(OR 20130127/1046)

Startdossier toegevoegde waarde

Dr. M. Ehren
m.ehren@ioe.ac.uk
m.c.m.ehren@utwente.nl

Juni 2013

Inhoudsopgave

1. Inleiding

Dit startdossier 'Toegevoegde waarde' biedt een overzicht van de wetenschappelijke, empirische stand van zaken van het onderzoek naar toegevoegde waarde in het primair en voortgezet onderwijs. De aandacht gaat uit naar de literatuur (ondersteund met internationale voorbeelden van 'good practice') over het gebruik van informatie over de toegevoegde waarde van scholen voor schoolverbetering (de feedbackfunctie), voor het beoordelen van, en toezicht houden op scholen (de accountabilityfunctie), en voor het informeren van stakeholders van de school (de informatiefunctie). Het startdossier geeft inzicht in de stand van zaken van het onderzoek op dit terrein en laat zien waar de gaten in kennis zitten; de resultaten bieden echter geen uitputtende review en bespreking van de literatuur, maar leveren de bouwstenen voor een advies van de Onderwijsraad over "toegevoegde waarde van scholen".

2. Definiëren leerwinst en toegevoegde waarde

Leerwinst en toegevoegde waarde zijn twee begrippen die veelal in elkaars verlengde worden gebruikt. Ze hebben echter betrekking op twee verschillende prestatiematen waarvoor verschillende statistische modellen worden gehanteerd. In de kamerbrief over de pilot leerwinst (25 november 2011, p.2/3) wordt leerwinst en toegevoegde waarde als volgt omschreven:

> Leerwinst: "de toename van vaardigheden of kennis van individuele leerlingen of groepen van leerlingen, gedurende (een bepaald deel van) de leerweg. De leerwinst wordt altijd bepaald door twee meetmomenten. Het verschil tussen de metingen maakt de ontwikkeling van de leerling of de groep zichtbaar. Het aantal meetmomenten is variabel, bijvoorbeeld door in het basisonderwijs de leerwinst elke twee jaar te meten, of door middel van een begin- en een eindtoets."
>
> "De toegevoegde waarde geeft aan welke bijdrage de school levert aan de ontwikkeling (of leerwinst) van alle leerlingen. Daarbij wordt nagegaan of de onderwijsopbrengst boven of onder het niveau ligt dat op basis van het beginniveau verwacht mocht worden. Als een gemiddeld resultaat van de leerlingen bijvoorbeeld hoger is dan verwacht, kun je spreken van een vorm van toegevoegde waarde. Een volgende stap is het vaststellen in hoeverre deze toegevoegde waarde is toe te schrijven aan de kwaliteit en inspanningen van de school en/of de leraar. Dit gebeurt door de toetsscore in meer of mindere mate te corrigeren voor andere factoren die kunnen bijdragen aan de leerwinst, zoals het opleidingsniveau van de ouders, de taal die thuis wordt gesproken, of de sociaalemotionele ontwikkeling van de leerling. Het is niet mogelijk om voor alle externe invloeden te corrigeren. De correcties die wel worden doorgevoerd, zijn gebaseerd op gemiddelden. De toegevoegde waarde is daarom een benadering van de werkelijke invloed die scholen hebben op de prestaties van hun leerlingen."

De leerwinst van groepen leerlingen kan dus worden gebruikt om de toegevoegde waarde van een school te meten wanneer de resultaten worden geaggregeerd naar schoolniveau en worden gecorrigeerd voor kenmerken van leerlingen waar de school geen invloed op heeft maar die de leerprestaties van leerlingen wel beïnvloeden.

2.1 Modellen voor toegevoegde waarde

In de literatuur worden verschillende modellen voor het meten van toegevoegde waarde uitgewerkt. (zie Hamilton en Koretz, in: Hamilton e.a., 2002; OECD, 2008; Onderwijsraad, 2003; Eecke, 2004; Bosker e.a., 2006). Hieronder volgt een indeling in modellen zoals beschreven door Hamilton en Koretz (in: Hamilton e.a., 2002).

Norm-referenced: resultaten van leerlingen in een school worden afgezet tegen die van een normpopulatie (vergelijkbare groep scholen) aan de hand van één van de volgende vier modellen:

- Contextualised attainment model/normgerelateerde standaard: gemiddeld eindresultaat van leerlingen op een school, gecorrigeerd voor leerlingkenmerken/gemiddelde niveau van vergelijkbare scholen minus een halve standaarddeviatie
- Percentile rank: percentage van een referentiegroep leerlingen die lagere scores behaalden dan de betreffende groep leerlingen in de school
- Standaard score: prestaties van een (groep) leerling(en) ten opzichte van een gemiddelde (groep) leerling(en) (uitgedrukt in z-scores)
- Grade equivalent: prestaties van een (groep) leerling(en) in vergelijking met de verwachte prestaties van de betreffende leergroep/klas met dezelfde leeftijd

Criterion/standards-referenced: prestaties van leerlingen in vergelijking met vaststaande standaarden (bijvoorbeeld referentieniveaus), *geen* vergelijking met prestaties van andere leerlingen

In deze modellen kunnen leerprestaties voor verschillende achtergrondkenmerken van leerlingen worden gecorrigeerd (bijvoorbeeld individuele kenmerken van leerlingen zoals SES, opleidingsniveau ouders, etnische achtergrond, type zorgleerling). Daarnaast moet in de berekening van zowel leerwinst als toegevoegde waarde ook rekening worden gehouden met onderbrekingen in de schoolloopbaan van leerlingen door verhuizing, verwijzing naar speciaal onderwijs etc.

In de pilot leerwinst PO worden op dit moment de volgende modellen uitgewerkt (zie Janssens, presentatie expertmeeting 17 mei 2013): zomervakantie-model, groeitempo-model, relatieve leerwinst-model, groepsoverzicht leerwinst, schooloverzicht leerwinst, vaardigheidsverschil-model, vaardigheidsgroei-model. Zie bijlage 1 voor een samenvatting.

Naast bovenstaande 'norm en criterion-referenced' modellen voor het berekenen van toegevoegde waarde onderscheiden Hamilton en Koretz (in Hamilton e.a., 2002) ook twee modellen voor het stellen van targets aan prestaties van scholen:
- Statusmodel: prestaties op 1 tijdpunt afzetten tegen een standaard (bijvoorbeeld minimum prestatie, gemiddeld prestatieniveau van vergelijkbare scholen of historisch gemiddelde)
- Groeimodel: prestaties vergelijken met prestaties in het verleden, met behulp van
  - Cross-sectionele methode: vergelijking van prestaties groep 8 leerlingen dit jaar met prestaties groep 8 leerlingen vorig jaar
  - Quasi-longitudinale methode: vergelijken van prestaties groep 8 leerlingen dit jaar met prestaties groep 7 leerlingen vorig jaar
  - Longitudinale methode: prestaties van individuele leerlingen worden vergeleken met hun eigen eerdere prestaties
  - Relatieve vooruitgang (werkelijke vooruitgang wordt afgewogen tov een voorspelde vooruitgang.

Dit kort overzicht laat zien dat ' de toegevoegde waarde' van een school geen eenduidig begrip is, maar op veel verschillende manieren kan worden berekend, waarbij de uitkomst van verschillende modellen bovendien verschillende waarden oplevert.

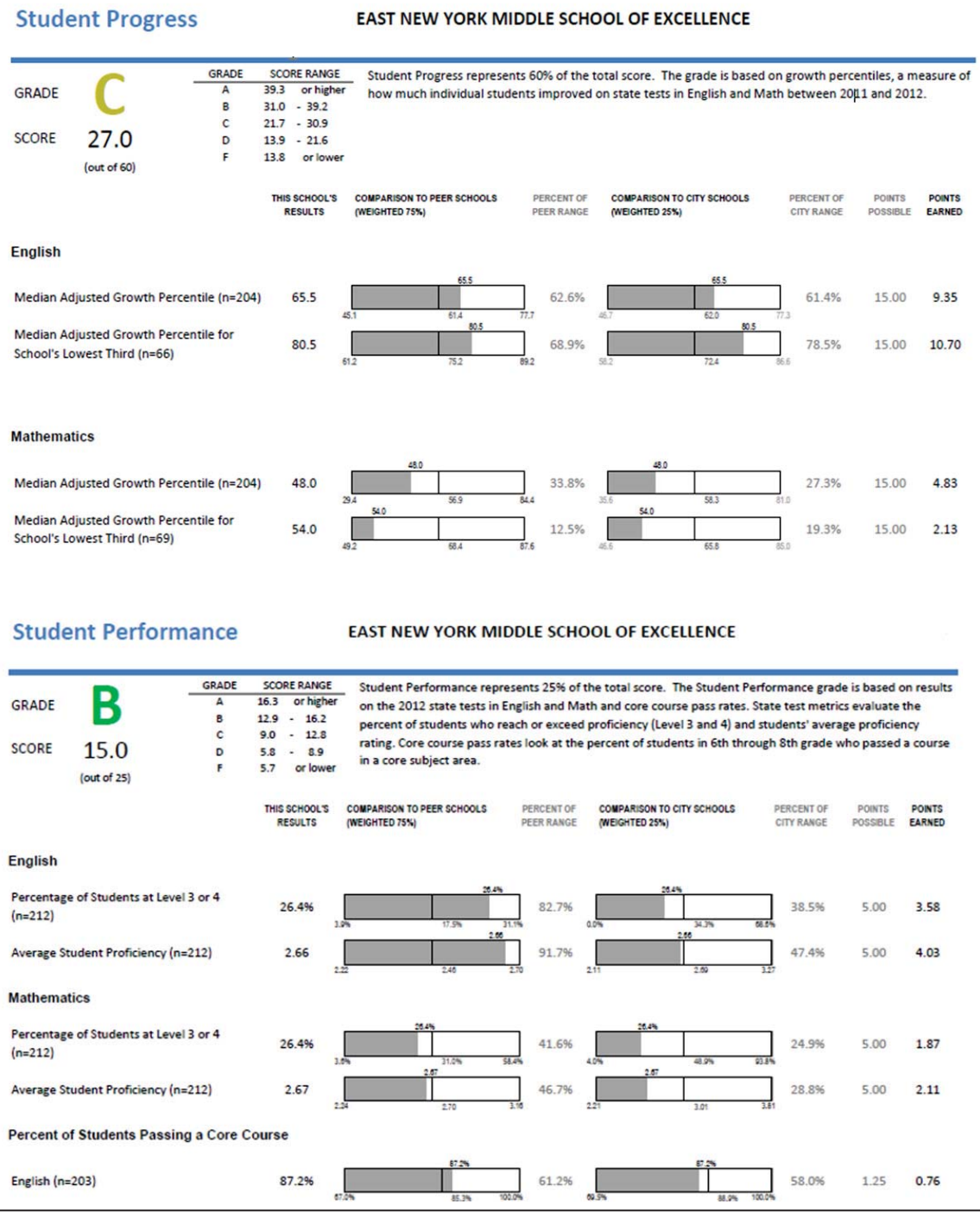
2.2 Rapporteren van toegevoegde waarde

De meeste modellen zijn gebaseerd op het meten van toegevoegde waarde van een school. Toegevoegde waarde kan echter ook worden gerapporteerd voor:
- Subgroep van leerlingen (regulier/hoogbegaafd; naar SES/etniciteit, geslacht)
- Vakken (taal/rekenen)
- Jaargroep/klas

Deze gedifferentieerde modellen geven meer gedetailleerde informatie over het functioneren van verschillende onderdelen van een school. Box 1 presenteert een voorbeeld van de rapportage van toegevoegde waarde-informatie over verschillende vakken en verschillende prestatieniveaus van leerlingen van een school in New York.

Vrijwel alle voorbeelden van rapportages van toegevoegde waarde hebben betrekking op taal en rekenen. De reden hiervoor is het ontbreken van robuuste methoden voor het meten van toegevoegde waarde op andere domeinen zoals sociale of maatschappelijke opbrengsten (persoonlijke communicatie Dijkstra, Dijkstra, 2012). Er is geen eenduidige beschrijving beschikbaar van wat er bijvoorbeeld onder sociale of maatschappelijke opbrengsten wordt verstaan en ook het toetsen van sociale opbrengsten levert allerlei problemen op; het is bijvoorbeeld niet bekend welke bijdrage de school aan sociale opbrengsten levert en er is onvoldoende bekend over hoe de leerlijn van leerlingen eruit ziet (leerlingen blijken bijvoorbeeld in leerjaar 1 van het voortgezet onderwijs beter te scoren dan enkele jaren later) , waardoor ruwe scores lastig gecorrigeerd kunnen worden voor het berekenen van een 'netto-effect'. Ook in andere domeinen zijn de mogelijkheden voor een valide en betrouwbare meting van toegevoegde waarde minder ver utigewerkt als voor taal en rekenen.

Box 1. Voorbeeld van rapportage toegevoegde waarde per vak en subgroep leerlingen (NY; http://schools.nyc.gov/OA/SchoolReports/2011-12/Progress_Report_2012_EMS_K678.pdf)



Student Progress

EAST NEW YORK MIDDLE SCHOOL OF EXCELLENCE

GRADE **C**

SCORE **27.0** (out of 60)

GRADE	SCORE RANGE
A	39.3 or higher
B	31.0 - 39.2
C	21.7 - 30.9
D	13.9 - 21.6
F	13.8 or lower

Student Progress represents 60% of the total score. The grade is based on growth percentiles, a measure of how much individual students improved on state tests in English and Math between 2011 and 2012.

	THIS SCHOOL'S RESULTS	COMPARISON TO PEER SCHOOLS (WEIGHTED 75%)	PERCENT OF PEER RANGE	COMPARISON TO CITY SCHOOLS (WEIGHTED 25%)	PERCENT OF CITY RANGE	POINTS POSSIBLE	POINTS EARNED
English							
Median Adjusted Growth Percentile (n=204)	65.5	45.1 – 65.5 – 77.7	62.6%	46.7 – 65.5 – 77.3	61.4%	15.00	9.35
Median Adjusted Growth Percentile for School's Lowest Third (n=66)	80.5	61.2 – 80.5 – 89.2	68.9%	58.2 – 80.5 – 86.6	78.5%	15.00	10.70
Mathematics							
Median Adjusted Growth Percentile (n=204)	48.0	29.4 – 48.0 – 84.4	33.8%	35.6 – 48.0 – 81.0	27.3%	15.00	4.83
Median Adjusted Growth Percentile for School's Lowest Third (n=69)	54.0	49.2 – 54.0 – 87.6	12.5%	46.6 – 54.0 – 85.0	19.3%	15.00	2.13

Student Performance

EAST NEW YORK MIDDLE SCHOOL OF EXCELLENCE

GRADE **B**

SCORE **15.0** (out of 25)

GRADE	SCORE RANGE
A	16.3 or higher
B	12.9 - 16.2
C	9.0 - 12.8
D	5.8 - 8.9
F	5.7 or lower

Student Performance represents 25% of the total score. The Student Performance grade is based on results on the 2012 state tests in English and Math and core course pass rates. State test metrics evaluate the percent of students who reach or exceed proficiency (Level 3 and 4) and students' average proficiency rating. Core course pass rates look at the percent of students in 6th through 8th grade who passed a course in a core subject area.

	THIS SCHOOL'S RESULTS	COMPARISON TO PEER SCHOOLS (WEIGHTED 75%)	PERCENT OF PEER RANGE	COMPARISON TO CITY SCHOOLS (WEIGHTED 25%)	PERCENT OF CITY RANGE	POINTS POSSIBLE	POINTS EARNED
English							
Percentage of Students at Level 3 or 4 (n=212)	26.4%	3.9% – 26.4% – 31.1%	82.7%	0.0% – 26.4% – 68.5%	38.5%	5.00	3.58
Average Student Proficiency (n=212)	2.66	2.22 – 2.66 – 2.70	91.7%	2.11 – 2.66 – 3.27	47.4%	5.00	4.03
Mathematics							
Percentage of Students at Level 3 or 4 (n=212)	26.4%	3.8% – 26.4% – 58.4%	41.6%	4.0% – 26.4% – 93.8%	24.9%	5.00	1.87
Average Student Proficiency (n=212)	2.67	2.58 – 2.67 – 3.16	46.7%	2.21 – 2.67 – 3.81	28.8%	5.00	2.11
Percent of Students Passing a Core Course							
English (n=203)	87.2%	67.0% – 87.2% – 100.0%	61.2%	69.5% – 87.2% – 100.0%	58.0%	1.25	0.76

3. Psychometrische randvoorwaarden

De psychometrische randvoorwaarden en potentiële problemen van het meten van toegevoegde waarde worden in de literatuur veelvuldig besproken (zie o.a. Eecke, 2004; Roeleveld e.a., 2011; Hamilton et al, 2002; Popham, 1999; Ladd en Walsh; Mooij e.a., 2012.; Bosker e.a., 2006; Van de Grift, 2009; Fitz-Gibbon and Tymms, 2002; Hanushek en Rivkin, 2010; Raudenbush, 2004; Rivkin, 2007; McCaffrey et al., 2004; zie ook: http://www.aaia.org.uk/category/pd/). De problemen en randvoorwaarden hebben enerzijds betrekking op de kwaliteit van toetsen voor het meten van leerprestaties, en anderzijds op het gebruiken van toetsscores voor het berekenen van leerwinst en toegevoegde waarde. In het eerste geval (kwaliteit van toetsen die ten grondslag liggen aan de toegevoegde waarde-maat) gaat het bijvoorbeeld over mogelijke measurement error in de toets wanneer de scores van leerlingen worden beïnvloed door andere factoren dan zijn/haar beheersing van de leerstof en vaardigheid; ook gaat het bijvoorbeeld om de mate waarin de toets een goede afspiegeling is van het inhoudsdomein dat wordt gemeten (bijvoorbeeld de referentieniveaus of de gebruikte onderwijsmethoden in een school).

Het ontwerp van de toets en de kwaliteitseisen die aan het ontwerp worden gesteld hangen in grote mate samen met het doel van de toets: een toets die bijvoorbeeld wordt gebruikt om een zak/slaag-beslissing te nemen over individuele leerlingen moet aan andere kwaliteitscriteria voldoen (bijvoorbeeld toetsitems die goed differentiëren rondom de zak-slaaggrens) dan een toets die bedoeld is om voortgang van leerlingen te meten (waarbij de items van twee toetsen op een zelfde vaardigheidsschaal worden geplaatst). Deze keuzes hebben ook consequenties voor de wijze waarop de toegevoegde waarde van een school kan worden bepaald. Toetsen die zijn ontworpen voor het nemen van een zak/slaag-beslissing zijn bijvoorbeeld niet geschikt voor het meten van toegevoegde waarde op basis van leerwinst (zie Woodworth e.a., 2012). Deze toetsen bevatten vaak weinig toetsitems die makkelijk of moeilijk zijn voor de 'doorsnee leerling' waardoor ze onbetrouwbaar zijn in het meten van beheersing en vaardigheid van leerlingen ver onder en boven de zak-slaaggrens.

Daarnaast bepaalt de keuze voor een toegevoegde-waarde maat en het niveau waarop wordt gerapporteerd ook de steekproef van leerlingen die wordt getoetst en de omvang van de toets. Voor een betrouwbare en valide rapportage op klas- of vakniveau is bijvoorbeeld een grotere steekproef en een langere toets (met meer items) nodig dan voor een rapportage op het niveau van de hele school. Volgens Hamilton en Koretz (in: Hamilton, 2002) is de kans op meetfouten groter wanneer de toegevoegde waarde wordt uitsplitst naar subgroepen of verschillende vakken. Het beperkt aantal leerlingen in een school, klas of vak zorgt er voor dat toevallige fluctuaties in toetsscores van een enkele leerling de gemiddelde score sterk beinvloedt waardoor de rapportage van toegevoegde waarde onbetrouwbaar is.

Voor een meer uitgebreide bespreking van mogelijke problemen en randvoorwaarden van verschillende maten van toegevoegde waarde verwijzen we graag naar eerder genoemde studies. Een belangrijke conclusie is in ieder geval dat het meten van toegevoegde waarde (ongeacht de maat die wordt gebruikt) een zorgvuldig data-managementsysteem vereist waarin bijvoorbeeld toetsscores van individuele leerlingen (met informatie over relevante achtergrondkenmerken) over de tijd kunnen worden gevolgd, aan de klas en school kunnen worden gekoppeld en eventueel per vak kunnen worden uitgesplitst.

4. Feedbackfunctie

De kamerbrief over de pilot leerwinst (25 november 2011) formuleert als eerste hoofddoel voor het meten van toegevoegde waarde: het stimuleren van opbrengstgericht werken van scholen en hun besturen. Inzicht in de leerwinst van leerlingen en de toegevoegde waarde van de school moet scholen en besturen informatie geven voor de eigen evaluatie van de kwaliteit van het onderwijs en daarmee schoolverbetering informeren. De aanname is dat betrouwbare en valide informatie over de output van de school doelgerichte (goal-oriented) schoolverbeterprocessen motiveren en informeren die leiden tot verbetering van leerprestaties van leerlingen. Het systematisch gebruik van opbrengstgegevens en gegevens over de toegevoegde waarde van de school zou 'deliberatie practice' kunnen vergroten waarbij scholen expliciet en weloverwogen beslissingen nemen over hun doelen, maatregelen nemen om deze doelen te bereiken en doelbereiking evalueren (zie Visscher en Ehren, 2011). Opbrengstgegevens en informatie over de toegevoegde waarde van de school functioneren in deze cyclus als prestatiefeedback waarbij de school een spiegel wordt voorgehouden die weergeeft 'hoe men het doet', bijvoorbeeld in vergelijking met de gemiddelde Nederlandse basisschool.
Verschillende meta-analyses (Black & William, 1998; Fuchs & Fuchs, 1986; Hattie, 2009; Hattie & Timperley, 2007; Kluger & DeNisi, 1996) beschrijven de effecten van prestatiefeedback en de wijze waarop leerlingen, leerkrachten en scholen deze feedback gebruiken voor het verbeteren van de gemeten prestaties. De literatuur over 'data use' en opbrengstgericht werken laat daarnaast zien hoe leerkrachten en scholen informatie over leerprestaties (kunnen) gebruiken om gefundeerder, bewuster en meer weloverwogen beslissingen te nemen over de inrichting en verbetering van het onderwijsproces en de schoolorganisatie (in plaats van op basis van intuïtie en subjectieve informatie). Deze literatuur over performance feedback en data use gebruiken we om hieronder de volgende thema's te bespreken:

- De wijze waarop scholen opbrengstgegevens, leerwinst en toegevoegde waarde (kunnen) gebruiken om schoolverbetering te informeren
- Kenmerken waaraan opbrengstgegevens, leerwinst en toegevoegde waarde moeten voldoen om schoolverbetering te informeren
- De randvoorwaarden in de (omgeving van de) school voor het gebruiken van opbrengstgegevens, leerwinst en toegevoegde waarde voor schoolverbetering

4.1 Gebruik van opbrengstgegevens en toegevoegde waarde voor schoolverbetering

Verschillende auteurs (Visscher en Coe, 2003; Verhaeghe, 2010; Hellrung en Hartig, 2013; Visscher en Ehren, 2011; U.S. Department of Education, 2010) beschrijven hoe scholen en leerkrachten opbrengstgegevens (kunnen) gebruiken om schoolverbetering en instructie te informeren. Opbrengstgegevens zijn zowel ongecorrigeerde (geaggregeerde) toetsscores van scholen, of informatie over de toegevoegde waarde van de school. Over het algemeen wordt verwacht dat vooral informatie over de toegevoegde waarde leidt tot schoolverbetering omdat deze indicator een waarheidsgetrouwer beeld geeft van het functioneren van de school.
In dit startdossier gaan we vooral in op het gebruik van opbrengstgegevens voor schoolverbetering; het gebruik van dergelijke gegevens voor het bijstellen en verbeteren van instructie door individuele leerkrachten (data-driven teaching) laten we buiten beschouwing.

Visscher en Coe (2003), Verhaeghe (2010) en Hellrung en Hartig (2013) beschrijven de volgende vormen van gebruik van prestatiefeedback, zoals opbrengstgegevens, voor schoolverbetering:

- Direct/instrumenteel gebruik: analyseren van opbrengstgegevens en beslissingen/acties nemen op basis van de analyse
- Conceptueel gebruik: opbrengstgegevens beïnvloeden de mindset van schoolleiders/docenten en indirect hun acties; bijvoorbeeld meer 'goal-oriented' gedrag
- Symbolisch gebruik: opbrengstgegevens worden gebruikt om beslissingen te rechtvaardigen en keuzes te legitimeren

Voorbeelden voor direct instrumenteel gebruik zijn vooral in de Verenigde Staten te vinden waar data use en opbrengstgericht werken al een aantal jaar op de agenda staat. Verschillende staten, districten

en lokale overheden hebben data-systemen en ondersteuningsprogramma's ingericht om opbrengstgericht werken en 'data-driven decision making' in scholen te stimuleren. Deze systemen en programma's, en de wijze waarop scholen data analyseren en gebruiken is veelal sterk gerelateerd aan de wijze waarop deze data voor school accountability worden gebruikt en de wijze waarop data aan scholen worden gerapporteerd (Means, Padilla en Gallagher, 2010). Externe accountability is een belangrijke motivatie om data te analyseren en te gebruiken voor schoolverbetering en het specifieke gebruik van deze data is in deze landen vaak niet los te zien van de functie die deze data heeft voor de externe beoordeling van de school.

De volgende voorbeelden van gebruik van opbrengstgegevens voor schoolverbetering komen uit de literatuur en aanvullende 'good practice' studies naar voren. Deze voorbeelden geven een indicatie van de wijze waarop opbrengstgegevens *kunnen* worden gebruikt door scholen. Verschillende auteurs laten echter zien dat scholen in een low stakes context opbrengstgegevens niet of beperkt gebruiken, zelfs niet wanneer deze (zoals in een pilot-project in Vlaanderen, zie Hulpia e.a, 2004; Van Petegem en Vanhoof, 2003) actief aan scholen worden verstrekt.

Voorbeelden van gebruik opbrengstgegevens voor schoolverbetering:

- Reguliere zelf-evaluaties waarbij doelen worden gesteld en geëvalueerd (bijvoorbeeld per vak, leerjaar, leerlingengroep) voor (verbetering van) prestaties (schoolbreed, per vak, per leerjaar, per klas, per leerlinggroep), waarbij prestaties met andere scholen worden vergeleken, sterke/zwakke punten worden geïdentificeerd, onderwijs(organisatie)processen worden bijgestuurd (bijvoorbeeld identificeren van en bijsturen van 'gaten' in het curriculum, financiële/personele middelen worden geheralloceerd, good practices in de school worden geïdentificeerd, verspreiden en opgeschaald.
- Evaluatie van innovaties/verbeterprogramma's. Verhaeghe (2010), Fitz-Gibbon en Tymms (2002) en Carlson e.a. (2011) beschrijven hoe verbeterprogramma's kunnen worden geëvalueerd door middel van een vergelijking van een beginmeting bij een generatie leerlingen (in een groep van deelnemende scholen) vóór de invoering van de innovatie met een nameting bij een navolgende generatie, bij voorkeur twee of drie jaar na de invoering van de innovatie.
- Planning van onderwijs(organisatie): projecties van prestaties van leerlingen gebruiken voor planning van onderwijsaanbod en inzet van personeel en financiële middelen, plaatsing van leerlingen in ondersteunende programma's en aanvullende begeleiding (r.t., zomerschool), plannen van het reguliere curriculumaanbod.
- Beoordelen personeel en inzet van professionalisering: evaluatie van functioneren van afdelingen/leerkrachten, analyseren van profielen van klassen om professionele ontwikkeling van leerkrachten te informeren (wanneer bijvoorbeeld prestaties van leerlingen in bepaalde vakken achterblijven).

Volgens Visscher en Coe (2002), Verhaeghe (2010) en Hellrung en Hartig (2013) wordt prestatiefeedback vaak niet gebruikt door scholen omdat scholen problemen ervaren met het interpreteren van de gepresenteerde statistische gegevens; het kost tijd om grafieken, tabellen en statistisch jargon te doorgronden en schoolleiders en docenten zijn niet in staat om bijvoorbeeld betrouwbaarheidsintervallen, schaalgemiddelden e.d. te interpreteren. Leesbare, overzichtelijke rapportages op klas en schoolniveau, aangevuld met externe ondersteuning in het analyseren en interpreteren van de overzichten en bepalen van interventies, zijn nodig om het effect van prestatiefeedback te verbeteren. Box 2 presenteert een voorbeeld uit New York van de wijze waarop prestatie informatie (zoals gepresenteerd in box 1) wordt gebruikt voor schoolverbetering.

Box 2. Data use in New York City
De Department of Education van New York City publiceert jaarlijks een 'progress report card' waarop de prestaties van scholen zichtbaar worden gemaakt. In de kaart worden prestaties en groei getoond, zoals onder andere het percentage leerlingen dat op niveau 1 en 2 van de gestandaardiseerde high stakes test presteert, en percentage leerlingen dat één leerjaar in prestaties vooruit is gegaan in vergelijking met vergelijkbare scholen, en in vergelijking met alle scholen in de stad (zie box 1). Scholen moeten jaarlijks een schoolverbeterplan opleveren waarin zij doelen opnemen om hun prestaties op deze maten te verbeteren. In dit schoolverbeterplan vermelden zij ook de acties om deze doelen te bereiken en de wijze waarop zij deze acties evalueren.
Voorbeelden van doelen zijn:

- ***ELA- 1 Year of Progress:*** *62.2% of our students performed at levels 3&4.* ***67%*** *of our students made at least 1+ year of progress, which is* ***79.5%*** *of the way from the lowest (46.8%) to the highest (72.2%) score relative to our Peer Horizon and* ***80.1%*** *of the way relative to our City Horizon.*
- ***MATH:*** ***87.7%*** *of our students performed at levels 3 &4.* ***66.4%*** *of our total student population made 1 year of progress which is* ***83.2%*** *of the way from the lowest* ***(42.7%)*** *to the highest* ***(71.2%)*** *score relative to our Peer Horizon and* ***62.2%*** *of the way relative to the City Horizon.*

Het opbrengstgericht werken in scholen wordt ondersteund met gestandaardiseerde formatieve benchmark assessments die zijn ontwikkeld in opdracht van de Department of Education; de Department of Education levert bovendien gestructureerde 'item-skills' overzichten aan waarin de prestaties van leerlingen en groepen leerlingen per vaardigheid worden gepresenteerd en op itemniveau kunnen worden geanalyseerd. Elke school heeft een 'inquiry team' waarin leerkrachten samenwerken om data te analyseren en acties afspreken om leerprestaties te verbeteren in het licht van de gestelde schoolbrede doelen. De Department of Education verwacht dat scholen deze activiteiten in een online platform 'Achievement Reporting and Innovation System' (ARIS) loggen, alwaar zij ook de data kunnen analyseren en ondersteunende materialen kunnen vinden voor het remediëren van achterblijvende prestaties. De quality reviewers (vergelijkbaar met onderwijsinspecteurs) evalueren en beoordelen deze vorm van opbrengstgericht werken tijdens hun bezoeken aan scholen (zie ook: Ehren en Hatch, submitted).

4.2 Kenmerken van opbrengstgegevens en toegevoegde waarde voor schoolverbetering

De wijze waarop opbrengstgegevens en informatie over toegevoegde waarde wordt gebruikt hangt sterk samen met de inhoud van toetsen en de wijze waarop toetsscores en informatie over de toegevoegde waarde van de school worden aangeleverd en kunnen worden geanalyseerd. Visscher en Coe (2003) onderscheiden technische, algemene en ethische randvoorwaarden waaraan prestatiefeedback moet voldoen om schoolverbetering te informeren.
Technische randvoorwaarden hebben betrekking op de kwaliteit van de gegevens (validiteit, betrouwbaarheid e.d.) en de tijdigheid waarmee ze beschikbaar zijn.
Algemene voorwaarden zijn vooral gericht op de mate waarin de prestatiefeedback overeenkomt met de informatiebehoefte van de school door bijvoorbeeld een antwoord te geven op de vraag hoe de school het doet met vergelijkbare scholen, en door (voor de school) geloofwaardige informatie te genereren. Volgens Cizek (2003) en Looney (2009) en Hellrung en Hartig (2013) zijn gestandaardiseerde toetsen vooral geschikt om een betrouwbaar beeld te geven van prestaties van leerlingen in brede inhoudsgebieden (zoals 'optellen en aftrekken' binnen het vak rekenen). Deze toetsen laten zien waarin de school sterk of zwak presteert, en de analyse van de toetsen kan bijvoorbeeld leiden tot aanpassingen in het curriculum, of in de geplande instructietijd. De scores op deze gestandaardiseerde toetsen kunnen echter niet altijd worden opgesplitst in gedetailleerde overzichten van de specifieke onderdelen van de leerstof waarop leerlingen zwak presteren, en gegevens komen veelal pas na een aantal weken beschikbaar. Deze toetsen zijn daarom vaak niet geschikt om beslissingen over instructie en lesplannen te informeren. Volgens Cizek (2003) en Looney (2009) zijn interne toetsen, zoals methodegebonden toetsen, en toetsen en opdrachten die docenten zelf samenstellen voor diagnostische activiteiten geschikter. Nederlands onderzoek (zie Ehren et al, in

prep) ondersteunt deze conclusie en laat zien dat basisscholen de gestandaardiseerde Cito-toets vooral gebruiken om na te gaan of de onderwijsmethoden alle getoetste onderdelen in voldoende mate behandelen (of dat aanvullende leerstof moet worden aangeboden), en dat de uitkomst van de Cito-toetsen worden gebruikt als benchmark om te bepalen of leerlingen op een voldoende niveau presteren. De Cito-eindtoets is, volgens docenten, beperkt bruikbaar om ook beslissingen te informeren over instructie en lesplannen omdat de toets aan het einde van de schoolloopbaan wordt afgenomen en scores niet meer kunnen worden gebruikt om het onderwijs aan de getoetste leerlingengroep bij te stellen.
Ethische randvoorwaarden hebben te maken met het voorkomen van mogelijke schade door bijvoorbeeld onbedoelde consequenties van benchmarking op een beperkte set van indicatoren.

Opbrengstgegevens en informatie over toegevoegde waarde zijn dus bruikbaar voor schoolverbetering wanneer ze voldoen aan de volgende kenmerken (OECD, 2008; Fitz-Gibbon and Tymms, 2002):

- Tijdigheid van de gepresenteerde informatie (wordt de informatie nog binnen het huidige schooljaar gepresenteerd zodat het onderwijs voor het huidige cohort nog kan worden bijgestuurd, of zodat de planning voor komend schooljaar kan worden geïnformeerd?)
- Mate waarin de toegevoegde waarde op het juiste aggregatieniveau wordt gepresenteerd om besluitvorming in de school te informeren (per leerjaar, klas, afdeling/vak, etc.)
- Mate waarin informatie over toegevoegde waarde laat zien of er bijgestuurd moet worden, en op welke onderdelen van het curriculum/instructie bijgestuurd moet worden (laat de informatie zien welke schooldoelen behaald zijn?; zijn er overzichtsrapportages beschikbaar over welke vakken/onderwerpen –groepen van- leerlingen niet beheersen?; geeft de informatie inzicht in welke onderdelen van de methode herhaald moet worden?; geeft de informatie inzicht in het functioneren van leerkrachten/klassen/afdelingen? Etc.)

Box 3 laat een voorbeeld uit Engeland zien van de wijze waarop toegevoegde waarde aan schoolleiders kan worden gepresenteerd voor evaluatie van vaksecties.

Box 3. Presentatie van TW-informatie in Engeland
http://www.cem.org/attachments/publications/CEMWeb009%20Feasibility%20Study%20Nat%20System%20VA%20Indicators.pdf

Fitz-Gibbon en Tymms (1997) tonen een tweetal overzichten die schoolleiders de mogelijkheid geeft voor een 'quick check' van de resultaten van verschillende vakken en afdelingen in hun school. Deze overzichten worden volgens hen gebruikt door schoolleiders in conversaties met sectieleiders.

Figure 4.3 Paired Bar graphs of Value-added average scores for each subject, with 'margin of uncertainty' on the prediction represented by a short line

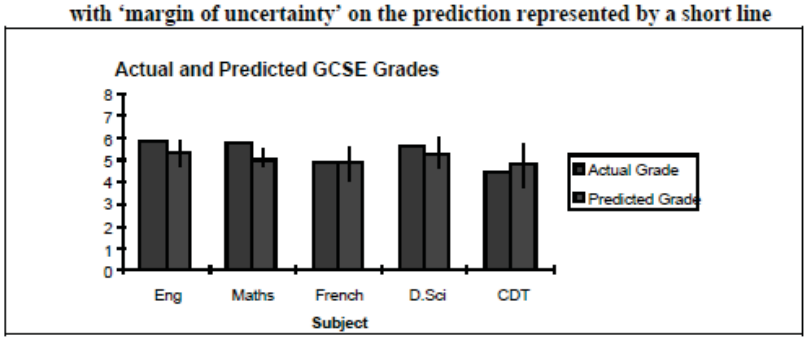
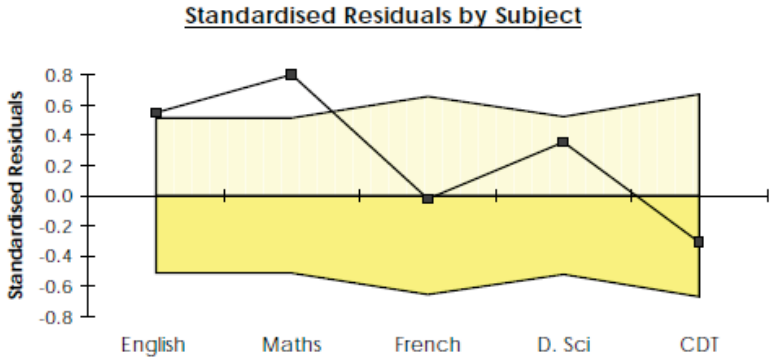


Figure 4.4 Statistical Process Control chart of the same data as in Figure 4.3



4.1 Externe factoren

Verschillende auteurs (o.a. Visscher en Coe, 2003; Visscher en Ehren, 2011) beschrijven aanvullende condities voor het gebruik van opgebrengstgegevens voor schoolverbetering, zoals situationele variabelen en persoonlijkheidskenmerken (e.g. schoolinterne condities: innovatiecapaciteit, opbrengstgerichte, prestatiegerichte schoolcultuur, veilig klimaat, lerende cultuur, data literacy) en de mate waarin scholen externe ondersteuning krijgen bij het interpreteren en gebruiken van data voor verbetering. Ook de complexiteit van de door te voeren verbetering verklaart de mate waarin opbrengstgegevens daadwerkelijk leiden tot een verbetering van leerprestaties. Volgens Hellrung en Hartig (2013) worden opgebrengstgegevens met name door rekenen/wiskunde docenten gebruikt omdat deze docenten meer bekend zijn met statische termen en beter in staat zijn data te interpreteren; bovendien leent de (hierarchische) structuur van het rekenen/wiskunde-onderwijs (met een duidelijkere structurering van afzonderlijke onderwerpen) zich beter voor het maken aanpassingen in het onderwijs op basis van opbrengstgegevens.

5. Accountabilityfunctie

De Minister van Onderwijs, Cultuur en Wetenschap formuleert in de kamerbrief over de pilot leerwinst (25 november 2011) als tweede hoofddoel van het meten van de toegevoegde waarde: 'een betere weging van de opbrengsten van een school in het externe inspectietoezicht'. Opbrengstgegevens en informatie over de toegevoegde waarde van de school krijgen daarmee ook een rol in de accountability van scholen. Belangrijke vragen in dat kader zijn:

- Hoe kan een maat voor toegevoegde waarde worden gebruikt voor het toezicht op/externe beoordeling van scholen (voor het meten van output van scholen en voor het classificeren van scholen)?
- Welke impact heeft deze invulling van het toezicht (en de accountabilityfunctie van toegevoegde waarde) voor schoolverbetering?

In onderstaande paragraaf bespreken we eerst hoe de onderwijsinspectie op dit moment opbrengsten van scholen beoordeelt en verbeteringen die in deze manier van beoordelen kunnen worden doorgevoerd. Vervolgens bespreken we de effecten en neveneffecten van toezicht op de toegevoegde waarde van scholen.

5.1 Toegevoegde waarde om scholen te beoordelen

In verschillende landen wordt de kwaliteit van scholen beoordeeld op basis van de toegevoegde waarde van de school. De wijze waarop opbrengstgegevens worden geanalyseerd en het statistische model om toegevoegde waarde te berekenen verschilt echter sterk (Timmermans, 2012). Variaties zijn er in de keuze voor een groei of statusmodel (zie eerdere paragraaf), en de keuze voor controlevariabelen.

Algemene problemen bij het gebruiken van toegevoegde waarde in het beoordelen van scholen

Verschillende auteurs (o.a. Mooij e.a., 2012; Wijnstra e.a., 2003; Verhaeghe, 2010) bespreken de problemen in het beoordelen van scholen op basis van hun toegevoegde waarde. Volgens Wijnstra e.a. (2003) is de specifieke bijdrage van scholen aan leeropbrengsten moeilijk te isoleren, ook bij het gebruik van een begin en eindmeting. Bijkomende factoren, zoals leeftijd, geslacht, sociaal milieu e.d. zullen behalve verschillen in de beginmeting ook tijdens het onderwijsproces doorwerken en tot verschillen in uitkomsten leiden. Het tijdsverschil tussen begin- en eindmeting in het basisonderwijs is bovendien erg groot en prestaties van leerlingen op begin en eindmeting hangen beperkt samen. Het aantal leerlingen is bovendien vaak klein waardoor de gegevens vaak weinig stabiel zijn. Ladd en Walsh (2002) en Timmermans (2012) laten zien dat verschillende indicatoren voor toegevoegde waarde (waarbij correcties voor verschillende achtergrondfactoren worden aangebracht) weliswaar samenhangen, maar toch tot sterk verschillende classificaties van individuele scholen als ineffectief, gemiddeld en effectief kunnen leiden. Met name voor de scholen rondom het gemiddelde vinden veel verschuivingen plaats, waarbij meer dan 50% van de scholen meer dan 10 plaatsen in de rangorde van scholen opschuiven bij het gebruik van een andere toegevoegde waarde-maat. Alleen de classificatie van de meest en minst effectieve scholen op basis van toegevoegde waarde is redelijk stabiel.

Toegevoegde waarde in het Nederlandse toezicht

De Nederlandse onderwijsinspectie gebruikt opbrengstgegevens van scholen in het risicogestuurde toezicht en in de kwaliteitsbeoordeling tijdens schoolbezoeken[1]. De kwaliteit van basisscholen wordt vooral bepaald aan de hand van de scores van leerlingen op de Cito-Eindtoets Basisonderwijs, waarbij leerlingscores worden geaggregeerd en worden gecorrigeerd voor het opleidingsniveau van de ouders van leerlingen. Om meetfouten te minimaliseren worden gegevens van drie achtereenvolgende jaren gebruikt, en voor scholen met minder dan 10 leerlingen per leerjaar voor vijf achtereenvolgende jaren.
In het voortgezet onderwijs wordt het advies van het basisonderwijs, doorstroomgegevens en resultaten van schoolonderzoeken en eindexamens gebruikt om het rendement in de onderbouw en bovenbouw, en het behaalde resultaat aan het eind van het voortgezet onderwijs en in de

[1] Zie http://www.onderwijsinspectie.nl/binaries/content/assets/Actueel_publicaties/2012/beoordeling-opbrengsten-basisonderwijs-2012.pdf

vervolgloopbaan te beoordelen. Naast opbrengstgegevens worden ook andere informatiebronnen en vormen van onderzoek gebruikt om scholen te beoordelen, zoals interviews met schoolbesturen, en observaties van lessen. De opbrengstgegevens vormen dus slechts één informatiebron en indicator voor de beoordeling van scholen.

Mooij e.a. (2012) onderzochten voor het primair onderwijs hoe de inspectie haar beoordeling van de kwaliteit van scholen met behulp van opbrengstgegevens kan worden verbeterd door alternatieve correctiefactoren (dan het opleidingsniveau van de ouders) te gebruiken. Zij ontworpen zeven modellen waarbij op verschillende manieren wordt gecorrigeerd voor verschillende leerlingkenmerken (opleiding ouders, etniciteit, huishoudinkomen, percentage uitkeringen, percentage lage inkomens, percentage niet-Westerse allochtoon), en drie soorten multiniveau regressie-analyses (o.a. covariaten op individueel of geaggregeerd niveau). Deze modellen werden getoetst met behulp van de landelijke COOL data, en gegevens die worden beheerd door het Centraal Bureau voor de Statistiek. In de multiniveau analyses werden 402 scholen en 8.561 leerlingen betrokken. De resultaten demonstreren dat twee correctiefactoren (gedetailleerde opleiding ouders; etniciteit) relatief betere resultaten geven dan de huidige inspectiebeoordeling waarbij gecorrigeerd wordt met behulp van leerlinggewicht. Timmermans (2012) vond daarnaast ook dat de toegevoegde waarde van scholen in verschillende cognitieve domeinen (taal/rekenen) onderling maar matig positief samenhangen en concludeert dat een combinatie van deze verschillende uitkomstmaten nodig zijn om de kwaliteit van een school te beoordelen. Beide onderzoeken laten zien dat de validiteit en betrouwbaarheid van de inspectiebeoordeling van de opbrengsten van de school kan worden verbeterd.

Gebruik van meerdere metingen en indicatoren in het beoordelen van de toegevoegde waarde
Suggesties uit de internationale literatuur voor het verbeteren van externe beoordeling van scholen op basis van toegevoegde waarde hebben veelal betrekking op het gebruik van meerdere metingen (bijvoorbeeld meerdere toetsen van dezelfde vaardigheden, of meerdere toetsen van verschillende vakdomeinen), waarbij de resultaten van verschillende meetinstrumenten worden gecombineerd in een beoordeling van de school (zie Koretz, 2003; Ehren, 2011). Hamilton en Koretz (in Hamilton e.a., 2002) en Gong en Hill (2001) beschrijven verschillende modellen om de resultaten te combineren tot een eindclassificatie van de kwaliteit van scholen:

- Compensatory model: onderpresteren op één van de (TW-)maten kan worden gecompenseerd met hoge prestaties op een andere (TW-)maat
- Conjunctive model: goede prestaties op alle (TW-)maten zijn nodig voor een positieve beoordeling
- Confirmatory model: prestaties op één van de (TW-)maten wordt gebruikt voor het valideren van de prestaties op een andere (TW-)maat
- Complentary model: goede prestaties op één van meerdere (TW-)maten is voldoende voor een voldoende beoordeling

Deze modellen leiden naar verwachting tot een meer valide en betrouwbare beoordeling en dat komt, volgens Ladd en Walsh (2002), de geloofwaardigheid van het externe toezicht ten goede, en daarmee ook de mate waarin zo'n systeem gewenste effecten heeft (Ladd en Walsh, 2002). Anderzijds verhogen deze modellen de complexiteit van het berekenen van de kwaliteit van een school; daarmee neemt ook de bruikbaarheid van deze informatie voor scholen en hun stakeholders af.

5.2 Effecten en neveneffecten van 'toegevoegde waarde'-toezicht

In de Verenigde Staten zijn opbrengstgegevens van scholen doorslaggevend bij het beoordelen van scholen. Een reviewstudie van Au (2007) naar de effecten van dergelijke 'test-based accountability' systemen laat zowel positieve, ambigue en negatieve effecten zien. Deze effecten ontstaan door het beoordelen en/of afrekenen van scholen op hun opbrengstgegevens; in deze studies wordt geen onderscheid gemaakt in externe beoordeling van scholen aan de hand van ruwe toetsscores, of aan de hand van de toegevoegde waarde van de school. Ladd en Walsh (2002), Nichols en Berliner (2005), Timmermans (2012) en de OECD (2008) verwachten echter dat externe beoordelingen op basis van toegevoegde waarde, in plaats van ongecorrigeerde opbrengstgegevens, minder neveneffecten veroorzaken, met name wanneer er sprake is van meerdere verschillende metingen van verschillende indicatoren (bijvoorbeelde verschillende toetsen van verschillende vakken).

Positieve effecten van 'test-based accountability' ontstaan volgens Stecher (in Hamilton, 2002) wanneer scholen hun onderwijsaanbod vergroten, of effectievere instructie geven wanneer zij worden gecontroleerd op hun opbrengstgegevens. Scholen verhogen bijvoorbeeld de instructietijd in de vakken waarin leerlingen zwak presteren (Koretz, McCaffrey and Hamilton, 2001). Dit soort reacties hebben een gunstige uitwerking op het leerproces van leerlingen en dragen ook bij aan de verbetering van hun prestaties.

Koretz, McCaffrey en Hamilton (2001) geven echter ook voorbeelden van *negatieve* reacties, zoals het frauderen met de toets of het uitsluiten van zwak presterende leerlingen van de toets. Jacob en Levitt (2003), Figlio en Getzler (2002) en Cullen and Reback (2006) beschrijven hoe docenten goede antwoorden voorzeggen, toetsopgaven de dag voor de toets met leerlingen doornemen dan wel zwakke leerlingen ertoe aanzetten om zich ziek te melden tijdens de toetsafname, om zo de toetsscores van de school te verhogen. Neveneffecten zijn volgens Ladd en Walsh (2002) en Nichols en Berliner (2005) ook dat scholen met lage scores problemen ervaren met het aantrekken van goede leerkrachten en een dalend moraal en motivatie onder leerkrachten.

Koretz et al. (2001) beschrijven ook ambigue reacties van scholen; de uitkomst van deze reacties kan zowel positief als negatief zijn, afhankelijk van de context waarbinnen ze ontstaan. Ambigue reacties betreffen bijvoorbeeld het herverdelen van middelen naar vakken en onderwerpen die in de toets aan bod komen ten nadele van vakken/onderwerpen die niet worden getoetst. Instructietijd, budget, instructiematerialen e.d. worden dan bijvoorbeeld vooral ingezet voor het onderwijs in taal en rekenen, en minder voor andere vakken. Ambigue reacties ontstaan ook wanneer docenten hun instructie richten op specifieke aspecten van de test (bijvoorbeeld leerlingen instrueren hoe ze multiple choice vragen kunnen beantwoorden), of wanneer scholen een selectief aannamebeleid gaan voeren en alleen leerlingen toelaten die naar verwachting goed gaan presteren, of wanneer de instructie wordt gericht op die leerlingen waarmee de meeste leerwinst valt te behalen.

Voorbeelden/vermoedens Nederland
Bovenstaande voorbeelden hebben vooral betrekking op de Verenigde Staten. Voor Nederland worden de volgende voorbeelden genoemd (Karsten, Visscher & Jong, 1999; Bosker e.a., 2006; Onderwijsraad, 2003; Vereniging van Leraren in Levende Talen, 2013; Ehren, Schildkamp en Gelderblom, in prep; Ehren en Swanborn, 2012; Vermeulen, 2012):

- Uitsluiten van leerlingen voor Eindtoets basisonderwijs, of voor het centraal schriftelijk eindexamen in het voortgezet onderwijs (van leerlingen met lage cijfers voor het schoolexamen)
- Leerlingen helpen bij afname toets
- Leerlingen buiten beschouwing laten bij berekenen schoolscore
- Selectie van leerlingen 'aan de poort'
- Af laten stromen van leerlingen in het VO om eindexamenrendement te verhogen
- Verwijzen van leerlingen naar het speciaal onderwijs
- Demotivatie leerkrachten
- Laag presterende scholen hebben problemen met vinden van leerkrachten
- Curriculumversmalling
- Afname samenwerkingsbereidheid scholen
- Toetstraining
- Teaching to the test
- Schoolonderzoeken VO inhoudelijk afstemmen op eindexamens
- Afname van samenwerkingsbereidheid tussen scholen
- Toename van segregatie van leerlingen.

Volgens Timmermans (2012) heeft een toegevoegde waarde-maat voor het meten van opbrengsten (in plaats van gebruik ruwe/ongecorrigeerde data) naar verwachting wel minder negatieve effecten, bijvoorbeeld omdat incentives voor het uitsluiten van zwakke leerlingen verdwijnen. Ook het gebruik van meerdere maten of aanvullende metingen van bijvoorbeeld niet cognitieve opbrengsten kan neveneffecten beperken volgens Koretz (2003). Het manipuleren van meerdere meetinstrumenten is

immers moeilijker dan het manipuleren of frauderen met één toets, zeker wanneer deze instrumenten en bijbehorende indicatoren regelmatig worden aangepast. De Inspectie voor de Gezondheidszorg (2012) doet in haar rapport bijvoorbeeld de aanbeveling om jaarlijks 20-25% van de indicatoren voor het beoordelen van ziekenhuizen te vervangen, en een indicator maximaal 4-5 jaar te gebruiken om strategisch gedrag te voorkomen. Een belangrijke conclusie uit het onderzoek naar 'test-based accountability' is in ieder geval dat een uitkomstmaat waar sancties aan zijn verbonden altijd tot strategisch gedrag zal leiden omdat scholen er groot belang bij hebben om hun score op deze maat te verbeteren.

5.3 Factoren die van invloed zijn op effecten en neveneffecten

Uit de internationale literatuur blijkt dat positieve, negatieve of ambigue reacties van scholen onder andere worden bepaald door:

- Kenmerken van de toets (voorspelbaarheid, narrowness of sampled items, mogelijkheden voor cheating tijdens afname)
- Targets en de wijze waarop scholen worden geclassificeerd als zeer zwak, zwak, voldoende, goed (strategisch gedrag rondom de grensscore)
- Consequenties en high stakes

Deze aspecten worden hieronder verder toegelicht (afgeleid van Visscher en Ehren, 2011).

Kenmerken van de (afname van de) toets

De kwaliteit van de toets waarmee scholen worden beoordeeld, bepaalt in belangrijke mate of de beoordeling effectief is of tot neveneffecten leidt. Met name de mate waarin een toets voorspelbaar is (het toetsen van dezelfde onderwerpen met behulp van hetzelfde soort vragen), en de timing van de toets zijn van invloed op de wijze waarop scholen opereren (Stecher; in Hamilton e.a., 2002). Daarnaast bepalen de afnamecondities (en de mogelijkheden voor fraude) of er negatieve effecten optreden.

Een toets is bedoeld om de vaardigheden van leerlingen in een bepaald domein (bijvoorbeeld taal) te meten. De uiteindelijke vragen in de toets representeren echter altijd slechts een deel van alle onderwerpen en vaardigheden in dat domein (bijvoorbeeld alleen schrijf- en leesvaardigheden en niet de mondelinge taalvaardigheid). De aanname is dat scores van leerlingen op de toets een beeld geven van hoe leerlingen presteren op het *hele* domein. Onderzoek in de Verenigde Staten laat echter zien dat toetsen waarmee scholen worden beoordeeld vaak dezelfde vaardigheden van leerlingen meten aan de hand van dezelfde item formats (bijvoorbeeld het multiple choice format, of open vragen waarbij een kort antwoord moet worden geformuleerd). Shepard en Cutts Dougherty (1991) vonden bijvoorbeeld dat leerkrachten hun leerlingen alleen leren om getallen op te tellen in een verticaal format wanneer dit format in de toets wordt gebruikt. Leerlingen bleken logischerwijs minder goed te presteren in het optellen van getallen in een horizontaal format. De voorspelbaarheid en de beperkte validiteit en betrouwbaarheid van de toets faciliteren in dit geval 'teaching to the test' waarbij onderwijs en instructie worden versmald naar de vakken, onderwerpen en vaardigheden die worden getoetst, en naar de item-formats waarin ze worden getoetst. Als gevolg hiervan wordt het leren van leerlingen ingeperkt, en geeft ook de score van een leerling op de toets een minder nauwkeurig beeld van de daadwerkelijke vaardigheid van leerlingen in het hele domein. Fuhrman (2003) verwacht dat geavanceerdere toetsen met open vragen rijkere instructie stimuleren (minder geavanceerde toetsen met daarin veel multiple choice vragen worden geassocieerd met instructie waarin het oefenen met werkbladen en het memoriseren van antwoorden centraal staan).

Targets en classificatie

Niet alleen de inhoud van de toets, maar ook de wijze waarop toetsscores worden geaggregeerd om scholen te beoordelen en de targets waar scholen aan moeten voldoen hebben implicaties voor schoolverbetering. Targets hebben betrekking op de benchmarks en grensscores die gebruikt worden om te beoordelen of scholen voldoende of onvoldoende presteren.

Hamilton en Koretz (2002) en Goertz en Duffy (2001) beschrijven twee typen opbrengsttargets: absolute of relatieve prestatietargets waarbij scholen respectievelijk een minimumscore moeten halen, dan wel waarbij wordt gekeken naar de positie van de school binnen een distributie van vergelijkbare

scholen, en naar een groeitarget, waarbij verwacht wordt dat scholen hun scores verbeteren ten opzichte van eerdere prestaties. Targets kunnen worden geformuleerd voor de school als geheel, voor afzonderlijke vakken, en voor verschillende subgroepen leerlingen (bijvoorbeeld leerlingen met verschillende sociaaleconomische en etnische herkomsten).
Hanushek en Raymond (2002) laten zien dat scholen die dicht bij de target presteren hun gedrag meer veranderen dan scholen die daar ver onder of boven presteren. Deze scholen zullen naar verwachting een grotere stimulans ervaren om opbrengstgericht te gaan werken, maar zullen ook meer strategisch gedrag vertonen. Jacob en Levitt (2003) en Stechter (2002) vonden bijvoorbeeld dat docenten in laag presterende scholen frauderen met de toets, of leerlingen excessief met de toets laten oefenen. Koretz en Baron (in Hamilton en Koretz, 2002) redeneren dat scholen en docenten die met onhaalbare targets worden geconfronteerd een sterke stimulans zullen ervaren om deze targets op oneigenlijke manieren te realiseren (vooral wanneer er zware consequenties aan de targets zijn verbonden).
De specifieke invulling van de target bepaalt dus hoe scholen reageren en hoe zij opbrengstgegevens gebruiken. Relevante indicatoren zijn de vakken, de leerjaren en de groepen leerlingen waarvoor targets worden geformuleerd. Targets op het gebied van taal en rekenen zullen scholen vooral motiveren om hun aandacht op deze vakken te richten. De onderwerpen die in de toets aan bod komen en de specifieke formats waarin ze worden getoetst zijn bepalend voor de inhoud van de instructie. Scholen in de V.S., waar targets voor specifieke subgroepen leerlingen (met verschillende sociaaleconomische en etnische herkomsten) worden gesteld, gebruiken opbrengstgegevens om de instructie op deze groepen af te stemmen.
Targets die bestaan uit minimale toetsscores blijken vooral te leiden tot het gebruik van opbrengstgegevens voor ‘educational triage’. Scholen richten hun aandacht, (extra) instructie, en begeleidingstijd dan vooral op leerlingen die rondom de grensscore presteren, met als doel deze leerlingen een toetsscore op of boven de minimale targetscore te laten halen. De leerlingen die ver onder de minimale target score presteren worden als ‘hopeloos’ beschouwd, en ook de leerlingen die al goed presteren krijgen geen extra aandacht. Deze vormen van accountability stimuleren scholen dus vooral om een zo hoog mogelijke overall score te halen in plaats van elke leerling de tijd, aandacht en begeleiding te geven die hij/zij nodig heeft.
Box 4 bevat een voorbeeld van de wijze waarop een school in New York prestatie informatie gebruikt.

Box 4. Voorbeeld strategisch gedrag en schoolverbetering New York City
De onderstaande teksten zijn overgenomen uit een jaarlijks schoolverbeterplan van een basisschool in New York. Scholen zijn verplicht om jaarlijks aan te geven hoe zij hun score op de progress report card (zie box 1) verbeteren.

1) By June 2010 all students inclusive of students in the LEP and SWD groups will demonstrate progress towards achieving state standards as evidenced by a 3% increase in student scoring at level 3 and 4, on the New York State ELA assessment.

Professional development:
On site Literacy Coach, mentor, consultants from Read Well in grades K-2 and Aussie Consultants in grades 3-5 will provide PD in the areas of literacy including: analyzing student data, comprehension strategies, differentiation, goal setting, collaborative team teaching and implementation of the reading and writing workshops. Opportunities for collaboration and inter-visitations will be provided. In order to improve writing instruction a Aussie consultant will provide PD to grades 3-5 by demonstrating best practices in third grade classrooms. Literacy coach will coordinate PD opportunities, provide demo lessons, meet with teachers during common planning, prepare schedules for consultant visits based on teachers' professional needs and conduct debriefing sessions. Common preparation periods will be used for professional development, planning, data analysis and inquiry team studies. The Read Well consultant will provide professional development on Read Well 2 to second grade teachers.

Academic intervention:
Small groups will be formed based on formal and informal assessments for students in grades 2-5. These groups will meet regularly and will be changed to meet the individual needs of our students. Read 180, a computer based literacy program, will be utilized for students in grades 3-5 who are approaching the standards. Our AIS push in/pull-out is designed for all level 1 and 2 students in grades 3-5 and at-risk students in grade 2. In addition, the SETTS teacher services at-risk students in K-5. SWDs and LEPs are included in the AIS pull out and Read 180 programs. ESL teachers provide additional small group instruction in test preparation and ELA skills to our LEP students.

Consequenties
Tot slot vormen ook de consequenties voor laag presterende scholen een belangrijke impuls voor zowel effecten als neveneffecten van vormen van 'test-based accountability'. Consequenties hebben betrekking op sancties (bijvoorbeeld boetes) en interventies (zoals intensieve monitoring) voor laag presterende scholen, en beloningen voor hoog presterende scholen. Ook docenten en leerlingen kunnen consequenties ondervinden van lage opbrengsten wanneer prestaties van leerlingen leidend zijn voor de beloning van leerkrachten, en wanneer leerlingen (zoals in de V.S.) bij lage prestaties in de zomer verplicht naar school moeten.
Sancties kunnen de focus van de school verplaatsen naar de targets en eisen van het accountability systeem. Wanneer deze targets betrekking hebben op opbrengsten zal daar de aandacht naar uitgaan. Sunderman (2009) legt uit dat sancties vooral effectief zijn wanneer ze gewenst gedrag (opbrengstgericht werken en hoge opbrengsten) uitlokken, wanneer ze onbehagen veroorzaken bij de relevante actoren, en wanneer ze daadwerkelijk kunnen en worden opgelegd. Sancties zijn niet effectief wanneer ze gekoppeld zijn aan ambigue, onzekere en moeilijk haalbare uitkomsten (bijvoorbeeld onrealistisch hoge opbrengsten voor scholen met een moeilijke leerlingpopulatie), en/of wanneer ze ongeloofwaardig zijn en gericht op diffuse actoren die weinig of geen invloed hebben op de uitkomsten.

6. Informatiefunctie

Een laatste functie van het meten van toegevoegde waarde is het informeren van stakeholders, zoals ouders, schoolbesturen en vervolgonderwijs, over de kwaliteit van een school. Van deze stakeholders wordt verwacht dat zij deze informatie gebruiken in de communicatie richting de school. Ouders zouden het inspectierapport bijvoorbeeld moeten gebruiken in hun schoolkeuze, terwijl schoolbesturen beleid ontwikkelen voor de verbetering van hun scholen. We bespreken hieronder de volgende vragen om meer zicht te krijgen op de mate waarin toegevoegde waarde deze informatiefunctie kan hebben:

- In welke mate maken ouders en andere stakeholders (schoolbesturen, vervolg-onderwijs) gebruik van informatie over de toegevoegde waarde van een school?
- Hoe maken zij gebruik van deze informatie?
- Waar moet een maat voor TW aan voldoen om bruikbaar te zijn voor stakeholders?
- Hoe leidt het informeren van stakeholders over de toegevoegde waarde van een school tot schoolverbetering?

6.1 Toegevoegde waarde en schoolkeuze van ouders

Onderzoek naar het gebruik van opbrengst-informatie door stakeholders laat vooral zien hoe ouders opbrengstgegevens van scholen (al dan niet) gebruiken in hun schoolkeuze ('choice'), om scholen aan te spreken op verbetering ('voice') en om hun kinderen op een beter presterende school te plaatsen ('exit'). Deze drie processen zouden tot verbetering van scholen en tot hoge prestaties van leerlingen moeten leiden (Karsten et al, 2010), bijvoorbeeld doordat ouders een school kiezen die goed past bij de behoeften van hun kinderen ('choice'), of doordat 'voice' en 'exit' de school aanzetten tot verbetering (OECD, 2008; Ladd en Walsh, 2002).

Choice en exit door/van ouders

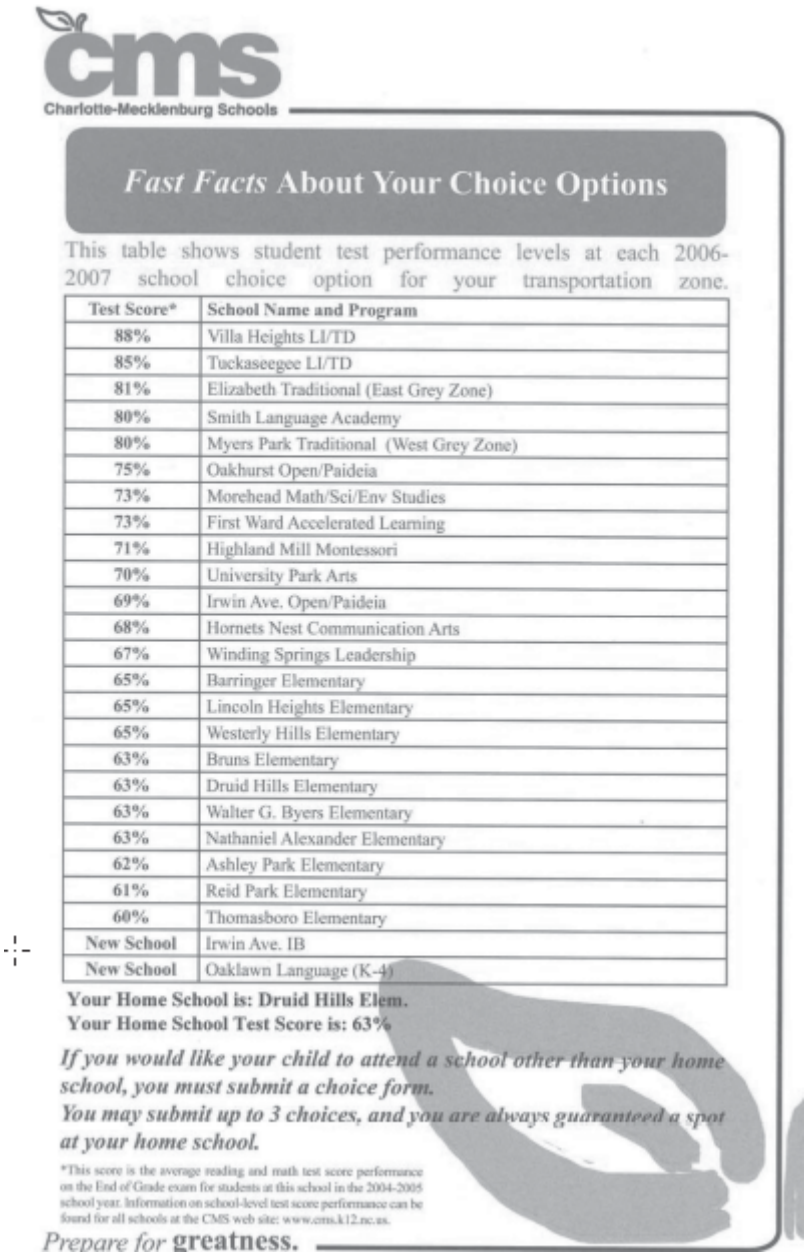
Met name naar choice en exit is veel onderzoek gedaan waarbij bijvoorbeeld is nagegaan hoe ouders uit verschillende sociale milieus ranglijsten van scholen en school-rapportkaarten (welke overigens niet altijd zijn gebaseerd op toegevoegde waarde van scholen, maar ook op ruwe data) gebruiken in hun schoolkeuze. Dit onderzoek levert een tegenstrijdig beeld op. Dronkers (1999) en Dronkers en Veenstra (2001) laten zien dat ranglijsten van scholen, zoals gepubliceerd in Trouw, maar weinig effect hebben op schoolkeuze door ouders; van alle waargenomen verschillen in het groeipercentage van scholen kan slechts 2% van de variantie verklaard worden door de publicatie van kwaliteitsgegevens in Trouw (zie ook andere studies die deze conclusies onderbouwen, Cullen, Jacob en Levitt, 2006; Cullen en Jacob, 2007). Ouders wegen andere argumenten mee, zoals nabijheid school, pedagogisch klimaat, reputatie van de school, aandacht voor sociale vaardigheden, leerlingbegeleiding, anti-pestbeleid, kunstklassen, sport, aansluiting op thuis (sfeer, opvoeding, levensbeschouwing) en schoolgrootte (Bosker en Scheerens, 2000; OECD, 2008; De Moor, 2009). Vooral ouders uit de middenklasse baseren hun schoolkeus op ranglijsten en kwaliteitskaarten, volgens Karsten e.a. (2010). De leesbaarheid van de informatie en de mate waarin deze aansluit bij hun informatiebehoefte speelt daarin een belangrijke rol.

Hastings en Weinstein (2008) vonden daarentegen (in een natuurlijk en veldexperiment) dat het verstrekken van toegankelijke informatie over prestaties van scholen in rapport-kaarten (zie hun voorbeeld in box 5) ook ouders met lage inkomens er toe kan aanzetten om een hoog presterende school te kiezen, met name wanneer zij meerdere hoog presterende scholen in de buurt hebben. Ook Koning en Van der Wiel (2010) vonden dat ouders en leerlingen uit verschillende sociale milieus de Trouw ranglijsten meewegen, vooral wanneer zij voor een VWO-school kiezen. De afstand tot huis en de mate waarin de school ook inzet op brede ontwikkeling van leerlingen speelde daarbij echter ook een rol.

Volgens Bell (2005) hebben ouders uit verschillende sociale milieus verschillende 'keuze-sets' op basis waarvan zij een school kiezen. Deze keuze-sets zijn het resultaat van een complexe interactie van hun beeld van academische capaciteiten van hun kinderen, hun eigen achtergrond en hun voorkeuren. Ook Cohen e.a. (2012) tonen aan dat ouders van leerlingen met verschillende adviezen voor VMBO/VWO andere keuzes maken en zich verschillend over schoolkeuze laten informeren. In ieder geval is voor alle groepen ouders het aanbod aan scholen in de buurt een belangrijke bepalende factor in de schoolkeuze van ouders (De Moor, 2009), waarbij ouders met een hogere opleidingsachtergrond

circa 200 meter verder willen reizen voor een school met een hogere Cito-score, en om een zwakke school te vermijden (Borghans, e.a.)
Bell (2005) toont tot slot aan aan dat het merendeel van de ouders (97%) hun kinderen niet van een laag presterende school afhaalt.

Box. Voorbeeld presentatie rapport kaarten aan ouders in Charlotte Mecklenburg
(zie Hastings en Weinstein, 2008)



cms
Charlotte-Mecklenburg Schools

Fast Facts About Your Choice Options

This table shows student test performance levels at each 2006-2007 school choice option for your transportation zone.

Test Score*	School Name and Program
88%	Villa Heights LI/TD
85%	Tuckaseegee LI/TD
81%	Elizabeth Traditional (East Grey Zone)
80%	Smith Language Academy
80%	Myers Park Traditional (West Grey Zone)
75%	Oakhurst Open/Paideia
73%	Morehead Math/Sci/Env Studies
73%	First Ward Accelerated Learning
71%	Highland Mill Montessori
70%	University Park Arts
69%	Irwin Ave. Open/Paideia
68%	Hornets Nest Communication Arts
67%	Winding Springs Leadership
65%	Barringer Elementary
65%	Lincoln Heights Elementary
65%	Westerly Hills Elementary
63%	Bruns Elementary
63%	Druid Hills Elementary
63%	Walter G. Byers Elementary
63%	Nathaniel Alexander Elementary
62%	Ashley Park Elementary
61%	Reid Park Elementary
60%	Thomasboro Elementary
New School	Irwin Ave. IB
New School	Oaklawn Language (K-4)

Your Home School is: Druid Hills Elem.
Your Home School Test Score is: 63%

If you would like your child to attend a school other than your home school, you must submit a choice form.
You may submit up to 3 choices, and you are always guaranteed a spot at your home school.

*This score is the average reading and math test score performance on the End of Grade exam for students at this school in the 2004-2005 school year. Information on school-level test score performance can be found for all schools at the CMS web site: www.cms.k12.nc.us.

Prepare for **greatness.**

Choice/schoolkeuze en verbetering van leerprestaties
Ook het onderzoek naar de relatie tussen enerzijds de schoolkeuze van ouders op basis van schoolpresaties, en anderzijds leerprestaties van leerlingen laat tegenstrijdige resultaten zien. Onduidelijk is of positieve relaties het gevolg zijn van schoolverbetering wanneer scholen onder druk staan van ‘dreigende’ daling in leerlingaantallen, of door een betere ‘fit’ van leerling en school wanneer ouders een school kiezen die het best past bij de kwaliteit en capaciteiten van hun kind. CPB-onderzoek laat een toename zien in prestaties van VO-scholen die laag in de Trouw-ranglijsten

presteren. Kinderen van laag opgeleide ouders die bewust kiezen voor een hoog presterende school, presteerden volgens Hastings en Weinstein (2008) ook beter dan de kinderen van ouders die geen geïnformeerde schoolkeuze hadden gemaakt. Verschillende auteurs (Smith, 1995; Koretz, 2003; Koning en Van der Wiel, 2010) wijzen echter op strategisch gedrag van scholen (bijvoorbeeld teaching to the test, uitsluiten van leerlingen) waardoor toetsscores verbeteren, terwijl prestaties van leerlingen gelijk blijven. Dergelijk strategisch gedrag kan de toename in prestaties in de ranglijsten ook verklaren.

Er is weinig informatie over het gebruik van opbrengst-gegevens door andere stakeholders (bijvoorbeeld vervolgonderwijs), of over de mate waarin ouders laag presterende scholen op verbetering aanspreken. De OECD (2008) verwacht dat stakeholders informatie over de toegevoegde waarde van een school gebruiken wanneer zij in staat zijn om deze informatie te interpreteren.

Referenties

Allen, R., & Burgess, S. (2013). Evaluating the provision of school performance information for school choice. *Economics of Education Review*.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational researcher*, *37*(2), 65-75.

Bambrick-Santoyo, Paul. "Data in the Driver's Seat." *Education Leadership* (2007): 43-46.

Berry, B., Turchi, L., Johnson, D., Hare, D., Duncan Owens, D. (2003). *The Impact of High-Stakes Accountability on Teachers' Professional Development: Evidence from the South.* A Final report

Bhola, D.S., Impara, J.C., Buckendahl, C.W. (2003). Aligning Tests with States' Content Standards: Methods and Issues. *Educational Measurement: Issues and Practice, ?,*21-29.

Borghans, L., Golsteyn, B.H.H., en Zölitz, U. (?). Parental Preferences for Primary School Characteristics.

Bosker, R. J., & Scheerens, J. (2000). Publishing school performance data. *European Education*, *32*(3), 12-30.

Bosker, R. J., & Jong-Heeringa, J. D. (2006). Leeropbrengsten van scholen. *Order*, *501*, 318.

Booher-Jennings, J. (2005). Below the Bubble: 'Educational Triage' and the Texas Accountability System. *American Educational Research Journal,* 42(2), 231-268.

Braun, H. (2004). Reconsidering the Impact of High-stakes Testing. *Education Policy Analysis Archives, 12(1), 1-43.*

Campbell, Carol, and Ben Levin. "Using data to support educational improvement." *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)* 21.1 (2009): 47-65.

Carlson, Deven, Geoffrey D. Borman, and Michelle Robinson. "A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement." *Educational Evaluation and Policy Analysis* 33.3 (2011): 378-398.

Chester, M.D. (2005). Introduction to the Special issue: Test Scores and State Accountability; Measuring the Impact of State Accountability Programs. *Educational Measurement: Issues and Practice, ?,* 3-4.

Chorny, V. and Webbink, D. (2010). The effect of accountability policies in primary education in Amsterdam. *CPB discussion paper, number 144.*

Cohen, L., de Jong, I., Jakobs, E., en Slot, J. (2012). *Het schoolkeuzeproces door de ogen van Amsterdamse ouders.* Gemeente Amsterdam: Bureau Onderzoek en Statistiek

Council of Chief State School Officers (CCSSO) (2004). A Framework for Examining Validity in State Accountability systems. Washington: CCSSO.

Courtney Bell (Oct. 2005). *All Choices Created Equal? How Good Parents Select "Failing" Schools*. National Center for the Study of Privatization in Education, Columbia University

Carnoy, M., Loeb, S. and Smith, T.L. (2001). Do Higher State Test Scores in Texas Make for Better High School Outcomes? *CPRE Research Report Series, RR-047.*

Carnoy, M. and Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Ecucational Evaluation and Policy Analysis,* 24(4), 305-331.

Cullen, J.B. and Reback, R. (2006). Tinkering toward accolades: school gaming under a performance accountability system. Cambridge: National Bureau of Economic Research. Working paper 12286. http://www.nber.org/papers/w12286.

Darling-Hammond, L. (2004). Standards, Accountability, and School Reform. *Teachers College Record, 106(6),* 1047-1085.

Datnow, Amanda, Vicki Park, and Priscilla Wohlstetter. "Achieving with data." *Los Angeles: University of Southern California, Center on Educational Governance* (2007).

Dijkstra, A., Dronkers, J., Karsten, S. (2004). *Private Schools as Public Provision for Education: School Choice and Market Forces in the Netherlands.* In: P.J. WOLF (ed), Educating Citizens. International Perspectives on Civic Values and School Choice, edited by P. J. Wolf [amp] S. Macedo, Washington DC : Brookings Institute, 2004, 67-90

Dijkstra, A.B. (2012). *Sociale opbrengsten van onderwijs. Rede in verkorte vorm*

uitgesproken bij de aanvaarding van het ambt van bijzonder hooleraar Toezicht & Socialisatie, scholen en onderwijsbestel vanwege de Inspectie van het Onderwijs (2012, June 14). Amsterdam: Vossiuspers UvA.

Downes, D. en Vindurampulle, O. (2007). *Value–added measures for school improvement.* Education Policy and Research Division Office for Education Policy and Innovation Department of Education and Early Childhood Development Melbourne: paper number 13.

Earl, L. (2005). From accounting to accountability: harnessing data for school improvement. Research conference 2005.

Ehren, M.C.M., Schildkamp, K., and Gelderblom, G. (in prep). High stakes testing and teachers' use of data.

Ehren, M.C.M. and Hatch, T. (submitted). Responses of schools to accountability systems using multiple measures; The case of New York City elementary schools. *Educational Assessment, Evaluation and Accountability*

Ehren, M.C.M. (2011). Risicogestuurd toezicht na wijziging van de Wet op het Onderwijstoezicht; een kritische reflectie. http://www.owinsp.nl/actueel/publicaties/Risicogestuurd+toezicht+na+wijziging+van+de+Wet+op+het+Onderwijstoezicht.html

Ehren, M.C.M. and Swanborn, M. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement 23(2),* 257-280.

Figlio, D.N. and Getzler, L.S. (2002). Accountability, ability and disability: gaming the system. *NBER working paper 9307.* http://www.nber.org/papers/w9307.

Figlio, D.N. And Lucas, M.E. (2004). Do high grading standards affect student performance? *Journal of Public Economics, 88,* 1815-1834.

Fitz-Gibbon, C.T. (1997). *The Value Added National project; final report Feasability Studies for a National System of Value-added indicators.* University of Durham: CEM

Fitz-Gibbon, C.T. & Tymms, P. (2002, January 16).Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives, 10(6).*

Gribben, M.A., Campbell, H.L. and Mathew, J. (2008). Are Advanced Students Advancing? Examining Achievement Trends Beyond Proficiency. *Paper presented at AERA 2008.*

Glass, G.V. (1972). The Many Faces of 'Educational Accountability'. *The Phi Delta Kappan*, 53(10), 636-639.

Goertz, M.E. and Duffy, M.C. (2001). Assessment and Accountability Across the 50 States. Policy brief RB-33. http://www.cpre.org/images/stories/cpre_pdfs/rb33.pdf

Gong, B., & Hill, R. (2001, March). *Some considerations* of *multiple measures in assessment and school accountability.* Presentation at the Seminar on Using Multiple Measures and Indicators to Judge Schools' Adequate Yearly Progress Under Title I (sponsored by CCSSO & US DOE), Washington, DC.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education.* Rand Corporation.

Hanushek, E.A. and Raymond, M.E. (2001). The Confusing World of Educational Accountability. *National Tax Journal,* 54(2), 365-384.

Hanushek, E.A and Raymond, M.E. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management, 24(2),* 297-327.

Hanushek, E.A. and Raymond, M.E. (2002). Lessons about the Design of State Accountability Systems. *Paper prepared for 'Taking Account of Accountabiliy: Assessing Policy and Politics', Harvard University.*

Hanushek, E.A., Raymond, M.E. (2004). Does School accountability lead to improved Student performance? *NBER Working Paper No. W10591*

Harlen, W., Gipps, C., Broadfoot, P., & Nuttall, D. (1992). Assessment and the improvement of education∗. *The Curriculum Journal*, *3*(3), 215-230.

Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, *123*(4), 1373-1414.

Heubert, J.P. and Hauser, R.M. (Eds.) (1999). High Stakes: Testing for Tracking, Promotion, and Graduation. Washington: National Academy Press.

Hill, R., Gong, B., Marion, S., DePascal, C., Dunn, J., & Simpson, M. (2005, November). Using value tables to explicitly value student growth. In *Conference on Longitudinal Modeling of Student Achievement. Dover, NH: The Center for Assessment*.

Hulpia, H. & Valcke, M. (2004): The Use of Performance Indicators in a School Improvement Policy: The Theoretical and Empirical Context, Evaluation & Research in Education, 18:1-2, 102-119

Inspectie voor de Gezondheidszorg (2004). *The Result Matters; Performance Indicators as independent measure of the quality of hospital care.* Den Haag: IGZ

Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics,* 89(5-6), 761-796.

Jacob, B.A. and Levitt, S.D. (2003). Rotten Apples: an investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics (august)*, 843-877.

Kane, T.J. and Staiger, D.O. (2001a). The promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, 16(4), 91-114.

Kane, T.J. and Staiger, D.O. (2001). Improving school accountability measures. NBER working paper series. http://www.nber.org/papers/w8156.

Kay Livingston & Jim McCall (2005): Evaluation: judgemental or developmental?, European Journal of Teacher Education, 28:2, 165-178

Karsten, S., Visscher, A.J., Dijkstra, A. en Veenstra, R. (2010). Towards standards for the publication of performance indicators in the public sector: the case of schools. *Public Administration*, *88*(1), 90-112.

Karsten, S., Roeleveld, J., Ledoux, G., Felix, C., Elshof., D. (2002). Schoolkeuze in een multi-etnische samenleving. SCO Kohnstamm Instituut, rapport 642

Koning, P. W. C., and K. M. Van Der Wiel. "Onderwijs-Kwaliteitsinformatie middelbare scholen maakt verschil." *Economisch Statistische Berichten* 95.4585 (2010): 294.

Koning, P., & Van der Wiel, K. (2010). *Ranking the schools: How quality information affects school choice in the Netherlands*. CPB.

Koning, P., & Van der Wiel, K. (2012). School responsiveness to quality rankings: An empirical analysis of secondary education in the Netherlands. *De Economist*, *160*(4), 339-355.

Koretz, D. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human resources: education, manpower and welfare,* 37(4), 752-777

Koretz, D.M. (2003). Using Multiple Measures to Address Perverse Incentives and Score Inflation. *Educational Measurement, 22(2),* 18-26.

Klein, S.P., Hamilton, L.S., McCaffrey, D.F. and Stecher, B.M. (2000). What do Test Scores in Texas Tell Us? *Education Policy Analysis Archives,* 8(49). http://epaa.asu.edu/epaa/v8n49/

Koretz, D.M., McCaffrey, D.F. and Hamilton, L.S. (2001). Towards a Framework for Validating Gains under High-Stakes Conditions. CRESST/Harvard Graduate School of Education: CSE Technical Report 551

Koretz, D.M. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human Resources (37(4),* 752-777.

Koretz, D.M. (2003). Using Multiple Measures to Address Perverse Incentives and Score Inflation. *Educational Measurement, 22(2),* 18-26.

Koretz, D.M. and Hamilton, L.S. (2003). *Teachers' Responses to High-Stakes Testing and the Validity of Gains: A Pilot Study*. CSE Report 610. Los Angeles/UCLA: Center for the Study of Evaluation. http://research.cse.ucla.edu/reports/R610.pdf

Ladd, H.F. (2007). Holding Schools Accountable Revisited. 2007 Spencer Foundation Lecture in Education Policy and Management.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education review*, *21*(1), 1-17.

Lane, S., Parke, C.S., Stone, C.A. (1998). A Framework for Evaluating the Consequences of Assessment Programs. *Educational Measurement: Issues and Practice,* 17(2), 24-28.

Lane, S. and Stone, C.A. (2002). Strategies for Examining the Consequences of Assessment and Accountability Programs. *Measurement: Issues and Practice, 21(1), 23-30.*

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *The Journal of Economic Perspectives*, *24*(3), 167-181.

Leithwood, K. and Earl, L. (2000). *Educational Accountability Effects: An International Perspective, 75(4),* 1-18.
Lee, J. (2007). Revisiting the impact of high-stakes testing on student outcomes from an international perspective. In L. Deretchin & C. Craig (Eds.), *International research on the impact of accountability systems* (pp. 65–82). Lanham, MD: Rowman & Littlefield.
Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems* (CSE Technical Report 539). Los Angeles: Center for the Study of Evaluation.
Linn, R.L. (2000). Assessments and Accountability. *Educational Researcher, 29(4),* 4-16.
Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255-270.
Mansell, W., and M. James. "Assessment Reform Group (2009) Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme. London, Economic and Social Research Council." *Teaching and Learning Research Programme*.
Meinhard, R., & Buckstein, S. (2001). *Choice Thinking*. Cascade Policy Institute
McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, *29*(1), 67-101.
McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability. Monograph*. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.
McCaffrey, D. F., & Hamilton, L. S. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project* (Vol. 506). Rand Corporation.
McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.
Means, B., Padilla, C., DeBarger, A., & Bakia, M. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. US Department of Education, Office of Planning, Evaluation and Policy Development.
Moe, T.M. (2002). The structure of school choice. *Hoover Digest*, no. 4 fall issue.
Mooij, T., Roeleveld, J., Fettelaar, D., & Ledoux, G. (2012). Kwaliteitsbeoordeling van scholen primair onderwijs: Het correctiemodel van de inspectie vergeleken met alternatieve modellen.
Moor, de A. (2009). Concurrentie en kwaliteit in het primair en voortgezet onderwijs. *TPEdigitaal, 3* 65-79.
Mortimore, P. , Sammons, P. & Thomas, s. (1994): School Effectiveness and Value Added Measures, Assessment in Education: Principles, Policy & Practice, 1:3, 315-332
Nichols, S.L. and Glass, G.V. and Berliner, D.C. (2006). High-stakes testing and student achievement: does accountability pressure increase student learning? *Education Policy Analysis Archives, 14(1),* Retrieved 14 November 2008 from http://epaa.asu.edu/epaa/v14n1.
Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Tempe, AZ: Education Policy Studies Laboratory, Arizona State University.
OECD (2008). Measuring Improvements; Best practices to assess value added of schools. Paris: OECD
Onderwijsraad (2003). Wat scholen toevoegen. Den Haag: Onderwijsraad
Plank, D.N. and Sykes, G. (Eds) (2003). Choosing Choice: School Choice In International Perspective Teachers College Press.
Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8-16.
Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics,* 92(5), 1394-1415.
Rivkin, S. G. (2007, November). *Value-added analysis and education policy.* Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf

Roeleveld, J., Mooij, T., Fettelaar, D. A. A. N., & Ledoux, G. (2011). *Correctiefactoren bij opbrengstmaten in het primair onderwijs: onderzoek ten behoeve van de Inspectie van het Onderwijs* (No. 868). Kohnstamm Instituut.

Roeleveld, J. (2003). Herkomstkenmerken en begintoets. *Secundaire analyses op het PRIMA-cohortonderzoek.[Background features and entry test]. Amsterdam: SCO-Kohnstamm Instituut*.

Sheldon, K.M. and Biddle, B.J. (1998). Standards, Accountability, and School Reform: Perils and Pitfalls. *Teachers College Record, 100(1),* 164-180.

Smith, P. (1995) On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, 18(2-3), 277-310.

Stevens, J., Estrada, S. and Parkes, J. (2000). Measurement Issues in the Design of State Accountability Systems. Paper presented at AERA.

Scheerens, J., & Hendriks, M. (2004). Benchmarking the quality of education. *European Educational Research Journal*, *3*(1), 101-114.

Stecher, B.M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices). Tests and their use in test-based accountability systems. In Hamilton, L.S., Stecher, B.M., Klein, S.P. (Eds.). *Making sense of Test-based Accountability in Education*. Santa Monica: Rand cooperation. http://www.rand.org/pubs/monograph_reports/MR1554/

Supovitz, Jonathon A., and Valerie Klein. "Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement." *Philadelphia: Consortium for Policy Research in Education* (2003).

Symonds, K. W. (2003). *After the test: How schools are using data to close the achievement gap*. Bay Area School Reform Collaborative.

Timmermans, A. (2012). Value added in educational accountability: Possible, fair and useful? Groningen: GION

U.S. Department of Education (2010). Use of Education Data at the Local Level; From Accountability to Instructional Improvement. U.S. Department of Education, Office of Planning, Evaluation, and Policy Development

Velden, van der, R. The proof of the pudding is in the eating. http://arno.unimaas.nl/show.cgi?fid=17166

Ver Eecke, E. (2004). Leerwinst als kwaliteitsindicator: een haalbare kaart of een brug te ver?. *Impuls*, *34*(3), 149-163.

Vereniging van Leraren in Levende Talen (2013). *Effecten van sturing op discrepantie tussen de cijfers van het centraal examen. Onderzoek naar de sturing door schoolleiders en de effecten daarvan op het taalonderwijs op havo en vwo en het schoolexamen bij de talen.* Utrecht: Vereniging van Leraren in Levende Talen

Verhaeghe, J. P. (2010). *Schoolfeedback als input voor interne kwaliteitszorg*. In: Handboek Reflectief Vermogen

Visscher, A. J. and Coe, R. (2003) 'School performance feedback systems : conceptualisation, analysis, and reflection.', School effectiveness and school improvement., 14 (3). pp. 321-349.

Visscher, A.J. en Ehren, M.C.M. (2011). De eenvoud en complexiteit van Opbrengstgericht Werken; Analyse in opdracht van de Kenniskamer van het Ministerie van Onderwijs, Cultuur en Wetenschap. http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/07/13/de-eenvoud-en-complexiteit-van-opbrengstgericht-werken.html

Walberg, Herbert J. (2007). School choice. The findings**.** Cato Institute

Wayman, J. C., Cho, V., & Johnston, M. T. (2007). The data-informed district: A district-wide evaluation of data use in the Natrona County School District.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). De toegevoegde waarde van de basisschool. *Verkenning van de mogelijkheden de schoolspecifieke bijdrage aan de onderwijsopbrengst in kaart te brengen met behulp van het Cito Leerlingvolgsysteem en de Eindtoets Basisonderwijs.* Arnhem: Citogroep.

Wolf, de, Inge F. and Janssens, Frans J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education*, 33(3),p. 379 — 396. http://www1.fee.uva.nl/scholar/wp/wp53-05.pdf

Woodworth, J. L., Lo, W. J., McGee, J. B., & Jensen, N. C.. (2012). The Impact of Selection of

Student Achievement Measurement Instrument on Teacher Value-added Measures. Retreived from www.invalsi.it [24 May 2013).
Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of education*, *25*(4), 469-483.
Zenisky, A. L. and Hambleton, R. K. (2012), Developing Test Score Reports That Work: The Process and Best Practices for Effective Communication. Educational Measurement: Issues and Practice, 31: 21–26. doi: 10.1111/j.1745-3992.2012.00231.x

Bijlage 1. Modellen pilot eerwinst en toegevoegde waarde

Source: Frans Janssens, presentatie expertmeeting toegevoegde waarde (17 mei 2013)

De UT ontwikkelt twee LW-modellen, nl het zgn Groeitempo-model en het Zomervakantie-model. In beide gevallen wordt de LW vergeleken met de gemiddelde groei van de leerlingen die bij de beginmeting hetzelfde vaardigheidsniveau hadden. Bij het vergelijken kan vastgesteld worden of de leerling zich gemiddeld heeft ontwikkeld voor zijn/haar niveau, meer of minder gegroeid is dan gemiddeld. Het verschil tussen beide modellen zit in de afname-momenten. Bij het Groeitempo-model gaat het om de leerwinst tussen twee of meer regulieren afname-momenten van het toetsen uit het Cito-LOVS en bij het zomervakantie-model gaat het om groei tijdens de zomervakantie, gemeten aan een toets die vlak voor en vlak na de zomervakantie is afgenomen.

Het Cito ontwikkelt het Relatieve Leerwinstmodel waarbij de groei per leerling wordt weergeven met een z- score. De z-score geeft aan hoe de vaardigheidsgroei van een leerling zich verhoudt tot de gemiddelde vaardigheidsgroei van een landelijke vergelijkingsgroep van leerlingen die bij de start hetzelfde vaardigheidsniveau hadden.

Het GION ontwikkelt twee TW-modellen: 1) het vaardigheidsverschil-model en 2) vaardigheidsgroei-model. De twee methoden verschillen van elkaar op enkele belangrijke punten. Het vaardigheidsverschil-model maakt gebruik van slechts twee vaardigheidsscores per leerling per rapportageperiode. Van elke leerling moet zowel een begin- (bijvoorbeeld M3) als eindmeting (bijvoorbeeld E7) van de vaardigheid beschikbaar zijn. Hiertussen wordt dan het verschil berekend; de individuele leerwinst of leervordering gedurende een bepaalde periode. In dit opzicht komt het vaardigheidsverschil-model overeen met het groeitempo-model van de UT en het Z-score benadering van het Cito. Bij berekeningen van de vaardigheidsgroei-modellen gebeurt nog iets extra's; alle beschikbare vaardigheidsscores in een rapportageperiode worden meegenomen. Dus niet alleen de begin- (M3) en eindmeting (E7) van een leerling maar ook de E3, M4, E4, M5, E5, M6 en E6 metingen, als die beschikbaar zijn.