnen.” (scholen, respondent 8)

Onduidelijk kader voor de onderwijsinspectie

Uit de gesprekken blijkt dat respondenten vragende partij zijn voor goede richtlijnen die de rol van de inspectie verduidelijken. Door het verplicht karakter van het leerwinstmonitoringssysteem lijkt het sommige respondenten vreemd dat de onderwijsinspectie geen automatisch inzagerecht heeft in de resultaten. Ook denken zij dat enkel scholen in ‘slechte papieren’ dit inzagerecht zullen weigeren, waardoor de inspectie des te meer redenen kan hebben om argwanend naar de school te kijken.

“Ja, en dat [kader voor de onderwijsinspectie] vonden we wel heel sympathiek, maar we hadden wel zo iets van ja maar een school gaat dat dan niet weigeren, naar inspectie toe.” (onderwijsverstrekkers, respondent 1)

Respondenten stellen zich ook de vraag op welke manier de deelname aan de metingen gecontroleerd zal worden en of het inzagerecht het ontwikkelingsgericht uitgangspunt van dit scenario kan ondermijnen wanneer de inspectie gebruik zal maken van de resultaten.

Betrouwbaarheid

De min of meer verspreide afname van de toetsen geeft, volgens een aantal inhoudsdeskundigen, een onbetrouwbaar beeld van de resultaten. Leerlingen krijgen immers net meer of minder tijd om bepaalde leerdomeinen in te oefenen waardoor er ruis op de data kan zitten. Een aantal respondenten hekelt de vrijheid in het keuzemoment en stellen dat de overheid meer sturend moet optreden in het geval van een dergelijk verplicht systeem.

Tot slot geven enkele respondenten aan dat de resultaten eveneens vertekend kunnen worden doordat digitale toetsen toch enkele beperkingen kennen tegenover de klassieke pen-en-papier toets. Voorbeelden die hierbij worden genoemd, betreffen de beperkte antwoordmogelijkheden bij een digitale toets, of de verschillen inzake ICT-vaardigheden van leerlingen die de betrouwbaarheid (en validiteit) van de meting kunnen verstoren.

2.6.4 Haalbaarheid

Uit de resultaten van Tabel 23 blijkt dat respondenten in het algemeen het scenario gemiddeld als eerder haalbaar inschatten ($M=3.42$). Een groot aantal respondenten vindt dat het systeem sterk aansluit bij bestaande koepelgebonden toetsen, zoals de OVSG-toetsen en interdiocesane toetsen.

Hoewel zij een meerwaarde zien in het valideren van deze toetsen en in het gebruik van leerwinstmeting, menen ze dat het verplichtend karakter van het leerwinstmonitoringssysteem het draagvlak bedreigt.

Andere respondenten zijn van mening dat dit leerwinstmonitoringssysteem naar het beeld van een “ideaal schoolfeedbackproject” neigt. Vooral in het secundair onderwijs zou de vraag naar een dergelijk systeem meer prangend aanwezig zijn, terwijl respondenten menen dat in het basisonderwijs de leerlingvolgsystemen voldoen aan de meeste verwachtingen. Echter zou de haalbaarheid voor dit systeem in het secundair onderwijs minder groot zijn omdat de verscheidenheid aan vakken, leerplannen en studierichtingen zeer groot is.

Het scenario benadrukt dat een zware eenmalige kost nodig is om de ontwikkeling van het leerwinstmonitoringssysteem te voltooien. Doordat de ontwikkeling gespreid kan worden en doordat de recurrente kostprijs van acceptabel geacht wordt, neemt de haalbaarheid volgens de respondenten toe.

2.7 Facultatieve toetsenbank

Tabel 24 - Wenselijkheid en haalbaarheid van scenario 7

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	2.88	3.83	3.13	4.20	3.51
Haalbaarheid	4.25	3.50	4.13	3.80	3.92

2.7.1 Wenselijkheid

Uit Tabel 24 kan worden afgeleid dat ook dit scenario door de respondenten gemiddeld als eerder wenselijk wordt beschouwd ($M=3.51$). Precies de helft van de respondenten schat het scenario (eerder) wenselijk in, terwijl vier respondenten hier minder positief tegenover staan. Acht respondenten schatten het scenario noch wenselijk, noch onwenselijk in. Het vrijblijvend karakter van de toetsen binnen dit scenario zorgt volgens hen voor zowel belangrijke voor- als nadelen.

2.7.2 Positieve effecten

Het facultatieve karakter van de toets zorgt volgens de respondenten voor enkele leerkansen bij scholen en leraren. Verschillende respondenten menen dat met dit scenario het beoogde doel - een positieve impuls in het kader van interne kwaliteitszorg - kan worden bereikt. Daarnaast kan dit scenario volgens sommige respondenten leiden tot een versterkt bewustzijn rond kwaliteitsvol evalueren.

Interne kwaliteitszorg

De uitgestrekte, vakoverschrijdende inhoud van de toetsen en de mogelijkheid om ook met op sociaal-emotionele aspecten gerichten observatielijsten aan de slag te gaan, worden goed onthaald door de respondenten. Doordat leraren en scholen op een geheel vrijblijvende manier toetsen kunnen ontlenuen aan de toetsenbank kan ook de vakgroepwerking een evaluatie-instrument in handen krijgen om de effectiviteit van de onderwijsprocessen in kaart te brengen. De resultaten van de toetsen kunnen op schoolniveau een grote meerwaarde vormen in het kader van interne kwaliteitszorgprocessen.

“En dat is een beetje mijn principe van schoolhouden, dus van laat ons nu eens wereldoriëntatie in de focus zetten. Laat ons de analyse maken van ons leerplan, van ons onderwijsaanbod, dekt dat mekaar en dan ook als we daaruit zijn, laat ons een keer gaan toetsen: hebben we de dingen bereikt die we wilden bereiken? Wat scoort er wel goed, wat scoort er niet goed? Hoe komt het dat dat niet goed scoort? Bij welke leerlingen scoort dat niet goed? Eigenlijk vind ik dat heel veel kansen hebben.”
(onderwijsverstrekkers, respondent 6)

Impuls voor het bewustzijn rond kwaliteitsvol evalueren

Zoals eerder vermeld, is een groot aantal respondenten van mening dat er in Vlaanderen een grote behoefte aan kwaliteitsvol evaluatiemateriaal bestaat. De toetsenbank kan een aanzet zijn om een degelijk evaluatiebeleid uit te bouwen. Een toetsenbank met goede (gevalideerde) toetsen kan een inspiratiebron kan zijn voor leraren om professionele instrumenten te leren ontwikkelen.

“Het is eigenlijk ontstellend hoeveel weinig echt goeie toetsen er in Vlaanderen zijn en ik denk dat als je een toetsenbank kan opbouwen waar echt wel degelijke toetsen zitten die betrouwbaar zijn en geëvalueerd zijn en je leraren er toe de mogelijkheid biedt om die in te werken in het onderwijs, dat je daar een mogelijke potentiële meerwaarde hebt, een mogelijke win-win situatie hebt.” (inhoudsdeskundigen, respondent 6)

Door de vrijblijvendheid van het systeem kunnen scholen en leraren experimenteren zonder dat misstappen gesanctioneerd worden, waardoor het bewustzijn rond zelfreflectie en evalueren in scholen vergroot wordt, aldus de respondenten.

“Dat vind ik wel een meerwaarde, omdat ik denk dat we eigenlijk wel, en dat we daar in de toekomst denk ik veel meer op zullen moeten inzetten denk ik, dat we in ons professionalisering en hier komt de lerarenopleiding veel bewuster moeten nadenken over evaluatie, toetsen en die meting en dit meer.” (onderwijsverstrekkers, respondent 3)

2.7.3 Kritiek en onwenselijke effecten

Respondenten vermoeden dat de kans op succes in sterke mate zal afhangen van de motivatie en het beleid in scholen. Net als in scenario 6 geeft het voorgestelde scenario mogelijk maar een beperkte prikkel aan de school tot kwaliteitsverbetering. Respondenten vinden het daarom wenselijk dat zowel onderwijsverstrekkers (inclusief hun pedagogische begeleidingsdiensten) als de overheid scholen en leraren aanmoedigen om gebruik te maken van deze toetsenbank. Daarnaast geven respondenten aan dat door de focus op enkel het mesoniveau en het gebrek aan leerwinstmeting en toegevoegde waarde-meting slechts beperkte informatie wordt aangeleverd. Ten slotte vragen enkele respondenten zich af of er effectief een behoefte bestaat aan dit scenario.

Beperkte stimulus voor scholen

Hoewel de autonomie van de scholen bewaard wordt, denkt een groot aantal respondenten dat de toetsen enkel zullen worden aangewend door scholen waar er voldoende “maturiteit aanwezig is”. Van scholen met een hoog beleidsvoerend vermogen verwachten respondenten dan ook dat ze toetsen zullen kiezen op basis van behoeften en doelen, terwijl scholen met een laag beleidsvoerend vermogen minder snel gebruik zullen maken van de toetsenbank. Sommige respondenten trekken

deze redenering door naar het niveau van de leraren en stellen dat het steeds dezelfde leraren zullen zijn die deze tools gaan aanwenden.

“Wacht wacht. Mijn goeie leraren nemen het af. En de leraren die het eigenlijk zouden nodig hebben, die 10%, het is er maar zo weinig. Die leraren die het eigenlijk zouden nodig hebben omdat ze na al die jaren nog geen les kunnen geven, en die vaak ook vast benoemd zijn van voor mijn tijd zelf nog hé en van veel ervoor... Die nemen dat niet af. Die nemen dat niet af. Die durven dat niet. En dan zit ik op een functioneringsgesprek te vragen: heb je dat al eens gedaan? En dan krijg ik het antwoord: nee, ik geloof daar niet in.” (scholen, respondent 1)

Omwille van bovenstaande redenen pleiten respondenten er dan ook voor dat de overheid en/of pedagogische begeleidingsdiensten scholen en leraren moeten sensibiliseren inzake het gebruik van de toetsenbank.

“Als zoiets wordt gelanceerd vanuit de overheid moet er daar echt een sensibiliseringcampagne bij komen van waarom doen we dat, wat is de meerwaarde daarvan, wat kunnen we daaruit leren, hoe gaan we daarmee aan de slag?” (beleidsmakers, respondent 4)

Beperkte informatie

Ondanks het feit dat dit monitoringssysteem veel informatie kan opleveren in de verschillende leerjaren en graden, vinden respondenten het jammer dat deze toetsenbank slechts resultaten oplevert die bijna uitsluitend op mesoniveau bruikbaar zijn. Doordat er geen sprake is van leerwinstmonitoring op zowel leerlingniveau en evenmin van een bepaling van toegevoegde waarde wordt het volgens respondenten moeilijk om de evolutie van leerlingen en leraren in kaart te brengen.

Bestaande nood?

Daarnaast geeft een groot aantal respondenten aan dat het scenario voor te weinig vernieuwing zorgt in het huidige onderwijslandschap. Zo kunnen basisscholen en secundaire scholen nu reeds in grote mate beroep doen op koepelgebonden toetsen, leerlingvolgsystemen, paralleltoetsen alsook op toetsen die uitgaan vanuit de diverse methodes. Deze evaluatie-instrumenten zouden gelijkaardige informatie genereren dan de toetsen uit de toetsenbank, hoewel respondenten menen dat dit scenario voor het secundair onderwijs eventueel een meerwaarde kan hebben. Respondenten vinden dit scenario dan ook meer interessant “als een surplus”, aansluitend bij andere scenario's.

2.7.4 Haalbaarheid

In het algemeen schatten respondenten de haalbaarheid van dit scenario eerder hoog in ($M=3.92$). Niet alleen de vrijblijvendheid zou hiertoe bijdragen, ook het ontwikkelingsgericht karakter van de toets en het behoud van de autonomie van scholen zorgen volgens respondenten voor een groot draagvlak bij onderwijsverstrekkers en scholen.

“Het is vrij goedkoop, het is vrij eenvoudig om te doen. (...). Je raakt ook niet aan de koepels, je raakt nergens aan. Iedereen behoudt zijn autonomie enzovoort. “ (Inhoudsdeskundigen, respondent 2)

Respondenten zijn echter minder overtuigd van de politieke bereidheid om dergelijke toetsenbank te financieren, ondanks de beperkte recurrente kost (950.000€), en geven aan dat dit scenario een te grote investering vraagt om daar een vrijblijvende output tegenover te stellen. Ook vragen ze zich af of de overheid wel bereid is om blijvend te investeren in een up-to-date toetsenbank met bijhorende ondersteuning.

2.8 Huidige trend wordt verdergezet

Tabel 25 - Wenselijkheid en haalbaarheid van scenario 8

	Beleidsmakers	Onderwijsverstrekkers	Scholen	Inhoudsdeskundigen	Gemiddelde
Wenselijkheid	2.00	4.40	2.00	3.50	2.98
Haalbaarheid	4.00	5.00	4.13	5.00	4.53

2.8.1 Wenselijkheid

De cijfers in Tabel 25 geven duidelijk aan dat de meningen van de verschillende categorieën van respondenten inzake de wenselijkheid van het achtste scenario vrij uiteenlopend zijn ($M=2.98$). 11 respondenten geven aan dat ze dit scenario enigszins tot helemaal onwenselijk vinden, terwijl slechts 9 respondenten dit scenario wel degelijk wenselijk vinden. Vier respondenten vinden dit scenario noch wenselijk, noch onwenselijk. Het blijken vooral de respondenten uit de categorie van 'beleidsmakers' en 'scholen' te zijn die het scenario (eerder) onwenselijk inschatten ($M=2.00$ voor beide groepen), terwijl de categorie van onderwijsverstrekkers dit scenario zeer wenselijk blijkt te vinden ($M=4.40$).

2.8.2 Positieve effecten

Voorstanders van dit scenario geven aan dat er een grote noodzaak is aan verdere professionalisering van leraren en scholen inzake datageletterdheid (en informatiegebruik) alsook op het vlak van evaluatiebeleid. De titel van het scenario 'huidige trend wordt vergezet', dekt volgens verschillende respondenten dan ook niet helemaal de lading.

"Ja maar ik vind dit niet "de huidige trend wordt verder gezet". Het zit er misschien in het woordje trend: je voegt hier heel veel aan toe vind ik. Ik vind dit niet "blijven bij de huidige situatie", ik vind dit een serieuze optimalisering van de huidige situatie."
(inhoudsdeskundigen, respondent 6)

Investeren in datageletterdheid

Uit de gesprekken blijkt dat er een grote vraag is naar het professionaliseren van scholen, leraren en ondersteunend personeel op het vlak van datageletterdheid. Eén respondent stelt echter vast dat deze behoefte ruimer is dan enkel het verwerken van statistische data en wijst erop dat 'informatiegebruik' een betere conceptualisering zou zijn.

Respondenten vinden dat deze ondersteuning zowel vanuit pedagogische begeleidingsdiensten als vanuit wetenschappelijke hoek mag komen. Het inzetten op deze vorm van professionalisering zou de onderwijskwaliteit doen toenemen, aldus de respondenten.

“Goh, ja voor een stuk kunnen de begeleidingsdiensten dat zeker opvangen, maar ik denk niet volledig, niet alleen omwille van tijdgebrek en zoveel andere taken, maar misschien omdat ze zelf ook niet altijd de nodige expertise in huis hebben dus ik denk dat ze daar zeker vanuit wetenschappelijke hoek ondersteuning in zouden moeten krijgen dan. Het is zeker niet iets dat zij alleen kunnen doen. En ja misschien ook de overheid daar dan ook maar meer middelen in voorzien, in een breed professionaliseringstraject.” (beleidsmakers, respondent 1)

Investeren in evaluatiebeleid

Zoals ook eerder vermeld (zie 2.7.3) geven respondenten aan dat met de huidige evaluatiemechanismen in het basisonderwijs al heel wat doelen bereikt kunnen worden, maar dat er vooral in het secundair onderwijs weinig instrumenten bestaan die de onderwijskwaliteit in kaart brengen. Door leraren en scholen verder te professionaliseren in het evalueren zou de beoordeling van leerlingen en onderwijskwaliteit meer betrouwbaar moeten worden, aldus de respondenten.

“Dat is een voordeel aan goed ontwikkelde toetsen, dat mensen dezelfde taal over zaken beginnen te spreken. Ze gaan, ze spreken over hetzelfde, het krijgt vulling.” (beleidsmakers, respondent 4)

De resultaten van betrouwbare toetsen leveren op hun beurt dan weer input om de interne kwaliteitszorgprocessen op de school verder te zetten.

Nadelen van gestandaardiseerde toetsen worden vermeden

Voorstanders van dit scenario vinden het vooral een voordeel dat de nadelen die in eerdere scenario's aan bod zijn gekomen, hier niet van toepassing zijn. Ze doelen dan op de afwezigheid van rankings, het wegvallen van de kans op *teaching to the test* en ander onwenselijk strategisch gebruik door scholen en leraren. Ook zijn ze tevreden van het huidig onderwijsklimaat waarin ruimte is voor vakoverschrijdende aspecten, het pedagogisch project van de school en het welbevinden van leerlingen.

2.8.3 Onwenselijke effecten

Ondanks bovenstaande positieve effecten van het huidige scenario geeft de meerderheid van de respondenten aan dat er een groot aantal onwenselijkheden in het huidige onderwijssysteem te vinden zijn. Zo zijn respondenten ervan overtuigd dat er grote (kwaliteits)verschillen tussen scholen bestaan, en dat de verantwoordingsgerichte initiatieven momenteel te vrijblijvend zijn, waardoor niet alleen sommige scholen, maar ook de onderwijsverstrekkers en de overheid niet altijd de vinger aan de pols weten te houden. Daarnaast menen heel wat respondenten dat het evaluatiebeleid in een aantal scholen op een niet-systematische manier wordt gevoerd. Ook zijn er een aantal factoren op het vlak van de schoolorganisatie die volgens respondenten belemmerend werken. Tot slot geven respondenten aan dat de huidige evaluatie-instrumenten weinig relevante informatie genereren.

“Nee, ik heb daar drie dingen bijgeschreven van het is te vrijblijvend, er is teveel willekeur en ik vind dat de verantwoordingsplicht, dat dat noodzakelijk is, zoals, ik heb het over in heel Vlaanderen nu, want iedereen doet maar waar hij eigenlijk zin in heeft.” (scholen, respondent 4)

Onzekerheid over de onderwijskwaliteit

In het huidige onderwijssysteem is het zicht op onderwijskwaliteit beperkt, aldus de respondenten. De koepelgebonden toetsen, die specifiek gericht zijn op een evaluatie op schoolniveau, worden pas op het einde van de basisschool afgenomen, waardoor er geen processen van de leerling of klassen in beeld worden gebracht. Respondenten vinden het bovendien jammer dat de peilingen en paralleltoetsen een wisselende focus hebben waardoor ook de resultaten van deze metingen een weinig systematisch beeld geven van de schoolwerking. De nood aan gestandaardiseerde toetsen om de onderwijskwaliteit te meten, is dan ook goed voelbaar tijdens de gesprekken.

Respondenten zijn er in grote mate van overtuigd dat er grote verschillen tussen de onderwijskwaliteit in verschillende scholen bestaan. Ze gaan ervan uit dat in scholen met een sterk beleidsvoerend vermogen vanuit een onderbouwde visie wordt gewerkt, maar dat er evengoed scholen zijn die nog steeds niet vanuit een bepaalde visie werken of met degelijke actieplannen aan de slag gaan. Het recht op kwalitatief goed onderwijs blijkt toch één van de grootste prioriteiten te zijn bij de respondenten.

“Maar de manier waarop nu een advies drie gegeven worden [door de onderwijsinspectie] of niet, ik heb op een bepaald moment gezegd... eigenlijk zouden we die school moeten sluiten, dat is een vorm van kindermishandeling... want ik vind kinderen hebben recht op goed onderwijs.” (scholen, respondent 1)

Niet-systematisch evaluatiebeleid in sommige scholen

Zoals eerder gesteld juichen de respondenten de investering in evaluatie sterk toe. Respondenten erkennen het gebrek aan inzicht in het opstellen van goede toetsen bij tal van leraren en scholen. Zo weten ze niet altijd welke norm gehanteerd wordt om leerlingen een A-attest toe te kennen, waardoor er geen objectieve vergelijking tussen de verschillende scholen mogelijk is. Daarnaast vermoeden ze ook dat sommige leraren en scholen uitspraken doen over leerwinstmeting op basis van niet-vergelijkbare toetsen. De toetsen die momenteel door leraren worden ontwikkeld, zullen volgens de respondenten niet betrouwbaar genoeg zijn om een goed zicht te hebben op de ontwikkeling van leerlingen aangezien ze niet altijd vanuit de leerplandoelen worden opgesteld. Respondenten hopen dan ook dat de overheid en/of onderwijsverstrekkers inzetten op het valideren van bestaand toetsmateriaal.

“Dat zou denk ik nog een een mogelijke piste zijn. Dat je zegt ok... De toetsen die er zijn, zijn zeker niet slecht, maar we gaan... misschien die toetsontwikkelaars nog meer ondersteuning bieden naar validering van de toetsen toe... Of we gaan mee zoeken hoe we inderdaad een heel uitgebreide digitale toetsenbank kunnen ontwikkelen.” (inhoudsdeskundige, respondent 5)

Huidige informatie is onvoldoende

Tot slot geeft een groot aantal respondenten aan dat de huidige evaluatie-instrumenten weinig meerwaarde bieden dan het louter vaststellen van de leerprestaties van de leerling. De bevraagde leraren geven dan ook aan dat ze op zoek zijn naar systemen die specifieke problemen kunnen detecteren, maar die eveneens een oplossing aanbieden voor de gedetecteerde leerproblemen.

Ook de bevroegde beleidsmakers geven aan dat er vraag is naar diepgaande informatie op macroniveau. Door de hoge mate van vrijblijvendheid is het moeilijk om zwakke plekken in het systeem bloot te leggen.

2.8.4 Haalbaarheid

De haalbaarheid van dit scenario wordt door de respondenten in het algemeen als zeer groot ingeschat ($M=4.53$). Respondenten vermoeden dat er weinig problemen in het onderwijsveld zullen optreden als men de huidige trend verder zet. Doordat het evaluatiebeleid een gevoelig thema is voor onderwijsverstrekkers, zal door het niet verplichtende karakter het draagvlak groot zijn. Respondenten binnen de groep van 'onderwijsverstrekkers' geven dan ook aan dat ze het gebruik van dergelijke toetsen niet zouden verplichten, maar dat ze het gebruik wel sterk zouden aanmoedigen indien de toetsen stroken met de visie. Enkele respondenten denken dat de overheid eveneens geen vragende partij is om veel te wijzigen aan het huidige onderwijslandschap door het huidige klimaat van besparingen.

“Natuurlijk dat is perfect haalbaar. Het loopt al jaren zo dus natuurlijk, dat is eigenlijk voor de overheid het goedkoopst. Ze kunnen nog goedkoper maar dat wil zeggen dat ze niet meer investeren in peilingen, die stap achteruit kunnen zij zich niet permitteren.”
(inhoudsdeskundigen, respondent 1)

Tegenstanders van dit scenario zijn er evenwel van overtuigd dat er nood is aan verandering. De nadelen van het huidige evaluatiebeleid zouden te bepalend zijn om geen voorstander te zijn van een nieuwe manier van evalueren.

“Ik denk dat we momenteel in zo een tijdsgeest leven, als ik zie wat er op Europees vlak gebeurt, dat niets doen, dat dat momenteel geen optie is, daar ben ik van overtuigd. Er moet iets meer gebeuren dan een soort van stand-still die hier beschreven wordt, maar waarbij dat er wel op professionalisering ingezet wordt.”
(beleidsmakers, respondent 3)

Andere respondenten zijn ervan overtuigd dat dit scenario geen alleenstaand scenario mag zijn, maar dat er nood is om ook in elk van de andere scenario's extra in te zetten op datageletterdheid en professionalisering van scholen en leraren.

DEEL 8 CONCLUSIE

In dit boek werden de resultaten van uiteenlopende onderzoekslijnen besproken om vanuit uiteenlopende invalshoeken een licht te schijnen op de vraag naar de wenselijkheid en haalbaarheid van de ontwikkeling en implementatie van een leer(winst)monitoringssysteem. In dit deel geven we een overzicht van de belangrijkste conclusies die uit de rijkdom van resultaten kunnen getrokken worden.

1 DOELEN VAN HET LEER(WINST)MONITORINGSSYSTEEM

Indien de Vlaamse overheid wenst te investeren in een leer(winst)monitoringssysteem, dient men in de eerste plaats een concrete set aan doelen te bepalen voor dit monitoringssysteem. Uit de internationale studie blijkt immers dat de impact van een leer(winst)monitoringssysteem het grootste is, wanneer aan elk van de toetsen een éénduidig doel wordt gesteld, ingebed in een transparante rationale. De rationale maakt duidelijk hoe elke gestandaardiseerde toets in het leer(winst)monitoringssysteem bijdraagt tot het vooropgestelde overkoepelende doel.

De literatuurstudie toont aan dat zowel verantwoordingsgerichte systemen, als ontwikkelingsgerichte systemen bepaalde positieve impulsen kunnen geven aan het onderwijs. Sommige studies wijzen uit dat deze positieve effecten versterkt worden indien de leer(winst)monitoring een high-stakes karakter heeft, maar deze bevinding wordt niet door alle studies gedeeld. Vanuit de literatuur komen de volgende impulsen het meest prominent in beeld op microniveau: het verwerven van een beter beeld van de leerprestaties of leerwinst van leerlingen, en daardoor het beoordelen van leerlingen volgens een 'gelijke lat', betere leerprestaties van leerlingen en een meer effectief onderwijs door leraren. Op mesoniveau wordt verwezen naar een betere interne kwaliteitszorg van scholen, terwijl men op macroniveau meestal een betrouwbaar algemeen inzicht in de leerprestaties van leerlingen in het hele onderwijssysteem beoogt. Verder toont de literatuurstudie aan dat een high-stakes leer(winst)monitoringssysteem vaker dan een low-stakes systeem aanleiding geeft tot verschillende onwenselijke neveneffecten, zoals een verhoogde emotionele impact op leerlingen en leraren, teaching to the test en andere vormen van onwenselijk strategisch gedrag in scholen. Om fraude tegen te gaan, dienen de gestandaardiseerde toetsen in een high-stakes leer(winst)monitoringssysteem bovendien telkenmale volledig te worden vernieuwd, waardoor de kostprijs voor de toetsontwikkeling sterk wordt opgedreven.

Uit de resultaten van de Delphi-studie blijkt evenwel dat low-stakes ontwikkelingsgerichte doelstellingen op micro- of mesoniveau in de Vlaamse onderwijscontext de voorkeur verdienen bij de ontwikkeling van een leer(winst)monitoringssysteem. Scenario's die uitgaan van dit ontwikkelingsgerichte perspectief worden hoger ingeschat inzake zowel wenselijkheid als haalbaarheid. Meer concreet dient een leer(winst)monitoringssysteem in Vlaanderen bijkomende betrouwbare en valide informatie te genereren op microniveau waardoor leraren in staat worden

gesteld om het onderwijsaanbod maximaal af te stemmen op de noden van de leerlingen alsook om leerlingen beter inzicht te geven in hun eigen voortgang. Op mesoniveau is de verwachting dat het leer(winst)monitoringssysteem een bijdrage levert aan een informatierijke omgeving voor de school door het systeem van interne kwaliteitszorg van betrouwbare en valide informatie over de leerprestaties of de leerwinst van leerlingen te voorzien. Om dit doel op mesoniveau te verwezenlijken dient de informatie over de school - volgens de opinie van verschillende respondenten in de Delphi-studie - in perspectief te worden geplaatst met output-informatie van scholen met een vergelijkbare input en context.

Het internationale onderzoeksluik bevestigt doorheen de vier cases het inzicht uit de literatuurstudie dat het onwenselijk is om zowel ontwikkelingsgerichte als verantwoordingsgerichte doelen te stellen aan één en dezelfde gestandaardiseerde toets. Door het verantwoordingsgerichte karakter wordt de inzet van de toets immers verhoogd, wat in elke case leidt tot indicaties van onwenselijk strategisch gedrag van scholen en waardoor de resultaten aan betrouwbaarheid en daarom aan nut verliezen voor gebruik in een ontwikkelingsgericht perspectief. Ook in de literatuurstudie werden resultaten gerapporteerd waaruit blijkt dat bij het toevoegen van een verantwoordingsgericht gebruik, de impact op de ontwikkelingsgerichte doeleinden vaak worden gemarginaliseerd.

2 LEERWINST EN TOEGEVOEGDE WAARDE

Indien in de Vlaamse onderwijscontext een monitoringssysteem zou worden geïmplementeerd, zou dit kunnen gebruikt worden om niet enkel leerprestaties van leerlingen, maar ook de leerwinst van leerlingen en de toegevoegde waarde van scholen in kaart te brengen. In deze sectie brengen we de conclusies van de drie onderzoeksluiken samen inzake de wenselijkheid en haalbaarheid van een monitoring van leerwinst en van toegevoegde waarde.

2.1 Leerwinst monitoren

In de eerste sectie werd besproken dat Vlaamse stakeholders de voorkeur geven aan scenario's waarin ontwikkelingsgerichte doelen voorop staan en waarin het monitoringssysteem gericht is op het microniveau van de leerlingen en de klas en/of het mesoniveau van de school. In het kader daarvan is het belangrijk dat dit monitoringssysteem voldoende nieuwe informatie kan genereren voor zowel scholen als voor individuele leraren. Hoewel in de internationale cases vaak bleek dat men de bevestiging van bestaande overtuigingen en oordelen ook als waardevol beschouwt, geven Vlaamse respondenten in de Delphi-studie aan dat men van dit monitoringssysteem verwacht dat het nieuwe inzichten aanreikt. Deze doelstelling kan het best worden bereikt indien het monitoringssysteem zich niet beperkt tot het aanleveren van informatie over de leerprestaties van leerlingen. Twee van de vier scenario's die in de derde ronde van de Delphi-studie een positieve inschatting krijgen inzake wenselijkheid, betreffen de scenario's waarin de leerwinst van leerlingen in kaart wordt gebracht binnen een ontwikkelingsgericht perspectief (met name scenario 2 en scenario 6). Van leerwinstmonitoring verwachten de respondenten in de Delphi-studie dat deze informatie een beter beeld kan opleveren van de mogelijkheden en beperkingen van de leerling waardoor het monitoringssysteem bijkomende informatie oplevert die leraren door middel van hun eigen toetsen niet hadden kunnen genereren. De resultaten van een leerwinstmonitoring kunnen eveneens een zinvolle bijdrage leveren tot het systeem van interne kwaliteitszorg in de school.

De literatuurstudie onderscheidt uiteenlopende manieren waarop de leerwinst van leerlingen in kaart kan worden gebracht. De verschillende conceptualiseringen werden uitgebreid beschreven. Uit de internationale studie blijkt dat leerwinstmonitoring op initiatief van de overheid eerder zeldzaam is, maar eerder ligt bij private aanbieders. Bij deze leer(winst)monitoringssystemen bleek in de praktijk een *gain-based model* (en met name het *gain score* model) het meest gebruikelijke te zijn. Bovendien wordt dit model positief gewaardeerd door de meeste gebruikers ervan. Bij dit model dienen alle items op een geijkte vaardigheidsschaal te worden ondergebracht, bijvoorbeeld via Item Response Theory (IRT). Ondanks de voordelen van de meer geraffineerde manier van leerwinstberekening in de internationale casestudie over het pilootproject 'Leerwinstmonitoring en toegevoegde waarde in Nederland', blijkt dat de gebruiksvriendelijkheid van relatief eenvoudige *gain score* modellen sterker wordt gewaardeerd in een ontwikkelingsgericht perspectief. Verder blijkt ook uit het internationale onderzoeksluik dat leerwinstmonitoring volgens een *gain score* model relevant is wanneer de leerlingen relatief frequent kunnen worden getoetst. Op die manier kan een betrouwbaar beeld worden gevormd over de leerwinst van leerlingen en krijgen leraren regelmatig zinvolle informatie om hun onderwijs bij te sturen. Daar staat tegenover dat zowel in de Delphi-studie als in de internationale studie wordt gewaarschuwd voor het overmatige toetsen van leerlingen waardoor een onderwijssysteem kan terechtkomen in een overdreven 'toetscultuur'.

2.2 Toegevoegde waarde monitoren

Een ontwikkelingsgerichte impact op mesoniveau houdt voor de meeste respondenten in de Delphi-studie in dat scholen de leerwinst van hun leerlingen kunnen vergelijken met de leerwinst van vergelijkbare scholen (scholen met een soortgelijke input en context). Er wordt dus duidelijk de voorkeur gegeven aan vergelijking van leerwinst boven vergelijking van leerprestaties. Bij de leerwinstberekening ten behoeve van deze vergelijking tussen scholen wordt idealiter verder gebouwd op de *gain-based* modellen die hierboven werden beschreven (zie 2.1). Meer complexe modellen van toegevoegde waarde, beschreven in de literatuurstudie, werden in de internationale studie niet geïdentificeerd, met uitzondering van het pilootproject 'Leerwinst en toegevoegde waarde in Nederland'. De evaluatie van de resultaten van dit project (zowel de evaluatie uitgevoerd door de projectgroep zelf, als de evaluatie op basis van onze eigen bevindingen) was evenwel niet van die aard om te concluderen dat deze modellen elders zomaar overgenomen dienen te worden. Zeker wanneer de focus op een ontwikkelingsgericht perspectief behouden blijft, kan er uit de Delphi-studie en het internationale onderzoeksluik geconcludeerd worden dat een relatief eenvoudig (en voor scholen makkelijk te begrijpen en te hanteren) model van toegevoegde waarde de voorkeur verdient.

Een complicatie hierbij is dat het vooralsnog niet evident is om te bepalen wat voor een school een 'vergelijkbare school' is. Deze bezorgdheid komt aan bod in de literatuurstudie en wordt tevens geuit door de respondenten in de Vlaamse onderwijscontext in de Delphi-studie. Uit het internationale onderzoeksluik blijkt dat verschillende onderwijssystemen verschillende parameters gebruiken bij het identificeren van vergelijkbare scholen en dat de gemaakte keuzes vaak door groepen van stakeholders worden bekritiseerd. Eén van de conclusies van het Nederlandse pilootproject 'Leerwinst en toegevoegde waarde' is dan ook dat bijkomend onderzoek vereist is om te bepalen met welke factoren rekening dient te worden gehouden om een eerlijke vergelijking van scholen mogelijk te maken.

3 OPENBAARHEID VAN DE RESULTATEN

In elk van de casestudies in het internationale onderzoeksluik werden resultaten van bepaalde gestandaardiseerde toetsen openbaar gemaakt. Hoewel de openbaarheid van resultaten in verschillende cases een relatief groot draagvlak heeft, worden in elke case aspecten ervan door verschillende stakeholders gecontesteerd. Er werden indicaties gevonden dat de openbaarheid van resultaten kan leiden tot verschillende onwenselijke neveneffecten. Het voorbeeld van Engeland - en meer bepaald het Centre of Evaluation and Monitoring - geeft duidelijk aan dat wanneer men een ontwikkelingsgericht doel beoogt, de resultaten enkel door de betrokken school zelf mogen worden gehanteerd.

Ook in de literatuur blijkt dat de openbaarheid van resultaten van een leer(winst)monitoringssysteem op schoolniveau kan leiden tot een verhoogde transparantie die door bepaalde stakeholders gewaardeerd wordt, maar dat ze bijna onvermijdelijk tot gevolg heeft dat rankings van scholen worden opgemaakt, waarbij een competitie tussen scholen kan ontstaan en waarbij scholen beoordeeld worden op een slechts beperkt deel van hun ruime opdracht. Dit kan onwenselijk strategisch gedrag en zelfs fraude in de hand werken.

De resultaten van de Delphi-studie wijzen uit dat het in de Vlaamse context niet als wenselijk wordt gezien dat resultaten van gestandaardiseerde toetsen in een leer(winst)monitoringssysteem openbaar zouden worden gemaakt. De twee scenario's waarbij openbaarheid van resultaten op schoolniveau werd vooropgesteld (scenario 4 en 5) werden elk door iedere categorie van respondenten gemiddeld sterk negatief beoordeeld (scores ver beneden de neutrale score '3'). De openbaarheid van de gegevens is daarbij de belangrijkste reden om beide scenario's af te keuren. Rankings, sociale segregatie en teaching to the test (als vernauwing van het curriculum) zijn in de Delphi-studie de meest prominent genoemde onwenselijke neveneffecten van de openbaarheid van resultaten.

Er dient dus bij de implementatie van een leer(winst)monitoringssysteem gekeken te worden naar de (juridische) randvoorwaarden waardoor het vermijden van de openbaarheid van resultaten kan worden vermijden. Het niet-openbaar zijn van de resultaten van het leer(winst)monitoringssysteem heeft meteen ook gevolgen voor de rol van de Onderwijsinspectie. Als besloten wordt om deze resultaten niet openbaar te maken, mag de inspectie de resultaten uit het leer(winst)monitoringssysteem noch publiceren, noch gebruiken in het oordeel over de kwaliteit van de school. Dit is in overeenstemming met de vrij grote consensus tussen Vlaamse stakeholders in de Delphi-studie dat de onderwijsinspectie de resultaten enkel kan gebruiken om na te gaan hoe de school deze resultaten heeft gebruikt binnen de processen van interne kwaliteitszorg op micro- en mesoniveau. Op die manier geeft het leer(winst)monitoringssysteem de onderwijsinspectie een extra instrument om een uitspraak te doen over de mate waarin, en de kwaliteit waarmee de school haar eigen kwaliteit bewaakt.

4 EIGENSCHAPPEN VAN DE TOETSEN IN EEN LEER(WINST)MONITORINGSSYSTEEM

4.1 Toetsenbank

Uit de resultaten van de derde ronde van de Delphi-studie kan geconcludeerd worden dat het voor Vlaanderen wenselijk is om niet enkel te investeren in één enkele gestandaardiseerde toets, maar eerder in een omvattend leer(winst)monitoringssysteem waaruit scholen verschillende toetsen kunnen ontlelen. De scenario's die in deze richting werden ontwikkeld (scenario 3 en 7 betroffen respectievelijk een toetsenbank gericht op microniveau en een toetsenbank gericht op mesoniveau), werden beide inzake wenselijkheid en haalbaarheid gemiddeld positief beoordeeld door de respondenten in de Delphi-studie. Een toetsenbank is dan een uitgebreide database waarin betrouwbare en gevalideerde gestandaardiseerde toetsen zijn opgenomen in uiteenlopende leerdomeinen (al dan niet uitgebreid met niet-cognitieve aspecten).

In de internationale studie worden verschillende voorbeelden van toetsenbanken gerapporteerd, zoals bij de private toetsontwikkelaars ACER en EAA in Australië, en bij CEM in Engeland. Hierbij worden telkens zeer uiteenlopende toetsen aangeboden. In de internationale studie werd evenwel geen voorbeeld gedocumenteerd waarin op initiatief van een overheid een toetsenbank werd opgesteld met toetsen in een breed aantal leerdomeinen. Ook in de Delphi-studie geven enkele respondenten aan dat men - hoewel men de ontwikkeling van een toetsenbank wenselijk acht - er niet van overtuigd is dat het de rol is van de overheid om hierin te investeren. Ook op basis van de literatuurstudie is het moeilijk te concluderen of een toetsenbank moet worden ontwikkeld: empirische studies zijn bijna uitsluitend gericht op de effectiviteit van één gestandaardiseerde toets en zelden wordt de impact van een toetsenbank nagegaan. Dit is ook methodologisch een zeer uitdagende opdracht.

4.2 Sterke psychometrische onderbouwing

In de literatuurstudie blijkt dat heel veel is geschreven omtrent het belang van een sterke psychometrische onderbouwing van gestandaardiseerde toetsen die deel uitmaken van een leer(winst)monitoringssysteem. Betrouwbaarheid en validiteit zijn essentiële vereisten van de gestandaardiseerde toetsen in zulk systeem.

In het internationale onderzoeksluik stellen we vast dat in elk van de bezochte onderwijssystemen met heel veel zorg wordt omgegaan met de betrouwbaarheid van de toetsen. Hiertoe worden zowel inhoudsdeskundigen en toetsexperten als leraren van het betreffende leerdomein bij het ontwikkelingsproces betrokken (met uitzondering van enkele high-stakes toetsen waarbij men de groep ontwikkelaars bewust klein houdt om de kans op lekken te minimaliseren). De perceptie van betrouwbaarheid en validiteit bij de gebruikers van het leer(winst)monitoringssysteem is ook cruciaal voor de aanvaarding van de gestandaardiseerde toetsen alsook voor het gebruik van de resultaten door leraren en scholen.

Uit zowel de literatuurstudie als in het internationale onderzoeksluik concluderen we tevens dat veel zorg moet besteed worden aan het definiëren van de manier waarop de cesuur wordt bepaald tussen verschillende vaardigheidsniveaus, of tussen 'slagen' en 'niet-slagen' voor een toets. Ondanks de aanbeveling uit de literatuur om enkel vanuit bepaalde standaarden uit te gaan (die vooraf worden

bepaald voor elk vaardigheidsniveau), zien we in dat in internationale voorbeelden ook rekening wordt gehouden met het groepsnorm-criterium, waardoor de definitieve *cut-off points* na de toetsafname worden vastgelegd.

Ondanks de grote zorg voor kwaliteit en psychometrische onderbouwing, staat soms vooral de validiteit van gestandaardiseerde toetsen ter discussie: de vele voorbeelden uit de internationale studie tonen inderdaad aan dat tot nog toe vaak een beperkt aspect van het hele gamma aan leerdoelen door middel van gestandaardiseerde toetsen wordt geëvalueerd. Dit wordt in de literatuurstudie '*construct underrepresentation*' genoemd. Om dit tegen te gaan, moeten inspanningen worden geleverd om via het leer(winst)monitoringssysteem ook zogenaamde 'hogere orde-vaardigheden' te toetsen.

4.3 Link met de eindtermen en ontwikkelingsdoelen

In de literatuurstudie worden enkele auteurs vernoemd die ervoor pleiten om bij de ontwikkeling van een leer(winst)monitoringssysteem te vertrekken van het curriculum - in het geval van Vlaanderen: de eindtermen en ontwikkelingsdoelen van de verschillende leerdomeinen. Hierdoor krijgen leraren en scholen een zicht op de mate waarin leerlingen bereikt hebben wat het onderwijsaanbod heeft beoogd. In het internationale onderzoeksluik werd evenwel geconcludeerd dat ook niet-curriculum gebonden gestandaardiseerde toetsen in gebruik zijn en gretig worden afgenomen door de scholen. Niet-curriculumgebonden toetsen hebben het voordeel dat meer algemene vaardigheden van leerlingen (zoals in het geval van CEM bijvoorbeeld algemene redeneervaardigheid) of vaardigheden die niet aan één leerdomein toe te schrijven zijn (zoals de lees- en rekenvaardigheden, die in Noorwegen beschouwd worden als een opdracht voor alle leerdomeinen) in kaart kunnen worden gebracht. In de twee andere casestudies (Nederland en Australië) wordt echter sterk vastgehouden aan het curriculum om gestandaardiseerde toetsen op te stellen.

Voor het Vlaamse onderwijssysteem kunnen we voorzichtig concluderen dat een leer(winst)-monitoringssysteem gebaseerd op de eindtermen en ontwikkelingsdoelen (eventueel leerplannen) de voorkeur geniet boven de systeem dat niet van deze doelen vertrekt. In zes van de acht scenario's die werden opgebouwd ter voorbereiding van de derde ronde van de Delphi-studie baseert het leer(winst)monitoringssysteem zich op de eindtermen en ontwikkelingsdoelen. In slechts één enkel scenario werd hiervan afgeweken en werd een leerwinstmonitoring van enkele essentieel geachte vaardigheden (die eindterm-overschrijdend zijn) vooropgesteld (zie Deel 6, 2.2). Hoewel dit scenario een hoge waardering krijgt, wordt de niet-curriculumgebonden inhoud van de toetsen door meerdere respondenten expliciet ter discussie gesteld.

Wel wees de internationale studie op enkele complicaties die worden veroorzaakt door de toetsen in een leer(winst)monitoringssysteem te enten op de eindtermen en ontwikkelingsdoelen. Zo zijn deze gestandaardiseerde toetsen onderhevig aan voortdurende wijzigingen en bijstellingen (onder meer de inschaling van toetsitems) als gevolg van wijzigingen aan de eindtermen, de ontwikkelingsdoelen en de leerplandoelen. Daarnaast blijkt dat in de Vlaamse onderwijscontext aan twee voorwaarden niet wordt, terwijl deze voorwaarden een impact hebben op de wenselijkheid en haalbaarheid van gestandaardiseerde toetsen op basis van eindtermen en ontwikkelingsdoelen. Ten eerste is de beschikbaarheid van precieze eindtermen op alle subniveaus een voorwaarde - een voorwaarde die voor de verschillende onderliggende niveaus in het basisonderwijs niet is

voldaan (aangezien enkel eindtermen aan het eind van het basisonderwijs zijn geformuleerd), en evenmin voor meer specifieke leerdomeinen in technische en beroepsgerichte opleidingen in het secundair onderwijs. Daardoor moet de toetsontwikkelaar zich in de Vlaamse context noodgedwongen baseren op leerplandoelen. Dat zou tot gevolg hebben dat rekening moet gehouden worden met verschillen in leerplannen tussen scholen (voor Vlaanderen zijn dat voornamelijk verschillen tussen onderwijsnetten). Ook in de Delphi-studie worden de gevolgen van de verschillende leerplannen voor de toetsontwikkeling als een complicatie met veel gewicht naar voren geschoven. Ten tweede wordt gesteld in de conclusie van het internationale onderzoeksluik dat de haalbaarheid sterk wordt verhoogd bij een comprehensief onderwijssysteem. Deze voorwaarde is voldaan voor het Vlaamse basisonderwijs, maar niet voor het secundair onderwijs: de curricula voor de verschillende studierichtingen in het secundair onderwijs zijn immers sterk uiteenlopend waardoor toetsen niet breed inzetbaar zouden zijn.

4.4 Brede toetsing

In de eerste en tweede ronde van de Delphi-studie werd door verschillende respondenten uitdrukkelijk gewezen op het belang van 'brede toetsing', waarbij zowel cognitieve als niet-cognitieve leerprestaties en zowel de leerdomeinen rekenen en taal als andere minder evident te toetsen leerdomeinen aan bod zouden komen. Hierdoor wordt de impact van het onwenselijke neveneffect van *teaching to the test* geminimaliseerd. De respondenten vinden het daarom belangrijk om te streven naar een brede evaluatie van de verschillende leerdomeinen, gericht op het in kaart brengen van leerprestaties of leerwinst inzake onder meer (beroeps)competenties, attitudes en talenten van leerlingen. Zoals eerder gesteld worden de twee scenario's waarin een toetsenbank met een ruim aantal toetsen over uiteenlopende leerdomeinen worden vooropgesteld, beide gemiddeld positief beoordeeld door de respondenten in de derde ronde van de Delphi-studie. Ook in andere scenario's die een brede toetsing van leerdomeinen en van zowel cognitieve als niet-cognitieve aspecten voorstellen, werd dit sterk gewaardeerd door de verschillende respondenten.

Bovendien werd in de literatuurstudie beschreven dat de resultaten op kennistoetsen (of toetsen van relatief eenvoudige vaardigheden) binnen de meest getoetste leerdomeinen (rekenen en taal) niet zonder meer kunnen veralgemeend worden naar andere vormen van output ('hogere vaardigheden', attitudes, competenties) of naar andere leerdomeinen. In verschillende high-stakes contexten worden inderdaad zeer uiteenlopende leerdomeinen getoetst (voorbeelden uit het internationale onderzoeksluik zijn het GCSE-examen in Engeland en het HSC-examen in Australië), maar wordt de toetsing van niet-cognitieve aspecten toch vaak aan de scholen zelf overgelaten (zonder standaardisering).

De ambitie om brede toetsing te realiseren, is evenwel niet eenvoudig. Er is dan ook verder onderzoek nodig naar de ontwikkeling van betrouwbare en valide toetsing van niet-cognitieve aspecten. Bovendien is er een bijkomende complicatie wanneer men tevens de ambitie heeft om het monitoringssysteem te richten op het in kaart brengen van leerwinst van leerlingen. In sectie 2.1 werd geconcludeerd dat het voor de Vlaamse onderwijscontext meest wenselijk is om - indien wordt overgegaan tot ontwikkeling van een leer(winst)monitoringssysteem - leerwinst te conceptualiseren als in een *gain score* model. Hierbij worden verschillende items van verschillende toetsen ingeschaald op een gemeenschappelijke vaardigheidsschaal. Dit heeft evenwel als nadeel dat leerstof over verschillende leerjaren niet altijd makkelijk kan getoetst worden: de succesvolle

modellen in de internationale cases beperken zich, met enkele uitzonderingen, tot lees- en rekenvaardigheid. De leerstof binnen andere leerdomeinen lijkt in verschillende gevallen gericht op verschillende entiteiten van kennis die naast elkaar staan, eerder dan in een 'groeijende lijn van vaardigheid' (bijvoorbeeld in het leerdomein geschiedenis waar in het ene leerjaar de klassieke periode en het volgende leerjaar de middeleeuwen aan bod komen). Bovendien lijkt dit in Vlaanderen moeilijker realiseerbaar in het secundair onderwijs door de sterke stratificatie van leerlingen in studierichtingen. De voorbeelden van leerwinstmonitoring via een *gain score* model in het secundair onderwijs uit het internationale onderzoeksluik betreffen ofwel leerwinstmonitoringssystemen met een zeer comprehensieve onderwijsstructuur ofwel leerwinstmonitoringssystemen waarvan de gestandaardiseerde toetsen niet gelinkt zijn aan het curriculum.

Uit elk van de drie onderzoeksluiken blijkt de meerwaarde van adaptieve toetsen voor toetsen die gebruikt worden in een ontwikkelingsgericht perspectief groot te zijn. Bij een adaptieve toets wordt het niveau van de aangeboden toetsitems maximaal aangepast aan het kennis- of vaardigheidsniveau van elk van de individuele leerlingen. Op basis van een beoordeling van de correctheid van het antwoord op alle vorige items wordt daaropvolgend een item aangeboden dat meer op het niveau van de leerling gericht is (en dus moeilijker of makkelijker kan zijn dan de voorgaande items). Voordelen zijn dat de toetsen een grotere uitdaging bieden voor elke leerling en dat het leer(winst)monitoringssysteem meer verfijnde informatie kan opleveren over de leerprestatie of leerwinst van elke leerling. Een voorwaarde hiervoor is dat de toets digitaal wordt afgenomen, wat door het doorgaans ontwikkelingsgericht karakter van deze toetsen evenwel geen probleem kan zijn (bij sterk verantwoordingsgerichte toetsen bleken klassieke pen-en-papier toetsen meer aangewezen om de (perceptie van) gelijkheid tussen leerlingen niet te bedreigen).

In het internationale luik vonden we voorbeelden van zowel zuivere adaptieve toetsen (waarbij een leerling na elk item wordt doorgestuurd naar een moeilijker/makkelijker item) als van getrapt adaptieve toetsen (waarbij subsets van items worden gebruikt). Uit de literatuurstudie blijkt dat deze getrapt adaptieve toetsen makkelijker (economischer) zijn in de ontwikkeling, terwijl slechts in geringe mate informatie verloren gaat ten opzichte van volledig adaptieve toetsen over de precieze vaardigheid van de leerlingen. In het internationale onderzoeksluik worden beide systemen door gebruikers sterk gewaardeerd. In de derde ronde van de Delphi-studie werd in één scenario gebruik gemaakt van getrapt adaptieve toetsen, wat door de respondenten als één van de grootste sterktes van dit scenario werd beoordeeld.

5 ONDERSTEUNING

Zowel in de literatuurstudie als in elk van de cases van het internationale onderzoeksluik werd benadrukt dat de ontwikkeling van een leer(winst)monitoringssysteem op zich niet voldoende is. De vooropgestelde impact kan maar bereikt worden indien bij de gebruikers (leraren en scholen) aan bepaalde voorwaarden is voldaan. Het ruime aanbod aan Vlaamse en internationale onderzoeksliteratuur in het schooleffectiviteitsonderzoek bevestigt dat er grote verschillen zijn tussen scholen inzake datagebruik. Een belangrijke verklaring voor deze verschillen ligt in het beleidsvoerend vermogen en de datageletterdheid van scholen. Om deze voorwaarden in te vullen, bleek zelfs gerichte ondersteuning van scholen noodzakelijk in onderwijssystemen waar scholen

reeds een langere traditie hebben in het omgaan met outputgegevens inzake de leerprestaties van leerlingen.

"It's taking a long time. We've been doing this testing for more than twenty years. We've grown into this. And probably for five to ten years, people didn't understand what it was. And how to read it and interpret it... it took a long time!" (interview, respondent AU13)

Ook in de eerste twee ronden van de Delphi-studie wordt het beleidsvoerend vermogen van scholen vaak expliciet genoemd als antwoord op de vraag welke bijkomende voorwaarde moet worden ingevuld wanneer het Vlaamse onderwijssysteem inzet op de ontwikkeling van een leer(winst)monitoringssysteem. In de literatuur blijkt dat dragers als 'reflectief vermogen', 'innovatief vermogen' en 'onderlinge ondersteunende relaties tussen personeelsleden' essentieel zijn. In de internationale studie zagen we dat leer(winst)monitoringsystemen die effectief een impact hebben op het micro- en/of mesoniveau gepaard gaan met allerlei mechanismen om het beleidsvoerend vermogen van scholen te ondersteunen. Als een meer specifiek onderdeel van het beleidsvoerend vermogen van scholen wordt de aanwezigheid van een voldoende ontwikkelde 'datageletterdheid van scholen' in de literatuurstudie als een cruciale voorwaarde benoemd opdat scholen effectief en correct gebruik kunnen maken van de resultaten die worden opgeleverd door het leer(winst)-monitoringssysteem. De bezorgdheid dat de datageletterdheid in de meeste Vlaamse scholen momenteel niet sterk genoeg ontwikkeld is (of dat scholen over onvoldoende tijd en ruimte beschikken om hun vaardigheden inzake datageletterdheid optimaal in te zetten) wordt ook in de eerste en tweede ronde van de Delphi-studie genoemd. Bijgevolg werd in zeven van de acht scenario's die ter beoordeling werden voorgelegd in de derde ronde van de Delphi-studie in zekere mate (financiële) ruimte voorzien om scholen extra te ondersteunen bij de betekenisgeving aan de resultaten. Deze inspanning wordt door geen enkele respondent in vraag gesteld - integendeel, er is eerder een vraag om deze inspanning ruimer in te vullen.

Het belang van het beleidsvoerend vermogen van scholen voor de effectiviteit van het leer(winst)monitoringssysteem houdt uiteraard een risico in: zo kan er sprake zijn van een 'Matheus-effect' waarbij enkel de scholen die over een sterk beleidsvoerend vermogen beschikken, baat hebben bij de ontwikkeling van een leer(winst)monitoringssysteem om de eigen kwaliteit nog sterker te maken. Om ook scholen met een minder sterk ontwikkeld beleidsvoerend vermogen reeds van bij aanvang te ondersteunen, lijkt het aangewezen om een stap verder te gaan. Waar dit praktisch mogelijk is en haalbaar binnen de grenzen van de autonomie van scholen, wordt in het internationale onderzoeksluik alsook in de tweede ronde van de Delphi-studie aangeraden zeer concrete ondersteunende materialen te voorzien die de individuele leraar inspiratie kunnen geven bij bepaalde profielen van leerlingenprestaties of leerwinst, of die schoolteams kunnen ondersteunen bij bepaalde profielen van geaggregeerde leerlingenprestaties of toegevoegde waarde. Op die manier kunnen ook de scholen met een minder sterk ontwikkeld beleidsvoerend vermogen acties ondernemen die een antwoord kunnen bieden op de resultaten die worden opgeleverd door het leer(winst)monitoringssysteem.

6 DRAAGVLAK CREËREN

Ten slotte kan uit de literatuurstudie en uit het internationale onderzoeksluik worden geconcludeerd dat er inspanningen nodig zijn om bij de uiteenlopende stakeholders een draagvlak te creëren voor de ontwikkeling en het gebruik van een leer(winst)monitoringssysteem. Om dit draagvlak te creëren is het raadzaam rekening te houden met de bovenstaande conclusie dat Vlaanderen zulk systeem in een verantwoordingsgericht perspectief momenteel wenselijk noch haalbaar acht. De aanvaarding van de toetsen zelf hangt volgens de resultaten van de literatuurstudie onder meer af van de betrokkenheid van leraren in de ontwikkeling van de toetsen en van de gepercipieerde betrouwbaarheid en validiteit.

Uit het internationale onderzoeksluik kan worden afgeleid dat een draagvlak voor een leer(winst)monitoringssysteem kan groeien naarmate gebruikers inzien in welke mate de resultaten ervan kunnen bijdragen tot de onderwijspraktijk.

“I think there are problems if people have been forced to do the test against their will. Then you find difficulties. It’s much better to find people wanting to do it and coming to you, and start with a small group. And then the heads telling the other teachers: ‘Oh, this is useful’. Go for that, and it’s spreading organically” (interview, respondent EN3)

Hierdoor wordt het belang van een focus op micro- en mesoniveau opnieuw bevestigd. Verschillende respondenten in de Delphi-studie pleiten er dan ook voor om relatief kleinschalig te beginnen met een groep van scholen die hier aan willen deelnemen, zodat de meerwaarde kan worden aangetoond en de impact kan worden geëvalueerd. Op basis hiervan kan een draagvlak worden gecreëerd. Ten slotte leren we ook uit het internationale onderzoeksluik dat het wellicht onmogelijk is om alle critici ervan te overtuigen dat de positieve effecten van een leer(winst)monitoringssysteem opwegen tegen de mogelijke onwenselijke neveneffecten.

DEEL 9 BELEIDSAANBEVELINGEN

Onderwijsonderzoek kan niet in de plaats van het politieke beleid treden. Het beleid neemt beslissingen en treft maatregelen. Daarbij laat het zich inspireren en leiden door een geheel van kennis, waarden en opvattingen inzake onderwijs, individu en samenleving. Wat beleidsondersteunend onderzoek doet is op een wetenschappelijk verantwoorde wijze gegevens verzamelen, ordenen en in samenhang presenteren. In deze studie betreft het gegevens die beleidsmakers kunnen inbrengen in hun beslissing of, en zo ja volgens welke modaliteiten, in het Vlaamse onderwijs een (verder) uitgebouwd systeem van leer(winst)monitoring opzetten wenselijk en haalbaar is.

Vertrekkende vanuit bovenstaande positionering willen we desalniettemin de kans nemen om ook als onderzoeksteam standpunt in te nemen. Dat standpunt hebben we vertaald in principes die in het uitdenken van een beleid inzake leer(winst)monitoring van wezenlijk belang zijn. In wat volgt beschrijven we vanuit het voorliggende onderzoek de uitgangspunten op basis waarvan leer(winst)monitoring(systemen) ons inziens van betekenis kan zijn in het Vlaamse onderwijs. Het realiseren van deze uitgangspunten is wat wij aanbevelen. Daarom hebben we ze als imperatief geformuleerd.

Gegeven de aard van de onderzoeksopdracht richten we ons in onze aanbevelingen in eerste instantie tot beleidsmakers in het Vlaamse onderwijs. Dat neemt niet weg dat wij schoolleiders en leerkrachten als de eerste verantwoordelijken beschouwen voor het behouden en verbeteren van onderwijskwaliteit. Voorbij gaan aan het schoolniveau zou er toe leiden dat we aanbevelingen in een vacuüm zouden presenteren. Hoewel we ons richten tot beleidsmakers zal de link met de school- en klaspraktijk dan ook een rode draad doorheen de aanbevelingen blijken te zijn.

De principes zijn niet zonder meer nieuw, en bouwen verder op empirische evidentie met betrekking tot leer(winst)monitoring(systemen) in verschillende onderwijscontexten, deze studie inclusief. Daarnaast spelen de aanbevelingen in op de inschatting van de wenselijkheid en haalbaarheid van mogelijke scenario's voor leer(winst)monitoring(systemen) voor Vlaanderen door een breed veld van stakeholders (cf. de Delphi-studie). Doelbewust richten we ons in de aanbevelingen nog op de eerder grote vraagstukken, wetende dat verdere uitwerking nodig is, dat meer verfijnde lessen uit deze studie mogelijk (en beschikbaar) zijn en dat antwoorden op de grote vraagstukken nieuwe vragen zullen oproepen. Het zijn echter precies de fundamentele aspiraties en mogelijkheden die in het beleidsdebat eerst dienen uitgeklaard te worden.

AANBEVELING 1: KIJK EN BESLIS VOORBIJ DE KARIKATUUR

Het voorliggende onderzoek heeft aangetoond dat de problematiek van leer(winst)monitoring niet eenduidig is. In de bestudeerde landen is het politieke en academische debat nog steeds actueel, en laten bestaande systemen nog steeds disfuncties zien. Het gevaar is dan ook groot dat het debat in Vlaanderen verengd zou blijven tot een discussie over de wellicht meest extreme vorm van

externe publicatie, namelijk in de vorm van rankings zoals die o.a. in de Britse context gangbaar is. Tegenstanders van vormen van centrale toetsing in Vlaanderen hebben het waardeoordeel over leer(winst)monitoringsystemen te vaak verengd tot zulke onwenselijke uitwassen in andere onderwijssystemen. Op die manier hebben ze een toon gezet die een serieus debat over voor- en nadelen bemoeilijkt heeft. Vanuit wetenschappelijk oogpunt pleiten wij voor het duidelijk kaderen van de problematiek in een bredere onderwijscontext en voor een analyse zonder a priori's. Een eerste evidente aanbeveling aan het Vlaamse onderwijsveld is dan ook goed kennis te nemen van de inhoud van dit onderzoeksrapport, om van daaruit tot geïnformeerde oordeelsvorming te komen. Dat vergt een bereidheid om taboedoorbrekend bestaande denkbeelden en overtuigingen te expliciteren, te confronteren met bestaande evidentie, in vraag te stellen en zo nodig te herformuleren. De variatie in verschijningsvormen en scenario's inzake leer(winst)monitoring maakt duidelijk dat het debat veel rijker en genuanceerder is dan het tot dusver in Vlaanderen gevoerd werd. Kortom: als het Vlaamse onderwijssysteem niet oplet, dan loopt ze het gevaar het goede en bruikbare te verwerpen omdat het niet onderscheiden wordt van het misleidende en onwenselijke.

Vlaanderen is vooralsnog terughoudend geweest in het uitwerken van systemen voor leer(winst)monitoring en van het openbaar maken van outputinformatie over scholen. Wij hanteren graag de invalshoek dat deze omzichtige houding Vlaanderen (internationaal) in een unieke context gebracht heeft. Het onderwijssysteem heeft de kans - en zonder meer ook de plicht - om te onderzoeken hoe de voordelen van een leer(winst)monitoringsysteem kunnen benut worden zonder de nadelen ervan te hoeven ondergaan. Vlaanderen dient de ambitie te hebben om de (zogenaamde) voorsprong van andere onderwijssystemen om te buigen tot een situatie waarin 'de wet van de remmende voorsprong' van toepassing is. Vanuit de in deze studie gedocumenteerde systemen, hun randvoorwaarden en (neven)effecten kan een systeem uitgedacht worden dat vanuit een Vlaams perspectief efficiënt en effectief is. Door middel van een sterk bewustzijn van de gedocumenteerde valkuilen kan een leer(winst)monitoringsysteem gecreëerd worden waarvan de meerwaarde en performantie algemeen erkend wordt. Dat vergt bereidheid om incrementeel te werken, bereidheid tot investering en de noodzaak tot het creëren van een juridisch kader dat mogelijke nadelen (met name inzake openbaarheid) onder controle houdt.

AANBEVELING 2: BOUW OP VANUIT DE BEOOGDE DOELSTELLINGEN

Eén van de conclusies van deze studie is dat de impact van een leer(winst)monitoringssysteem het grootst is wanneer aan de verschillende onderdelen een éénduidig doel wordt gekoppeld. Het expliciteren van de rationale en assumpties achter mogelijke systemen moet dan ook het vertrekpunt zijn van het beleidsdebat. Eerst dient op een transparante wijze geduid te worden wat men wil bereiken om pas vervolgens de vraag te stellen hoe dat kan bereikt worden. We toonden aan dat een breed pallet aan doelstellingen ten gronde kan liggen aan leer(winst)monitoringsystemen. Het spreekt voor zich dat de uitrol anders is wanneer men bijvoorbeeld gelijkwaardigheid van diploma's overheen scholen wil garanderen dan wanneer men over concrete onderwijs- en leerprocessen in de klas wil informeren. We toonden aan dat zowel verantwoordingsgerichte als ontwikkelingsgerichte systemen bepaalde positieve impulsen kunnen geven aan onderwijskwaliteit. Over de mogelijkheden om beide doelstellingen te combineren in één systeem zijn buitenlandse lessen echter eerder pessimistisch. We willen aanbevelen om met verschillende partners in kaart te brengen over welke doelstellingen overeenstemming kan bereikt worden. Vervolgens dient te

worden nagegaan of bestaande systemen deze doelen realiseren en wat in termen van leer(winst)monitoring verder dient uitgebouwd te worden. Wij zien in het Vlaamse onderwijs te vaak beargumenteerd worden dat bestaande systemen volstaan zonder grondige analyse van de doelstellingen die deze systemen al dan niet realiseren. Zo is het bijvoorbeeld de vraag in welke mate koepelgebonden initiatieven scholen en leraren momenteel in staat stellen om de leerwinst van hun leerlingen in kaart te brengen en over jaren heen op te volgen. Het Vlaamse onderwijssysteem dient zich ervan bewust te zijn dat het behoud van de status quo ook nadelen heeft. In dat geval kiest men er immers voor om waardevolle doelstellingen niet te realiseren.

De vraag naar de wenselijke doelstellingen van en leer(winst)monitoringsystemen is geenszins een geïsoleerd vraagstuk. De vraag is gekoppeld aan andere beleidsdossiers die op elk op hun eigen manier deze discussie kunnen beïnvloeden, maar ook op hun beurt onvermijdelijk beïnvloed worden door de beslissingen die inzake de implementatie van een leer(winst)monitoringssysteem zullen worden genomen. Daarbij denken we bijvoorbeeld aan het debat rond eindtermen, implementatie M-decreet, loopbaandebat, en het debat over de professionalisering van schoolbesturen.

AANBEVELING 3: STEL ONTWIKKELING CENTRAAL, NIET EEN ENGE FOCUS OP VERANTWOORDING

De onderzoeksresultaten leren dat zowel schoolontwikkeling als verantwoording een uitgangspunt kunnen zijn voor leer(winst)monitoring. Onze aanbeveling is evenwel om in het uitwerken van systemen zonder meer de nadruk te leggen op leerprocessen en de ontwikkeling die daarmee gepaard gaat. De focus op individueel (leerlingen), professioneel (leraren) en organisationeel (scholen) leren is van cruciaal belang. Leer(winst)monitoring met een te sterke focus op controle kan er toe leiden dat het leerpotentieel op diverse niveaus niet ten volle benut wordt. Over het algemeen blijkt bij diverse stakeholders in de Vlaamse onderwijscontext een positieve houding ten aanzien van ontwikkelingsgerichte leer(winst)monitoring. Het is dan ook aangewezen deze positieve houding verder te koesteren en te bewaken dat schoolleiders en leerkrachten autonoom gemotiveerd blijven om informatie te gebruiken in het kader van hun beleidsvoering en onderwijspraktijk.

We willen benadrukken dat gestandaardiseerd toetsen en een zogenaamde hoge inzet (high-stakes) niet hoeven samen te gaan. Een directe eenzijdige koppeling van de uitkomsten van gestandaardiseerde toetsen aan beslissingen over voortgang/zittenblijven van leerlingen, toegang tot vervolgstudies, salarissen of bonussen voor personeelsleden maar ook beloningen of (publieke) sancties voor scholen dient vermeden te worden. Ons advies is dus om leer(winst)informatie vooral formatief in plaats van (enkel) summatief te gebruiken. In de leergerichte context die we beogen, wordt informatie in eerste instantie verzameld om het leren van de leerling, leraar en de school te informeren in een proces van (school)ontwikkeling en niet (enkel) om tot een (eind)oordeel te komen. Dat betekent ook dat in het ter verantwoording roepen van scholen en leraren ons inziens best gefocust wordt op de wijze waarop gebruik van leer(winst)monitoring hen tot (school)ontwikkeling brengt, eerder dan verantwoording te vragen voor de (output)resultaten die in het kader van het monitoringsysteem in kaart worden gebracht.

Wanneer scholen aan informatie uit gestandaardiseerde toetsen op leerlingniveau een rol willen toekennen in attestering en advisering, dan kan dit ons inziens enkel in een context waarin tevens

de inbreng van de professionele leraar en van de delibererende klassenraad gegarandeerd blijft. Dat neemt niet weg dat we scholen willen stimuleren om - complementair aan de bestaande aanpak - van gestandaardiseerde schooloverstijgende toetsen gebruik te maken als referentiekader om aansluiting te (blijven) vinden bij de van leerlingen verwachte minimumprestaties.

AANBEVELING 4: BOUW (VERDER) AAN EEN INFORMATIERIJKE OMGEVING VOOR SCHOLEN (MET INBEGRIIP VAN LEERWINSTINFORMATIE)

De vergelijking met cases in de landenstudie leert dat er in het Vlaamse onderwijs - ondanks recente initiatieven zoals 'Dataloop' - relatief weinig informatie beschikbaar is. Dat geldt bij uitstek voor informatie over leerwinst. Op basis van onze bevraging van Vlaamse stakeholders concludeerden we dat men rekent op een leer(winst)monitoringssysteem om op microniveau bijkomende betrouwbare en valide informatie te genereren waardoor leraren in staat worden gesteld om tot gelijkgerichte inschattingen van leerlingprestaties te komen, om het onderwijsaanbod maximaal af te stemmen op de noden van de leerlingen maar ook om leerlingen (en hun ouders) zelf beter inzicht te geven in hun eigen voortgang. Op mesoniveau is de verwachting dat het leer(winst)monitoringssysteem een bijdrage levert aan een informatierijke omgeving voor de school door het systeem van interne kwaliteitszorg van betrouwbare en valide informatie over de leerprestaties of de leerwinst van leerlingen te voorzien. Om dit doel op mesoniveau te verwezenlijken, dient de informatie over de school in perspectief te worden geplaatst met outputinformatie van scholen met een vergelijkbare input en context. We stellen een eerder grote consensus vast rond deze idee dat ijkpunten voor informatie meer betekenis geven aan informatie, en rond de vaststelling dat vergelijkende informatie momenteel (te) schaars is. Door alle stemmen aan het woord te brengen, toont deze studie aan dat de gedragenheid voor deze ideeën in het Vlaamse onderwijsveld globaal genomen aanzienlijk groter is dan beleidsdebatten nogal eens laten uitschijnen.

Indien we scholen willen stimuleren tot informatiegebruik, dan strekt het tot aanbeveling om voor scholen een meer informatierijke omgeving uit te bouwen. Het is zaak om op verschillende niveaus te investeren in relevante, valide en betrouwbare, en gebruiksvriendelijke leer(winst)informatie. De ontwikkeling, verspreiding en het faciliteren van gebruik van gestandaardiseerd toetsmateriaal zou met andere woorden een collectief beleidsstreven moeten zijn.

AANBEVELING 5: VERPLICHT HET PRINCIPE VAN LEER(WINST)MONITORING, NIET DE KEUZE VOOR EEN SYSTEEM

Met het 'Decreet betreffende de kwaliteit van onderwijs' werd benadrukt dat instellingen zelf de eerste verantwoordelijken zijn voor de eigen kwaliteit en daartoe een zelfstandig beleid moeten voeren. Het decreet verplicht scholen tot interne kwaliteitszorg maar blijft terughoudend in de invulling ervan. Het kwaliteitsdecreet beperkt zich tot de verplichting voor elke instelling om op systematische wijze haar eigen kwaliteit te onderzoeken en te bewaken en legt geen richtlijnen op over de organisatie ervan. Het luidt dat de instelling zelf kiest op welke wijze ze dit doet. Wij adviseren de overheid om een gelijkaardige terughoudendheid aan de dag te leggen in het uitvaardigen van inhoudelijke en organisatorische voorschriften inzake leer(winst)monitoring. Dat

betekent dat scholen decretaal zonder meer de opdracht kunnen krijgen om leerprestaties en leerwinst van leerlingen in kaart te brengen, te onderzoeken en te bewaken, maar dat deze verplichting zich niet vertaalt in het opleggen van het gebruik van deze of gene systeem. Ondertussen beschouwen wij het als de opdracht van de overheid om in overleg met de onderwijsverstrekkers het instrumentarium ter beschikking te stellen dat scholen hierbij kunnen aanwenden en om vervolgens toezicht te houden op de mate waarin scholen hun decretale verwachting vorm geven.

AANBEVELING 6: VOORZIE IN EEN BREED AANBOD DAT RECHTDOET AAN DE VEELZIJDIGHEID VAN HET LEREN VAN LEERLINGEN EN VAN DE ONDERWIJSOPDRACHT VAN SCHOLEN

Uit de verschillende onderdelen van deze studie blijkt het belang van 'brede toetsing', waarbij zowel cognitieve als niet-cognitieve leerprestaties en zowel de leerdomeinen rekenen en taal als andere minder evident te toetsen leerdomeinen aan bod komen. Verder leest men in deze studie een pleidooi voor aandacht voor bijvoorbeeld 'leren leren' en sociale vaardigheden maar ook voor de uitbouw aan socio-emotionele vragenlijsten (welbevinden). Gestandaardiseerde toetsen hoeven niet beperkt te zijn (c.q. blijven) tot strikt cognitieve prestaties en hebben best ook betrekking op vaardigheden en attitudes. Dit vergt echter substantiële investeringen, omwille van de noodzaak aan verder onderzoek naar mogelijkheden om deze vaardigheden en attitudes op betrouwbare en valide manier in kaart te brengen, omwille van de kost ten behoeve van het ontwikkelen van vaardigheidstoetsen, alsook omdat de afname van zulke toetsen arbeidsintensiever is. Scenario's die een ruim geheel van toetsen over uiteenlopende leerdomeinen vooropstellen, worden inzake wenselijkheid positief beoordeeld. Een volwaardige inschatting van bereikte leerresultaten van leerlingen en van de gerealiseerde kwaliteit van scholen maakt ook ons inziens dat een beperking tot cognitieve leerprestaties een onwenselijke verenging van het streefbeeld zou zijn. Bovendien is een breed aanbod dé manier om tegemoet te komen aan één van de meest gehoorde kritieken op bestaande systemen, met name de verenging van het onderwezen curriculum. Wij adviseren om deze brede benadering van bij de start centraal te stellen en zijn ondertussen beducht voor een beleidsscenario waarbij in de drang om snel tot een opstap tot leer(winst)monitoring te komen, toch gewerkt zal worden met een beperkte set van cognitieve indicatoren.

AANBEVELING 7: ZET WAAR MOGELIJK IN OP LEERWINST

In elke van deze aanbevelingen spreken we van een leer(winst)monitoringssysteem, om aan te duiden dat het zowel kan gaan om het documenteren van leerprestaties als van leerwinst. De vele gedocumenteerde voordelen inzake leerwinst, onder meer de verwachting dat hierdoor een veel grotere impact kan bereikt worden op het niveau van (de begeleiding van) het individuele leerproces van elke leerling, leiden ons evenwel tot de aanbeveling om, waar mogelijk, in te zetten op het in kaart brengen van de leerwinst van leerlingen. Uit de respons van de Vlaamse stakeholders blijkt dat het draagvlak van een monitoringssysteem door leerwinstinformatie sterk wordt vergroot. Het scenario waarin het meest expliciet wordt ingezet op de productie van leerwinstinformatie in een ontwikkelingsgericht perspectief werd door de Vlaamse respondenten sterk gewaardeerd. Een beperking van dit scenario was evenwel dat het niet strookt met de voorgaande aanbeveling, met name dat het aanbod geen brede toetsing inhield. Ervaringen uit de literatuur en het buitenland,

maar ook van eerdere pogingen in de Vlaamse onderwijscontext, leren evenwel dat genereren van leerwinstinformatie in uiteenlopende leerdomeinen erg moeilijk kan zijn door het ontbreken van een graduele groeilijn in de kennis en vaardigheden van leerlingen in verschillende domeinen. Ten behoeve van de integratie van aanbeveling 6 en 7 dient daarom sterk geïnvesteerd te worden in debat omtrent welke vormen van leerwinstinformatie bij welke leerdomeinen beschikbaar zouden moeten zijn en in onderzoek en ontwikkeling om dit vervolgens mogelijk te maken.

AANBEVELING 8: VOORZIE IN EEN PALLET AAN INITIATIEVEN VOOR DESKUNDIGHEIDSONTWIKKELING

Het voorliggende onderzoek maakt duidelijk dat het noodzakelijk is te werken aan het menselijk en sociaal kapitaal van scholen met betrekking tot het gebruik van leer(winst)monitoringsystemen. Het internationale onderzoeksluik toont aan dat elders te vaak voorbij gegaan werd aan de nochtans onontbeerlijke expertise-ontwikkeling en expertisedeling bij gebruikers. Hoewel scholen hier zelf een wezenlijke rol in te vervullen hebben, willen we de overheid een aantal aanbevelingen meegeven waarbij we de beoogde professionalisering in een ruimer kader plaatsen. Het gebruik van leer(winst)monitoringsystemen past immers in een ruimer verhaal van gekende thema's als kwaliteitszorg, informatiegebruik en datageletterdheid, maar ook in een verhaal van ethisch datagebruik door alle betrokkenen binnen (en buiten) het onderwijsveld. Deskundigheidsontwikkeling wordt in eerste instantie best geïntegreerd in initiatieven die zich op deze thema's richten. Toch zijn ook meer gefocuste strategieën behulpzaam. Zo zou de overheid kunnen overwegen om (tijdelijk) middelen te reserveren in het werkpakket van scholen, specifiek voor interne kwaliteitszorg en gericht op gebruik van leer(winst)monitoringsystemen. Laat deze middelen dan na verloop van tijd deel uitmaken van de reguliere financiering. Verder lijkt ook het sterker integreren van informatiegebruik en werken met (resultaten van) gestandaardiseerde toetsen in het curriculum van opleidingen van (toekomstige) schoolleiders en leerkrachten verstandig. Prille ervaringen inzake schoolfeedbackgebruik toonden in Vlaanderen reeds de noodzaak en potentie aan om tezamen met een aanbod van gestandaardiseerde toetsen te voorzien in professionaliseringsactiviteiten voor verschillende doelgroepen. Hetzelfde geldt voor het aan scholen en schoolbesturen ter beschikking stellen van experts, waarop zij specifiek voor ondersteuning met betrekking tot gebruik van het leer(winst)monitoringsysteem een beroep kunnen doen. Dat kan bijvoorbeeld in de schoot van de pedagogische begeleidingsdiensten gebeuren. Met het oog op de transfer van inzichten met betrekking tot gebruik van leer(winst)monitoringsystemen is het ook aangewezen om de uitbouw van voor scholen interne en externe thematische netwerken rond leer(winst)monitoring te ondersteunen.

AANBEVELING 9: WERK MET PILOOTPROJECTEN EN MONITOR DE WERKING VAN HET SYSTEEM

Als de overheid overweegt om een leer(winst)monitoringsysteem op te zetten verdient het aanbeveling eerst een degelijk pilootproject uit te voeren. Het opzetten van een leer(winst)monitoringsysteem is een complexe onderneming die aan verscheidene facetten van de organisatie van scholen en leren raakt en waarbij zeer diverse stakeholders betrokken zijn. Kenmerken van het functioneren van het systeem en (beleefde) (neven)effecten laten zich niet

(eenduidig) voorspellen. Bestaande (buitenlandse) systemen zijn onmogelijk zonder meer te transfereren naar een Vlaamse context. Wij adviseren de overheid daarom om van het opzetten van een leer(winst)monitoringsysteem een groeitraject te maken. Op die manier kan gezocht worden naar garanties om het systeem (in groeiende mate) aan de verwachte kenmerken en randvoorwaarden te laten voldoen. In deze groeifase kan vervolgens ook ingezet worden op informerende en draagvlakcreërende communicatie. Een incrementele aanpak is ons inziens aangewezen. Het gebruik van een leer(winst)monitoringsysteem helpt namelijk zichzelf 'verkopen' indien het aanbod en de omkadering van bij het moment van globale uitrol kwaliteitsvol ingevuld wordt. Helaas kan het gebruik zichzelf bij wijze van spreken ook bemoeilijken tot onmogelijk maken indien dit niet het geval is. De regel 'baat het niet dan schaadt het niet' gaat niet op. Daarvoor is het prille draagvlak dat we in deze studie blootlegden te kwetsbaar en kostbaar.

Van bij de start dient verder een strategie voor monitoring naar de werking van het systeem ingebouwd te worden. Vanuit een overkoepelend kwaliteitszorgidee adviseren wij om vertrekkende vanuit de beoogde doelstellingen een set van prestatie- en perceptie-indicatoren te ontwikkelen en op te volgen. Op die manier kan het verbeterbare geïdentificeerd worden, kunnen mogelijke neveneffecten vroegtijdig getraceerd worden en kan verdere ontwikkeling van het leer(winst)monitoringsysteem geïnformeerd worden.

LITERATUURLIJST

- 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>
- 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
- 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
- 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). Toekomst 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.
- 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, 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.
- 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.
- 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.
- 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.
- 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*.
- 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
- 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? Discrepancies 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.
- 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.
- 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
- 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.
- 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>
- Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 385-443.
- Goldstein, H., & Leckie, G. (2008). School league tables: what can they really tell us? *Significance*, 5(2), 67-69.
- 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.
- 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 de 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.
- 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. (2003). "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).
- 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.
- 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.
- 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. (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
- 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 ijkijgstest. 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>
- Parkes, J., & Stevens, J. J. (2003). Legal issues in school accountability systems. *Applied Measurement in Education*, 16, 141-158.
- 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. (2006). Panoptic performativity and school inspection regimes: disciplinary mechanisms and life under special measures. *Journal of Education Policy*, 21, 147-161.
- 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).
- 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.
- 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). *Sammons Key Characteristics of Effective Schools 1995*.
- 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. (2012). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, 24(1), 1-38.
- 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
- 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.
- 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: inconcistencies 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, 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., Sway Hoover, 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.
- 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.
- 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.
- 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.
- 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 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.
- 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 amoricum 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.
- 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).
- Wiliam, 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.

- 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.