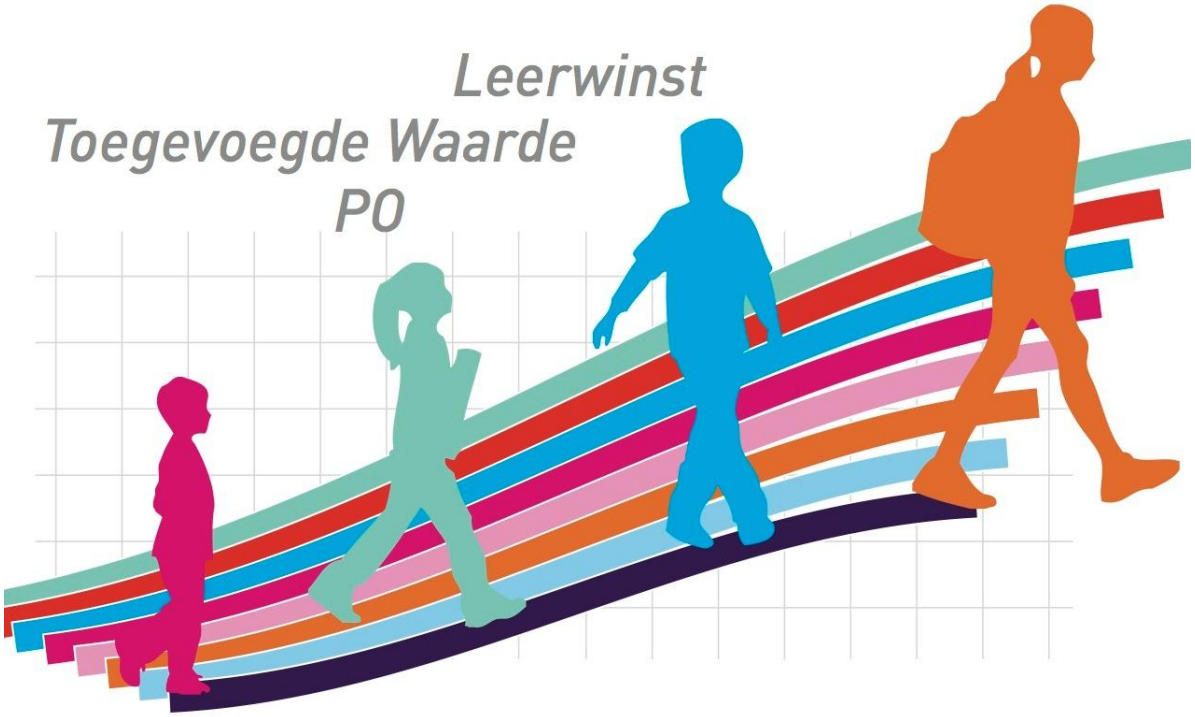
Leerwinst en toegevoegde waarde in het primair onderwijs



Eindrapportage

Frans J.G. Janssens

Lyset Rekers-Mombarg

Ellen Lacor

Ministerie van Onderwijs, Cultuur en Wetenschap

Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

CED groep
educatieve diensten

UNIVERSITEIT TWENTE.

rijksuniversiteit groningen
gion, gronings instituut voor onderzoek van onderwijs

Cito zeker weten

Leerwinst en toegevoegde waarde in het primair onderwijs

Eindrapportage

Datum: Januari 2014
Kenmerk: U/O&I/14 002
Afdeling: CED-Groep, Onderzoek & Innovatie
Auteurs: Frans J.G. Janssens, Lyset Rekers-Mombarg, Ellen Lacor
Project: LTW-PO
Versie: Definitief

Met bijdragen van:
Remco Feskens (Cito)
Jos Keuning (Cito)
Renske de Leeuw (Universiteit Twente)
Hans Luyten (Universiteit Twente)
Ilse Papenburg (Cito)
Carla Versteeg (CED-Groep)
Johan Wesseling (CED-Groep)

Inhoud

Samenvatting

In het actieplan 'Basis voor Presteren' heeft de toenmalige minister van OCW aangekondigd in het schooljaar 2011-2012 voor het primair onderwijs te starten met een pilot Leerwinst en Toegevoegde Waarde. Hierin worden de verschillende aspecten en werkwijzen voor het bepalen van leerwinst van leerlingen en toegevoegde waarde van scholen uitgewerkt en beproefd. De pilot bestaat uit een samenwerking tussen onderwijspraktijk, wetenschap, de Inspectie van het Onderwijs en het Ministerie van OCW.

De belangstelling voor leerwinst en toegevoegde waarde komt mede voort uit de nodige onvrede over de wijze waarop de leerprestaties van basisscholen worden betrokken in de beoordeling van de kwaliteit van het onderwijs (zie bijv. De Wolf, 2012). De gangbare praktijk is dat vooral eindtoetsen, afgenomen in groep 8, de basis vormen voor het beoordelen van de opbrengsten. Scholen met veel zorgleerlingen en met veel leerlingen met taal- en ontwikkelingsachterstanden voelen zich benadeeld. Scholen met een meer bevoorrechte populatie scoren vaker hoger dan gelijkwaardige scholen met een minder bevoorrechte schoolbevolking (Raudenbusch, 2004; Koretz, 2008; Rothstein, 2009).

Definities

Leerwinst en toegevoegde waarde zijn begrippen die in het onderwijs geen eenduidige betekenis hebben. *Leerwinst* wordt in voorliggend rapport gedefinieerd als de vergelijking van twee toetsprestaties op twee verschillende momenten waarbij een verschil gezien wordt als positieve of negatieve 'leerwinst'. De leerwinst kan bijvoorbeeld elk leerjaar gemeten worden of alleen door middel van de afname van een begin- en een eindtoets.

Onder *toegevoegde waarde* wordt verstaan de bijdrage die de school levert aan de leerwinst van alle leerlingen. De invloed van de school wordt statistisch berekend door rekening te houden met de verschillende factoren die de leerwinst kunnen hebben beïnvloed, maar die buiten de invloedssfeer van de school liggen. Zo'n factor is bijvoorbeeld de samenstelling van de schoolbevolking: etniciteit, geslacht, socio-economische status en type zorgleerlingen. Door de leerwinst te corrigeren voor niet-schoolse factoren wordt het effect van de school zichtbaar. Maar omdat we niet weten welke specifieke schoolfactoren dit effect hebben veroorzaakt, is de toegevoegde waarde te beschouwen als een indicatie van de invloed die scholen hebben op de prestaties van hun leerlingen.

Er zijn twee doelen die met leerwinst en toegevoegde waarde worden nagestreefd. Dat is in de eerste plaats dat ze een rol gaan spelen in het opbrengstgericht werken. Dit doel wordt in het voorliggende rapport omschreven als leerwinst en toegevoegde waarde in *schoolverbeteringsperspectief.*
Hieronder wordt verstaan dat leerwinst en toegevoegde waarde gebruikt kunnen worden voor zowel de afstemming van het onderwijs op de leerlingen als voor het beoordelen van de leerprestaties door de school zelf in het kader van de kwaliteitszorg.
Het tweede doel van leerwinst en toegevoegde waarde is dat deze betrokken kunnen worden in de verantwoording over de leerprestaties door de scholen aan de ouders en aan de Inspectie van het Onderwijs. De inspectie op haar beurt kan leerwinst en toegevoegde waarde betrekken in de

beoordeling van de leerprestaties van scholen. Omdat hier rekenschap en openbaarmaking een belangrijke rol spelen, wordt dit doel in dit rapport omschreven als leerwinst en toegevoegde waarde in *accountabilityperspectief*.

Conclusies

Uit de pilot is gebleken dat, vanuit het schoolverbeteringsperspectief gezien, maten voor leerwinst en toegevoegde waarde een waardevolle bijdrage leveren aan de verdere ontwikkeling van het opbrengstgericht werken in scholen en aan de beoordeling van hun eigen leerprestaties. De groei van de prestaties van alle leerlingen is nauwkeuriger te bepalen en te beoordelen. Ook kunnen scholen hun leerwinst vergelijken met die van andere scholen, waardoor een nieuw referentiepunt beschikbaar komt op basis waarvan binnen het systeem voor kwaliteitszorg conclusies getrokken kunnen worden over de organisatie en inrichting van het onderwijs.

Leerwinst

Gebleken is dat de bestaande toets- en schoolinformatiesystemen die op scholen in gebruik zijn, geschikt gemaakt kunnen worden om de leerwinst op leerling-, groeps- en schoolniveau te bepalen. Het bleek ook mogelijk de gemiddelde leerwinst van een school te vergelijken met de gemiddelde leerwinst van andere scholen. Daarvoor moeten wel een aantal technische maatregelen getroffen worden door de aanbieders van toets- en schoolinformatiesystemen, zoals het beschikbaar stellen van normgegevens en de mogelijkheid om op flexibele wijze cohorten samen te stellen waarvoor de leerwinst berekend kan worden. Ook moeten op schoolniveau afspraken gemaakt worden over het gebruik van toetsen en de invoer van toetsgegevens om tot bruikbare leerlingbestanden te komen op basis waarvan de leerwinst kan worden bepaald.

Toegevoegde waarde

De pilot heeft ook laten zien dat het mogelijk is om de toegevoegde waarde van scholen te schatten op basis van leerwinstgegevens, die gecorrigeerd worden voor relevante niet-schoolse factoren. De bepaling van toegevoegde waarde is echter nog te complex om deze in bestaande toets- en schoolinformatiesystemen toe te passen. In de pilot zijn twee modellen ontwikkeld die van elkaar verschillen in het aantal toetsmomenten dat in het model wordt opgenomen en daardoor ook in hun gebruiksmogelijkheden verschillen.

1. Het *vaardigheidsverschil-model* is gebaseerd op de vaardigheidsscores op twee LVS-toetsen uit hetzelfde toetsdomein die als begin- en eindmeting dienst doen. Met dit model zou de toegevoegde waarde geschat kunnen worden over een langere onderwijsperiode waarbij slechts gebruik wordt gemaakt van gegevens van twee toetsmomenten, bijvoorbeeld met als beginmeting groep 3 en als eindmeting groep 8. Technisch gezien is zo'n vergelijking mogelijk, maar de vraag is echter of dit tot een betekenisvolle interpretatie van toegevoegde waarde leidt. Dit geldt met name als dit model door scholen gebruikt wordt voor opbrengstgericht werken.

2. Het *vaardigheidsgroei-model* is gebaseerd op vaardigheidsscores van meer toetsmomenten uit hetzelfde toetsdomein. In het geval een school voor alle leerstofgebieden halfjaarlijks alle LVS-toetsen afneemt, kan op deze wijze voor elk leerstofgebied de toegevoegde waarde over nagenoeg alle tussenliggende leerjaren uit de hele basisschoolperiode worden bepaald.

Ofschoon de scholen de uitkomsten van beide modellen op hun waarde weten te schatten, gaat de voorkeur uit naar het vaardigheidsgroei-model. Dit model biedt zowel de mogelijkheid om terug te kijken naar de geleverde toegevoegde waarde, als vooruit te kijken naar de mogelijke ontwikkeling van de toegevoegde waarde. Het vaardigheidsgroei-model kan dus in dienst staan van zowel het schoolverbeterings- als het accountabilityperspectief.

Betere beoordeling van opbrengsten

Leerwinst en toegevoegde waarde zijn niet alleen een waardevolle aanvulling op de beoordeling van de opbrengsten door de scholen zelf, maar kunnen ook een bijdrage leveren aan een betere beoordeling van de opbrengsten van scholen door de Inspectie van het Onderwijs. Op dit moment geschiedt die beoordeling vooral op basis van gecorrigeerde schoolgemiddelden op eindtoetsen. Omdat deze vorm van beoordeling is gebaseerd op de leerprestaties van één groep leerlingen wordt deze in de literatuur ook wel statusmeting of 'test score snapshot' genoemd. Deze beoordelingsvorm heeft één belangrijke tekortkoming: er wordt geen rekening gehouden met de beginsituatie van leerlingen en ook niet met de inspanning van de school om leerlingen zo goed mogelijk te laten presteren. De ene school krijgt nu eenmaal leerlingen die een hoger initieel niveau hebben of van wie de ouders meer mogelijkheden hebben om de schoolloopbaan van de leerlingen te stimuleren dan de andere school. Dat zijn factoren waarop scholen geen invloed hebben.

Maten voor leerwinst bieden de mogelijkheid bij de beoordeling van de leerprestaties van scholen wel rekening te houden met de beginsituatie en met de prestatiegroei van hun leerlingen ongeacht hun beginsituatie, prestatieniveau of hun achtergronden. Ook voor hoogbegaafden, zorgleerlingen en leerlingen met taal- en ontwikkelingsachterstanden kan de leerwinst zichtbaar worden gemaakt.
Toegevoegde waarde-schattingen bieden vervolgens de mogelijkheid de bijdrage van de school aan die leerwinst in beeld te brengen door te corrigeren voor niet-schoolse factoren. Door hiervoor te corrigeren wordt het deel van de prestatie, dat aan de school en niet aan de kenmerken van de leerlingen kan worden toegeschreven, transparanter.

Hoe verder?

Leerwinstberekeningen kunnen door scholen zelf worden uitgevoerd als de toetspraktijk en de verschillende toets- en schoolinformatiesystemen daarop worden aangepast. De projectgroep adviseert de staatssecretaris van OCW daarover een aantal afspraken te maken met de scholen en met de aanbieders van toets- en schoolinformatiesystemen. Voor een betrouwbare bepaling van de leerwinst van een leerstofgebied is het van belang dat er toetsen worden gebruikt die gebaseerd zijn op een vaardigheidsschaal. Deze toetsen dienen op vaste momenten te worden afgenomen en de uitkomsten moeten op een systematische en nauwkeurige wijze in een

schoolinformatiesysteem te worden geregistreerd. De verschillende toets- en schoolinformatiesystemen die in het primair onderwijs worden gebruikt om toetsscores te registreren en te verwerken, dienen te beschikken over de technische mogelijkheden om leerwinst te kunnen berekenen. Toetsaanbieders moeten normtabellen beschikbaar stellen om de leerwinst op leerling-, groeps- en schoolniveau te beoordelen.
Aanbieders van toets- en schoolinformatiesystemen dienen aandacht te hebben voor de wijze waarop scholen leerwinst kunnen implementeren vanuit een schoolverbeteringsperspectief.

Voor het gebruik van leerwinst en toegevoegde waarde in een accountabilityperspectief geldt dat daaraan nog de nodige haken en ogen zitten. Dit betreft zowel technische als ethische kwesties. De projectgroep adviseert de staatssecretaris van OCW om op basis van de conclusies en de aanbevelingen van de pilot met de PO-Raad, de Inspectie van het Onderwijs en wetenschappers op korte termijn een plan te ontwikkelen:

1. voor de wijze waarop de toegevoegde waarde van scholen zo nauwkeurig mogelijk voor niet-schoolse factoren kan worden gecorrigeerd,
2. hoe deze gegevens teruggekoppeld kunnen worden aan de scholen,
3. op welke wijze leerwinst en toegevoegde waarde schattingen door scholen gebruikt kunnen worden om extern verantwoording af te leggen over hun opbrengsten en
4. op welke wijze de Inspectie van het Onderwijs leerwinst en toegevoegde waarde in het toezicht kan betrekken.

Daarbij is van belang om na te gaan welke instantie de noodzakelijke gegevens over toegevoegde waarde van scholen zou kunnen verzamelen en in beeld kan brengen. Gezien de complexiteit van de modellering van toegevoegde waarde ligt het voor de hand wetenschappelijke instituten te betrekken bij de verdere ontwikkeling van de toepassing van toegevoegde waarde-schattingen.

1. Inleiding

1.1. Aanleiding

Of kinderen voldoende op school leren is een vraag die vaak wordt gesteld. Voor de leerkracht is het antwoord op deze vraag van belang om instructie en verwerking op de leerlingen af te stemmen. Voor de ouders is het belangrijk te weten of de schoolvorderingen van hun kinderen naar wens verlopen. Voor het schoolbestuur en de inspectie is het antwoord van belang om de opbrengsten van de school te kunnen beoordelen. Maar hoe voor de hand liggend deze vraag ook is, het is lastig om daarop een duidelijk antwoord te geven. Het meten van groei in schoolprestaties, de leerwinst, is nu eenmaal minder eenvoudig dan het meten van bijvoorbeeld de lichaamsgroei. De kwestie is dan ook hoe leerwinst gedefinieerd moet worden en welke maat daarvoor gebruikt zou kunnen worden. Beleidsmakers, onderwijsinspecteurs, schoolbesturen, leraren en ouders zijn geïnteresseerd in verschillende vragen omtrent de groei in leerprestaties waarop zo'n maat antwoord moet geven. De één is geïnteresseerd in de absolute prestatiegroei, terwijl de ander geïnteresseerd is in de vorderingen van een leerling in vergelijking tot leeftijdgenoten. Ook doet zich de vraag voor of bij de beoordeling van de opbrengsten van een school de leerwinst of toegevoegde waarde betrokken moet worden om zodoende meer recht te doen aan de inspanningen van een school. Nu worden scholen voornamelijk beoordeeld op hun eindopbrengsten en wordt te weinig recht gedaan aan het startniveau van leerlingen.

Om het opbrengstgericht werken te ondersteunen en bij de beoordeling van de opbrengsten de inspanningen van de school te betrekken, heeft in 2010 de toenmalige regering Rutte-I besloten dat de leerwinst van individuele en groepen leerlingen beter in beeld gebracht moet worden. Onder leerwinst verstaan we de toename van vaardigheden of competenties van leerlingen, zoals die tijdens de schoolloopbaan op verschillende meetmomenten kan worden vastgesteld. Naast de leerwinst is de toegevoegde waarde van een school een belangrijke indicator van de opbrengsten van een school. Deze indicator geeft aan welke bijdrage de school als geheel levert aan de ontwikkeling van de leerlingen. Het belang daarvan neemt alleen maar toe met de invoering van passend onderwijs. Scholen moeten in het onderwijs nog beter aansluiten bij verschillen tussen leerlingen. De ontwikkeling van een faire en eenvoudig hanteerbare werkwijze om de leerwinst van leerlingen en de toegevoegde waarde van de school in beeld te brengen, kan scholen daarbij behulpzaam zijn. Daarom startten OCW en de onderwijsinspectie in het schooljaar 2011-2012 voor het primair onderwijs een pilot, waarin de verschillende aspecten en werkwijzen voor het bepalen van leerwinst van leerlingen en toegevoegde waarde van scholen worden beproefd. Een dergelijke pilot is in 2012 ook van start gegaan voor het voortgezet onderwijs.

De pilot voor het primair onderwijs vindt plaats in een samenwerkingsverband van scholen, wetenschappers en experts en heeft vooral een ontwikkelingsgericht karakter. Een belangrijk onderdeel van de pilot is nagaan vanaf welk moment in de schoolloopbaan de leerwinst en toegevoegde waarde kan worden vastgesteld en welke leerstofgebieden daarvoor geschikt zijn. De vraag die zich hier in het bijzonder voordoet is of de eerste drie jaren van het basisonderwijs hierin betrokken kunnen worden. Het is ook van belang na te gaan hoeveel meetmomenten

nodig zijn om de leerwinst en toegevoegde waarde goed te kunnen vaststellen. Kan dat op basis van twee meetmomenten, bijvoorbeeld in groep 3 en in groep 8, of zijn er ook tussenliggende momenten van belang?

Het is bekend dat de interpretatie van gegevens uit leerlingvolgsystemen niet voor alle leerkrachten gesneden koek is (Visscher & Ehren, 2011; Van der Kley, 2013). Daarom wordt in de pilot veel aandacht besteed aan de presentatie en gebruiksvriendelijkheid van maten voor leerwinst en toegevoegde waarde. Ook is er veel aandacht voor de mogelijkheden om deze maten te gebruiken voor de verbetering van het opbrengstgericht werken en voor de verantwoording over de opbrengsten door de scholen zelf.

Voorliggende rapportage bevat het eindverslag van de pilot. Het eindverslag bestaat uit een overzicht van thans in Nederland gebruikte maten voor leerwinst en toegevoegde waarde, de gevolgde werkwijze om tot nieuwe maten voor leerwinst en toegevoegde waarde te komen en de opvattingen van de deelnemende scholen over de bruikbaarheid van deze maten voor opbrengstgericht werken en de beoordeling van leerprestaties. Het verslag wordt afgesloten met conclusies over en aanbevelingen voor de invoering van leerwinst en toegevoegde waarde in het primair onderwijs.

1.2. Begripsomschrijving

Leerwinst en toegevoegde waarde zijn begrippen die in het onderwijs geen eenduidige betekenis hebben. Het begrip leerwinst wordt soms gebruikt om aan te geven dat een leerling in een bepaald leer- en vormingsgebied vooruit is gegaan en daarvoor bijvoorbeeld een hoger rapportcijfer heeft gekregen. Soms wordt leerwinst gebruikt om aan te geven dat de vorderingen van een leerling boven verwachting zijn. Toegevoegde waarde wordt in het ene geval gebruikt om aan te geven dat de school op een bepaalde terrein iets extra's te bieden heeft, zoals een bijzonder goede leerlingzorg of veel aandacht voor creatieve vakken in het onderwijs. In andere gevallen wordt toegevoegde waarde gebruikt om de bijdrage aan te geven van wat de school toevoegt aan datgene wat leerlingen van huis meekrijgen of buiten de school opsteken. Het komt echter ook voor dat de begrippen leerwinst en toegevoegde waarde door elkaar worden gebruikt.

Leerwinst wordt in voorliggend rapport gedefinieerd als de toename van kennis en vaardigheden van individuele of groepen leerlingen blijkend uit de resultaten op toetsen waarmee de vorderingen onderling kunnen worden vergeleken (Harris, 2011). In de kern is leerwinst de vergelijking van twee toetsprestaties op twee verschillende momenten waarbij een positief of negatief verschil gezien wordt als 'leerwinst' (Gong, Perie & Dunn, 2006). Het verschil tussen de metingen maakt de ontwikkeling van de leerling of de groep zichtbaar. Het aantal meetmomenten is variabel. De leerwinst kan bijvoorbeeld elk leerjaar gemeten worden of alleen door middel van de afname van een toets aan het begin en aan het eind van een langere onderwijsperiode.

Onder *toegevoegde waarde* wordt verstaan de bijdrage die de school levert aan de leerwinst van alle leerlingen. Het zou natuurlijk fantastisch zijn als we precies zouden kunnen bepalen welke

schoolse factoren de leerwinst hebben veroorzaakt, maar zo ver is de wetenschap niet. Daarvoor zijn de factoren die van invloed zijn op de leerprestaties te complex, te talrijk en lastig te meten. Wat wel mogelijk is, is de invloed van de school statistisch te berekenen door rekening te houden met zoveel mogelijk factoren die de leerwinst kunnen hebben beïnvloed, maar die buiten de invloedssfeer van de school liggen. Zo'n factor is bijvoorbeeld de samenstelling van de schoolbevolking met variabelen zoals etniciteit, geslacht, socio-economische status, intelligentie en type zorgleerlingen. Door de leerwinst zo goed mogelijke te corrigeren voor niet-schoolse factoren wordt in principe het effect van de school zichtbaar. We weten echter niet welke specifieke schoolfactoren dit effect hebben veroorzaakt. Ook kunnen we niet voor alle relevante niet-schoolse factoren corrigeren, omdat we ze niet kennen of ze niet goed hebben kunnen meten. Daarom is de toegevoegde waarde te beschouwen als een indicatie van de invloed die scholen hebben op de prestaties van hun leerlingen (Harris, 2011; Bosker, 2012; Timmermans, 2012).

Er zijn twee doelen die met leerwinst en toegevoegde waarde kunnen worden nagestreefd. Dat is in de eerste plaats dat ze een rol gaan spelen in het opbrengstgericht werken door scholen. Dit doel wordt in het voorliggende rapport omschreven als leerwinst en toegevoegde waarde in *schoolverbeteringsperspectief.* Hieronder wordt verstaan dat leerwinst en toegevoegde waarde gebruikt kunnen worden voor zowel de afstemming van het onderwijs op de leerlingen als voor het beoordelen van de leerprestaties door de school zelf in het kader van de kwaliteitszorg. Het tweede doel van leerwinst en toegevoegde waarde is dat deze betrokken kunnen worden in de verantwoording over de leerprestaties door de scholen aan de ouders en aan de Inspectie van het Onderwijs. De inspectie op haar beurt kan leerwinst en toegevoegde waarde betrekken in de beoordeling van de leerprestaties van scholen. Omdat hier rekenschap en openbaarmaking een belangrijke rol spelen, wordt dit doel in dit rapport omschreven als leerwinst en toegevoegde waarde in *accountabilityperspectief*.

1.3. Beleidsontwikkelingen

In 2007 is in Nederland in een advies van prof. dr. J. Peschar 'Over leerwinst als stelselindicator' voor het eerst aandacht geschonken aan leerwinst als maat voor de prestatiegroei van leerlingen en de mogelijkheid de inspanningen van de school zichtbaar te maken (Peschar, 2007). Op basis van dit advies voerde het SCO-Kohnstamm Instituut in 2008 een verkenning uit naar de mogelijkheden om leerwinst en toegevoegde waarde te berekenen op de longitudinale data van het PRIMA-cohortonderzoek en op het COOL-cohortonderzoek (Roeleveld, Van der Veen & Ledoux, 2008). Deze verkenning kreeg geen onmiddellijk vervolg.

In 2010 wordt in het Regeerakkoord 'Vrijheid en verantwoordelijkheid' van het kabinet Rutte I van leerwinst en toegevoegde waarde een speerpunt gemaakt. De introductie van een verplichte begintoets samen met een centrale eindtoets creëert twee meetmomenten waarmee de leerwinst binnen scholen voor primair onderwijs kan worden vastgelegd. Dit regeerakkoord geeft hiermee al richting aan de wijze waarop de toegevoegde waarde van scholen moet worden bepaald, namelijk met behulp van een groeimodel.

De aanleiding om leerwinst en toegevoegde waarde van basisscholen in beeld te brengen is het streven om de beoordeling van de onderwijsopbrengsten door de inspectie te verbeteren. De inspectie beoordeelt de opbrengsten van basisscholen op basis van de score op een eindtoets. Deze score moet aan een bepaalde norm voldoen. De inzet van het regeerakkoord is dat bij de beoordeling van de opbrengsten “de toegevoegde waarde (leerwinst) zwaarder gaat wegen”, en dat “er wordt geïnvesteerd in de centrale en/of uniforme toetsing op het PO (…) zodat de leerwinst objectief kan worden gemeten door de onderwijsinspectie”. Dit moet mede gezien worden in het licht van de maatschappelijke en politieke vraagtekens die worden geplaatst bij de huidige beoordeling. In de huidige situatie worden de opbrengsten gecorrigeerd voor kenmerken van de leerlingpopulatie, in het bijzonder op basis van de zogenaamde gewichtenregeling. Deze correctie is onvoldoende onderscheidend voor de groep scholen met veel leerlingen die extra ondersteuning behoeven of voor de grote groep scholen met weinig of geen gewichtenleerlingen.

Naar aanleiding van dit regeerakkoord verschijnt in mei 2011 voor het primair onderwijs het actieplan 'Basis voor presteren'[1]. In dit plan kondigt de toenmalige minister aan in het schooljaar 2011-2012 te starten met een pilot Leerwinst en Toegevoegde waarde, waarin de verschillende aspecten en werkwijzen voor het bepalen van leerwinst van leerlingen en toegevoegde waarde van scholen wordt uitgewerkt en beproefd. Deze pilot moet eind 2013 leiden tot een uitgewerkt voorstel voor de vormgeving en de invoering van een werkwijze om de toegevoegde waarde van scholen te bepalen en de functie daarbij van een begintoets, de tussentijdse toetsen uit een leerlingvolgsysteem (LVS-toetsen) en de centrale eindtoets.

In 2011 brengt De Onderwijsraad een advies uit over het actieplan 'Basis voor presteren' getiteld ‘Een stevige basis voor iedere leerling’ (Onderwijsraad, 2011). De raad formuleert vier aanbevelingen voor verbetering van het actieplan. Een ervan heeft betrekking op de toetsing in het primair onderwijs. De raad is van mening dat één (eind)toets niet voldoet als maat voor toegevoegde waarde van een school, omdat deze vooralsnog met drie grote onzekerheden gepaard gaat:

1. Niet met zekerheid kan bepaald worden wat de verdienste van de school of leerkracht is bij de prestaties over de jaren heen. Ouders, huiswerkbegeleiding, vrienden en dergelijke hebben invloed op het leren. Corrigeren voor deze invloeden is moeilijk, want er zijn veel factoren die een rol kunnen spelen. Deze factoren zijn bovendien niet allemaal bekend en verder kan hun invloed door de tijd heen variëren.
2. Statistische onzekerheidsmarges worden groter door bij de berekeningen met dergelijke factoren rekening te houden. De praktische bruikbaarheid van een score wordt daardoor kleiner. Op kleine scholen speelt deze moeilijkheid in nog sterkere mate.
3. Het is onduidelijk op welk moment in de schoolloopbaan de toegevoegde waarde van een school gemeten zou kunnen worden. Wellicht komt deze pas later in de (school)loopbaan tot uiting in resultaten.

[1] Tweede Kamer, vergaderjaar 2010-2011, 32 500-VIII, nummer 207.

Om bovenstaande redenen kunnen de verplichte eindtoetsen voor taal en rekenen niet voldoen als grondslag voor de beoordeling van prestaties van scholen. Bovendien is de raad van oordeel dat een beoordeling van de prestaties van scholen de algehele kwaliteit van een basisschool als instelling voor funderend onderwijs moet weerspiegelen. Deze is breder dan de leerprestaties van leerlingen voor taal en rekenen.

De raad onderschrijft het streven scholen te stimuleren tot hogere prestaties. Hij adviseert de minister daarbij vooral in te zetten op het stimuleren van het lerend vermogen van scholen. In dat licht is volgens de raad de ontwikkeling wenselijk om binnen een pilot het begrip toegevoegde waarde te laten onderzoeken. De pilot kan waardevolle informatie opleveren over het bepalen van leerwinst en toegevoegde waarde, gericht op het bevorderen van opbrengstgericht werken van de basisscholen om zo de kwaliteit van basisscholen te verbeteren. Tevens verdient het aanbeveling na te gaan in hoeverre de bepaling van toegevoegde waarde voor scholen hoge administratieve lasten met zich meebrengt.

In november 2011[2] stuurt de minister een brief over de pilot aan de Tweede Kamer. Rekening houdend met het advies van de Onderwijsraad, formuleert de minister twee doelen die met de pilot bereikt moeten worden. Het eerste hoofddoel is het opbrengstgericht werken van scholen en hun besturen verder te stimuleren. Het in beeld brengen van leerwinst en toegevoegde waarde biedt scholen en besturen beter inzicht in de vorderingen van de leerlingen en de bijdrage die de school hieraan levert. Het verschaft scholen waardevolle informatie voor de eigen evaluatie van de kwaliteit van het gegeven onderwijs.

In het verlengde hiervan ligt het tweede hoofddoel van de pilot: nagaan in hoeverre een werkwijze om leerwinst en toegevoegde waarde te bepalen bijdraagt aan een betere interpretatie van de opbrengsten van een school. Daarmee kan het een instrument zijn dat niet alleen door de school en zijn bestuur kan worden gebruikt, maar ook door de Inspectie van het Onderwijs bij de beoordeling van de onderwijsopbrengsten.
In de pilot wordt ook nagegaan in hoeverre beide doelen goed met elkaar verenigbaar zijn. Daarbij wordt gekeken of zich ongewenste neveneffecten voordoen, zoals *teaching to the test* of een te eenzijdige aandacht voor taal en rekenen.

Ook bij het wetsvoorstel Centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs wordt leerwinst en toegevoegde waarde betrokken. In zijn brief aan de Tweede Kamer (15 maart 2013), ter voorbereiding op het wetsvoorstel, wijst staatssecretaris Dekker erop dat hij er aan hecht om de leerwinst in kaart te brengen om zo het beste uit iedere leerling te kunnen halen. Ook vindt hij het belangrijk om de toegevoegde waarde van scholen inzichtelijk te maken, zodat in beeld kan worden gebracht wat een school met haar leerlingen weet te bereiken. Inzicht in leerwinst en toegevoegde waarde is hierbij primair van belang voor de zelfevaluatie van de school, maar ook voor een versterkte professionele dialoog tussen het schoolbestuur en de inspectie, waarbij de opbrengsten in het juiste perspectief worden

[2] Tweede Kamer, vergaderjaar 2011–2012, 31 293, nr. 126.

geplaatst. De staatssecretaris benadrukt dat hij inzicht in de resultaten niet wil benutten voor prestatiebekostiging van scholen.

Tijdens het debat over de centrale eindtoets op 21 maart 2013 bleek opnieuw dat de pilot Leerwinst en Toegevoegde waarde de belangstelling heeft van de Tweede Kamer. De pilot is erop gericht aan te tonen dat de beschikbaarheid van toetsresultaten van belang is om leerwinst en toegevoegde waarde te kunnen bepalen. Door het gebruik van een leerlingvolgsysteem verplicht te stellen en door de invoering van een verplichte eindtoets wordt aan deze randvoorwaarde voldaan. Ook is gebleken dat de kwaliteit en de vergelijkbaarheid van de gegevens een belangrijke voorwaarde is. In dit kader is het amendement van het Kamerlid Straus[3] relevant: dit amendement houdt in dat de gegevens in het leerling- en onderwijsvolgsysteem met elkaar moeten kunnen worden vergeleken.

1.4. Opdrachtformulering

Op 16 november 2011 krijgt een projectgroep bestaande uit onderzoekers van de Universiteit Twente, de Rijksuniversiteit Groningen (GION) en Cito de opdracht om in samenwerking met scholen voor basis- en speciaal (basis)onderwijs in een pilot maten voor leerwinst en toegevoegde waarde te ontwikkelen. De CED-Groep biedt praktische ondersteuning tijdens de pilot en is het eerste aanspreekpunt voor de deelnemende scholen. De projectleiding is in gezamenlijke handen van de onderwijsinspectie en het ministerie van Onderwijs, Cultuur en Wetenschap. De tussen- en de eindresultaten worden voorgelegd aan een breed samengestelde klankbordgroep.
De pilot had in eerste instantie een looptijd van november 2011 tot oktober 2013. Vanwege aanloopproblemen is een verlenging verleend, waardoor het eindrapport in de loop van januari 2014 is opgeleverd.

1.5. Leeswijzer

Hoofdstuk 2 beschrijft de opzet van de pilot. Hierin komen het doel van de pilot, de werving van de scholen, de organisatie van de pilot en de uitvoering van de werkzaamheden aan de orde.

In hoofdstuk 3 wordt de theorie rond leerwinst en toegevoegde waarde behandeld, waaronder de verschillende manieren waarop deze kunnen worden berekend en de factoren die bepalend zijn voor de nauwkeurigheid van maten voor leerwinst en toegevoegde waarde.

Leerwinst en toegevoegde waarde zijn onderwerpen die vooral in de VS en in Engeland in de belangstelling staan. Toch hebben we in Nederland ook ervaringen opgedaan met leerwinstbepaling en de schatting van toegevoegde waarde. Hoofdstuk 4 geeft een overzicht van Nederlands onderzoek op dit terrein en van enkele concrete toepassingen in het basisonderwijs.

In hoofdstuk 5 worden de resultaten van de pilot behandeld. Dat zijn natuurlijk in de eerste plaats de ontwikkelde en in de pilot uitgeprobeerde modellen voor leerwinst en toegevoegde waarde. Daarnaast wordt ook aandacht besteed aan de ervaringen van de deelnemende scholen

[3] Tweede Kamer, vergaderjaar 2012–2013, 33 157, nr. 30.

met de rapportages over de verschillende modellen en met de bruikbaarheid ervan voor de onderwijspraktijk. Aan het eind van dit hoofdstuk worden conclusies getrokken over de bruikbaarheid van leerwinst en toegevoegde waarde voor het basisonderwijs. Ook wordt een overzicht gegeven van de voorwaarden waaraan voldaan moet worden, wil leerwinst en toegevoegde waarde van praktische betekenis zijn voor het basis- en speciaal basisonderwijs in Nederland.

Hoofdstuk 6 bevat een slotbeschouwing waarin de conclusies nog eens de revue passeren. Daarin worden ook aanbevelingen gedaan gericht aan aanbieders van toets- en schoolinformatiesystemen, de scholen, de Inspectie van het Onderwijs en het ministerie van OCW om:

- Leerwinstberekeningen op schoolniveau mogelijk te maken vanuit een schoolverbeteringsperspectief.
- Leerwinst en toegevoegde waarde te betrekken in het toezicht.
- Aanvullend onderzoek te doen naar de toepassing van toegevoegde waarde vanuit een accountabilityperspectief.

2. Opzet en uitvoering pilot

2.1. Inleiding

In dit hoofdstuk komen achtereenvolgens de volgende onderwerpen aan de orde. Allereerst worden in paragraaf 2.2 de doelen beschreven die ten grondslag liggen aan de pilot en de specifieke vragen die zich voordoen op het terrein van leerwinst en toegevoegde waarde. Aan de pilot zijn met het oog op een brede implementatie in het basisonderwijs enkele voorwaarden gesteld die van betekenis waren voor de werving van scholen die aan de pilot konden meedoen. Deze worden beschreven in paragraaf 2.3.
In paragraaf 2.4. komt de werving aan de orde van de scholen die in de pilot hebben geparticipeerd. Omdat uitgegaan moest worden van de bestaande toetspraktijk in het basisonderwijs en omdat alle deelnemende scholen toetsen uit het leerlingvolgsysteem van Cito gebruiken, wordt in paragraaf 2.5 toegelicht op welke wijze deze toetsen gebruikt behoren te worden. Paragraaf 2.6 geeft een beschrijving van de organisatie van de pilot.
In paragraaf 2.7 komt de werkwijze aan de orde die gevolgd is om scholen te informeren over de uitkomsten van de verschillende leerwinst en toegevoegde waarde modellen.
Paragraaf 2.8 beschrijft de opzet van de evaluatie die is uitgevoerd om de satisfactie van de deelnemende scholen te peilen met het gebruik van leerwinst en toegevoegde waarde maten in de onderwijspraktijk.

2.2. Doelen en onderzoeksvragen

Het eerste doel van de pilot is het opbrengstgericht werken van scholen en hun besturen verder te stimuleren. Het in beeld brengen van leerwinst en toegevoegde waarde biedt scholen en besturen beter inzicht in de vorderingen van de leerlingen en de bijdrage die de school hieraan levert. Het verschaft scholen nieuwe informatie voor de eigen evaluatie van de kwaliteit van het onderwijs als aanvulling op de informatie die de verschillende toets- en schoolinformatiesystemen thans leveren. Leerwinst en toegevoegde waarde worden in dit verband beschouwd vanuit een *schoolverbeteringsperspectief*.

In het verlengde hiervan ligt het tweede doel van de pilot: nagaan in hoeverre een werkwijze om leerwinst en toegevoegde waarde te bepalen zou kunnen bijdragen aan een betere beoordeling van de opbrengsten van een school. Daarmee kan het een instrument zijn dat niet alleen door de school en zijn bestuur, maar ook door de Inspectie van het Onderwijs wordt gebruikt bij de beoordeling van de onderwijsopbrengsten. Hier wordt leerwinst en toegevoegde waarde bezien vanuit een *accountabilityperspectief*.

Rondom het ontwikkelen en gebruiken van een instrument om leerwinst en toegevoegde waarde te bepalen speelt een aantal vraagstukken. Zowel de PO-Raad[4] als de Onderwijsraad (Onderwijsraad, 2011, p. 31 e.v.) wijzen daarop. Deze vragen zijn opgenomen in de eerder genoemde brief waarmee de Tweede Kamer is geïnformeerd over het doel en de opzet van de

[4] http://www.poraad.nl/content/kanttekeningen-bij-wetsvoorstel-toetsing

pilot (november 2011). In de pilot is bij het ontwikkelen en beproeven van maten voor leerwinst en toegevoegde rekening gehouden met de volgende specifieke vragen:

1. Op welk moment en op welke manier kan in de basisschoolperiode een beginmeting op valide en betrouwbare wijze bij jonge leerlingen worden afgenomen?

   Hierbij spelen drie subvragen:

   1.1 Zijn toetsen bedoeld voor leerlingen uit groep 3 geschikt voor een beginmeting?

   1.2 Maakt het daarbij uit of leerlingen uit groep 3 hebben deelgenomen aan voor- en vroegschoolse educatie?

   1.3 In hoeverre kunnen de basisvaardigheden van leerlingen uit groep 3 vergeleken worden met de basisvaardigheden van leerlingen aan het einde van de basisschoolperiode?
2. Welke toetsen van welke toetsaanbieders zijn bruikbaar voor de bepaling van leerwinst en toegevoegde waarde?
3. Kan voor verschillende groepen leerlingen binnen de school (zwak, gemiddeld, excellent) de leerwinst en toegevoegde waarde afzonderlijk worden bepaald?
4. Kan leerwinst en toegevoegde waarde een bijdrage leveren aan het gebruik van het ontwikkelingsperspectief voor individuele leerlingen in het speciaal (basis)onderwijs?

Tot slot is nagegaan in hoeverre beide hoofddoelen goed met elkaar verenigbaar zijn. In het bijzonder doet zich de vraag voor of het gebruik van leerwinst en toegevoegde waarde ten behoeve van het toezicht op het onderwijs ongewenste neveneffecten in de hand werkt, zoals strategisch gedrag van scholen om zo goed mogelijk uit de bus te komen.

2.3. Voorwaarden

Met het oog op een brede implementatie in het primair onderwijs zijn twee voorwaarden geformuleerd waaraan de uitkomsten van de pilot moeten voldoen.

1. Er dient zoveel mogelijk aangesloten te worden op de bestaande toetspraktijk, waardoor de toetsdruk op de pilotscholen niet toeneemt. Om de leerwinst te kunnen bepalen is afgesproken dat de scholen die aan de pilot mee kunnen doen een serie gestandaardiseerde toetsen gebruiken, waarmee de vorderingen van leerlingen gedurende een langere periode kunnen worden geïnterpreteerd. Dit is bijvoorbeeld mogelijk met de toetsen uit het Cito Volgsysteem primair en speciaal onderwijs.
2. Er dient ook gebruik te worden gemaakt van de toets- en schoolinformatiesystemen die scholen gebruiken om toetsscores op te slaan, zoals ParnasSys, Dotcomschool, ESIS web-based en het computerprogramma Cito-LOVS. Aansluiting bij deze systemen beperkt de administratieve last voor de scholen.

De toetsen die in de pilot in aanmerking komen voor het bepalen van leerwinst en toegevoegde waarde zijn toetsen voor taal (spelling, woordenschat, technisch en begrijpend lezen) en rekenen-wiskunde. De keuze voor het accent op taal en rekenen-wiskunde is vooral praktisch van aard. In de eerste plaats is de toetstechnische kwaliteit van toetsen bij deze leergebieden

hoog. Dat betekent dat in principe de betrouwbaarheid en validiteit van deze toetsen geen beperkende factor behoeven te zijn in de berekening van leerwinst en toegevoegde waarde.

In de tweede plaats lag in het actieplan 'Basis voor Presteren' de focus op het verbeteren van de prestaties voor Nederlandse taal en rekenen-wiskunde. Daarom ligt ook binnen de pilot het accent op het beproeven van werkwijzen voor het bepalen van leerwinst en toegevoegde waarde van scholen voor deze kernvakken. Dit sluit overigens niet uit dat er, afhankelijk van de mogelijkheden en behoeften van de deelnemende scholen, ook aandacht is geweest voor andere leer- en vormingsgebieden, zoals voor de sociaal-emotionele ontwikkeling van leerlingen.

2.4. Werving scholen

Er hebben 45 scholen - deels op uitnodiging, deels op eigen initiatief - meegedaan aan de pilot. Het betreft een gevarieerde groep van scholen. Bij de werving is niet alleen rekening gehouden met de vertegenwoordiging van verschillende onderwijssoorten (basis- en speciaal (basis)onderwijs), maar ook met verschillende onderwijskundige en pedagogisch-didactische richtingen van scholen, zoals Dalton-, Montessori- en Vrije Scholen en scholen vanuit verschillende denominaties. Verder is gekeken naar een spreiding in geografische ligging en naar de samenstelling van de schoolpopulatie, waaronder het percentage gewichtenleerlingen.

Vooraf zijn criteria opgesteld waaraan de scholen binnen de pilot zouden moeten voldoen. Nadat scholen zich hadden opgegeven heeft er een intakegesprek plaatsgevonden met een adviseur van de CED-Groep. Tijdens deze intake zijn scholen bevraagd op de selectiecriteria (zie bijlage 1). Tevens is geïnventariseerd welke ervaringen de scholen hadden met opbrengstgericht werken en in welke mate en op welk niveau dit werd toegepast. Verder is er uitvoerig stilgestaan bij de toetspraktijk van de school. Uiteindelijk is er voor gekozen om alle aangemelde scholen aan de pilot te laten meedoen. De complete lijst van deelnemende scholen is in bijlage 2 opgenomen.

2.5. LVS-toetsen van Cito

Alle deelnemende scholen gebruiken toetsen uit het Cito Volgsysteem primair en speciaal onderwijs (LVS-toetsen van Cito). Deze toetsen hebben onder andere als doel om het vaardigheidsniveau van een leerling te bepalen, deze te vergelijken met dat van leerlingen in eenzelfde jaargroep en bij meerdere metingen de vaardigheidsontwikkeling van een leerling te volgen in de tijd. De wijze waarop deze toetsen dienen te worden gebruikt wordt hieronder beschreven.

Van toetsscore naar vaardigheidsscore

Om te kunnen bepalen wat het niveau is van een leerling wordt de toetsscore omgezet in een zogenaamde vaardigheidsscore ofwel een score op een onderliggende vaardigheidsschaal. Hoe hoger de vaardigheidsscore, des te hoger de vaardigheid van de leerling is.
De vaardigheidsschaal maakt het mogelijk om zowel de resultaten van een leerling op verschillende toetsmomenten met elkaar vergelijken als de resultaten van leerlingen in dezelfde groep die verschillende toetsen hebben gemaakt. De vaardigheidsscores tussen verschillende leerstofgebieden, zoals rekenen-wiskunde en spelling, zijn niet te vergelijken. De verschillende

toetsen meten namelijk andere vaardigheden. Daarom heeft Cito voor elke toets een andere vaardigheidsschaal ontwikkeld. Vergelijk dit bijvoorbeeld met lengte en gewicht: voor het ene wordt de meter gehanteerd, voor het andere is de gram de meeteenheid. Deze verschillende vaardigheidsschalen maken het voor het onderwijs soms lastig om te communiceren over de vaardigheidsscore. Immers, een vaardigheidsscore van 54 bij rekenen-wiskunde betekent iets anders dan 54 bij begrijpend lezen. Een van de manieren waarop Cito het mogelijk maakt om toetsresultaten toch goed te kunnen interpreteren zijn de zogenaamde *vaardigheidsniveaus*.

Vaardigheidsniveaus

Alle toetsen in het Cito Volgsysteem primair onderwijs zijn genormeerd. Deze normen zijn bedoeld om de vaardigheid van leerlingen te vergelijken met die van andere leerlingen in eenzelfde jaargroep.
Op basis van de resultaten van de leerlingen in normeringonderzoeken is een indeling in niveaugroepen gemaakt: de vaardigheidsniveaus. Cito hanteert twee verschillende niveau-indelingen[5]:

- een indeling in niveaus I tot en met V en
- een indeling in niveaus A tot en met E.

De niveaugroepen (zie tabel 1) geven aan hoe een leerling scoort ten opzichte van een landelijke steekproef leerlingen. Vaardigheidsniveaus betreffen relatieve scores en zijn bedoeld om de behaalde vaardigheidsscore (een absolute score) te kunnen interpreteren.

Tabel 1 Verdeling over de groepen bij de niveau-indeling in A tot en met E en bij I tot en met V

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

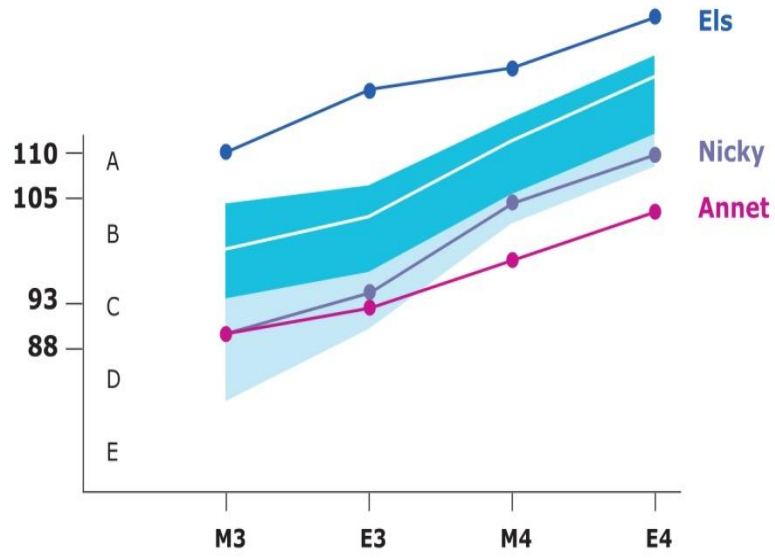
[5] Van oorsprong bestond voor de LVS-toetsen uitsluitend de niveau-indeling A-E. In de tweede generatie toetsen (vanaf 2005) is daar de niveau-indeling I-V aan toegevoegd. De belangrijkste redenen hiervoor zijn 1) het ontbreken van een gemiddelde groep in de verdeling A-E, 2) afstemming met de Entreetoets waar de niveau-indeling I-V gehanteerd wordt. De gebruiker kan kiezen welke niveau-indeling hij hanteert.

Toetsen op maat

De LVS-toetsen zijn vaardigheidstoetsen. Een toets is daarom samengesteld met opgaven van uiteenlopende moeilijkheid. Voor leerlingen die een reguliere ontwikkeling doormaken past meestal de toets die voor dat moment ontwikkeld is. Bij leerlingen met een afwijkende ontwikkeling en/of speciale onderwijsbehoefte adviseert Cito om op maat te toetsen. Dit kan door een toets van een hoger of lager niveau voor te leggen of de toetsen Speciale leerlingen af te nemen. Doordat alle toetsen van een leergebied (ook de toetsen voor Speciale leerlingen) op eenzelfde vaardigheidsschaal staan, zijn de behaalde vaardigheidsscores vergelijkbaar. Hierbij dient opgemerkt te worden dat uitsluitend vaardigheidsscores die behaald zijn op eenzelfde generatie toetsen (1e generatie/oud of 2e generatie/nieuw) vergelijkbaar zijn.

Voldoende vooruitgang

Wanneer in de schoolloopbaan van de leerlingen meerdere toetsen van een bepaald leergebied worden afgenomen, kan de ontwikkeling van een leerling in de tijd worden gevolgd.
Deze gegevens kunnen uitgezet worden in een grafiek (zie figuur 1):



Figuur 1 Leerwinst van drie leerlingen in de periode M3-E4

In de grafiek zijn de vaardigheidsschaal, de afgenomen toetsen en het vaardigheidsniveau te zien. De grafiek laat zien dat voor alle leerlingen geldt dat er sprake is van leerwinst, namelijk hun groei in vaardigheid. De vraag is echter of voor iedere leerling geldt dat de leerwinst voldoende of naar verwachting is. Leerlingen met eenzelfde startpositie kunnen een andere groei doormaken en qua leerwinst van elkaar verschillen (vergelijk leerling Nicky en Annet). En ook voor Els geldt de vraag of haar leerwinst voldoende is. Ze haalt weliswaar hoge vaardigheidsscores, maar deze ontwikkelen zich niet altijd in een rechte lijn.

2.6. De organisatie van de pilot

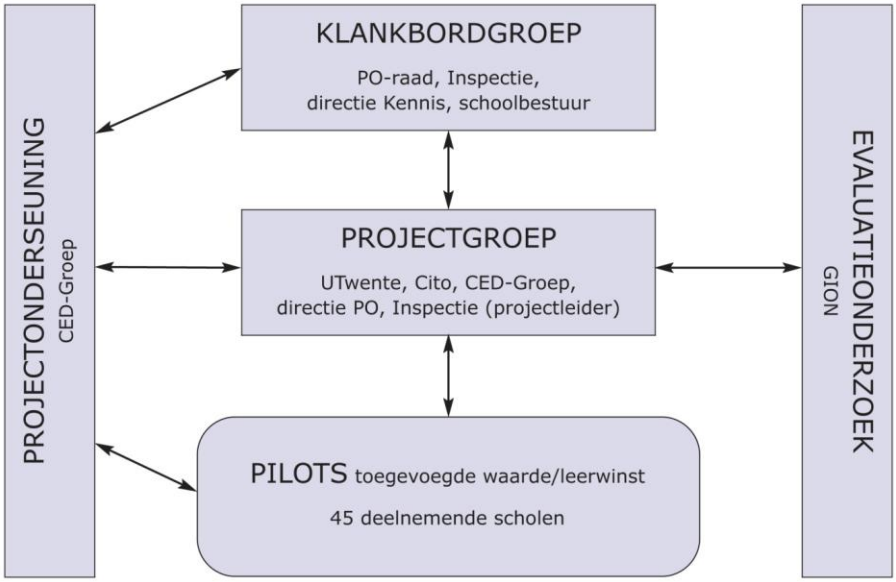
De pilot is gericht op het ontwikkelen en uitproberen van verschillende maten voor leerwinst en toegevoegde waarde op de 45 deelnemende scholen. Daar waren verschillende partijen bij betrokken (zie figuur 2). Het ontwikkelonderzoek is uitgevoerd door onderzoekers van Cito en van de universiteiten van Groningen (GION) en Twente. De onderzoekers stelden op basis van toetsgegevens van de 45 deelnemende scholen schoolspecifieke rapporten op voor de verschillende maten voor leerwinst en toegevoegde waarde.

Naast ontwikkelwerk zijn door GION ook schoolrapportages gemaakt waarin de vaardigheid van sbo-leerlingen in de pilot wordt vergeleken met die van sbo-leerlingen in het COOLspeciaal onderzoek. Ook heeft het GION een evaluatie-onderzoek uitgevoerd onder de deelnemende scholen naar de inzichtelijkheid en de bruikbaarheid van de rapportages voor zowel opbrengstgericht werken als de beoordeling van de leerprestaties.

De CED-Groep verleende ondersteuning bij de werving en selectie van de scholen, de bespreking van de leerwinst- en toegevoegde waarderapportages met en de begeleiding van de scholen bij het opbrengstgericht werken en de extra dataverzameling die nodig was voor de COOLspeciaal schoolrapportages.

De werkzaamheden stonden onder leiding van een projectgroep die bestond uit de onderzoekers van de meewerkende instituten en uit medewerkers van de CED-Groep, OCW en van de Inspectie van het onderwijs. De algehele projectleiding was in gezamenlijke handen van de inspectie en het ministerie van OCW. De projectondersteuning was in handen van de CED-Groep (zie bijlage 3).

De tussen- en de eindresultaten zijn voorgelegd aan een klankbordgroep, die bestond uit vertegenwoordigers van de PO-Raad, de Algemene Vereniging Schoolleiders (AVS), de Inspectie van het Onderwijs, de directie Kennis OCW, de gemeente Rotterdam (dJ&O) en het project Kwaliteitsaanpak Basisonderwijs Amsterdam.



Figuur 2 Organisatie van de pilot

2.7. Werkwijze op de scholen

De 45 scholen zijn door een groep van elf adviseurs van de CED-Groep met ruime ervaring op het gebied van opbrengstgericht werken begeleid (zie bijlage 3). Tevens hebben de adviseurs een belangrijke rol gespeeld in de landelijke bijeenkomsten met de deelnemende scholen. De adviseurs hebben steeds contact gehad met dezelfde school vanaf de intake tot aan het laatste gesprek in november 2013.

Intake

Na de ontwikkeling van een intakeformulier hebben de adviseurs vervolgens de scholen bezocht. Tijdens dit bezoek zijn de gegevens van het intakeformulier besproken en is er uitleg gegeven over de verschillende leerwinst- en toegevoegde waarde modellen die in de pilot gebruikt zouden kunnen worden (zie bijlage 4). De deelnemende scholen hebben op basis van de informatie die ze de adviseurs hebben gegeven, een keuze gemaakt uit de verschillende modellen. Verder zijn afspraken gemaakt over de wijze waarop de scholen toetsgegevens konden aanleveren.

Schoolbezoeken

Sinds de start van de pilot in november 2011 hebben vier ronden schoolrapportages plaatsgevonden. Voorafgaand aan elke ronde hebben de deelnemende scholen de meest recente toetsdata en administratiegegevens van hun leerlingen ge-upload naar de Toetsservice van de CED-Groep. Na de nodige databewerkingen door de Toetsservice zijn door de onderzoekers van Cito, de Universiteit Twente en het GION data-analyses uitgevoerd. Bij elke rapportageronde zijn de schoolspecifieke rapportages over hun leerwinst en toegevoegde waarde op een afgeschermde website geplaatst (Sharepoint). Alleen de schooladviseur, de betreffende school en de onderzoekers hadden toestemming deze rapportages in te zien. Bij drie van de vier rapportageronden hebben schoolbezoeken plaatsgevonden. Bij de derde ronde, kort voor de zomervakantie, is ervoor gekozen om het gesprek over de rapportages telefonisch te voeren. Als voorbereiding op de schoolgesprekken over de rapportages hebben de onderzoekers deze eerst met de adviseurs besproken.

Bij een schoolbezoek zat een adviseur van de CED-Groep met de directie van de school - vaak de ib-er en de (adjunct)schooldirecteur – om tafel om de schoolrapporten te bespreken. Vaak was ook een lid van het projectteam aanwezig. De schooladviseur gaf uitleg over de schoolrapporten en besprak de resultaten. De scholen is nadrukkelijk gevraagd om feedback op de modellen te geven, zowel inhoudelijk als vormtechnisch. De uitkomsten van deze gesprekken zijn in verslagen vastgelegd en teruggekoppeld aan de onderzoekers en de projectgroep. De feedback van de scholen is gebruikt om de schoolrapporten in de volgende ronde te verbeteren. De terugkoppeling naar de scholen was dus niet alleen gericht op het gebruik van de schoolrapportages, maar ook op het opsporen van eventuele tekortkomingen in en verbetering van de rapportages.
Een steeds terugkerend probleem bleek de juistheid en volledigheid van de gegevens die de scholen hebben aangeleverd. Soms bleken gegevens niet correct of werden leerlingen in de rapportages gemist. De adviseurs zijn door de onderzoekers ingezet om de oorzaken van incomplete data op de scholen te achterhalen en indien mogelijk alsnog aan te vullen.

Landelijke bijeenkomsten

Gedurende de pilot zijn er met de participerende scholen drie bijeenkomsten geweest. Deze vonden plaats aan het begin, halverwege en aan het eind van de pilot. Omdat bij deze bijeenkomsten de aanwezigheid van alle deelnemende scholen gewenst was, zijn de scholen actief door de adviseurs benaderd. Bij de opzet en inhoud van deze bijeenkomsten is rekening gehouden met de wensen van de scholen. De adviseurs waren bij deze bijeenkomsten aanwezig om als aanspreekpunt voor hun scholen te dienen. De bijeenkomsten werden steeds door tenminste 95% van de scholen bezocht.

De eerste bijeenkomst (december 2011) was vooral gericht op het toelichten van de werkwijze van de pilot en op de wijze waarop de scholen daarbij betrokken worden. Ook is tijdens die bijeenkomst de website van de pilot geïntroduceerd en de wijze waarop onderling web-based informatie kon worden uitgewisseld.
De tweede bijeenkomst (september 2012) stond in het teken van het onderling uitwisselen van ervaringen met de rapportages en van de wijze waarop er op de verschillende scholen mee werd gewerkt. Ook is aandacht besteed aan wenselijk geachte verbeteringen van de rapportages.

De laatste bijeenkomst (december 2013) draaide om de inhoud van het concept van het eindverslag van de pilot met het oog op de implementatie van leerwinst en toegevoegde waarde in het primair onderwijs. Daarbij was ook staatssecretaris Dekker aanwezig. Naast de conclusies die op basis van de pilot getrokken kunnen worden, zijn de scholen ook actief betrokken in een discussie over de aanbevelingen die noodzakelijk zijn om de uitkomsten van de pilot op een bredere schaal te kunnen verspreiden. Daarmee is in deze eindrapportage rekening gehouden.

Sbo-scholen

Onder de deelnemende basisscholen in de pilot bevinden zich scholen voor speciaal basisonderwijs (sbo). Ofschoon ook deze scholen al enige tijd LVS-toetsen bij hun leerlingen afnemen, is tijdens de pilot gebleken dat hun toetspraktijk noodzakelijkerwijs afwijkt van die van de reguliere basisscholen. Op sbo-scholen kan de leerwinst maar voor een kleine groep leerlingen worden berekend omdat veel leerlingen pas instromen vanaf groep 5. Standaard zijn in de pilot de M-toetsen in groep 3 als startmeting voor het berekening van leerwinst en toegevoegde waarde gebruikt.

Het vergelijken van de leerwinst van sbo-leerlingen onderling, waarbij uitsplitsingen worden gemaakt naar de problematiek van de leerlingen, zou een alternatief kunnen zijn. Maar gezien het kleine aantal deelnemende sbo-scholen bleken de mogelijkheden hiertoe echter beperkt. Om de sbo-scholen uit de pilot toch feedback te geven op hun toetsprestaties is buiten de pilot om een oplossing gevonden door gebruik te maken van de zogenaamde 4D-index van de CED-Groep (zie daarvoor par. 4.3.3) en door de toetsprestaties van hun leerlingen te vergelijken met die van leerlingen van sbo-scholen die participeerden in het onderzoek COOLspeciaal (zie daarvoor par. 5.5).

2.8. Evaluatie gebruik schoolrapportages

Na de laatste rapportageronde (oktober-november 2013) is samen met de scholen de balans opgemaakt over de uiteindelijke vorm die de rapportages hebben gekregen. De schoolrapportages zijn in de eerste plaats ontwikkeld vanuit een schoolverbeteringsperspectief. Naast vragen over het praktische nut en de inzichtelijkheid van de rapportages, is het daarom ook van belang om te evalueren of scholen deze inzetten voor opbrengstgericht werken. Een verstandig gebruik van toetsgegevens uit een leerlingvolgsysteem, waarna de instructie voor leerlingen op maat wordt gegeven uitgaande van de te bereiken doelen, is immers de kern van opbrengstgericht werken (Inspectie van het Onderwijs, 2010; Ledoux, Blok, & Boogaard, 2009; Schildkamp & Kuiper, 2010; Visscher & Coe, 2003). In de tweede plaats kan de informatie uit de rapportages ook gebruikt worden om op schoolniveau de leerwinst en toegevoegde waarde als een opbrengst van de school te beoordelen. Deze informatie kan door de school gebruikt worden om verantwoording af te leggen aan de ouders, het bestuur of aan de Inspectie van het Onderwijs.

Een belangrijke vraag is dan óf en in hoeverre de pilot erin is geslaagd modellen en daarop gebaseerde rapportages te ontwikkelen die zowel het opbrengstgericht werken als een goede beoordeling van de opbrengsten bevorderen. Zijn er hanteerbare rapportages gemaakt voor de scholen? Vinden de scholen dergelijke rapportages inzichtelijk en bruikbaar om het onderwijs aan hun leerlingen te verbeteren? Hebben ze meerwaarde ten opzichte van andere leerling- en groepsrapportage over toetsprestaties? En kunnen deze rapportages ook een rol spelen in het kader van het accountabilityperspectief?

Methode

Om bovenstaande vragen over de schoolrapportages te beantwoorden, is in oktober en november 2013 een gepersonaliseerde internet-enquête bij 313 respondenten uit de pilot uitgezet. Dat waren leerkrachten van de groepen 3, 5 en 7, de intern begeleiders (ib-ers) en de (adjunct)directeuren van alle deelnemende scholen. De leerkrachten hebben deels andere vragen gekregen dan de ib-ers en de directeuren. Alleen aan de ib-er en de directeur zijn vragen gesteld over de bruikbaarheid en inzichtelijkheid van de meest recente versie van de leerwinst– en toegevoegde waarde rapportages. Voor de leerkrachten waren deze vragen niet relevant omdat bij de schoolbezoeken is gebleken dat de schoolrapportages lang niet altijd worden besproken met de leerkrachten.

De enquête bestond uit 21 vragen voor de leerkrachten en 12 vragen voor de ib-er en de directeur. Iedereen beantwoordde de vragen vanuit zijn eigen professionele perspectief. Het invullen duurde gemiddeld 16 minuten. De respondenten zijn drie keer per email benaderd om de enquête in te vullen en één keer telefonisch door de CED-Groep via de contactpersoon van de school. Per school hebben uiteindelijk gemiddeld bijna zeven personen de enquête ingevuld. Onder leerkrachten (n=213) was de bruikbare respons 54% en onder directeuren/ib-ers (n=89) was dit 93%. Ten behoeve van het eindrapport is alleen gebruik gemaakt van de antwoorden van de directeuren en ib-ers.

Op 40 van de 45 pilotscholen heeft tenminste één respondent - ib-er en/of directeur - alle vragen over het nut en de bruikbaarheid van de schoolrapportages volledig ingevuld. De antwoorden zijn geaggregeerd naar schoolniveau om zo een valide beeld te krijgen van de mening van de relevante personen op een school over de meest recente versie van de schoolrapportages (rapportageronde 4). Aan de hand van rapportcijfers en stellingen voorzien van een vijf-puntsschaal ('helemaal mee oneens' tot en met 'helemaal mee eens') is nagegaan wat de mening is van de respondenten over de inzichtelijkheid en bruikbaarheid, de meerwaarde en toekomstige bruikbaarheid van de definitieve schoolrapportages.

3. Leerwinst en Toegevoegde waarde – theoretisch kader

3.1. Inleiding

In 1987 dook ineens het begrip 'leerlingvolgsysteem' (LVS) op en het verspreidde zich zeer snel in het Nederlandse onderwijs. Thans behoort het tot de standaardinventaris van bijna iedere basisschool (Inspectie van het Onderwijs, 2013). Onder een LVS wordt een systeem verstaan dat bestaat uit:

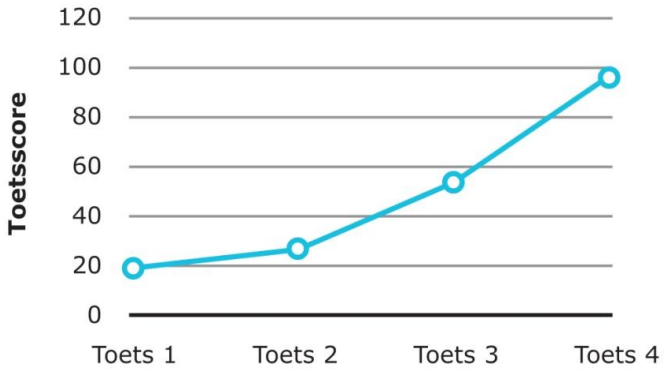
1. een serie toetsen die gebaseerd is op longitudinale leerlijnen uitgezet over meerdere aaneensluitende leerjaren;
2. een registratiesysteem waarin de vorderingen per leerling en per groep kunnen worden vastgelegd;
3. normen voor de beoordeling van de vorderingen.

De kracht van zo'n systeem moet gezocht worden in de mogelijkheid om, ondanks dat opeenvolgende toetsen inhoudelijk en qua moeilijkheidsgraad van elkaar verschillen, de voortgang van een individuele leerling gevolgd en beoordeeld kan worden. Daarvoor moeten de toetsresultaten tot een gemeenschappelijke vergelijkingsbasis worden terug gebracht (zie Fisher & Twing, 2006; Gong, Perie, & Dunn, 2006). In het geval van het LVS van Cito is dit een zogenaamde vaardigheidsschaal waarop de opgaven van gemakkelijk naar moeilijk geordend zijn (zie ook par. 2.5). Op zo'n schaal is de leerwinst van een leerling af te lezen wanneer twee toetsprestaties met elkaar op die schaal worden vergeleken (Kolen & Brennan, 2004). Maar de vraag of er voldoende leerwinst is geboekt, is niet eenvoudig te beantwoorden. De kwestie is dan ook hoe leerwinst gedefinieerd moet worden, welke maat daarvoor het beste gebruikt kan worden en wanneer er sprake is van voldoende leerwinst. Datzelfde geldt ook voor toegevoegde waarde.

In dit hoofdstuk wordt nader ingegaan op wat leerwinst en wat toegevoegde waarde is, hoe ze gemeten kunnen worden en welke factoren van belang zijn voor de kwaliteit ervan.

3.2. De bijdrage van de school aan de leerprestaties

In de kern is leerwinst de vergelijking van twee toetsprestaties van een individuele leerling op twee verschillende momenten, waarbij een positief verschil gezien wordt als leerwinst (Gong, Perie & Dunn, 2006). Door vervolgens het gemiddelde van de individuele leerwinsten te nemen, kan de gemiddelde leerwinst van een groep leerlingen of van de betreffende school worden verkregen (zie figuur 3 op de volgende pagina).



Figuur 3 Prestaties van een groep leerlingen op 4 toetsen

Maar prestatieverschillen tussen toetsen kunnen niet zomaar geïnterpreteerd worden als 'leerwinst'. Van belang is dat de toetsen om de leerwinst te berekenen betrekking hebben op hetzelfde inhoudelijke domein of - nog liever - op dezelfde vaardigheid. Als toetsprestaties uit figuur 3 betrekking hebben op bijvoorbeeld in het ene geval technisch lezen en in het andere geval begrijpend lezen, is er natuurlijk geen sprake van groei of leerwinst. Dat zijn immers verschillende vaardigheden. Dat de toetsen onderling inhoudelijk vergelijkbaar zijn en dezelfde cognitieve vaardigheid meten is een essentiële voorwaarde om leerwinst te kunnen berekenen (Stevens & Zvoch, 2006; Hamilton, McCaffrey, & Koretz, 2006).

Er zijn verschillende manieren waarop leerwinst kan worden berekend en deze staan ook verschillende interpretaties van de uitkomsten toe (Center for Public Education, 2007; Ligon, 2008). Zo zijn er modellen die antwoord geven op vragen als 'hoeveel leerwinst is geboekt en is dit voldoende?'. Deze modellen beschrijven de feitelijke leerwinst en worden al dan niet vergezeld van een maat om deze te beoordelen. Voorts zijn er modellen die de toekomstige leerwinst voorspellen op basis van de huidige en eerdere prestaties van leerlingen of op basis van hun intelligentie. En tot slot zijn er modellen die ook interpretaties toelaten over de vraag of de leerwinst aan de school is toe te schrijven. Deze worden toegevoegde waarde modellen genoemd (Castellano & Ho, 2013).

Leerwinst wordt in de onderwijspraktijk, maar ook in beleidskringen, vaak gezien als een maat voor de toegevoegde waarde van een school. In strikte zin laat de leerwinst alleen zien hoeveel vooruitgang één of meer leerlingen hebben geboekt gedurende een bepaalde periode, maar daaruit is niet af te leiden in welke mate de leerwinst op het conto van de school geschreven kan worden. Het is onwaarschijnlijk dat alle groei in leerprestaties volledig is toe te schrijven aan het gevolgde onderwijs. Leerlingen leren immers niet alleen op school. Ook buiten school – in de thuisomgeving, bij vrienden of vriendinnen, op de sportclub – doen ze kennis, vaardigheden en ervaringen op die mede de leerprestaties op school bepalen (Bosker, 2012). Voorts zijn er op leerling-, klas- en schoolniveau heel veel factoren die een bijdrage leveren aan de leerprestaties, die ook nog eens van school tot school kunnen verschillen, zoals de persoonskenmerken van leerlingen, de didactisch aanpak, de samenstelling van klassen, enzovoort (Hattie & Anderman, 2013). Het is bovendien aannemelijk dat de bijdrage van niet-schoolse factoren aan de leerwinst niet op elk leer- en vormingsgebied even groot is. Dat komt doordat er vaardigheden zijn die ook door invloeden van buiten de school 'groeien', zoals woordenschat, terwijl andere

vaardigheden vooral op school aan de orde gesteld worden, zoals bewerkingen met kommagetallen. Zo spelen niet-schoolse factoren (zoals de thuissituatie, maar ook kenmerken van de buurt waarin een kind opgroeit) waarschijnlijk een grotere rol bij de verwerving van taalvaardigheden dan bij de verwerving van rekenvaardigheden.

Gezien de impact van interne en externe factoren op het leren op school heeft de toegevoegde waardebepaling van een school dan ook de nodige beperkingen. Bij de berekening ervan wordt vrijwel altijd gecorrigeerd voor *fairness variables*. Dat zijn leerlingkenmerken, zoals sociaal economische status (SES) en etniciteit en schoolcontextfactoren, zoals urbanisatiegraad en percentage gewichtleerlingen. Uit wetenschappelijk onderzoek weten we dat dergelijke kenmerken van invloed zijn op de leerprestaties van leerlingen, maar dat ze los staan van de kwaliteit van het geleverde onderwijs (OECD, 2008; p. 126). De aanname is dat wanneer de leerwinst wordt gecorrigeerd voor niet-schoolse factoren, de invloed van de schoolse factoren op de leerwinst zichtbaar wordt. Dit wordt ook wel schooleffect genoemd. Door deze correctie wordt geprobeerd de vergelijking van toegevoegde waarde tussen scholen zo eerlijk mogelijk te maken. In hoeverre op deze manier daadwerkelijk een faire vergelijking wordt gemaakt, hangt sterk af van de mate waarin men kan controleren voor niet-schoolse factoren. Enerzijds is het praktisch niet mogelijk om alle relevante achtergrond- en contextkenmerken te meten en te betrekken in de berekeningen. Anderzijds moet worden vermeden dat voor kenmerken wordt gecorrigeerd die scholen wel degelijk kunnen beïnvloeden, zoals bijvoorbeeld motivatie om te leren of ondersteuning door ouders.

3.3. Leerwinstbepaling

Er zijn verschillende modellen in omloop om de leerwinst te meten, variërend van eenvoudige vergelijkingen van toetsprestaties tot en met complexe statistische analyses, waarbij ook wordt nagegaan of de leerwinst toe te schrijven is aan het gegeven onderwijs. In dat verband worden de begrippen 'leerwinst' en 'toegevoegde waarde' wel eens door elkaar gebruikt, ofschoon er duidelijke verschillen zijn voor wat betreft het doel waarvoor de modellen ontwikkeld zijn en de statistische analyses die worden toegepast. Toegevoegde waarde modellen zijn bedoeld om de bijdrage van de school aan de leerwinst zichtbaar te maken en daarvoor zijn complexe statistische analyses noodzakelijk. Bij leerwinst gaat het alleen om de vraag of er prestatiegroei is en of deze in overeenstemming is met bepaalde verwachtingen.

Groei kan uitgedrukt worden in een absolute of in een relatieve maat. Een voorbeeld van een absolute maat is de wijze waarop de fysieke groei van kinderen wordt uitgedrukt. Dat gebeurt bijvoorbeeld in centimeters door de lichaamsgroei op twee momenten te vergelijken. Ook voor de groei in leerprestaties is een absolute maat mogelijk. Deze kan bepaald worden door de prestatie-ontwikkeling tussen twee meetmomenten vast te stellen. De absolute leerwinst is dan het verschil tussen de twee scores die op beide meetmomenten zijn behaald, bijvoorbeeld de scores op een vaardigheidsschaal (Betebenner & Linn, 2009). Het is echter lastig om op basis van absolute leerwinst een norm te bepalen waaruit afgeleid kan worden of de leerwinst 'voldoende', 'naar behoren' of 'naar verwachting' is (DePascale, 2006).

Bij het vaststellen van de relatieve leerwinst wordt de prestatiegroei van een leerling vergeleken met de prestaties van overeenkomstige leerlingen (Martin, 1985; Smith & Yen, 2006). Dat kan bijvoorbeeld op basis van een verdeling van leerwinstscores van een representatieve steekproef van leerlingen, zoals leerjaargenoten of leeftijdsgenoten, die dienst doet als referentiegroep. De leerwinst per leerling kan dan worden uitgedrukt in de afwijking van de gemiddelde leerwinst van de referentiegroep of via een percentielscore die de relatieve plaats van een leerling aangeeft in de totale verdeling van leerwinstscores binnen de referentiegroep. In het eerste geval gaat het om de vraag in hoeverre de leerwinst boven of onder het gemiddelde van de referentiegroep ligt. In het tweede geval geeft de percentielscore een indicatie van het percentage leerlingen uit de referentiegroep die dezelfde of een betere leerwinst heeft behaald.

3.4. Toegevoegde waarde bepaling

De wetenschappelijke literatuur over toegevoegde waarde is niet altijd even scherp over wat daar nu precies onder verstaan moet worden. Er zijn ruwweg twee manieren waarop toetsresultaten worden gebruikt om de bijdrage van de school aan de leerprestaties te beoordelen. Dat zijn aan de ene kant modellen waarbij de schoolprestaties worden beoordeeld op basis van één enkele gecorrigeerde toetsprestatie van één groep leerlingen. Dit wordt in de literatuur wel s*tatusmeting* of *'test score snapshot'* genoemd (Harris, 2011, p. 43). Aan de andere kant zijn er modellen waarin meer toetsprestaties van hetzelfde leerlingcohort worden betrokken, al dan niet gecorrigeerd voor leerling- en/of schoolkenmerken. Deze worden in de literatuur aangeduid als *groei-* of *toegevoegde waarde modellen* (Braun, Chudowsky & Koenig, 2010). De Engelse inspectie gebruikt de term *value added* voor situaties waar ze een begintoets en een eindtoets aan elkaar kunnen relateren. De term *contextual value added* wordt gebruikt als naast een relatering van een begin- aan een eindtoets ook gecorrigeerd wordt voor leerlingkenmerken. Kenmerkend voor deze toegevoegde waarde modellen is het gebruik van tenminste een begin- en een eindtoets afgenomen bij dezelfde groep leerlingen. De basis voor deze modellen is dus leerwinst.

De meeste modellen schatten de toegevoegde waarde van een school op basis van het verschil tussen de waargenomen leerwinst en de leerwinst die verwacht zou mogen worden na correctie voor leerlingkenmerken. Het gaat daarbij om leerlingkenmerken die van invloed kunnen zijn op de vooruitgang in prestaties maar waarop de school zelf geen invloed heeft.

De belangrijkste kritiek op toegevoegde waarde modellen is dat, ondanks het gebruik van meerjarige toetsgegevens van dezelfde groep leerlingen en ondanks de correctiefactoren, er geen *causaal* verband gelegd kan worden tussen de leerwinst en het effect van het onderwijs daarop (Raudenbush, 2004[6]). Er is extra informatie nodig over de invloed van bijvoorbeeld onderwijsprocessen om de feitelijke bijdrage van de school aan de leerwinst te onderzoeken. (Braun, Chudowsky & Koenig, 2010; Baker et al., 2010; Harris, 2009). Toegevoegde waarde-schattingen geven daarom op zijn best een indicatie van de mogelijke bijdrage van de school aan de leerprestaties.

Toegevoegde waarde op basis van statusmetingen

Er zijn modellen waarin alleen de status van de prestaties op één toetsmoment betrokken wordt, bijvoorbeeld een eindtoets in het hoogste leerjaar (Harris, 2011). Bij deze aanpak wordt de gemiddelde schoolscore op een enkele toets 'aangepast' door rekening te houden met verschillen tussen leerlingpopulaties. Op scholen met bijvoorbeeld relatief veel leerlingen van hoog opgeleide ouders - een gunstige leerlingpopulatie - wordt de gemiddelde schoolscore naar beneden bijgesteld. Op scholen met relatief veel leerlingen van laag opgeleide ouders, gebeurt het omgekeerde. Deze aanpak is in Nederland - recent - door professor J. Dronkers (Dronkers, 2013) toegepast op basis van eindtoetsgegevens over drie afzonderlijke jaren (zie ook par. 4.3.4). In de literatuur wordt deze aanpak gezien als een eerste stap in de richting van een toegevoegde waarde model, omdat er nog geen rekening wordt gehouden met de leerwinst over een langere periode of met de uitgangssituatie van dezelfde groep leerlingen. Dit zijn echter wel belangrijke voorwaarden voor het bepalen van de toegevoegde waarde van een school. Het gaat daarbij immers om de vraag naar de bijdrage van de school aan de leerprestaties.

Toegevoegde waarde op basis van leerwinst

De meeste toegevoegde waarde modellen gaan uit van tenminste twee toetsprestaties van dezelfde groep leerlingen (Harris, 2011; Timmermans, 2012). De basis voor de schatting van toegevoegde waarde zijn gegevens over de leerwinst van leerlingen van een bepaalde school, die vervolgens wordt vergeleken met de leerwinst van leerlingen van vergelijkbare scholen. In de meeste modellen wordt de leerwinst gecorrigeerd voor school- en leerlingkenmerken, zoals bijvoorbeeld de samenstelling van de schoolbevolking (etniciteit, geslacht, socio-economische status en type zorgleerlingen). Door deze correctie wordt het effect van de school zichtbaar.

Er zijn ook toegevoegde waarde modellen waarbij geen correcties plaatsvinden voor achtergrondkenmerken van leerlingen of scholen. De reden daarvoor is dat de leerlingen, vanwege het gebruik van tenminste twee toetsgegevens van dezelfde groep, in feite dienst doen als hun eigen controle op eerdere prestaties. In de beginmeting is namelijk al verdisconteerd dat

[6] "Value-added indicators correct some of the problems of indicators based on mean proficiency. They hold schools accountable for the learning that a student exhibits while under the care of the school. This has a strong intuitive appeal, and yet the value-added approach is also open to cogent criticism. Thus both methods - those based on mean proficiency and those based on value added - produce estimates with considerable uncertainty and some unknown bias. The logical thing to do in the presence of uncertainty is to seek more information. It is plausible to assume that parents and educators would like to know both how much their children know at a given time and how fast they are learning, based on the best available tests. Yet decisive and effective action to improve schools requires more information, including information gleaned from expert judgment" (Raudenbush, 2004, p. 36-37).

de prestaties zijn beïnvloed door de kenmerken van die leerlingen. Omdat in deze modellen verder niet gecorrigeerd wordt voor niet-schoolse factoren, is het de vraag of deze modellen de bijdrage van de school aan de leerwinst voldoende zichtbaar maken.

Wel of geen correctie?

Er is in de literatuur veel discussie over het nut van het betrekken van leerling- en andere kenmerken in de toegevoegde waarde modellen (McCaffrey, Lockwood, Koretz, Louis & Hamilton, 2004). Dat kunnen basale demografische kenmerken zijn, zoals geslacht, etnische herkomst en socio-economische status, maar ook meer geavanceerde kenmerken zoals begaafdheid, handicaps of andere cognitieve aspecten die een rol spelen in het leren op school. Ofschoon het betrekken van zoveel mogelijk achtergrondkenmerken in een toegevoegde waarde model van dienst kan zijn om de relatie tussen onderwijs en leren zichtbaar te maken, geldt ook hier dat overdaad schaadt. Daarom is er veel studie gedaan naar de vraag welke achtergrondkenmerken ertoe doen.

Sommige wetenschappers (zie Castellano & Ho, 2013) gaan ervan uit dat demografische kenmerken, zoals geslacht of etnische herkomst, een constante en directe invloed hebben op de toetsresultaten en dat daarvoor dus niet gecorrigeerd moet worden. Deze kenmerken beïnvloeden namelijk de hoogte van de toetsprestatie. Als deze toetsprestatie vervolgens in een analysemodel wordt gecorrigeerd voor demografische kenmerken, dan gebeurt dat als het ware twee keer, namelijk eerst in het toetsresultaat en daarna in het model. Een neveneffect van correctie is dat er feitelijk verschillende prestatieverwachtingen voor verschillende leerlingpopulaties worden gecreëerd, zoals voor leerlingen met een hoge en voor leerlingen met een lage socio-economische status. Tegenstanders beschouwen dit als ongewenst.

Anderen betogen (zie Castellano & Ho, 2013) dat scholen voor verschillende (deel)populaties verschillende opbrengsten produceren en dus ook op dit punt qua effectiviteit van elkaar kunnen verschillen. Door dit per (deel)populatie zichtbaar te maken (en er dus niet voor te corrigeren) wordt ook duidelijk of de effectiviteit van scholen voor bepaalde groepen leerlingen verbeterd kan worden. Daarom wordt in sommige gevallen niet statistisch gecorrigeerd voor achtergrondkenmerken, maar worden wel de verschillen per (deel)populatie gerapporteerd.

3.5. Kwaliteitsfactoren

Berekeningen met betrekking tot leerwinst of toegevoegde waarde zijn alleen dan zinvol als ze zijn gebaseerd op valide gegevens. Dit houdt in dat er kwalitatief goede toetsen volgens de voorschriften op de juiste momenten worden afgenomen bij alle leerlingen waarvoor de toets bedoeld is. Hierbij komt nog dat de meting van de leerwinst en toegevoegde waarde betrouwbaarder wordt naarmate deze is gebaseerd op meer dan twee metingen per leerling verspreid in de tijd. Deze metingen moeten dan wel inhoudelijk goed vergelijkbaar zijn.
Verder zijn er nog andere kwesties die de betrouwbaarheid en de validiteit van leerwinst en toegevoegde waarde in gevaar kunnen brengen, zoals de selectie van correctiefactoren, kleine aantallen leerlingen waarover de leerwinst of toegevoegde waarde wordt berekend, de (in)stabiliteit van de uitkomsten en ontbrekende leerlinggegevens vanwege tussentijdse in- en uitstroom (zie bijv. Center for Public Education, 2007). Hieronder wordt puntsgewijs nader

ingegaan op factoren die van invloed zijn op de kwaliteit van maten voor leerwinst en toegevoegde waarde (zie ook Willms, 2008).

Meetfouten

Hoe professioneel toetsen ook zijn ontwikkeld, er dient bij de interpretatie altijd rekening te worden gehouden met meetfouten en andere onzekerheden. De lengte van kinderen kunnen we tot op de millimeter nauwkeurig meten. We hebben daarvoor goede meetapparatuur tot onze beschikking. De meting van de lengte is in het algemeen betrouwbaar en valide. Dat geldt echter niet voor het meten van leerprestaties. Toetsen hebben nu eenmaal een bepaalde mate van imperfectie. De leerling kan een 'slechte dag' hebben gehad, het taalgebruik in de toets kan te moeilijk zijn, de toets dekt niet precies de inhoud van het gegeven onderwijs, enzovoort. Zelfs bij de meest betrouwbare toets is sprake van een bepaalde meetfout waardoor de behaalde score niet precies de 'echte' score hoeft te zijn. De invloed van de meetfout op de toetsprestaties is overigens niet voor alle leerlingen hetzelfde: deze is groter voor leerlingen met een lage of een hoge score dan voor leerlingen met een gemiddelde score (National Research Council & National Academy of Education, 2010).

Met deze meetfout, maar ook met andere onzekerheden, moeten leerwinst- en toegevoegde waarde modellen rekening houden. Veel onderzoekers lossen dat op door de standaardmeetfout te melden of door de uitkomsten in betrouwbaarheidsintervallen (de boven- en ondergrens waarbinnen de werkelijke uitkomst waarschijnlijk ligt) te publiceren, zoals het geval is bij een meerdaagse weersvoorspelling. Sommige modellen houden rekening met de standaardmeetfout waardoor de leerwinst nauwkeuriger kan worden berekend.

Controle voor achtergrondkenmerken

De nauwkeurigheid van schattingen van de toegevoegde waarde van scholen wordt in hoge mate bepaald door de correctiefactoren die in het model worden betrokken. Zo kan het voorkomen dat ambitieuze ouders eerder een school voor hun kind kiezen die kwalitatief goed aangeschreven staat dan minder ambitieuze ouders. Een deel van de toegevoegde waarde heeft dan te maken met dat ambitieniveau, dat op zijn beurt verweven is met de kwaliteit van de school. Bij de berekeningen van de toegevoegde waarde van een school zal voor de belangrijkste buitenschoolse invloeden op de leervorderingen gecorrigeerd moeten worden (fairness-kenmerken). In de periode van acht jaar die een modale leerling nodig heeft om de basisschool te doorlopen kunnen allerlei gebeurtenissen optreden waar de school geen invloed op heeft, maar die wel een sterke invloed kunnen hebben op de leerprestaties van de leerlingen en daarmee op de gerealiseerde leerwinst of de toegevoegde waarde. Hierbij kan bijvoorbeeld gedacht worden aan verandering in de gezinssituatie, van de samenstelling van de wijk of buitenschools leren, zoals zomerscholen. Dit zou in ieder geval verdisconteerd moeten worden bij de beoordeling van de toegevoegde waarde van een school. Het toevoegen van extra gegevens over leerling- en contextkenmerken in het model kan de nauwkeurigheid van de schatting verhogen, maar kan ook leiden tot meer complicaties zoals ontbrekende gegevens (OECD, 2008).

Kleine scholen

Een andere uitdaging die te maken heeft met meetfouten en onzekerheden wordt veroorzaakt door de aantallen leerlingen waarvoor de leerwinst en toegevoegde waarde wordt berekend. Als het aantal leerlingen van een school daalt of de eenheid van analyse wijzigt (bijvoorbeeld van schoolniveau naar groepsniveau), neemt de onzekerheid van de leerwinstmaat of de toegevoegde waardeschatting toe (Schochet & Hanley, 2010). Het gevolg is dat de leerwinst op groepsniveau of de toegevoegde waarde op schoolniveau per jaar sterk kan wisselen. Onderzoek naar de precisie van toegevoegde waarde modellen laat zien dat bij scholen met een klein aantal leerlingen de standaardmeetfout vaak zo groot is dat de toegevoegde waarde van deze scholen niet zichtbaar is (McCaffrey & Lockwood, 2008).

Stabiliteit van schattingen

Uit onderzoek naar toegevoegde waarde modellen blijkt dat de uitkomsten niet stabiel zijn. Alle modellen produceren schattingen die per school van jaar tot jaar verschillen (McCaffrey, Sass & Lockwood, 2008). Veel van deze modellen zijn thans nog in ontwikkeling en leveren niet steeds dezelfde uitkomsten (Goldschmidt et al., 2005; Reardon, & Raudenbush, 2009; Harris, 2011; Timmermans (2012). Dit roept de vraag op of hier sprake is van een artefact van het model of dat de prestaties van scholen van jaar tot jaar daadwerkelijk verschillen. Dit verschijnsel tast het vertrouwen aan in toegevoegde waarde modellen, omdat leraren en ook onderzoekers er vanuit gaan dat de kwaliteit van het onderwijs langzaam verandert in plaats van jaar tot jaar. Als oplossing voor dit probleem wordt er vaak voor gekozen om gegevens over meerdere cohorten te middelen.

Ontbrekende gegevens

Idealiter hebben alle leerlingen voor wie de leerwinst berekend moet worden een complete toetsgeschiedenis. Maar in de praktijk is dit lang niet altijd het geval: leerlingen stromen tussentijds in of uit, hebben niet altijd mee gedaan aan alle toetsafnames of de school is overgegaan op een nieuwe versie van de toetsen waardoor vergelijking met de oude toetsresultaten vaak niet meer mogelijk is. Daarnaast vorderen niet alle leerlingen op gelijke wijze: sommige leerlingen doubleren en maken dezelfde toets een jaar later nog een keer. Dit heeft de nodige gevolgen voor de leerwinstberekening als grondslag voor een schatting van de toegevoegde waarde[7]. Op leerlingniveau is het dan niet altijd mogelijk om tot leerwinstberekening te komen. Dat geldt ook voor de berekening van de leerwinst op groeps- of schoolniveau. Enerzijds kan er sprake zijn van te geringe aantallen om tot een betrouwbare maat te komen. Anderzijds is het de vraag of de inspanningen van de school op een juiste wijze gemeten kunnen worden als niet alle leerlingen hun schoolloopbaan op dezelfde school zijn begonnen of hebben voltooid.

Incomplete leerlinggegevens moeten wel zoveel mogelijk worden meegenomen bij de berekeningen van de leerwinst en toegevoegde waarde. Dat dergelijke gegevens ontbreken is een belangrijke beperking van de validiteit van toegevoegde waarde, omdat dat vaak de gegevens van zwakkere leerlingen betreft (Rubin et al., 2004; Timmermans, 2012). Het schatten

[7] Het vaardigheidsverschil-model is daar gevoeliger voor dan het vaardigheidsgroei-model.

van toegevoegde waarde op basis van enkele leerlingen met volledige gegevens resulteert daardoor in een overschatting van de prestaties van de leerlingen en de school (Thomas et al., 1997; Timmermans, 2012). Daarom wordt bij de modellen gebruik gemaakt van bepaalde beslisregels en procedures. Soms leiden ontbrekende gegevens tot uitsluiting van de betreffende leerlingen, soms worden de ontbrekende gegevens geschat en vervangen door een beredeneerde score.

Strategisch gedrag

Alom bestaat de vrees dat naarmate leerwinst en toegevoegde waarde een rol gaan spelen in de externe beoordeling van scholen, scholen hun gedrag hierop gaan afstemmen. De ideale situatie is natuurlijk dat alle scholen op een integere wijze de vorderingen van hun leerlingen monitoren. We kunnen onze ogen echter niet sluiten voor het risico dat scholen strategisch gedrag gaan vertonen of zelfs fraude plegen, zeker als indicatoren voor leerwinst en toegevoegde waarde gebruikt worden voor accountability-doeleinden. Scholen zouden met name dan in de verleiding kunnen komen om ervoor te zorgen dat hun leerlingen op de begintoets erg lage scores halen. Onderstaande uitspraak van Donald Campbell (Campbell, 1976), ook bekend als Campbell's law, lijkt bij uitstek van toepassing op het gebruik van gestandaardiseerde toetsen om scholen en leerkrachten af te rekenen op de prestaties van hun leerlingen: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor".

Ervaringen in het binnen- en buitenland laten zien dat zich diverse vormen van strategisch gedrag kunnen voordoen als scholen en leerkrachten zich sterk onder druk gezet voelen om goede resultaten te produceren. Zo bestaat de vrees dat scholen geen toetsen afnemen bij leerlingen waarvan verwacht wordt dat ze de schoolprestaties negatief beïnvloeden. Dit verschijnsel wordt 'reshaping the testpool' genoemd.

Uit onderzoek in de VS is bekend dat slecht presterende leerlingen met een lage sociaal-economische status vaker worden 'vrijgesteld' van toetsen na introductie van een accountabilitysysteem. Ook is geconstateerd dat het preventief zittenblijven toenam als gevolg van de introductie van het accountabilitybeleid. Een ander verschijnsel dat gericht is op het realiseren van zo hoog mogelijke leerprestaties, is het overaccentueren van of beperken van het onderwijs tot de toetsinhouden. Dit verschijnsel staat bekend als 'teaching to the test'. Bovenstaande verschijnselen doen zich vooral voor bij toetsen waarvan de uitslagen openbaar worden gemaakt en in een context waarin scholen op die prestaties worden 'afgerekend' (De Wolf & Janssens, 2007).

3.6. Conclusies

Leerwinst heeft betrekking op het meten van de prestatiegroei gedurende een bepaalde periode in de schoolloopbaan van een leerling of van een groep leerlingen. Om deze te meten dient gebruik te worden gemaakt van toetsen waarmee de prestatiegroei ook als leerwinst in een bepaald leerstofgebied kan worden geïnterpreteerd. De meeste eenvoudige vorm van leerwinstbepaling is de berekening van het verschil tussen de scores van twee toetsafnames. Dit wordt *absolute leerwinst* genoemd. Het verschil tussen deze twee scores geeft wel groei aan, maar beantwoordt niet de vraag of deze groei voldoende is of niet. Bij gebrek aan absolute normen die aangeven hoeveel groei verwacht mag worden, wordt in de meeste gevallen uitgeweken naar een vorm van *relatieve leerwinst*. Hiermee wordt bedoeld dat de groei van een leerling of van een groep leerlingen wordt vergeleken met die van gelijkwaardige leerlingen. Deze relatieve vergelijking wordt bijvoorbeeld ook toegepast bij de beoordeling van de lichaamsgroei en het lichaamsgewicht van kinderen.

Ofschoon in de wetenschappelijke literatuur doorgaans onder toegevoegde waarde de bijdrage van een school aan de leerprestaties wordt verstaan, is er niet altijd overeenstemming over de wijze waarop deze kan worden vastgesteld. Er zijn modellen die uitgaan van een enkele toetsprestatie van een groep leerlingen en er zijn aanpakken die ook eerdere toetsprestaties in het model opnemen. Daarnaast zijn er modellen waarbij de toetsprestaties worden gecorrigeerd voor relevante leerling- en schoolkenmerken en er zijn modellen waarin dat niet gebeurt.

Om uitspraken mogelijk te maken over de bijdrage van de school aan de leerprestaties van hun leerlingen, moet op zijn minst bekend zijn wat het instroom- en het uitstroomniveau van de leerlingen is. Om de toegevoegde waarde van een school zo eerlijk mogelijk te schatten is het van belang dat de leerprestaties worden gecorrigeerd voor factoren waarop de school geen invloed heeft, zoals de kenmerken van de schoolbevolking. Modellen die enerzijds uitgaan van de prestatiegroei van een cohort en anderzijds corrigeren voor leerlingkenmerken van dat cohort, voldoen dus aan twee belangrijke voorwaarden om uitspraken te doen over de invloed van de school op de leerprestaties. Door meer meetmomenten dan alleen een begin- en een eindmeting in het model op te nemen, zijn dergelijke modellen ook beter bestand tegen allerlei factoren die de betrouwbaarheid en validiteit van de uitkomsten kunnen bedreigen, zoals instabiliteit en ontbrekende gegevens.

4. Nederlandse toepassingen van leerwinst en toegevoegde waarde

4.1. Inleiding

In de afgelopen decennia zijn met name in de Angelsaksische landen modellen voor leerwinst en toegevoegde waarde ontwikkeld. Deze spelen voornamelijk een rol in verschillende accountabilitysystemen, omdat deze indicatoren de belofte in zich dragen een eerlijke vergelijking mogelijk te maken tussen de prestaties van verschillende scholen. Voorbeelden van toegevoegde waarde indicatoren die op dit moment worden gebruikt zijn Contextualized Value Added (Ray, 2006; Ofsted, 2010), Tennessee Value Added Assessment System (Sanders et al., 1994; Sanders, 2003) en het Colorado Growth Curve Model (Betebenner, 2007; Betebenner, 2009).

Ook in Nederland zijn verschillende pogingen ondernomen om leerwinst en toegevoegde waarde een praktische invulling te geven. Er is in ons land eerder onderzoek gedaan naar de wijze waarop leerwinst en toegevoegde waarde kan worden berekend. Dit onderzoek wordt in paragraaf 4.2 samengevat. Toepassingen van leerwinst en toegevoegde waarde in Nederland worden in paragraaf 4.3 beschreven. Dat zijn enerzijds leerwinstberekeningen op basis van bijvoorbeeld schoolgemiddelde vaardigheidsscores, vaardigheidsniveaus en intelligentie. Deze kunnen in het onderwijs worden gebruikt. Anderzijds hebben we in ons land ook ervaring opgedaan met toegevoegde waardeberekeningen zoals die van prof. J. Dronkers. Op basis daarvan publiceerde RTL Nieuws in 2013 een ranglijst van basisscholen.
In paragraaf 4.4 worden de bestaande toepassingen afgezet tegen de leerwinst- en toegevoegde waardematen die in de pilot worden ontwikkeld.

4.2. Onderzoek

Op basis van een notitie uit 2007 van prof. dr. J. Peschar (Peschar, 2007) waarin wordt ingegaan op verschillende manieren om de opbrengsten van het onderwijs te meten, voerde het SCO-Kohnstamm Instituut een verkenning uit naar de mogelijkheden om leerwinst en toegevoegde waarde te berekenen, gebruikmakend van toetsen uit het LVS van Cito (Roeleveld, Van der Veen & Ledoux, 2008). Men gebruikte daarvoor de longitudinale data van het PRIMA-cohortonderzoek en bespraken de mogelijkheden die het COOL-cohortonderzoek[8] biedt. Leerwinstberekeningen vonden plaats op basis van verschillen in vaardigheidsscores op taal- en rekentoetsen bij een cohort basisschoolleerlingen, gebruikmakend van vaardigheidsscores behaald in de groepen 4, 6 en 8. De toegevoegde waarde (in het rapport Groeimodellen genoemd) is op drie manieren berekend. Deze manieren verschillen van elkaar in de mate

[8] Tussen 1994 en 2005 hebben het ITS en het SCO-Kohnstamm Instituut het PRIMA-cohortonderzoek uitgevoerd, waarmee het Nederlandse basisonderwijs periodiek in kaart is gebracht. Op de circa 600 scholen die aan PRIMA deelnamen, werden om de twee jaar taal- en rekentoetsen afgenomen bij de leerlingen in de groepen 2, 4, 6 en 8, en werden vragenlijsten voorgelegd aan de leerlingen en hun ouders, leerkrachten en directeuren. In totaal zijn zes PRIMA-metingen uitgevoerd. De meeste PRIMA-scholen hebben aan meerdere metingen deelgenomen, waardoor het mogelijk was om ontwikkelingen in de tijd te volgen. Vanaf schooljaar 2007-2008 is PRIMA opgevolgd door een nieuw cohortonderzoek: COOL5-18. Dit onderzoek volgt leerlingen van 5 tot 18 jaar in hun schoolloopbaan door het primair en voortgezet onderwijs en het mbo.

waarin de leerwinst werd gecorrigeerd voor etnische herkomst, sekse, opleidingsniveau van ouders en leeftijd.

De onderzoekers komen tot de conclusie dat voor de relatief eenvoudige vraag “wat kunnen en kennen Nederlandse leerlingen op een bepaald moment?” referentieniveaus, PPON-metingen of indicatoren uit internationale onderzoekingen als PISA en TIMMS gebruikt worden. Voor meer complexe vragen als de bijdrage van het onderwijs aan de ontwikkeling van leerlingen moeten ook complexere dataverzamelingen en analysemethoden gebruikt worden. Men pleit er voor longitudinale cohortdata te gebruiken. Daarmee kan de groei van alle of van onderscheiden groepen leerlingen in het onderwijs worden gemodelleerd om de bijdrage van het onderwijs aan die groei zo goed mogelijk te onderscheiden van allerlei buitenschoolse factoren.

Timmermans (2012) onderzocht de mogelijkheden van toegevoegde waarde-indicatoren voor het Nederlandse onderwijstoezicht. Hoewel toegevoegde waarde in verschillende landen al wordt gebruikt als indicator voor opbrengsten binnen onderwijstoezicht, zijn er vele verschillende manieren waarop de toegevoegde waarde-indicatoren binnen deze toezichtsystemen worden geoperationaliseerd. De verschillen in operationalisering komen met name tot uiting in de keuze van het statistische model en de keuze van de controlevariabelen die worden gebruikt voor het maken van een eerlijke vergelijking tussen scholen.

De toegevoegde waarde-indicatoren die op dit moment binnen internationale accountabilitysystemen worden gebruikt verschillen zeer sterk in de keuze voor controlevariabelen. In het Tennessee Value Added Assessment System, een indicator die in verschillende staten van de Verenigde Staten wordt toegepast, wordt alleen gecontroleerd voor verschillen tussen leerlingen in beginniveau. Bij Contextual Value Added, dat wordt gebruikt door de Engelse onderwijsinspectie, wordt daarentegen een veel uitgebreidere set van controlevariabelen gebruikt. In dit model worden het beginniveau, achtergrondkenmerken van de leerlingen en kenmerken van de schoolcompositie opgenomen als controlevariabelen.

Het onderzoek van Timmermans (2012) was gericht op de ontwikkeling van toegevoegde waarde-indicatoren voor verschillende onderwijssectoren en op de bruikbaarheid ervan voor een eerlijke vergelijking van de effectiviteit van scholen. In het onderzoek zijn drie typen data gebruikt (cohortdata, onderwijsnummer-data en data uit leerlingvolgsystemen). Op basis van deze drie typen data bleek het mogelijk om de toegevoegde waarde van scholen te schatten. Echter, hiervoor moet data beschikbaar zijn op het niveau van de leerling en moet er minimaal een indicatie van het begin- en eindniveau aanwezig zijn. De drie typen data die in de verschillende studies zijn gebruikt hebben elk voor- en nadelen en leiden soms tot verschillende interpretaties van de toegevoegde waarde-indicator. De data afkomstig uit leerlingvolgsystemen is geschikt voor het schatten van de toegevoegde waarde op basis van leerwinst. In deze data zijn er op verschillende momenten indicaties van de prestaties van leerlingen gemeten op eenzelfde vaardigheidsschaal. Dit biedt de mogelijkheid om de groei van leerlingen in kaart te brengen. Toegevoegde waarde op basis van een groeimodel kan daardoor geïnterpreteerd worden als een indicator voor de relatieve groei/ontwikkeling van leerlingen in een school in vergelijking tot andere scholen.

Een tweede algemene bevinding is dat de schatting van de toegevoegde waarde van een school samengaat met een relatieve grote onbetrouwbaarheid. Door deze grote onbetrouwbaarheid kan men slechts drie groepen scholen van elkaar onderscheiden, namelijk ineffectieve, gemiddelde en effectieve scholen. Deze onbetrouwbaarheid is volgens Timmermans (ibid, p. 198) niet uniek voor toegevoegde waarde, maar geldt voor alle prestatie-indicatoren van scholen.

Een terugkerend probleem in de verschillende studies van Timmermans was het ontbreken van gegevens bij verschillende controlevariabelen. Dat gegevens ontbreken is een belangrijke beperking voor de validiteit van toegevoegde waarde, omdat dat vaak de gegevens van zwakkere leerlingen betreft. Het schatten van toegevoegde waarde op basis van leerlingen met volledige gegevens resulteert daardoor in een overschatting van de prestaties van de leerlingen en de school.

4.3. Praktische toepassingen

Er is niet alleen in Nederland onderzoek gedaan naar de toepasbaarheid van leerwinst en toegevoegde waarde, er zijn ook toepassingen beschikbaar voor het onderwijsveld. Deze zijn als volgt in te delen:

1. leerwinst op basis van schoolgemiddelde vaardigheidsscores;
2. leerwinst op basis van vaardigheidsniveaus;
3. leerwinst op basis van verwachtingen;
4. toegevoegde waarde op basis van Eindtoetsscores.

4.3.1. Leerwinst op basis van schoolgemiddelde vaardigheidsscores

De gemeente Amsterdam heeft in 2012 voor de derde keer de Kwaliteitswijzer Basisonderwijs Amsterdam uitgegeven. Op de online versie van de Kwaliteitswijzer is het mogelijk om de ontwikkeling van de scholen over het schooljaar 2010-2011 te zien. Verder wordt per school een overzicht gegeven van de beoordeling van de onderwijsinspectie, de score op de Cito-eindtoets en de adviezenverdeling in groep 8 en de niveauverdeling van de tussentijdse taal- en rekentoetsen van groep 2 tot en met groep 6. Van de 209 basisscholen in de stad leverden 205 scholen de benodigde informatie aan. De papieren versie van de Kwaliteitswijzer is verspreid onder 10.000 ouders die op het punt staan een basisschool voor hun kind te kiezen. In de kwaliteitswijzer is ook informatie te vinden over het ‘gemiddelde resultaat van vergelijkbare school’ bij de Cito-eindtoets, de adviezen en de beoordelingen van tussentijdse taal- en rekentoetsen. Hiermee wil men een indicatie geven van de toegevoegde waarde van een school voor deze leerprestaties (Kwaliteitswijzer Basisonderwijs Amsterdam 2010-2011, p 5).

Berekeningswijze gemiddelde resultaat van vergelijkbare school

In de Kwaliteitswijzer Basisonderwijs Amsterdam van het schooljaar 2011-2012 is ook het ‘gemiddelde resultaat van vergelijkbare school’ als indicator opgenomen. Deze is bepaald door de scholen in te delen in vijf schoolgroepen op basis van een zestal achtergrondkenmerken van de leerlingen: opleiding ouders, inkomen ouders, herkomst, soort woning (huur/koop), ozb-waarde van de woningen, vve-indicatie. Dit vond men later een te grove indeling. Daarom is voor de Kwaliteitswijzer van 2012-2013 gekozen voor een verfijnde aanpak.

Voor de Kwaliteitswijzer Basisonderwijs Amsterdam 2012-2013 zijn door het CBS van alle leerlingen op alle Amsterdamse basisscholen gegevens verstrekt aan het bureau Onderzoek en Statistiek van de gemeente Amsterdam. Het gaat om opleidingsniveau ouders, huishoudeninkomen, doelgroepleerling voor vve en etnische herkomst. Dit wordt gekoppeld aan de geboortedatum van de leerling en de schoolidentificatiegegevens. Met behulp van een factoranalyse is op basis van de leerlingkenmerken een gemeenschappelijk onderliggend kenmerk berekend: de factor 'context'. Deze factor geeft de gemiddelde context van leerlingen op Amsterdamse basisscholen weer, waarbij rekening gehouden is met verschillen in belangrijkheid tussen de leerlingkenmerken (factorladingen). Vervolgens is de gemiddelde factorscore per school bepaald. Dit nieuwe schoolkenmerk is als voorspeller van leerprestaties meegenomen in lineaire regressieanalyses op het niveau van scholen. Zo is een voorspelling van het schoolgemiddelde op de Cito-eindtoets en de schooladviezen in groep 8 verkregen, gegeven de schoolcontext. Door het feitelijke schoolgemiddelde op bijvoorbeeld de Cito-eindtoets te vergelijken met het voorspelde schoolgemiddelde op Cito-eindtoets, wordt inzicht verkregen in de kwaliteit van de basisschool. Een goede basisschool is een school die het beter doet dan verwacht. Omgekeerd, een zwakke basisschool doet het slechter dan verwacht. Op deze manier kunnen de opbrengsten van Amsterdamse basisscholen op een betere manier met elkaar vergeleken worden dan wanneer geen rekening gehouden zou worden met de sociaal-etnische achtergrond van de leerlingen.

Leerwinstberekening

In de Amsterdamse Kwaliteitswijzer 2012-2013 wordt, naast de gecorrigeerde gemiddelde schoolscore op de Eindtoets en de schooladviezen van groep 8, voor de eerste keer de leerwinst van een school op de tussentijdse taal- en rekentoetsen gepresenteerd. Deze is niet gebaseerd op individuele vaardigheidsscores, maar op de gemiddelde vaardigheidsscore van een groep leerlingen. De leerwinst wordt berekend door van twee opeenvolgende schooljaren het verschil in toetsresultaat (gemiddelde vaardigheidsscore op een bepaald leerstofgebied) te berekenen. Dit geeft aan wat de groei is die de groep leerlingen heeft doorgemaakt. Vervolgens wordt deze groei vergeleken met de gemiddelde groei op die toets van alle Amsterdamse scholen. Is deze (veel) hoger dan gemiddeld dan krijgt de school voor die toets een opwaartse pijl. Ligt de groei rond het gemiddelde in Amsterdam dan krijgt de school een horizontale pijl voor die toets. Is de groei (veel) lager dan gemiddeld dan wordt een neerwaartse pijl weergegeven.

4.3.2. Leerwinstberekeningen op basis van vaardigheidsniveaus

De Loos Monitoring (www.deloos.net) biedt aan scholen voor primair onderwijs een service waarbij tegen betaling leerwinstberekeningen worden samengesteld. De leerwinstmethode van De Loos Monitoring is bedoeld voor scholen die gebruik maken van de LVS-toetsen van Cito. De leerwinst wordt berekend op basis van verschillen en verschuivingen in zogenaamde vaardigheidsniveaus van leerlingen (zie paragraaf 2.5).

De leerwinst van De Loos Monitoring wordt vastgesteld op basis van de samenvoeging van toetsen voor verschillende leerstofdomeinen tot drie onderdelen: begrip (begrijpend lezen, woordenschat, studievaardigheden), taal (technisch lezen, taal, luisteren en spreken) en

rekenen-wiskunde. De leerwinstberekening is gebaseerd op de optelsom in de wisseling in vaardigheidsniveaus van de drie genoemde onderdelen van een cohort leerlingen van een bepaalde school, afgezet tegen de vaardigheidsniveau-indeling van Cito. Tot een cohort wordt gerekend de jaargroep leerlingen die op enig moment gelijktijdig in groep 8 kunnen zitten (zie voor details De Loos Monitoring, zj).

4.3.3. Leerwinstberekeningen op basis van verwachtingen

In beginsel zijn er twee soorten leerwinstmodellen. Er zijn modellen die de leerwinst tussen het ene en het andere toetsmoment beschrijven, zoals bijvoorbeeld bij de hierboven beschreven aanpakken van Amsterdam en De Loos Monitoring. En er zijn modellen die een feitelijk toetsresultaat afzetten tegen het resultaat dat op basis van een bepaald referentiekader verwacht zou mogen worden. Deze laatste leerwinstmodellen worden in de literatuur wel 'growth prediction models' genoemd (Castellano & Ho, 2013). De leerwinst wordt in deze modellen dus niet beschreven in termen van prestatiegroei tussen twee toetsmomenten, maar in termen van wat de prestatie op een toets van een leerling had kunnen zijn, gezien bijvoorbeeld de intelligentie van die leerling. Er zijn in Nederland twee voorbeelden van leerwinstmodellen die deze aanpak volgen.

Leerwinst op basis van intelligentie

Een mogelijkheid om de effecten van onderwijs te beoordelen is door de leerprestaties te vergelijken met de leercapaciteit of intelligentie van leerlingen. Door een groeiende groep Nederlandse basisscholen wordt hiervoor de *Niet-Schoolse Cognitieve Capaciteiten Test* (NSCCT) gebruikt (Van Batenburg & Van der Werf, 2004), bijvoorbeeld door Enschedese scholen van de schoolbesturen VSO en Consent. De NSCCT is een 'quick screenings' instrument voor algemene intelligentie van leerlingen in de groepen 4, 6 en 8 van het basisonderwijs. Deze test wordt klassikaal in een uur afgenomen. De test bestaat, in vergelijking tot een intelligentietest als de WISC, uit een kleinere selectie van items uit het intelligentiedomein en is ontwikkeld om gebruikt te worden als controle variabele in cohortonderzoek. De NSCCT wordt door scholen gebruikt voor twee verschillende toepassingen (Van Batenburg, 2012):

1. het objectiveren van de leerkrachtoordeel over de algemene intelligentie van een leerling;
2. het vergelijken van de algemene intelligentiescore met de vaardigheidsscore op LVS-toetsen van een leerling.

De eerste toepassing van de NSCCT is gericht op de leerkracht om leerlingen op te sporen die meer kunnen dan verwacht. Voor de testafname geeft de leerkracht een inschatting van de verwachte intelligentiescore van de leerling. Deze wordt vergeleken met de latere testuitslag. Is de testuitslag van de leerling veel hoger dan de inschatting van de leerkracht, dan heeft de leerkracht te lage verwachtingen van deze leerling. De leerling laat blijkbaar in de klas minder zien dan op grond van zijn capaciteiten verwacht zou kunnen worden en wordt daarom als onderpresteerder aangemerkt. De leerkracht geeft zijn inschatting in vijf in intelligentie oplopende categorieën (5= 70-85; 4= 86-95; 3= 96-103; 2= 104-110; 1= 111-130), de testuitslagen kunnen in dezelfde categorieën worden weergegeven. Door zijn inschatting met de uitslag te vergelijken kan de leerkracht direct zien in hoeverre zijn verwachting juist is.

Voor de tweede toepassing wordt de NSCCT-score getransformeerd naar de niveau-indelingen van het Cito LVS (A t/m E en I t/m V, zie par. 2.5). Zo kunnen discrepanties tussen de intelligentie van de leerling en zijn/haar prestaties op de LVS-toetsen aan het licht komen. Het gaat hierbij niet om het voorspellen van een toekomstige vaardigheidsscore van een leerling op basis van zijn eerdere intelligentiescore, maar om het vergelijken van het vaardigheidsniveau van een leerling op een bepaald moment met het vaardigheidsniveau van leerlingen van gelijke intelligentie op datzelfde moment. Deze informatie kan de school benutten om de onderwijsaanpak van individuele leerlingen te verbeteren en zo hun additionele waarde te verhogen..

Leerwinst op basis van de 4D-index

De CED-Groep ontwikkelde een integraal model voor schoolverbetering op basis van data, duiden, doelen, doen, het zogenaamde 4D-model (zie www.opbrengstgerichtwerken4d.nl). Binnen een schoolverbeteringstraject worden de antwoorden op uiteenlopende vragen gezocht die in de kern steeds gaan over de vergelijking van leerprestaties. Om dit soort vragen te kunnen beantwoorden, is een eenduidige maat nodig. Hiervoor is door de CED-Groep de 4D-index ontwikkeld. Alle toetsdata van elk leerstofgebied van elk niveau worden genormaliseerd naar een verdeling met een gemiddelde van 200 en een spreiding van 15.

Leerwinst wordt in deze index door de CED-Groep gedefinieerd als het verschil tussen een feitelijke ontwikkeling met de normale verwachte ontwikkeling tussen twee of meer toetsmomenten. Als een leerling een normale verwachte ontwikkeling doormaakt, is de 4D-index van deze leerling op elk toetsmoment hetzelfde. Een verschil in 4D-index tussen twee toetsmomenten duidt dan op (positieve of negatieve) leerwinst. Op een zelfde wijze kan de leerwinst voor een groep of school worden bepaald.

Scholen werken opbrengstgericht met het 4D-model om hun onderwijs te verbeteren en daarmee hun resultaten te verhogen. Door met behulp van de 4D-index antwoorden te zoeken op bovenstaande vragen en deze te relateren aan de onderdelen van het onderwijsleerproces (leerstofaanbod, leertijd, didactisch handelen, pedagogisch handelen, schoolklimaat), krijgt de school een helder beeld van de punten waarop zij kan verbeteren. Uit de antwoorden op de vragen kan blijken dat de gerealiseerde leertijd voor zwakke rekenaars onvoldoende is, of bijvoorbeeld het didactisch handelen voor spelling in de middenbouw verbeterd moet worden. Op basis van deze inzichten wordt een geschikte interventie gekozen, dat kan een professionaliseringstraject zijn of bijvoorbeeld de aanschaf van nieuwe materialen. Adviseurs van de CED-Groep ondersteunen op deze manier scholen voor regulier, speciaal en voortgezet onderwijs.

4.3.4. Toegevoegde waarde op basis van eindtoetsscores

Op 15 september 2013 publiceerde RTL Nieuws een ranglijst met de toegevoegde waarde van Nederlandse basisscholen gebaseerd op hun scores op eindtoetsen afgenomen in groep 8. De toegevoegde waarde is uitgedrukt in een schoolcijfer, waardoor aan de scholen een rangordening gegeven kon worden. De berekening van de toegevoegde waarde is gedaan door prof. dr. J. Dronkers (zie: http://www.schoolcijferlijst.nl/HOME_choice.html). Daartoe zijn in de eerste plaats de gemiddelden van verschillende soorten eindtoetsen (zoals de Eindtoets Basisonderwijs, de Drempeltest of het Schooleindonderzoek) vergelijkbaar gemaakt. In de tweede plaats is het effect van niet-deelnemende leerlingen op de gemiddelde scores van alle scholen geschat. Omdat de data die aan RTL zijn geleverd geen informatie bevatte over de aantallen leerlingen die, om de een of andere reden, niet in de eindtoetsafnames zijn betrokken, is voor alle scholen het schoolgemiddelde daarvoor gecorrigeerd. Bij die correctie is er vanuit gegaan dat niet-deelnemende leerlingen overwegend zwakke leerlingen zijn.

Op basis van deze gereconstrueerde schoolscore heeft prof. Dronkers de toegevoegde waarde van iedere school geschat (Dronkers, 2013). Uitgangspunt bij deze schatting is dat leerlingkenmerken van invloed zijn op de score. Zo kan een school een lagere prestatie hebben dan andere scholen omdat de schoolpopulatie uit relatief veel achterstandsleerlingen bestaat. Bij de bepaling van de toegevoegde waarde is gecorrigeerd voor drie factoren: leerlinggewicht, het niet in Nederland geboren zijn en voor de gemiddelde sociale status van de wijken waaruit de leerlingen van een school afkomstig zijn. Op basis van deze factoren is een nieuwe gemiddelde score berekend die, gezien deze factoren, verwacht zou mogen worden. Omdat de samenstelling van de scholen verschilt, krijgen we op grond van deze berekening voor iedere school een andere verwachting die hoger of lager kan uitpakken dan de oorspronkelijke score. Tot slot wordt die verwachte schoolscore gerelateerd aan de behaalde schoolscore met de formule: Toegevoegde waarde = behaalde schoolscore – verwachte schoolscore.

4.4. Relatie van de bestaande toepassingen met de pilot

In de pilot is de individuele score van een leerling op één vaardigheidsschaal de basis voor de bepaling van de leerwinst. Het verschil tussen twee vaardigheidsscores van dezelfde leerling levert een absolute leerwinstmaat op. Door dit verschil te vergelijken met dat van leerlingen met dezelfde beginvaardigheidsscore is in de pilot een relatieve leerwinstnorm ontwikkeld (zie par. 5.3). Bij de bepaling van de toegevoegde waarde wordt in de pilot uitgegaan van de groei in vaardigheidsscores per leergebied die vervolgens wordt gecorrigeerd voor factoren die weliswaar van invloed zijn op de leerwinst, maar waar de school zelf geen invloed op heeft (zie par. 5.4.). De aanpak in de pilot verschilt op verschillende onderdelen van de hierboven beschreven bestaande berekeningen van leerwinst en toegevoegde waarde. We zullen enkele verschillen toelichten.

De Amsterdamse Kwaliteitswijzer

In de Amsterdamse Kwaliteitswijzer 2012-2013 wordt de leerwinst van een school, in tegenstelling tot de pilot, niet gebaseerd op individuele vaardigheidsscores, maar op de gemiddelde vaardigheidsscore van een groep leerlingen van een school. Het verschil tussen de gemiddelde vaardigheidsscore voor bijvoorbeeld begrijpend lezen in groep 6 van schooljaar

2012-2013 wordt vergeleken met de gemiddelde vaardigheidsscore voor begrijpend lezen van groep 5 van het voorafgaande schooljaar. Het gaat dan om een vergelijking van grotendeels dezelfde leerlingen. Het verschil in gemiddelde vaardigheid wordt als de 'leerwinst' van de betreffende school beschouwd. Door dit te vergelijken met dat van alle Amsterdamse scholen (referentiegroep) wordt inzichtelijk gemaakt of de gemiddelde leerwinst van de school boven, gelijk of onder gemiddeld is. In tegenstelling tot de berekening van toegevoegde waarde in de pilot wordt er bij de Amsterdamse aanpak voor leerwinstbepaling geen rekening gehouden met de sociaal-etnische achtergrond van de leerlingen omdat, volgens de gemeente Amsterdam, uit onderzoek blijkt dat dit niet van invloed is op de leerwinst.

De Loos Monitoring

Bij de aanpak van De Loos Monitoring is een aantal kanttekeningen te plaatsen. In de eerste plaats gaat de leerwinstberekening niet uit van de absolute leerwinst van leerlingen in termen van vaardigheidsscores, maar van de relatieve groei in de positie die een leerling inneemt ten opzichte van een landelijke steekproef in vaardigheidsniveaus. In de tweede plaats combineert De Loos Monitoring de leerwinst van verschillende leergebieden op vaardigheidsniveau. Er wordt daarbij geen rekening gehouden met het feit dat de vaardigheidsschalen per leergebied van elkaar verschillen. De vraag is echter wat de inhoudelijke betekenis is van het combineren van prestaties op het gebied van spelling, technisch lezen, taal, luisteren en spreken.

De leerwinstmethode rekent met de vaardigheidsniveau-indeling van Cito en transformeert deze naar een numerieke waarde om ermee te kunnen rekenen. Over de beperking hiervan meldt De Loos Monitoring (z.j., p. 15) dat deze transformatie niet empirisch is onderbouwd en dat een vaardigheidsniveau een onnauwkeurige maat is. Immers ondanks dat de vaardigheidsscores van leerlingen stijgen betekent dit niet automatisch een wisseling van vaardigheidsniveau. Zo kan een leerling die bij de eerste toetsafname in het laagste niveau valt, ondanks een betekenisvolle stijging van zijn of haar vaardigheidsscore, bij volgende toetsafnames nog steeds tot hetzelfde vaardigheidsniveau behoren, namelijk de 20% laagst scorende leerlingen. De eventuele leerwinst is dan niet zichtbaar.

Leerwinst op basis van NSCCT

Bij het meten van de effecten van onderwijs moet rekening worden gehouden met het beginniveau van de leerling. In de pilot wordt daarvoor steeds de vaardigheidsscore van de midden groep 3 meting gebruikt, of – indien het een toets voor kleuters betreft – de midden groep 1 meting. Er zijn situaties waarbij het niet altijd mogelijk of zinvol is om hiervoor een leerprestatietoets te gebruiken, bijvoorbeeld wanneer leerlingen een volledig nieuw domein leren, waarvan zij nog niets weten. Hoe snel leerlingen een nieuw domein leren hangt van hun intelligentie af. De NSCCT is een praktisch alternatief voor meer uitgebreide intelligentietests vanwege de korte duur en de klassikale afname.

Bij de bepaling van de toegevoegde waarde van scholen gaat het erom de effecten van leerkracht en schoolvariabelen op de ontwikkeling in leerprestaties zo zuiver mogelijk vaststellen. Vanuit deze optiek wordt in sommige toegevoegde waarde modellen wel voor intelligentie gecorrigeerd. Uitgangspunt bij de pilot is geweest dat de deelnemende basisscholen

geen extra data bij hun leerlingen hoefden te verzamelen. De berekeningen van de leerwinst en toegevoegde waarde zijn derhalve gebaseerd op de gegevens die beschikbaar zijn in de schoolinformatiesystemen van de deelnemende scholen. De intelligentiescore van leerlingen behoort echter niet tot de standaardgegevens. Daarom is dit kenmerk in de pilot buiten beschouwing gebleven. De bepaling van leerwinst op basis van de NSCCT is overigens wel een bruikbare werkwijze voor scholen die dit instrument gebruiken.

Toegevoegde waarde op basis van eindtoetsgegevens

Het toegevoegde waarde model van prof. Dronkers is niet gebaseerd op meerjarige leerwinstberekeningen van een cohort leerlingen van een basisschool, maar op het gemiddelde van drie aansluitende schoolscores op eindtoetsen van een school van drie verschillende cohorten (zie ook par. 3.4). In de literatuur (Castellano & Ho, 2013, p. 21) staat deze aanpak bekend als het zogenaamde 'conditional status model' waarbij de status wordt gecorrigeerd met aanvullende informatie en dat leidt tot een opbrengst van een school die je zou mogen verwachten. Het voorspellen van een te verwachten toetsscore is dus niet hetzelfde als leerwinst.

In de toegevoegde waarde schatting heeft prof. Dronkers een aantal zaken buiten beschouwing gelaten. Wanneer alleen met de gegevens van een enkele eindtoets wordt gewerkt is niet duidelijk met hoeveel vaardigheid de leerlingen op een gegeven moment de school zijn ingestroomd. Dat wordt pas duidelijk als naast een eindmeting ook een beginmeting heeft plaatsgevonden. Bij het gebruik van alleen een eindmeting is het lastig een beeld te krijgen van de invloed van de inspanningen van de school op de leerprestaties. Daaruit zou de toegevoegde waarde moeten blijken en dit sluit ook beter aan bij de intuïtieve betekenis van toegevoegde waarde. De interpretatie van het begrip toegevoegde waarde op de wijze zoals toegepast door prof. Dronkers hangt in hoge mate af van de keuze van de correctiefactoren. Bij een andere keuze van de correctiefactoren of wanneer deze inhoudelijk wijzigen, veranderen ook de te verwachten toetsscores. Bovendien maakt hij geen onderscheid naar leerstofgebied waar de (meeste/minste) toevoegde waarde wordt geboekt. Dit is voor de scholen wel belangrijk om te weten.

5. Resultaten

5.1. Inleiding

Dit hoofdstuk bevat een beschrijving van de resultaten van de pilot. Omdat toetsgegevens van de deelnemende scholen uit hun schoolinformatiesystemen zijn geëxporteerd, wordt in paragraaf 5.2 beschreven op welke wijze de gegevensbestanden zijn gemaakt waarmee de onderzoeksinstituten hun modellen konden berekenen.

Paragraaf 5.3 bevat een beschrijving van de leerwinstmodellen die ten behoeve van de pilot zijn ontwikkeld. Zowel Cito als de Universiteit Twente hebben leerwinstberekeningen ontwikkeld. De twee toegevoegde waarde modellen die in de pilot door GION zijn ontwikkeld worden beschreven in paragraaf 5.4. Zowel de verschillen als de overeenkomsten van deze modellen worden nader toegelicht.

De toetspraktijk in sbo-scholen wijkt af van de andere basisscholen in de pilot. Daardoor was het niet goed mogelijk om voor deze scholen op dezelfde wijze rapportages over de leerwinst- en toegevoegde waarde op te stellen als voor de reguliere scholen. Om deze scholen toch feedback te geven op de leerprestaties is ervoor gekozen hen te vergelijken met sbo-scholen die participeren in COOLspeciaal. Zie daarvoor paragraaf 5.5.

Paragraaf 5.6. geeft een beeld van de opvattingen van de deelnemende scholen over de bruikbaarheid van de modellen en de daarop gebaseerde schoolrapportages. De gegevens daarvoor zijn verzameld via een internet-enquête onder de deelnemende scholen.

In paragraaf 5.7 worden tenslotte conclusies getrokken over de gebruikswaarde van leerwinst en toegevoegde waarde in het basisonderwijs voor schoolverbetering en accountability. Ook worden de voorwaarden toegelicht waaronder leerwinst en toegevoegde waarde in het onderwijs kunnen worden geïmplementeerd.

5.2. Databestanden

Verzamelen

De 45 scholen die data hebben geleverd voor de pilot, hebben hun toetsresultaten geëxporteerd uit hun digitaal toets- en schoolinformatiesysteem met behulp van de ingebouwde exportmogelijkheden in ParnasSys, ESIS web-based, LOVS computerprogramma van Cito of Dotcomschool. Eén school gebruikt het schoolinformatiesysteem LARS (http://www.clb-lars.be). Het bleek niet mogelijk om uit LARS op een gestandaardiseerde wijze de gevraagde toets- en leerlinggegevens te exporteren, maar de school was in staat om de gevraagde gegevens in een excelbestand te zetten, zodat de gegevens van deze school samen met de gegevens van de andere scholen verwerkt konden worden.

De dataverzameling vond uiteindelijk in zes ronden plaats:

- ronde 1:	april 2012	actuele (medio 2011-2012) en historische toetsresultaten van alle leerlingen die nu (= medio 2011-2012) op school zitten
- ronde 2:	juni 2012	toetsresultaten van eind 2011-2012 voor de scholen die niet deelnemen aan het seizoensgebonden leerwinstmodel
- ronde 3:	september 2012	toetsresultaten van eind 2011-2012 en de resultaten van de seizoensgebonden afname (start 2012-2013) voor de scholen die deelnemen aan het seizoensgebonden leerwinstmodel
- ronde 4:	februari 2013	toetsresultaten van medio 2012-2013
- ronde 5:	juni 2013	toetsresultaten van eind 2012-2013
- ronde 6:	september 2013	resultaten van de seizoensgebonden afname (start 2013-2014) voor de scholen die deelnemen aan het seizoensgebonden leerwinstmodel

Data cleaning

Voor de berekening van leerwinst en toegevoegde waarde was het nodig van iedere leerling per toetsreeks één resultaat per afnameperiode in het databestand te hebben. Dat is conform de opzet van de planning en afname van de LVS-toetsen van Cito. Echter, de onderwijspraktijk volgt deze niet altijd. Het is gebruikelijk dat leerlingen met een leerachterstand getoetst worden met een toets die past bij hun niveau (op maat toetsen). Daarnaast schrijft Cito bij een enkele toets voor om deze in bepaalde gevallen nogmaals af te nemen. Ook komt het voor, dat leerlingen buiten de reguliere afnameperioden getoetst worden als de leerkracht tussentijds meer informatie over de ontwikkeling van de leerling(en) nodig heeft.

Om één toetsresultaat per afnameperiode te kunnen selecteren, moesten de toetsresultaten eerst in afnameperioden ingedeeld worden. Dat is gedaan op basis van de afnamedatum, omdat niet iedere leerling de toets heeft gemaakt die op basis van leerjaar en afnameperiode gemaakt zou moeten worden. Echter, de afnamedatum bleek binnen deze pilot geen betrouwbaar gegeven.
Alle digitale toets- en schoolinformatiesystemen eisen dat er een afnamedatum ingevoerd wordt bij elk toetsresultaat. De systemen zorgen ervoor dat er altijd een afnamedatum ingevoerd wordt, door standaard een afnamedatum, de datum van het moment van invoeren, in te vullen. De gebruiker kan deze datum altijd wijzigen in de werkelijke afnamedatum, maar dat wordt vaak niet gedaan. In veel gevallen is de ingevoerde afnamedatum daardoor niet de datum waarop de toets is afgenomen, maar de datum waarop de toets is ingevoerd.
Als de toetsresultaten kort na de toetsafname ingevoerd worden, is het geen probleem als niet de werkelijke afnamedatum, maar de invoerdatum ingevoerd wordt. Maar het komt regelmatig voor dat de toetsresultaten geruime tijd na de afname worden ingevoerd en dat daardoor de ingevoerde afnamedatum (= invoerdatum) ook (ruim) buiten de voorgeschreven afnameperiode valt. Kennis hebben van de juiste afnamedatum is wel van belang voor de juiste samenstelling van de groep leerlingen waarvoor de leerwinst berekend moet worden.

Dit probleem is grotendeels ondervangen door het hele schooljaar op te delen in drie afnameperioden:

- begin (B) van 1 september t/m 30 november
- medio (M) van 1 december t/m 31 maart
- eind (E) van 1 april t/m 31 augustus

Ten behoeve van het seizoensgebonden leerwinstmodel (= eindafnamen worden in de tweede week na de zomervakantie herhaald) is er vanaf eind schooljaar 2011-2012 een vierde afnameperiode toegevoegd:

- begin (B) van 1 oktober t/m 30 november
- medio (M) van 1 december t/m 31 maart
- eind (E) van 1 april t/m 31 juli
- seizoensgebonden afname (S) van 1 augustus t/m 30 september

Uitgaande van deze afnameperioden is er voor iedere leerling per toetsreeks één resultaat per afnameperiode geselecteerd. Als er van een leerling twee of meer toetsresultaten in één afnameperiode waren, is telkens de laatste afname geselecteerd.

Oude en nieuwe LVS-toetsen

Bij de toetsreeksen is telkens onderscheid gemaakt tussen de oude LVS-toetsen van Cito (van vóór 2006) en de nieuwe toetsen (vanaf 2006). De oude en de nieuwe toetsen voor een bepaald leerstofgebied hebben niet dezelfde vaardigheidsschaal. Daardoor is het niet mogelijk om de leerwinst te bepalen uit het verschil tussen een vaardigheidsscore behaald op een oude toets en een vaardigheidsscore behaald op een nieuwe toets. Cito heeft de nieuwe toetsen telkens voor één leerjaar uitgebracht, beginnend bij de toetsen voor groep 3. Daarbij heeft Cito geadviseerd om de nieuwe toetsen telkens vanaf groep 3 in te voeren, zodat de leerlingen niet halverwege hun schoolloopbaan overstappen van de oude naar de nieuwe toetsen.
De Drie-Minuten-Toets (DMT) vormt hierop een uitzondering. Cito heeft deze toets in 2009 voor groep 3 tot en met 8 vernieuwd. Het belangrijkste verschil met de oude DMT is, dat de toets niet meer een aparte score voor iedere leeskaart oplevert, maar één vaardigheidsscore voor alle gelezen leeskaarten samen.

De meeste scholen zijn ook in zijn geheel, voor groep 3 tot en met 8, overgestapt op de nieuwe DMT. Daardoor zijn er veel leerlingen die tijdens hun schoolloopbaan zijn overgestapt van de oude naar de nieuwe DMT. Omdat de leeskaarten niet inhoudelijk zijn gewijzigd, is het mogelijk om de scores op de oude DMT om te zetten naar een vaardigheidsscore op de nieuwe DMT, mits de juiste combinatie van leeskaarten is gelezen. Cito heeft waar mogelijk de verzamelde scores van de oude DMT omgezet naar vaardigheidsscores op de nieuwe DMT.

Opbouw bestand

In de op te leveren databestanden zijn de toetsresultaten van iedere leerling gerangschikt op didactische leeftijd (DL) van de leerling op het moment van toetsen, uitgedrukt in toetsmomenten (M3 (= medio leerjaar 3), E3 (= eind leerjaar 3), M4, enz.). Daarnaast is voor iedere leerling een variabele 'cohort' toegevoegd waarbij het cohort het schooljaar is waarin de leerling begint in groep 3 (DL = 0).

In de aangeleverde exportbestanden zit geen informatie over de DL van de leerling: geen DL op het moment van toetsen en geen gegevens over het moment van starten in groep 3. De exportbestanden bevatten wel informatie over het leerjaar of de jaargroep op het moment van toetsen. Voor leerlingen met een normale schoolloopbaan (= geen leerjaar doubleren of overslaan) is het in theorie mogelijk om uit het leerjaar of de jaargroep de DL af te leiden. Echter, de aangeleverde leerjaren/jaargroepen zijn onvoldoende betrouwbaar om daar voor alle leerlingen de DL uit af te leiden. Op enkele (sbo-)scholen zitten alle leerlingen in jaargroep 0 of 9 en een aantal leerlingen zitten, volgens de aangeleverde data, binnen één schooljaar in meerdere leerjaren. Bovendien ontbrak bij een aantal leerlingen een deel van de toetsresultaten, waardoor voor deze leerlingen niet was af te leiden of zij vóór het oudste beschikbare toetsresultaat een leerjaar hadden gedoubleerd of overgeslagen.

Besloten is om het cohort en de DL te bepalen op basis van de (kalender)leeftijd van de leerling. Daarbij is het referentiepunt de leeftijd op 1 oktober. Het cohort van een leerling is dan het schooljaar waarin de leerling op 1 oktober 6 jaar is. Vervolgens wordt er voor elk toetsresultaat een tijdstip, het moment in de schoolloopbaan van de leerling, bepaald waarop de toets is gemaakt (= variabele "afn_code_berekend"). Het toetsresultaat van medio het schooljaar waarin de leerling op 1 oktober 6 jaar was, krijgt het tijdstip M3, het toetsresultaat van eind het daaropvolgende schooljaar krijgt het tijdstip E4, enz.
Voor leerlingen die in groep 3 zijn begonnen in het schooljaar waarin zij op 1 oktober 6 jaar waren, en die geen leerjaar hebben gedoubleerd of overgeslagen, komen de zo berekende tijdstippen overeen met de werkelijke toetsmomenten. Bij leerlingen met een vertraging of een versnelling verschuiven de tijdstippen ten opzichte van de toetsmomenten.

Achtergrondgegevens

Voor het bepalen van toegevoegde waarde zijn, naast de toetsresultaten, ook gegevens over de achtergrond van de leerlingen nodig (onder andere opleiding en herkomst ouders en gegevens over het type zorgleerling). Een deel van deze achtergrondgegevens is automatisch mee geëxporteerd met de toetsresultaten. Om ook de achtergrondgegevens met de toetsresultaten mee te kunnen exporteren uit ParnasSys, is een aangepaste exportmogelijkheid in ParnasSys toegevoegd.

De via de toetsresultatenexports verzamelde achtergrondgegevens zijn middels een excelbestand aan de scholen voorgelegd met het verzoek deze achtergrondgegevens aan te vullen met gegevens die al op school aanwezig zijn. Het was nadrukkelijk niet de bedoeling om de ontbrekende gegevens bij ouders op te vragen. Bij de verzamelde en aangevulde achtergrondgegevens ontbraken bij veel leerlingen de gegevens over de herkomst van de

ouders. Sinds de invoering van de nieuwe gewichtenregeling registreren veel scholen de herkomst van de ouders niet meer, omdat het leerlinggewicht daar niet meer van afhankelijk is (de nieuwe gewichtenregeling kijkt alleen nog naar de opleiding van de ouders).

5.3. Leerwinstmodellen

5.3.1. Inleiding

Omdat alle scholen uit de pilot gebruik maken van LVS-toetsen van Cito is voldaan aan twee belangrijke voorwaarden om leerwinst en toegevoegde waarde te kunnen berekenen.

In de eerste plaats staan deze toetsen interpretaties toe in termen van groei op een vaardigheidsschaal (Yen, 2007). Deze vaardigheidsschaal maakt het mogelijk om enerzijds de resultaten van een leerling op verschillende toetsmomenten met elkaar te vergelijken.

Anderzijds kunnen met deze schaal ook de resultaten van leerlingen in dezelfde groep worden vergeleken, die verschillende toetsen uit hetzelfde leerstofgebied hebben gemaakt. Ieder leergebied waarvoor toetsen beschikbaar zijn, kent zijn eigen vaardigheidsschaal. Daarom kan de leerwinst alleen per leerstofgebied en niet over de verschillende leerstofgebieden heen worden uitgerekend.

In de tweede plaats dekken de LVS-toetsen belangrijke leerstofgebieden over meerdere leerjaren van het basisonderwijs, zoals rekenen-wiskunde en technisch en begrijpend lezen. Daarmee zijn deze toetsen 'gevoelig' voor de kwaliteit van de geboden instructie en verwerking. Daarom kan er met deze toetsen een relatie gelegd worden tussen de hoogte van de scores en de kwaliteit van het gegeven onderwijs (Popham, 2007, p. 146-147[9]). Dit is met name een belangrijke voorwaarde voor de bepaling van toegevoegde waarde, want via een toegevoegde waarde model wordt immers geprobeerd een schatting te maken van de bijdrage van de school aan de leerprestaties.

Met LVS-toetsen kan de absolute leerwinst vastgesteld worden in termen van groei op een vaardigheidsschaal. Maar daarmee is niet de vraag beantwoord of de groei van een bepaalde leerling of van een groep leerlingen voldoende is of niet. Niet voor alle leerlingen is de groei hetzelfde (Koedel & Betts, 2009; Tong & Kolen, 2007; Luyten & Ten Bruggencate, 2011). De ene leerling groeit schoksgewijs, terwijl een andere leerling een meer vloeiende vooruitgang boekt. Sommige leerlingen groeien in een bepaalde periode meer dan andere leerlingen uit dezelfde groep. En ondanks dat leerlingen op eenzelfde niveau beginnen wil dat niet zeggen dat ze dezelfde groei doormaken. In de pilot is daarom gezocht naar een norm om de leerwinst te kunnen beoordelen. Door de leerwinst van leerlingen te vergelijken met die van vergelijkbare leerlingen ontstaat een referentiepunt dat als norm kan worden gebruikt. We noemden dit eerder relatieve leerwinst (zie par. 3.3). In de pilot is bij de berekening van de leerwinst deze vorm van normering gevolgd.

[9] "An instructionally sensitive test would be capable of distinguishing between strong and weak instruction by allowing us to validly conclude that a set of students' high test scores are meaningfully, but not exclusively, attributable to effective instruction. . . . In contrast, an instructionally insensitive test would not allow us to distinguish accurately between strong and weak instruction".

In de pilot zijn twee verschillende modellen ontwikkeld om de leerwinst over een langere periode te bepalen. Cito ontwikkelde het zogenaamde *Z-score model* en de Universiteit Twente het *Groeitempo-model*. De Universiteit Twente heeft op basis van het Groeitempo-model ook een model gemaakt om de leerwinst gedurende de zomervakantie in kaart te brengen. Dit model is in de pilot het *Seizoensgebonden leerwinstmodel* gaan heten. De verschillende modellen worden hieronder toegelicht. Maar eerst wordt aandacht besteed aan de onderwijsperiode waarover in de pilot de leerwinst is berekend.

5.3.2. Leerwinstperioden

Voor het op de juiste wijze bepalen van de leerwinst en toegevoegde waarde is het van belang een beslissing te nemen over de onderwijsperiode waarover en over de groep leerlingen (cohort) waarvoor deze worden berekend. In de paragraaf 5.2 is uitgelegd dat het van groot belang is dat van alle leerlingen die tot het cohort gaan behoren bekend is op welk moment welke toetsen zijn afgenomen, in welke leerjaar ze zitten en hoe lang ze op de huidige school onderwijs hebben ontvangen.

Alle scholen uit de pilot ontvangen op basis van hun eigen gegevens rapportages over de leerwinstmodellen van Cito en de Universiteit Twente. Om de scholen in de gelegenheid te stellen de modellen met elkaar te vergelijken is er afgesproken om in de rapportages hetzelfde startpunt voor de leerwinstberekening te nemen voor dezelfde groepen leerlingen. Als startpunt is gekozen voor medio groep 3: M3 (zie tabel 2). Voor begrijpend lezen is eind groep 3 (E3) het startpunt, omdat er in groep 3 geen mediotoets beschikbaar is voor dit vak (zie tabel 3).

Daarnaast is er gekozen om te werken met leerlingcohorten. In welk cohort een leerling komt wordt bepaald aan de hand van de leeftijd van de leerling. Voor de pilot is uitgegaan van de regel “Wanneer de leerling 6 jaar oud is, zit hij of zij in groep 3”. Voor deze regel is gekozen, omdat het niet bekend is vanaf welk moment iedere leerling is gestart in groep 3.
In de rapportages aan de scholen worden de resultaten echter niet per cohort getoond, maar per groep omdat dit voor de school een betekenisvolle eenheid is.

Tabel 2

Overzicht leerwinstperiode en cohort (rekenen-wiskunde, spelling, technisch lezen en woordenschat)

Periode	Cohort
M3-M5/E5	2010
M3-M6/E6	2009
M3-M7/E7	2008
M3-M8	2007

Tabel 3

Overzicht leerwinstperiode en cohort (begrijpend lezen)

Periode	Cohort
E3-M5/E5	2010
E3-M6/E6	2009
E3-M7/E7	2008
E3-M8	2007

5.3.3. Groeitempo-model

De Universiteit Twente heeft het *Colorado Growth Curve Model* (zie http://www.schoolview.org/ColoradoGrowthModel2.asp) als inspiratie gebruikt om een leerwinstmodel te ontwikkelen. Het Colorado Growth Curve Model plaatst leerlingen in één van de volgende drie prestatiecategorieën, laag, gemiddeld of hoog, om vervolgens de ontwikkeling over een bepaalde periode te volgen. De ontwikkeling van de leerprestaties wordt in het Colorado-model als volgt gelabeld:

- *Catching up* als een leerling progressie vertoont van de lage naar een hogere categorie;
- *Keeping up* als een leerling uit de gemiddelde of hogere categorie zich in die groepen weet te handhaven of
- *Moving up* wanneer de leerling opschuift van de gemiddelde naar de hoge categorie.

Het leerwinstmodel uit Colorado wordt vanwege deze labeling ook wel het CuKuMu-model genoemd. Het daarop gebaseerde leerwinstmodel dat de Universiteit Twente ontwikkelde, heeft de naam *Groeitempo-model* (GTM) gekregen.

De Universiteit Twente heeft het GTM ontwikkeld vanuit de gedachte dat de methodiek achter het model zo 'doorzichtig' mogelijk moet zijn. Om de ontwikkeling van de leerlingen voor bepaalde vakken over een langere periode in kaart te brengen is hun prestatieniveau aan het begin en het einde van een bepaalde periode vergeleken, bijvoorbeeld rekenen-wiskunde in groep 3 en groep 5.

De begin- en eindmeting is vastgesteld in termen van vaardigheidsscores. Deze vaardigheidsscores zijn vervolgens omgezet naar de zogenoemde *vaardigheidsniveaus* van Cito, waarbij de indeling A tot en met E of de nieuwere indeling I tot en met V is aangehouden. De groei van een leerling wordt vergeleken met die van de leerlingen die bij de beginmeting hetzelfde vaardigheidsniveau hadden. De leerlingen worden op basis van hun beginsituatie ingedeeld in een van de vijf vaardigheidsniveaugroepen. Vervolgens wordt voor elke leerling nagegaan hoe groot de leerwinst is in vergelijking met de andere leerlingen in dezelfde groep. Zodoende wordt rekening gehouden met de mogelijkheid dat leerlingen met verschillende aanvangsniveaus niet allemaal evenveel leerwinst boeken. Voor leerlingen met een hoog aanvangsniveau is de gemiddelde groei vaak iets kleiner dan voor leerlingen die met een lage score beginnen. Voor die leerlingen is er wat meer ruimte om winst te boeken.

Wanneer de leerwinst is bepaald worden de leerlingen in drie categorieën van gelijke grootte ingedeeld:

1. de 33,3% leerlingen met de hoogste groei (bovengemiddeld);
2. de middelste 33,3% (gemiddeld);
3. de 33,3% met de laagste groei (ondergemiddeld).

Doordat de leerlingen binnen hun vaardigheidsniveau ingedeeld worden in een van de drie categorieën, kan de groei van leerling met een hoog vaardigheidsniveau (bijvoorbeeld A of I) afgezet worden tegen andere hoog presterende leerlingen. Als een leerling van

vaardigheidsniveau A of I tot de categorie 'laagste 33,3%' behoort, betekent dit dat deze leerling binnen dat niveau tot de 33,3% leerlingen hoort die de minste groei hebben getoond. Met andere woorden leerlingen kunnen in een van de drie categorieën terecht komen ongeacht hun vaardigheidsniveau bij de start in groep 3. Van leerlingen uit het hoogste vaardigheidsniveau (A of I) kan dus blijken dat ze weinig leerwinst laten zien, terwijl leerlingen uit het laagste niveau (E of V) veel leerwinst kunnen boeken en daardoor in de categorie 'hoogste groei' vallen.

In de schoolrapportages worden de uitkomsten per leerling gegeven. Om ook een beeld te geven van de leerwinst op groeps- of leerjaarniveau, wordt eveneens gerapporteerd hoeveel leerlingen er uit de groep of het leerjaar in elk van de drie categorieën zitten.

In de praktijk zal blijken dat een 'gemiddelde' school in elke categorie ongeveer evenveel leerlingen heeft zitten (33,3% per groep). Het zal vrijwel nooit voorkomen dat op een school de verdeling over de drie groepen precies gelijk is. Op de meeste scholen zal het percentage leerlingen in de drie categorieën afwijken van het gemiddelde. Vaak gaat het om kleine afwijkingen die te wijten zijn aan toevallige schommelingen. Soms zijn de afwijkingen echter te groot om aan het toeval toe te schrijven. Het aantal leerlingen in de hoogste categorie op een school kan soms groot zijn. Ook het omgekeerde is natuurlijk mogelijk: het percentage leerlingen met lage groei is opvallend groot. Aan de scholen in de pilot is gerapporteerd of er opvallende afwijkingen zijn.

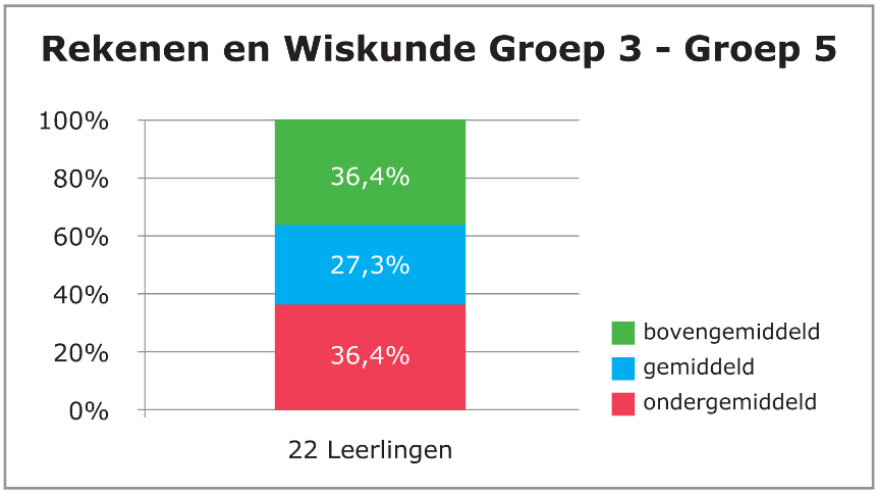
Tabel 4 en figuur 4 geven een voorbeeld uit een rapportages over het Groeitempo-model. Het gaat hier om de vooruitgang in rekenvaardigheid tussen de mediometing in groep 3 (januari-februari) en de eindmeting in groep 5 (mei-juni) van 22 leerlingen.

Tabel 4 Leerwinst rekenen-wiskunde in de periode groep 3 tot en met groep 5

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

Over de individuele leerwinst van de 22 leerlingen (zie tabel 4) valt het volgende te melden. Er zijn in die groep twee leerlingen met een hoog aanvangsniveau (vaardigheidsniveau I) en zij hebben een gemiddelde vooruitgang geboekt. Van de zeven leerlingen met het één na hoogste aanvangsniveau (vaardigheidsniveau II) hebben er zes een bovengemiddelde leerwinst geboekt en één leerling een leerwinst die lager is dan gemiddeld. Voor vier van de vijf leerlingen met een gemiddeld aanvangsniveau (vaardigheidsniveau III) ligt de leerwinst lager dan gemiddeld en heeft er één een gemiddelde leerwinst laten zien. Van de zeven leerlingen met het één na laagste aanvangsniveau (vaardigheidsniveau IV) laten er drie een ondergemiddelde leerwinst zien, drie

een gemiddelde en één een bovengemiddelde leerwinst. Eén leerling is op het laagste niveau begonnen, maar deze heeft wel een bovengemiddelde leerwinst geboekt.



Figuur 4 Categorie-indeling Groeitempo-model periode groep 3 tot en met 5 rekenen-wiskunde

Uit figuur 4 blijkt dat van deze groep leerlingen uiteindelijk 36,4% een bovengemiddelde leerwinst heeft geboekt, 27,3% een gemiddelde en 36,4% een ondergemiddelde leerwinst. Hoewel de verdeling over deze drie groepen niet exact gelijk is, zijn in dit geval de verschillen te klein om er veel betekenis aan te hechten. De verschillen zijn waarschijnlijk te wijten aan toevallige schommelingen.

5.3.4. Seizoensgebonden leerwinstmodel

Bij de aanvang van de pilot is de deelnemende scholen ook aangeboden de leerwinst te bepalen tijdens de zomervakantie (zie bijlage 4). De ontwikkeling van leerlingen tijdens de zomervakantie laat zien hoeveel vooruitgang leerlingen nog boeken als ze een lange periode niet naar school gaan. Dit plaatst informatie over leerwinst en toegevoegde waarde in een nieuw perspectief. Ook buiten schooltijd doen leerlingen kennis en vaardigheden op. Voor sommige leerlingen geldt dat tijdens de zomervakantie hun prestaties achteruit gaan. Om die reden geeft een E-meting afgenomen aan het einde van het schooljaar geen nauwkeurige en actuele informatie over de startsituatie van leerlingen in het nieuwe schooljaar. Bovendien wordt de volgende serie LVS-toetsen pas een half jaar later weer afgenomen. Er zit dus een flinke periode tussen een E- en M-afname. Daarom is voor scholen informatie over de ontwikkeling van leerlingen tijdens de zomervakantie uiterst relevant. Vandaar dat een aantal scholen uit de pilot ervoor heeft gekozen ook over deze periode de leerwinst te berekenen.

Binnen de pilot hebben de scholen die belangstelling hadden voor dit model, twee weken na de zomervakantie opnieuw de E-versie van de reguliere eindmeting van het vorig leerjaar afgenomen. Dit werd gedaan voor de vakken rekenen-wiskunde, spelling en technisch lezen. De vaardigheidsscores van voor de vakantie werden afgetrokken van de vaardigheidsscores van na de vakantie. Dit verschil, de groei tijdens zomervakantie werd vergeleken met de groei die de leerlingen hebben geboekt tijdens het schooljaar.

Figuur 5 geeft een voorbeeld van de rapportage over de leerwinst van Thomas Janssens tijdens de zomervakantie.

			Vaardigheids-score M3	Niveau E3	Vaardigheids-score na de vakantie	Niveau E5	Verschil vaardigheids-score
Niveau I	Gemiddelde groei: 2,0	Gemiddeld	38	I	40	I	2

Figuur 5 Voorbeeldrapportage leerwinst seizoensgebonden model

Thomas Janssens heeft op de toets rekenen-wiskunde E3 een vaardigheidsscore van 38 (gemiddeld niveau) en op de toets na de vakantie een vaardigheidsscore van 40 (gemiddeld niveau). Zijn groei is 2 punten. De gemiddelde groei voor leerlingen uit het hoogste niveau op E3 is twee punten. De groei van Thomas tijdens de zomervakantie is dus gemiddeld.

Het seizoensgebonden leerwinstmodel onderscheidt zich van de andere modellen die in de pilot zijn ontwikkeld. Net zoals bij de andere leerwinstmodellen wordt de omvang van de leerwinst in kaart gebracht, maar in dit geval wordt de individuele leerwinst vergeleken met die van leerlingen als ze geen onderwijs krijgen van hun leerkrachten. In zekere zin levert dit model op een bijzondere manier een indicatie van de toegevoegde waarde van een school. Tijdens de zomervakantie heeft de school geen directe invloed op de vaardigheidsgroei van leerlingen. Mocht er tijdens de zomervakantie toch sprake zijn van groei, dan kan deze dus niet worden toegeschreven aan de school, maar bijvoorbeeld wel aan de inspanningen van ouders. En in het geval er geen groei of zelfs terugval is, dan is dat een indicatie dat onderwijs effect heeft.

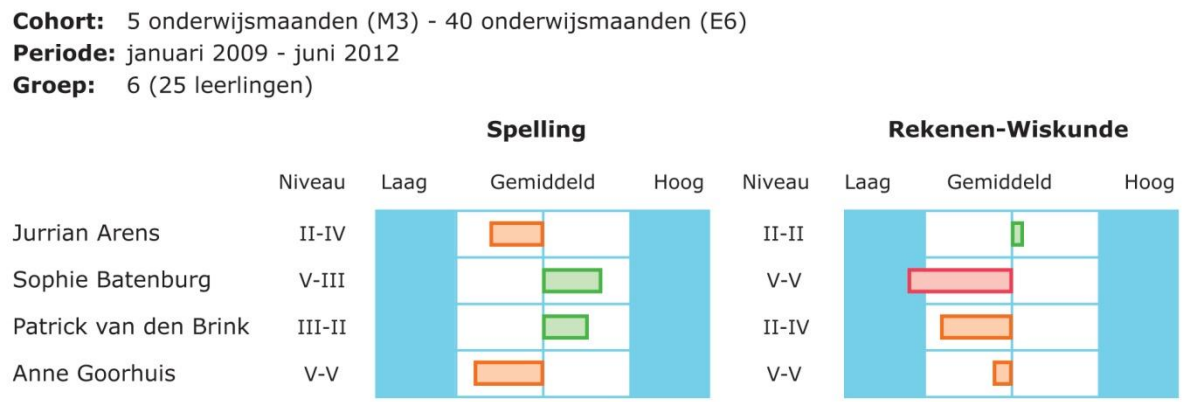
Een belangrijk verschil met een toegevoegde waarde model is dat daarin getracht wordt door middel van statistische correcties zoveel mogelijk rekening te houden met de achtergronden van de leerlingen en de schoolomgeving. Een voordeel van het seizoensgebonden leerwinstmodel is dat er geen aanvullende achtergrondinformatie nodig is over de leerlingen en de schoolcontext. Het vereist wel een extra toetsafname vlak na de zomervakantie. Een probleem is wel dat voor dit doel geen specifieke toetsen voorhanden zijn. Tijdens de pilot is dezelfde toets gebruikt als ongeveer 8 weken daarvoor. Hierdoor kan er sprake zijn van een leereffect. Tevens zijn de waardes (bovenste 33,3%, middelste 33,3% en onderste 33,3%) bepaald op basis van de metingen binnen het project en bestaan er nog geen gevalideerde groei bepalingen voor de aanvangsperiode na de zomervakantie. Een andere beperking is dat niet alle toets- en schoolinformatiesystemen het toe laten om extra toetsscores in te voeren.

5.3.5. Z-score model

Cito heeft een relatieve maat voor leerwinst ontwikkeld op basis van zogenoemde z-scores. Een technische verantwoording van dit model is te vinden in bijlage 5. De z-score geeft aan hoe de vaardigheidsgroei van een (groep) leerling(en) zich verhoudt tot de gemiddelde vaardigheidsgroei van een landelijke vergelijkingsgroep. De 0-lijn (zie figuur 7) stelt daarbij het landelijke gemiddelde voor. De werkwijze om van individuele scores te komen tot oordelen over de mate van leerwinst via het z-score model wordt hieronder toegelicht.

Score op leerlingniveau

De eerste stap is dat voor iedere leerling voor één bepaalde periode de groei in vaardigheidspunten is berekend. Daartoe is de vaardigheidsscore van de medio afname in groep 3 (M3) afgetrokken van de vaardigheidsscore van het laatste afnamemoment, in dit geval de eindafname in groep 6 (E6) van juni 2012 (zie figuur 6). Vervolgens is bepaald hoe de groei van de leerling zich verhoudt tot de groei van leerlingen met hetzelfde startniveau. Ofwel, er wordt gekeken of de leerling meer, gelijk of minder gegroeid dan andere leerlingen met dezelfde vaardigheidsscore op afnamemoment M3.

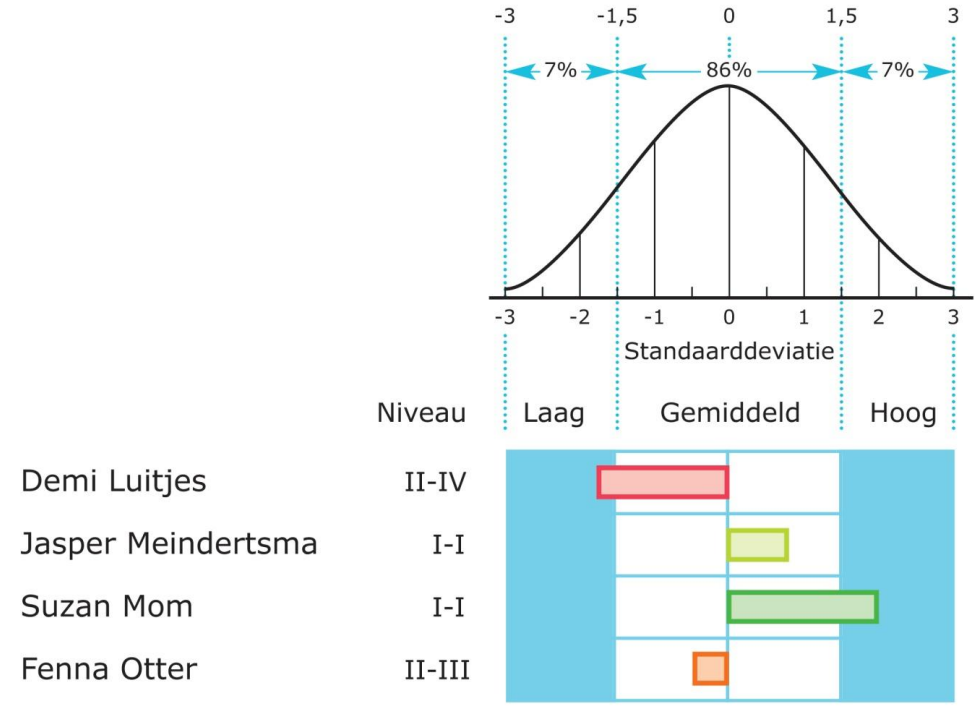


Figuur 6 Groepsoverzicht z-score model

Betekenis z-score

De z-score is een relatieve maat die als volgt te interpreteren is (zie figuur 7):

- Een z-score van 0 wijst op een gemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau;
- Een z-score > 0 wijst op een bovengemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau;
- Een z-score < 0 wijst op een benedengemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau.



Figuur 7 Categorieënindeling leerwinst in relatie tot z-score

Een z-score die exact gelijk is aan 0 zal niet zo vaak voorkomen. Kleine afwijkingen naar boven en beneden wel. Daarom kent het Groepsoverzicht Leerwinst een scoreverdeling in drie categorieën: laag, gemiddeld en hoog. De z-scores tussen -1.5 en +1.5 worden aangemerkt als gemiddeld. Naar verwachting behaalt 86% van de leerlingen een score tussen deze grenzen (zie figuur 7). Ongeveer 7% van de leerlingen behaalt een lagere score en ongeveer 7% een hogere score.

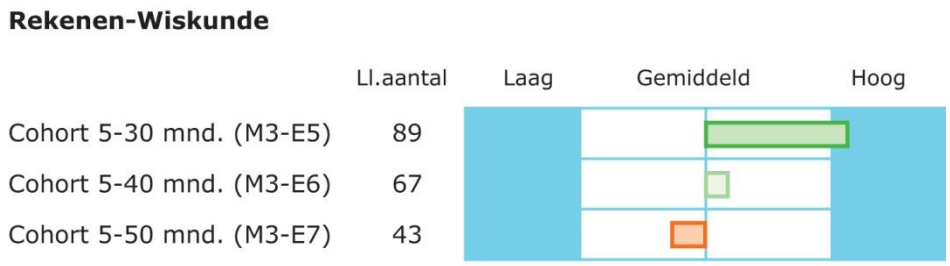
Een score in de categorie laag (rood balkje) betekent dat de leerling behoort tot de 7% leerlingen die het minst gegroeid zijn, vergeleken met leerlingen met hetzelfde startniveau. Een score in de categorie hoog (donkergroen balkje) betekent dat de leerling behoort tot de 7% leerlingen die de meeste groei hebben gerealiseerd, vergeleken met leerlingen met hetzelfde startniveau.

Leerlingen waarbij de z-score in de categorie 'gemiddeld' valt, laten in principe geen afwijkende groei zien. Toch is er voor gekozen om binnen deze categorie twee kleuren te hanteren, lichtgroen voor de leerlingen waarbij de z-score groter dan 0 is en oranje bij de leerlingen waarbij de z-score kleiner dan 0 is. In het algemeen geldt hoe langer het gekleurde balkje, hoe meer de groei afwijkt van de (gemiddelde) groei van leerlingen met hetzelfde startniveau.

Score op schoolniveau

De Schoolrapportage Leerwinst toont de gemiddelde groei van een cohort leerlingen per vakgebied. De gemiddelde vaardigheidsgroei per cohort is berekend door de z-score van alle individuele leerlingen op te tellen en te delen door het aantal leerlingen. Het betreft dus een gemiddelde z-score: *de schoolgemiddelde leerwinst*. Deze gemiddelde z-score is vergeleken met

de gemiddelde z-scores van andere scholen in Nederland (zie figuur 8). Omdat de betrouwbaarheid van een groepsgemiddelde afhangt van de grootte van de groep, is bij de berekening van de groepsgemiddelden rekening gehouden met de groepsgrootte.



Figuur 8 Schoolrapportage leerwinst op schoolniveau

Interpretatie van de gemiddelde z-score

De interpretatie van het resultaat op schoolniveau is identiek aan de interpretatie van het resultaat op leerlingniveau. Een score tussen -1.5 en +1.5 kan dus beschouwd worden als gemiddeld. Bij een hogere of lagere score is de groei die een cohort leerlingen gemiddeld laat zien op de school opvallend groter of kleiner dan de groei die leerlingen gemiddeld laten zien op andere scholen. In de Schoolrapportage Leerwinst op schoolniveau zijn dezelfde kleurcoderingen gebruikt als bij de Schoolrapportage Groepsoverzicht Leerwinst.

In de schoolrapportage bevat elk cohort (voor zover data aanwezig) meerdere 'sub'cohorten. Zo bestaat het Cohort 5-30 mnd. (M3-E5) in figuur 8 niet alleen uit de groep leerlingen die in juni 2012, 30 onderwijsmaanden hadden, maar ook de leerlingen die de voorafgaande schooljaren in juni 30 onderwijsmaanden hadden. Het aantal leerlingen dat in de berekening van de gemiddelde z-score is meegenomen, staat steeds genoemd onder 'Ll. aant.'.
Om praktische redenen is binnen de pilot zowel de periode (M3-laatste afnamemoment) en het aantal leergebieden beperkt. In principe is het Z-score model toe te passen op iedere willekeurige periode (M3-E3, E5-M6, enzovoort) en alle leergebieden.

5.4. Toegevoegde waarde modellen

5.4.1. Inleiding

In de pilot zijn twee modellen voor de bepaling van toegevoegde waarde ontwikkeld: het vaardigheidsverschil-model en het vaardigheidsgroei-model. De overeenkomst van deze modellen met de eerder beschreven leerwinstmodellen is dat ook de groei in vaardigheidsscores op LVS-toetsen, de leerwinst, de basis vormt voor de bepaling van de toegevoegde waarde. Dat betekent ook dat de toegevoegde waarde uitsluitend per vaardigheidsschaal kan worden bepaald.

Zoals eerder toegelicht zijn LVS-toetsen gevoelig voor instructie. Daarmee is voldaan aan twee belangrijke voorwaarden voor de schatting van de toegevoegde waarde van scholen: leerwinstgegevens op basis van een vaardigheidsschaal en gebruik van toetsen waarmee een relatie gelegd kan worden met de kwaliteit van het gegeven onderwijs. Op basis van deze

voorwaarden kan de toegevoegde waarde van een school opgevat worden als een indicatie van de kwaliteit van het gegeven onderwijs in een specifiek leerstofgebied.

Een belangrijk verschil met de leerwinstmodellen is dat de toegevoegde waarde uitsluitend op schoolniveau wordt vastgesteld. Daartoe wordt per leerstofgebied de leerwinst van een leerlingcohort op een school gemiddeld. Bij toegevoegde waarde zijn we immers op zoek naar de bijdrage van de school aan prestatiegroei van alle leerlingen binnen een leerstofgebied. Om die bijdrage in beeld te brengen zijn correcties nodig om de invloed van niet-schoolse factoren op de prestatiegroei zo goed mogelijk uit te zuiveren.

De twee in de pilot ontwikkelde toegevoegde waarde modellen verschillen hoofdzakelijk van elkaar in het aantal toetsmomenten dat in het model is opgenomen en daardoor ook in hun gebruiksmogelijkheden. Hierna worden beide modellen kort toegelicht. Een uitgebreide uitleg van de modellen wordt gegeven in bijlage 6.

Vaardigheidsverschil-model

Het vaardigheidsverschil-model is gebaseerd op de vaardigheidsscores op twee LVS-toetsen van hetzelfde leerstofgebied die als begin- en eindmeting dienst doen. Met dit model zou de toegevoegde waarde geschat kunnen worden over een langere onderwijsperiode waarbij slechts gebruik wordt gemaakt van twee toetsmomenten, bijvoorbeeld met als beginmeting groep 3 en als eindmeting groep 8. Technisch gezien is zo'n vergelijking mogelijk, maar de vraag is echter of dit tot een betekenisvolle interpretatie van toegevoegde waarde leidt. Dit geldt met name als dit model door scholen gebruikt wordt voor opbrengstgericht werken.

Ook al worden voor eindmeting en beginmeting toetsen gebruikt uit het LVS van Cito waaraan een vaardigheidsschaal ten grondslag ligt, dan is het nog maar de vraag of bijvoorbeeld de rekenvaardigheid van leerlingen uit groep 3 vergeleken kan worden met die van dezelfde leerlingen uit groep 8. De bandbreedte waarbinnen de groei in vaardigheid van een bepaald leergebied zinvol kan worden geïnterpreteerd, is beperkt tot enkele aansluitende leerjaren. Dit heeft te maken met de inhoud en de verdeling van de leerstof over de verschillende leerjaren. Een leerling uit groep drie krijgt bijvoorbeeld om die reden in toetsen geen opgaven met kommagetallen voorgelegd, maar een leerling uit groep 8 wel. Om inhoudelijke redenen ligt eerder een vergelijking van de groei tussen groep 3 en 5 of tussen groep 5 en groep 8 voor de hand, dan een vergelijking van de groei tussen groep 3 en 8. De rekenvaardigheid van een leerling uit groep 3 is nu eenmaal een andere dan die van een leerling uit groep 8. Bij de interpretatie van de toegevoegde waarde, maar ook van de leerwinst, dient daarmee rekening te worden gehouden.

Een beperking van het vaardigheidsverschil-model is dat het terugkijkt op de toegevoegde waarde van een school (vanaf eindmeting op beginmeting) en dus voor de school weinig houvast biedt om tijdig bij te sturen. Het vaardigheidsverschil-model is eerder geschikt voor de beoordeling van de eindopbrengsten (accountabilityperspectief) en in beduidend minder mate voor opbrengstgericht werken (schoolverbeteringsperspectief). Dat geldt niet voor het vaardigheidsgroei-model, waarop we later zullen terugkomen.

Voorbeeldrapportage vaardigheidsverschil-model

Hieronder wordt een voorbeeld gegeven van de resultaten van het vaardigheidsverschil-model zoals deze in een schoolrapportage aan een school zijn gepresenteerd.

Cohort 2010: Onderwijsmaanden 5 tot 30 (M3-E5)

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

Figuur 9 Voorbeeldrapportage van het vaardigheidsverschil-model

Figuur 9 toont de resultaten van het vaardigheidsverschil-model. Het is gebaseerd op één beginmeting in groep 3 (M3 of E3) en één eindmeting in groep 5 (E5) voor de leerstofgebieden rekenen-wiskunde, spelling, technisch lezen en woordenschat. Tussen die begin- en eindmeting is per leerstofgebied over de hele groep leerlingen de gemiddelde groei in vaardigheidsscores berekend. In dit voorbeeld gaat het om cohort 2010, dat zijn de leerlingen die op 1 oktober 2010 zes jaar oud waren. Per leerstofgebied wordt in de tweede kolom (= Pilot-gemiddelde) de gemiddelde bruto- en netto leerwinst van de pilotscholen in termen van vaardigheidsscores gegeven. De bruto leerwinst van de pilotscholen is de gemiddelde ongecorrigeerde leerwinst van alle leerlingen op alle pilotscholen. De netto leerwinst de gemiddelde leerwinst gecorrigeerd voor fairness-kenmerken. Voor rekenen-wiskunde is in de pilot de bruto-gemiddelde groei 52,8 en - na correctie - de netto-gemiddelde groei 47,9.

De derde kolom (= Uw school (BI)) geeft het gemiddeld bruto (48,4) en netto (47,3) groeiresultaat van de school voor rekenen-wiskunde. Tussen haakjes staat het 95% betrouwbaarheidsinterval (BI) waarin naar alle waarschijnlijkheid het 'echte' bruto-gemiddelde (48,4) valt. Dat ligt dus in werkelijkheid ergens tussen 44,6 en 52,3. Voor het netto-gemiddelde (47,3) ligt dat naar alle waarschijnlijkheid tussen 43,5 en 51,2. In beide gevallen valt het pilot-gemiddelde ook binnen beide betrouwbaarheidsintervallen. Dat levert in de laatste kolom de

conclusie op dat de gemiddelden voor rekenen-wiskunde van deze school niet afwijken van het pilot-gemiddelde en dus naar alle waarschijnlijkheid 'gelijk' zijn. Deze school heeft, ook na correctie, een gemiddelde leerwinst voor rekenen-wiskunde die te vergelijken is met de gemiddelde leerwinst van alle scholen. De school levert dus volgens het vaardigheidsverschil-model geen betekenisvolle toegevoegde waarde voor rekenen-wiskunde. Voor de andere leerstofgebieden ligt dat anders.
Voor spelling geldt dat de school het 'bruto' (25,0) beter doet dan het pilot-gemiddelde (22,4). Het pilot-gemiddelde valt net buiten het betrouwbaarheidsinterval (22,8 - 27,2). Maar na correctie is dit verschil verdwenen en valt het netto-gemiddelde van de school (21,7) en van de pilot (19,9) binnen de marge van het betrouwbaarheidsinterval (19,5 – 23,9). De school levert dus - als we naar het netto resultaat kijken - voor spelling geen betekenisvolle toegevoegde waarde. Voor technisch lezen levert de school dat wel omdat het netto-resultaat (57,2) significant beter is dan het pilot-gemiddelde (49,3).

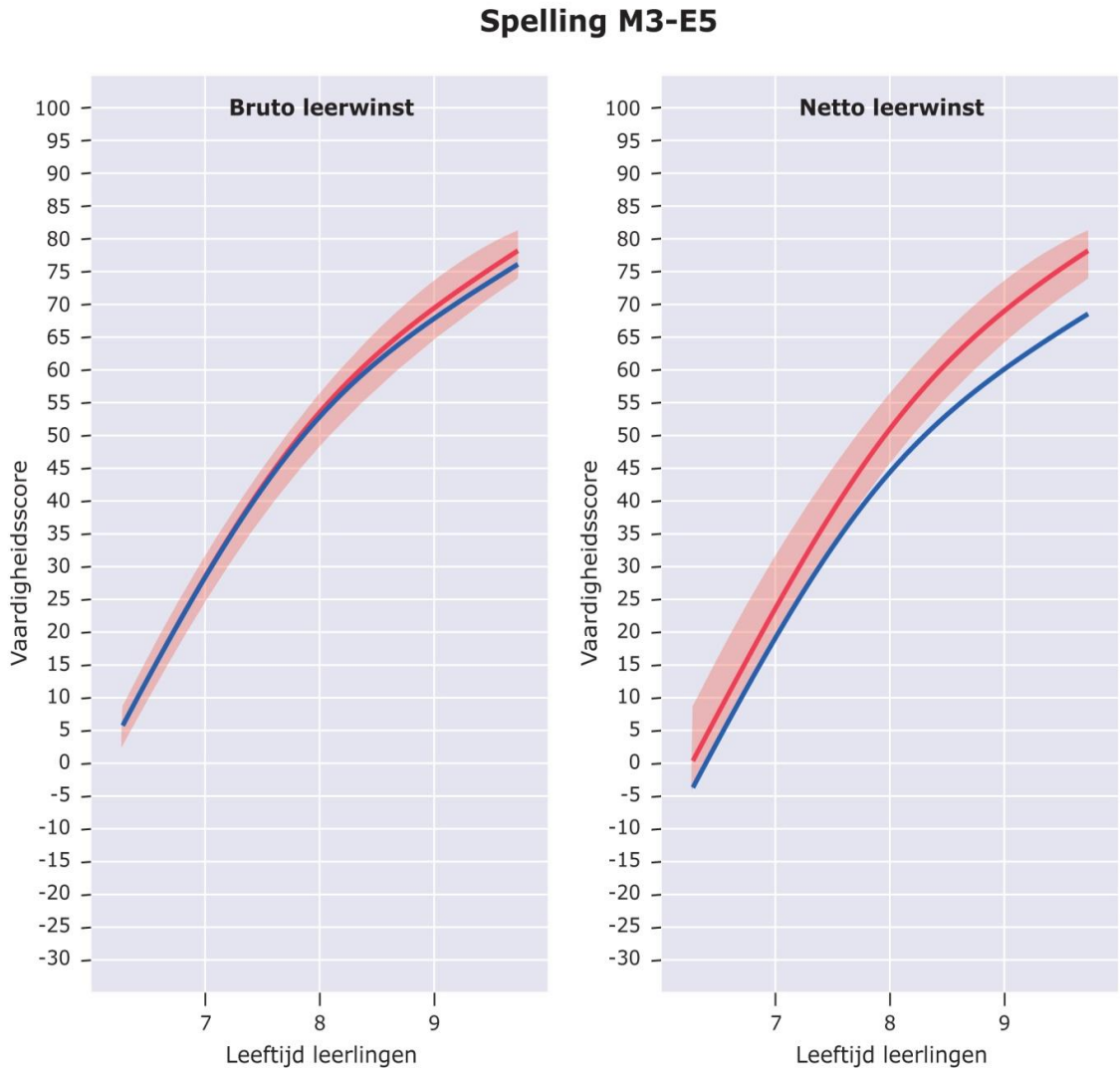
Vaardigheidsgroei-model

Het vaardigheidsgroei-model is gebaseerd op vaardigheidsscores van meer toetsmomenten uit hetzelfde leerstofgebied. Er kunnen zelfs alle toetsmomenten uit een hele serie LVS-toetsen voor hetzelfde leerstofgebied in worden betrokken. In het geval een school voor alle leerstofgebieden halfjaarlijks alle LVS-toetsen afneemt, kan op deze wijze voor elk leerstofgebied de toegevoegde waarde over nagenoeg alle tussenliggende meetmomenten uit de hele basisschoolperiode worden bepaald. Dat biedt kansen voor sturing van het onderwijs van een cohort dat nog op school zit. Zo kan met dit model, bijvoorbeeld te beginnen in groep 3, (half)jaarlijks van een cohort de toegevoegde waarde geschat worden. Daaraan kunnen dan ook (half)jaarlijks consequenties voor eventuele bijsturing worden verbonden. Op deze manier kunnen gaandeweg de schoolloopbaan van leerlingen, naast gegevens over de ontwikkeling van de leerwinst, ook gegevens over de toegevoegde waarde van de school nog een rol spelen in maatregelen bedoeld om het onderwijs te verbeteren. Door met dit model de toegevoegde waarde-schatting met enige regelmaat in opeenvolgende leerjaren te herhalen, kan het vaardigheidsgroei-model niet alleen in dienst staan van het accountability- maar ook van het schoolverbeteringsperspectief.

Voorbeeldrapportage Vaardigheidsgroei-model

De bevindingen van het vaardigheidsgroei-model worden in het schoolrapport in de vorm van figuren met groeilijnen gepresenteerd. In figuur 10 staat een voorbeeld van deze groeilijnen.

Figuur 10 laat een schoolrapportage zien op basis van het vaardigheidsgroei-model voor spelling over de periode medio groep 3 tot en met medio groep 5. De grafieken laat de gemiddelde leerwinst een groep leerlingen van een bepaalde school zien in de vorm van een rode groeilijn, terwijl de gemiddelde leerwinst voor rekenen-wiskunde van alle scholen over dezelfde periode uit de pilot in een blauwe lijn is weergegeven. Het 95% betrouwbaarheidsinterval dat geldt voor de rode lijn, is aangegeven in de vorm van een lichtere kleur rood.



Figuur 10 Voorbeeldrapportage van het Vaardigheidsgroei-model

De linker grafiek geeft het bruto-resultaat, dus zonder correctie voor niet-schoolse factoren. Omdat de gemiddelde groei van de pilotscholen (de blauwe lijn) binnen het betrouwbaarheidsinterval valt van de betrokken school (het rode gebied), kan gesteld worden dat de bruto-leerwinst van deze school niet betekenisvol afwijkt van alle scholen uit de pilot.

De rechter grafiek (netto leerwinst) geeft de gemiddelde leerwinst van de school en de pilotscholen weer nadat de correctie heeft plaatsgevonden voor niet-schoolse factoren. Duidelijk is te zien dat de school (rode lijn) vlak voordat de leerlingen 8 jaar werden, het beter is gaan doen dan de pilotscholen. De blauwe lijn (het pilot-gemiddelde) raakt niet langer de ondergrens van het betrouwbaarheidsinterval. Na het 9e levensjaar van de groep leerlingen doet de school het zelfs beduidend beter. Op basis van de schatting van het vaardigheidsgroei-model kan geconcludeerd worden dat deze school rond het 8ste levensjaar van de groep leerlingen, dus ergens in groep 4, relatief meer toegevoegde waarde is gaan leveren. De grafiek laat ook zien dat de school medio groep 3 geen afwijkende toegevoegde waarde levert en dus actie kon ondernemen om de prestatie van de school te verbeteren.

5.4.2. Overeenkomsten tussen de modellen

De werkwijze om de toegevoegde waarde van een school te schatten is voor beide modellen grotendeels gelijk. Het uitgangspunt is steeds de leerwinst van individuele leerlingen. De leerwinst van leerlingen is op te delen in een schools deel en een niet-schools deel. Onder het schoolse deel wordt verstaan de veronderstelde bijdrage van het onderwijs aan de leerwinst, terwijl het bij het niet-schoolse deel om de bijdrage van buitenschoolse factoren gaat. Bij de toegevoegde waarde gaat het erom de omvang van het schoolse deel van de leerwinst vast te stellen. In beide modellen wordt dit gedaan door de invloed van het niet-schoolse deel op de leerwinst zo goed mogelijk uit te sluiten. Hiervoor is gebruik gemaakt van statistische modellen die volgens de internationale wetenschappelijke literatuur geschikt zijn om de toegevoegde waarde van een school te bepalen (zie ook bijlage 6).

Zoals gezegd, de toegevoegde schatting heeft steeds betrekking op de leerwinst van een school als geheel, dus op de prestaties van een hele groep leerlingen (cohort) op een bepaald leerstofgebied. Dit in tegenstelling tot de in paragraaf 5.3 beschreven leerwinstmodellen. Deze zijn vooral gericht op beoordeling van de groei in prestaties van individuele leerlingen. Uiteraard zijn voor de schatting van de toegevoegde waarde van een school ook de vaardigheidsscores van de individuele leerlingen nodig. De individuele groei in vaardigheidsscores gedurende een bepaalde periode - de leerwinst - wordt op schoolniveau gemodelleerd met een statistisch meerniveau regressiemodel. In het model wordt rekening gehouden met de school waarop de leerling zit, dat de leerlingen zijn gegroepeerd in jaargroepen en dat jaargroepen zijn gegroepeerd in scholen. Dit is belangrijk omdat zo wordt voorkomen dat bij het vergelijken van scholen de gevonden verschillen tussen scholen in toegevoegde waarde te makkelijk statistisch significant worden bevonden.

Gemeenschappelijk voor beide modellen is dat de toegevoegde waarde in twee stappen is geschat. De eerste stap houdt in dat de gemiddelde leerwinst wordt bepaald zonder deze te corrigeren voor niet-schoolse factoren. Deze *schoolgemiddelde leerwinst* is te beschouwen als het bruto-schooleffect ofwel de bruto leerwinst[10]. Tijdens de tweede stap in de modellering is in beide gevallen de schoolgemiddelde leerwinst gecorrigeerd voor dezelfde niet-schoolse factoren, zoals het opleidingsniveau van ouders en de aanwezigheid van bepaalde typen zorgleerlingen (add, adhd, dyslexie, dyscalculie, enzovoort). Dit levert het netto-schooleffect of de netto leerwinst op. De netto leerwinst is te beschouwen als een schatting van de *toegevoegde waarde*. Hieronder wordt het verschil tussen bruto- en netto leerwinst nader toegelicht.

Schoolgemiddelde leerwinst

Het resultaat van de eerste stap in het modelleren van de toegevoegde waarde is de bepaling van de gemiddelde leerwinst die de leerlingen op een school hebben weten te realiseren in een bepaalde rapportageperiode (zie figuur 9 en 10). Het is de optelsom van het schoolse en niet-schoolse deel van de leerwinst van leerlingen, ook wel bruto leerwinst genoemd. Omdat geen rekening gehouden is met de invloed van interne en externe factoren op de leerwinst, geeft deze

[10] Harris (2011, p. 69-71) noemt dit 'basic value-added' en reserveert 'advanced value-added' voor modellen waarin rekening wordt gehouden met de invloed van niet-schoolse factoren.

maat geen uitsluitsel over de mate waarin de school deze beïnvloed heeft of niet. En geeft dus ook geen antwoord op de vraag of we over de leerwinst van een school tevreden kunnen zijn of niet.

Om toch een oordeel te kunnen geven over de bruto leerwinst van een school is gebruik gemaakt van een relatieve norm door deze te vergelijken met de gemiddelde bruto leerwinst van alle pilotscholen. In de schoolrapportages wordt daarom zowel het schoolgemiddelde als het pilotgemiddelde van de leerwinst gepresenteerd. Zo wordt duidelijk of de groei van de leerwinst van een groep leerlingen van een school in een bepaalde periode beter, gelijk of slechter is dan van even oude leerlingen op de andere basisscholen binnen de pilot. Een voordeel van het tonen van de schoolgemiddelde bruto leerwinst in schoolrapportages is dat deze met hand nagerekend kan worden. Dit bevordert de inzichtelijkheid voor de scholen. Wel wordt hierbij in het midden gelaten welke invloed schoolse- en niet schoolse factoren daarop hebben gehad. Daarvoor is nog een tweede stap in de modellering nodig.

Toegevoegde waarde

De tweede stap bij het modelleren van de leerwinst van leerlingen is de bepaling van de toegevoegde waarde van een school. Dit is de leerwinst die de basisschool gerealiseerd heeft, rekening houdend met de 'moeilijkheidsgraad' van de leerlingpopulatie. Voor basisscholen met een moeilijke leerlingpopulatie – bijvoorbeeld veel leerlingen van laag opgeleide ouders of veel zorgleerlingen – is het moeilijker om een bepaalde leerwinst te realiseren dan voor scholen die een makkelijke leerlingpopulatie hebben. De toegevoegde waarde van een school wordt in beide modellen berekend door het statistische model voor de totale leerwinst uit te breiden met kenmerken die een 'faire' vergelijking tussen scholen in de weg staan. Op deze manier wordt gecorrigeerd voor de invloed van kenmerken waar de school 'niets aan kan doen' maar die - volgens de wetenschappelijke literatuur - wel van invloed zijn op de leerprestaties van leerlingen. Deze kenmerken worden ook wel *fairness-kenmerken* genoemd. Het resultaat van het berekenen van de toegevoegde waarde kan gezien worden als een herziening van de (bruto) schoolgemiddelde leerwinst van een school in een bepaalde rapportageperiode. Deze netto leerwinst geeft de omvang weer van het effect van het schoolse deel op de leerwinst van leerlingen. De niet-schoolse invloeden zijn er nu, zo goed als mogelijk, uitgezuiverd. Hiermee wordt een goede indicatie verkregen van de toegevoegde waarde: het is de schatting van de toegevoegde waarde van de school.

Norm

Ook bij een toegevoegde waarde-schatting doet zich de vraag voor naar een norm om het niveau van de toegevoegde waarde te kunnen beoordelen. Dit gebeurt op eenzelfde manier als bij de schoolgemiddelde leerwinst, zij het dat deze in de tweede stap voor alle pilotscholen ook gecorrigeerd wordt voor niet-schoolse factoren. Zoals gezegd, wordt door deze correctie zichtbaar wat de bijdrage is van de schoolse factoren aan de leerwinst. Daarmee geeft dit een indicatie van de toegevoegde waarde van de school. Maar, net zoals dat het geval is bij de schoolgemiddelde leerwinst, blijkt hieruit niet of we daarover tevreden kunnen zijn. Daarom wordt ook in dit geval in de pilot gebruik gemaakt van een relatieve norm. De toegevoegde waarde van een school, de schoolgemiddelde netto leerwinst, is vergeleken met het gemiddelde

van alle basisscholen uit pilot[11]. Uit deze vergelijking blijkt of de toegevoegde waarde van een school beter, gelijk of slechter is dan de toegevoegde waarde van de andere basisscholen uit de pilot. Deze vergelijking geeft een beeld van de relatieve positie van de school. Dat is een waardevolle toevoeging op de leerwinstberekening van de school, omdat door de vergelijking duidelijk wordt hoe de leerwinst van een school zich verhoudt tot die van een groep scholen met een vergelijkbare leerlingpopulatie.

De rol van fairness-kenmerken in de toegevoegde waarde modellen

In de pilot is de keuze van mogelijke fairness-kenmerken beperkt. Volgens afspraak hoefden de scholen geen extra gegevens te verzamelen over of bij hun leerlingen. Er is uitgegaan van gegevens over leerlingkenmerken die de scholen standaard in hun schoolinformatiesystemen verzamelen of van schoolkenmerken die bij een intakegesprek met de schooldirecteur achterhaald konden worden, respectievelijk op de site van de school te vinden waren. De toegevoegde waarde berekeningen van beide modellen zijn op dezelfde fairness-kenmerken gebaseerd. Een overzicht van de fairness-kenmerken is te vinden in paragraaf 5.4.4.

De fairness-kenmerken die aan de toegevoegde waarde modellen zijn toegevoegd, worden ook wel covariaten genoemd. Dit zijn variabelen die in de berekeningen worden betrokken omdat correctie ervoor op de toetsprestaties wenselijk is. Doordat ze aan de modellen zijn toegevoegd, wordt de invloed van deze kenmerken op de leerwinst geneutraliseerd. Een voorbeeld kan dit wellicht verduidelijken.

Stel dat niet-westerse allochtone leerlingen gemiddeld genomen een leerwinst voor spelling behalen die 4 (vaardigheids)punten lager is dan die van autochtone leerlingen. Westerse allochtone leerlingen daarentegen scoren gemiddeld 2 punten lager. In de berekeningen wordt hiervoor gecorrigeerd door bij de niet-westerse allochtone leerlingen 4 punten op te tellen bij de werkelijk behaalde leerwinst voor spelling. Bij westerse allochtone leerlingen komen er 2 punten bij. Met andere woorden, we voegen als het ware bij iedere allochtone leerling een stukje leerwinst toe die bepaald wordt door zijn etnische herkomst. Als dit een autochtone leerling zou zijn geweest dan had hij naar verwachting respectievelijk 4 en 2 punten hoger gescoord. Zo wordt per leerling een gecorrigeerde leerwinstscore berekend die vrij is van de invloed van etniciteit. Er zijn vanwege deze correctie - als het ware - alleen maar autochtone leerlingen op de pilot-scholen.

Dit gebeurt op een vergelijkbare wijze ook voor de overige vijf fairness-kenmerken. Als naast etniciteit ook de andere fairness-kenmerken aan het model zijn toegevoegd, is de behaalde leerwinst niet alleen gezuiverd van de invloed van etniciteit, maar tegelijkertijd ook van de invloed van het opleidingsniveau van de ouders en de invloed van zorgleerlingen (dyslexie met indicatie, dyscalculie met indicatie, adhd of add met indicatie en autisme/ass/pdd-nos met indicatie). Door de correctie verkrijgt iedere leerling een voor fairness-kenmerken gecorrigeerde leerwinst. Door vervolgens de gecorrigeerde leerwinsten van de leerlingen te middelen per school, wordt de gemiddelde netto leerwinst van een school verkregen. Het is dat

[11] In de pilot diende voor de toegevoegde waarde bepaling het gemiddelde van de deelnemende scholen uit de pilot als referentiepunt, maar in feite zou dat een landelijk gemiddelde moeten zijn.

deel van de leerwinst voor spelling dat met enige zekerheid aan de school is toe te schrijven. De netto leerwinst geeft zo een indicatie van de toegevoegde waarde van een school.

Onzekerheidsgrenzen

De toegevoegde waarde van een school zoals berekend met de twee modellen, is in feite een zo goed mogelijke schatting van de werkelijke waarde. Er is sprake van een schatting omdat er veel onzekerheden zijn die de uitkomst kunnen hebben bepaald, maar waar we geen weet van hebben, zoals meetfouten (zie daarvoor par. 3.5). Toch willen we graag inzicht krijgen in de werkelijke waarde. Om daar met wat meer zekerheid iets over te kunnen zeggen, wordt gebruik gemaakt van zogenaamde onzekerheidsgrenzen, ook wel betrouwbaarheidsintervallen genoemd (zie figuur 9 en 10). Dergelijke intervallen worden bijvoorbeeld ook gebruikt bij de lange termijn weersvoorspellingen om aan te geven hoe zeker we kunnen zijn van neerslag of van de temperatuur.

Betrouwbaarheidsintervallen geven in ons geval de grenzen aan van de mate waarin we zeker kunnen zijn van de schoolgemiddelde bruto leerwinst of van de toegevoegde waarde (netto leerwinst) van een school. Door deze intervallen te gebruiken kan met 95% zekerheid worden gezegd dat werkelijke gemiddelde leerwinst of toegevoegde waarde van de school ligt tussen de gepresenteerde ondergrens en bovengrens (zie bijvoorbeeld figuur 9). Met andere woorden, er is maar een hele kleine kans - 5% om precies te zijn – dat de werkelijke waarde voor de gemiddelde totale leerwinst of toegevoegde waarde niet in dit interval ligt.

Met de onzekerheidsgrenzen kan ook bepaald worden of het schoolgemiddelde van de leerwinst of toegevoegde waarde statistisch significant afwijkt van het pilot-gemiddelde. Dat is het geval als het pilot-gemiddelde buiten de onzekerheidsgrenzen van het schoolgemiddelde ligt. Dit wordt in schoolrapportages weergeven met de woorden ‘hoger’ (de school doet het beter dan het pilot-gemiddelde) of ‘lager’ (de school doet het slechter dan het pilot-gemiddelde). Is het schoolgemiddelde niet significant afwijkend van het pilot-gemiddelde dan wordt dit aangegeven met ’gelijk’ (zie figuur 9).

5.4.3. Verschillen tussen de modellen

Naast overeenkomsten tussen de twee toegevoegde waarde modellen zijn er ook verschillen. Kennis hierover is van belang om een eventuele keuze uit één van de twee modellen te kunnen onderbouwen. Het meest in het oog springende verschil tussen beide modellen is dat in het verschil-model slechts twee toetsmomenten een rol spelen (zie figuur 9), terwijl het bij het groei-model gaat om zoveel mogelijk toetsmomenten (zie figuur 10).

Bij de berekening van het verschil-model worden alleen leerlingen betrokken die een vaardigheidsscore hebben op zowel de begin- als eindmeting van een bepaalde toetsdomein. Bij het groei-model worden alle beschikbare vaardigheidsscores in een rapportageperiode van de leerlingen op een school meegenomen. Dus niet alleen de begin- en eindmeting, maar ook alle tussenliggende M- en E-metingen van de leerlingen. Ook instromende en uitstromende leerlingen in een bepaalde rapportageperiode doen mee met de berekeningen. Het is ook geen probleem als een leerling een keer een toets heeft gemist. In principe doet elke leerling met

minimaal twee bruikbare vaardigheidsscores per rapportageperiode op een bepaalde toets mee in de berekeningen. In zijn algemeenheid betekent dit dat het groei-model op meer leerlingen en op meer meetmomenten per leerling is gebaseerd dan het verschil-model. Het groei-model geeft zo een completer beeld van de ontwikkeling in leerprestaties van leerlingen dan het verschil-model.

Een tweede belangrijk verschilpunt is dat bij het groei-model wordt uitgegaan van de exacte leeftijd van een leerling op het moment van de toetsafname (zie figuur 10), terwijl bij het verschil-model wordt uitgegaan van onderwijstijd (zie figuur 9). Bij het groei-model moet daarom de geboortedatum van de leerlingen en de datum van de toetsafname bekend zijn. Dat bij het verschil-model – en ook bij de leerwinstmodellen – de leerwinst van een leerling wordt afgezet tegen het aantal maanden onderwijs dat een leerling heeft gehad, heeft een zekere onnauwkeurigheid tot gevolg. In de modellen wordt medio groep 3 als startpunt van de berekening van de leerwinst genomen. Er wordt aangenomen dat alle leerlingen dan 6 maanden onderwijs hebben gehad hebben. In werkelijkheid is dit voor een aanzienlijke groep leerlingen niet het geval. Kleuterbouwverlenging, - versnelling of doubleren in groep 3 - komt regelmatig voor, maar kan niet goed worden verwerkt in de huidige toets- en schoolinformatiesystemen. Hierdoor is het niet mogelijk om nauwkeurig vast te stellen wanneer een leerling voor de eerste keer in groep 3 zit en dus precies 6 maanden onderwijs heeft gehad. Het groei-model is hier veel minder gevoelig voor doordat het gebaseerd is op de kalenderleeftijd. Het probleem van de juiste bepaling van het startpunt - bij precies 6 maanden onderwijs – speelt in dit model alleen een rol bij het vaststellen van de juiste vergelijkingsgroep (cohort).

5.4.4. Ontwikkeling van de rapportages

De mening van de scholen over de begrijpelijkheid en inzichtelijkheid van de toegevoegde waarde modellen en discussies hierover binnen de projectgroep zijn de voornaamste drijfveer geweest voor de ontwikkeling die deze rapportages gedurende de looptijd van het project hebben doorgemaakt. In de kern zijn de statistische meerniveau regressiemodellen steeds gelijk gebleven. De set van fairness-kenmerken is wel aan verandering onderhevig geweest. Hetzelfde geldt voor de uitleg en weergave van de resultaten van de toegevoegde waarde modellen en de lay-out van het Schoolrapport Toegevoegde waarde.

In de eerste rapportageronde (september 2012) is de meest uitgebreide set van fairness-kenmerken toegepast om de netto toegevoegde waarde van een school te bepalen. Tabel 5 geeft een overzicht van de oorspronkelijke set kenmerken met hun categorie-indeling. In het eerste Schoolrapport Toegevoegde waarde werd per kenmerk uit tabel 5 de gemiddelde waarde of procentuele verdeling van de betreffende school én die van alle pilot-scholen samen gepresenteerd. Dit stelde de school in de gelegenheid om de correctheid van hun fairness-kenmerken te checken. Voor enkele scholen was dit aanleiding om uit eigen beweging opnieuw data te verzamelen en de gegevens in hun schoolinformatiesystemen te herzien.

Tabel 5 Overzicht van de toegepaste 'fairness'-kenmerken uit de eerste rapportageronde

'Fairness'-kenmerk	**Categorieën**
Schoolkenmerken	
Aantal leerlingen	--
Denominatie	RK/PC/Bijzonder Neutraal/Openbaar/Hindoe
Type basisschool	SO/SBO/regulier
Schoolconcept (%)	Jenaplan/Dalton/Montessori/Vrije school/Freinet
VVE deelname	Ja/nee
Leerlingkenmerken	
Sekse	Jongen/Meisje
Gewichtenregeling:	0/0,3/1,2
Opleidingsniveau ouder 1	12 categorieën*
Opleidingsniveau ouder 2	12 categorieën*
Hoogste opleidingsniveau* ouders	12 categorieën*
Etniciteit leerling (CBS definitie)	Autochtoon/ Niet-westers allochtoon/Westers allochtoon
Eenoudergezin	Ja/nee
Gescheiden ouders	Ja/nee
Dyslexie	Ja, met indicatie/Vermoeden/ Nee
Dyscalculie	Ja, met indicatie/Vermoeden/ Nee
ADHD of ADD	Ja, met indicatie/Vermoeden/ Nee
Autisme, ASS of PDD-NOS	Ja, met indicatie/Vermoeden/ Nee
Andere stoornis	Ja, met indicatie/Vermoeden/ Nee

* Het opleidingsniveau van de ouders/verzorgers is ingedeeld in 12 categorieën:
1 = geen onderwijs gevolgd
2 = 1-3 jaar basis onderwijs
3 = 4-6 jaar basis onderwijs/svo (=categorie 1 gewichtenregeling)
4 = 1-2 jaar lbo/vmbo bbl-kbl/(i(vbo)
5 = 3-4 jaar lbo/vmbo bbl-kbl/i(vbo) (=categorie 2 gewichtenregeling)
6 = 1-2 jaar mavo/vmbo tl-gl
7= 3-4 jaar mavo/vmbo tl-gl (=categorie 3 gewichtenregeling)
8 = 1-3 jaar havo/vwo
9 = 4-6 jaar havo/vwo
10 = mbo/leerlingwezen
11 = hbo
12 = universiteit

In de tweede schoolrapportageronde is een aantal fairness-kenmerken uit tabel 5 niet meer teruggekomen. De feedback van scholen en discussies binnen de projectgroep hebben hier aanleiding toe gegeven. De redenen waren divers: de schoolkenmerken bleken niet (statistisch) relevant, er was een beter alternatief kenmerk voorhanden (hoogste opleiding ouders is nauwkeuriger gemeten dan de gewichtregeling) en correctie voor 'sekse' en 'overige (leer)stoornissen' was niet gewenst. Dit resulteerde in een beperkte set van fairness-kenmerken

die tot en met de laatste rapportageronde is toegepast: het hoogste opleidingsniveau van de ouders, de etnische herkomst van de leerling en het type zorgleerling (dyslexie, dyscalculie, adhd/add en autisme/ass/pdd-nos). De categorisering van deze fairness-kenmerken is gelijk gebleven aan de eerste rapportageronde, met uitzondering van de vier items over het type zorgleerling. De antwoordcategorie 'vermoeden' is niet langer gespecificeerd, omdat dit door de scholen te subjectief bevonden werd. Daardoor resteerde een tweedeling: een leerling heeft een indicatie voor een bepaald type zorg of niet.

Vanaf de tweede rapportageronde is een andere belangrijke verbetering doorgevoerd. Dit betreft de figuren met de resultaten van het vaardigheidsgroei-model. In deze figuren werden naast de gemiddelde groeilijnen van de school en alle pilotscholen samen ook de onzekerheidsgrenzen van beide groeilijnen weergegeven in de vorm van dunne stippellijnen. De scholen vonden dit slecht leesbaar en bovendien lastig te begrijpen. Om dit te verbeteren zijn de resultaten van de vaardigheidsgroei-modellen in een andere statistisch pakket met een betere grafische ondersteuning ingelezen en bewerkt. De onzekerheidsgrenzen zijn vereenvoudigd. Er is alleen nog een gekleurde balk rondom de gemiddelde groeilijn van de school geprint. In hun feedback na de tweede rapportageronde hebben veel scholen aangegeven de figuren beduidend beter te vinden. Op enkele uitzonderingen na heeft men er geen negatief commentaar meer op.

In de derde rapportageronde is veel aandacht besteed aan het verbeteren van de uitleg van de modellen, de beschrijving van de resultaten en de lay-out van het Schoolrapport Toegevoegde waarde. Dit heeft geresulteerd in een aparte algemene toelichting op de toegevoegde waarde modellen voor de scholen met daarin een volledig uitgewerkt voorbeeld van het verschil- en het groei-model. Daardoor kon het schoolspecifieke Schoolrapport Toegevoegde waarde compacter worden. Voor elke rapportageperiode en elk toetsdomein zijn de resultaten van het verschil- en het groei-model van een school direct onder elkaar geplaatst. Ook is weergegeven op hoeveel leerlingen de gepresenteerde resultaten zijn gebaseerd. Dit was in vorige versies van het schoolrapport niet het geval. Zo is het voor de betreffende school gemakkelijker geworden om de resultaten van de twee modellen te vergelijken en kan men beter beoordelen welk model het meest bruikbaar en inzichtelijk is.

In de vierde en laatste rapportageronde is een update van de data uit schoolinformatiesystemen gebruikt bij de berekeningen van de toegevoegde waarde van de scholen. De toelichting op het schoolrapport Toegevoegde waarde is ter beoordeling voorgelegd aan een communicatiemedewerker van de CED-Groep. Dit heeft geleid tot nog enkele aanpassingen.

5.5. Sbo-scholen

Onder de deelnemende basisscholen in de pilot bevinden zich enkele sbo-scholen. Deze scholen nemen al enige tijd LVS-toetsen af bij hun leerlingen. Daarom zouden er genoeg gegevens moeten zijn om de leerwinst van sbo-leerlingen of toegevoegde waarde van sbo-scholen te kunnen berekenen. Op de sbo-scholen doen zich echter enkele specifieke problemen voor. Ten eerste kon de leerwinst maar voor een kleine groep sbo-leerlingen berekend worden, omdat veel leerlingen pas instromen vanaf groep 5. Standaard worden in de pilot de medio toetsen in groep 3 als startmeting voor het berekening van leerwinst en toegevoegde waarde gebruikt. Voor sbo-scholen is dit niet zinvol, omdat hierdoor veel van hun leerlingen buiten de berekeningen vallen.

Ten tweede is een vergelijking van de leerwinst of toegevoegde waarde van sbo-scholen met die van reguliere basisscholen uit de pilot mogelijk ook van beperkte waarde. In het algemeen is de afstand in leerprestaties van sbo-leerlingen tot even oude reguliere basisschoolleerlingen dusdanig groot dat deze prestaties voor de sbo-school geen reëel doel vormen. Het vergelijken van de leerwinst van sbo-leerlingen onderling en daarbij de uitsplitsing naar de problematiek van de leerlingen, kan wel waardevolle extra inzichten opleveren. Maar daarvoor waren binnen de pilot de mogelijkheden te beperkt, gezien het kleine aantal deelnemende sbo-scholen.

Een voor de sbo-scholen uit de pilot acceptabele oplossing is gevonden door hun leerlingen per toetsmoment - en niet naar leerwinst van het ene moment naar het andere - te vergelijken met leerlingen van sbo-scholen die hebben deelgenomen aan de startmeting van het onderzoek COOLspeciaal. COOLspeciaal is een grootschalig cohortonderzoek naar schoolloopbanen in het speciaal onderwijs en speciaal basisonderwijs (Ledoux, Roeleveld, Van Langen & Smeets, 2012). Het doel van dit cohortonderzoek is om de cognitieve en sociaal-emotionele ontwikkeling van leerlingen in deze typen onderwijs in Nederland in beeld te brengen. In schooljaar 2010-2011 is voor de eerste keer een meting uitgevoerd bij so- en sbo-leerlingen met het geboortejaar 1998 en 2001. Zij waren toen ongeveer 10 en 13 jaar oud, hetgeen overeen komt met de leeftijd van groep 5- en 8-leerlingen in het reguliere basisonderwijs. Bij deze leerlingen zijn de LVS-toetsen voor speciale leerlingen afgenomen voor de leergebieden rekenen-wiskunde (M5 en M8), begrijpend lezen (M5 en M8) en technisch lezen (M5 en M8). Verder heeft de leerkracht over elke leerling een zorgprofiel-vragenlijst ingevuld. De vragenlijst bestaat uit 38 vragen over type en ernst van de problemen of beperkingen. Hiermee is een indeling gemaakt van de meest voorkomende vormen van problematiek of combinaties van types problematiek (Ledoux, Roeleveld, Van Langen & Paas, 2012. Aan de eerste meting van COOLspeciaal hebben 33 sbo-scholen met ruim 1.200 leerlingen deelgenomen. De resultaten hieruit kunnen als representatief gezien worden voor de landelijke populatie van so- en sbo-leerlingen.

Op vijf sbo-scholen uit de pilot hebben de leerkrachten ook de COOLspeciaal zorgprofiel-vragenlijst ingevuld, voor 141 leerlingen met het geboortejaar 2000 en 2003. Deze leerlingen waren in schooljaar 2012-2013 ongeveer 10 en 13 jaar oud. De zorgprofiel-informatie is gekoppeld aan de vaardigheidsscores van de verschillende M-toetsen uit de periode 2012-2013 van deze leerlingen. De leerlingen zijn op basis van hun zorgprofiel ingedeeld in veertien verschillende categorieën van problematiek, conform de indeling in COOLspeciaal.

Op basis van deze data zijn speciale schoolrapporten samengesteld waarin de vaardigheidsscore op de LVS-toetsen rekenen-wiskunde, begrijpend lezen en technisch lezen van elke individuele sbo-leerling uit de pilot in 2012-2013 is vergeleken met de gemiddelde score op dezelfde toets van alle sbo-leerlingen in COOLspeciaal in 2010-2011. Aangegeven is of de betreffende leerling uit de pilot ten opzichte van die gemiddelde COOLspeciaal score beter, gelijk of slechter scoort. De score van elke sbo-leerling uit de pilot is ook nog eens vergeleken met de gemiddelde score op dezelfde toets van alle leerlingen uit COOLspeciaal die dezelfde combinatie van soorten problematiek hebben als de betreffende leerling. Die combinatie is daarom bij elke leerling aangegeven. De vergelijking gaat op dezelfde manier als die met het algemeen gemiddelde van COOLspeciaal. Aangegeven wordt of de leerling ten opzichte van de gemiddelde COOLspeciaal-score per probleemcombinatie beter, gelijk of slechter scoort.

In tabel 6 wordt een uitgewerkt voorbeeld uit het COOLspeciaal schoolrapport gegeven voor een deelnemende sbo-school. Tabel 6 laat de vaardigheidsscore zien van leerling ZZ op begrijpend lezen op een leeftijd die leerlingen in het reguliere basisonderwijs doorgaans hebben in het midden van groep 5. Deze leerling heeft een externaliserende en internaliserende problematiek, in combinatie met communicatieve en verstandelijke problemen. Zijn vaardigheidsscore op moment M5 bedraagt -1. Dat is 'slechter' dan de gemiddelde sbo-leerling op dezelfde leeftijd, die 2,4 scoort, maar 'gelijk' aan de gemiddelde sbo-leerling met dezelfde probleemcombinatie, die 1,5 scoort.

Tabel 6 COOLspeciaal leerlingspecifieke rapportage voor de LVS-toets begrijpend lezen - M5

school	groep	leerling	problematiek	Vaardigheids-score leerling op M5	Gemiddelde vaardigheids-score op M5 in COOL^{speciaal}	leerling scoort:	Gemiddelde vaardigheids-score op M5 in COOL^{speciaal} bij zelfde problematiek	leerling scoort:
X	Y	ZZ	ext+int+comm +verstand	-1	2,4	slechter	1,5	gelijk

Op deze manier is te zien hoe de individuele leerlingen van school X scoren in vergelijking met de gemiddelde sbo-leerling en met de gemiddelde sbo-leerling met dezelfde problematiek. Om het rapport overzichtelijk te houden, zijn in de schoolrapporten de onder- en bovengrenzen van de betrouwbaarheidsintervallen niet vermeld. Zij zijn in een bijlage geplaatst. Alleen het geschatte gemiddelde is vermeld in de leerlingrapportage. Een vergelijking van gehele scholen of klassen met scholen of klassen in het COOLspeciaal onderzoek, of met andere scholen of klassen in de pilot, is niet uitgevoerd. De aantallen leerlingen per klas en school in pilot zijn te gering om betrouwbare uitspraken te kunnen doen. Dit geldt des te sterker bij de gemaakte uitsplitsingen naar de voorkomende soorten problematiek.

5.6. Evaluatie gebruik schoolrapportages

Zoals vermeld in paragraaf 2.8 is na de laatste rapportageronde (oktober-november 2013) via een enquête geëvalueerd wat de deelnemende scholen vonden van de inzichtelijkheid en de bruikbaarheid van de schoolrapportages over leerwinst en toegevoegde waarde voor schoolverbetering en voor accountabilitydoeleinden.

Deze paragraaf bevat een overzicht van de belangrijkste resultaten van de internet-enquête onder directeuren en ib-ers van de pilotscholen. Er wordt onderscheid gemaakt tussen reguliere (n=33) en sbo-scholen (n=7), omdat op voorhand al duidelijk was dat de schoolrapportages meer toegesneden zijn op het reguliere onderwijs dan op het speciaal onderwijs.

Waardering voor de rapportages

Aan de respondenten is gevraagd een rapportcijfer (een geheel getal tussen de 1 en 10) te geven voor zowel de inzichtelijkheid als de bruikbaarheid van de laatste versie van de schoolrapportages. Daarbij moet vermeld worden dat Cito en de Universiteit Twente een gezamenlijke laatste versie van de leerwinstrapportage hebben gemaakt. Per school is het schoolgemiddelde rapportcijfer bepaald. Tabel 7 toont hiervan de resultaten.

Tabel 7 Het schoolgemiddelde rapportcijfer voor de inzichtelijkheid en bruikbaarheid van de rapportages over leerwinst en toegevoegde waarde

		Reguliere scholen (n=33)				Sbo-scholen (n=7)
		gemiddeld	% onvoldoende (<5)	% neutraal (5-7)	% goed (>7)	gemiddeld
Leerwinst-schoolrapport	Inzichtelijkheid	6,9	9%	24%	67%	5,9
	Bruikbaarheid	6,4	15%	33%	52%	5,4
Toegevoegde waarde schoolrapport	Inzichtelijkheid	6,5	15%	30%	55%	5,4
	Bruikbaarheid	6,3	24%	27%	49%	5,1

Uit tabel 7 is af te lezen dat de rapportcijfers variëren tussen 5,1 (zie laatste kolom) en 6,9 (zie derde kolom). De directeuren en ib-ers van reguliere scholen geven gemiddeld hogere rapportcijfers voor de schoolrapportages dan de directeuren en ib-ers van sbo-scholen. Binnen elk schooltype krijgen de schoolrapportages over leerwinst in het algemeen een hoger rapportcijfer dan de rapportages over toegevoegde waarde. En binnen elk type schoolrapportage geven de scholen een hoger rapportcijfer voor de inzichtelijkheid dan voor de bruikbaarheid.

Hoewel de schoolgemiddelde rapportcijfers van reguliere scholen niet bijzonder hoog zijn (voldoende tot ruim voldoende), laten de onderliggende verdelingen van de rapportcijfers een genuanceerder en positiever beeld zien. Van de reguliere scholen beoordeelt 67% en 55% (zie

zesde kolom) de inzichtelijkheid van de rapportages over leerwinst respectievelijk toegevoegde waarde met een goed cijfer (= schoolgemiddelde cijfer 7 of hoger). De helft van de scholen beoordeelt de bruikbaarheid van beide rapportages als goed (zie kolom 6: leerwinst rapportages 52% en toegevoegde waarde rapportage 49%).

Motieven voor de waardering

Tabel 8 Motieven voor een onvoldoende of goed rapportcijfer voor de Leerwinst en Toegevoegde waarde schoolrapportages van reguliere pilotscholen

		gemiddeld rapportcijfer	motieven
Leerwinst-schoolrapport	Inzichtelijkheid	Onvoldoende	'Gegevens kloppen niet' 'Leerlingen staan niet per groep, maar door elkaar'
		Goed	'Duidelijke en leesbare rapportage' 'Geeft goed inzicht in ontwikkeling en leerwinst' 'Met de juiste toelichting zijn ze goed te lezen'
	Bruikbaarheid	Onvoldoende	'Aantal leerlingen te klein om conclusies te trekken' 'Moeten eerst analyseren wat we lezen' 'Krijgen geen info die we niet ook al uit ParnasSys kunnen halen' 'Bruikbaar, maar eigen overzichten en analyses beter'
		Goed	'Na uitleg konden we wel kijken waar onze sterke en zwakke punten liggen' 'Bruikbaarheid staat en valt met bijhouden van gegevens in ParnasSys' 'Rapporten worden met de leerkrachten besproken en de plannen worden afgestemd op niveaus van de leerlingen'
Toegevoegde waarde schoolrapport	Inzichtelijkheid	Onvoldoende	'Te weinig gegevens' 'Toelichting van adviseur noodzakelijk' 'Grafische weergave is duidelijk, maar interpretatie niet' 'Verschil tussen Toegevoegde waarde modellen niet duidelijk'
		Goed	'Veel tabellen. Die zijn op zich duidelijk' 'De grafiek geeft een vlot beeld. De getallen in de tabel zijn lastig te interpreteren' 'Het is goed om te zien hoe onze school scoort t.o.v. scholen met gelijke populatie'
	Bruikbaarheid	Onvoldoende	'Niet echt bruikbaar door verkeerd opleidingsniveau ouders' 'Dit deel bevat nog teveel valkuilen om een realistisch beeld te kunnen geven' 'Erg statistisch; er is nog teveel toelichting op de resultaten nodig'
		Goed	'Duidelijk' 'Het geeft een genuanceerde/eerlijker beeld van de prestaties van de leerlingen' 'Het gegeven dat de opbrengsten worden afgezet tegen scholen van een gelijke leerlingpopulatie zegt wat over de schoolopbrengsten en kan aanleiding geven tot beleidswijzigingen.'

Om meer inzicht te krijgen in de motieven voor de waardering voor de rapportages is aan de ib-ers en directeuren gevraagd om een toelichting te geven bij elk rapportcijfer. Het merendeel van hen heeft dat gedaan. Tabel 8 geeft een overzicht van de motieven van reguliere basisscholen om een onvoldoende (5 of lager) of juist een goed (7 of hoger) rapportcijfer te geven.

Opvallend is dat enkele motieven die een onvoldoende rapportcijfer onderbouwen ook terugkomen bij de motieven voor een goed rapportcijfer. Het gaat dan om de uitleg die nodig is om de schoolrapportages goed te begrijpen of te interpreteren. Voorbeelden hiervan zijn 'we moeten eerst analyseren wat we lezen', 'toelichting van de schooladviseur is noodzakelijk' en 'na uitleg konden we wel kijken waar onze sterke en zwakke punten liggen'. Voor de ene school is dit een reden om een onvoldoende te geven en voor een andere school dus niet. Verder worden bij een onvoldoende rapportcijfer motieven genoemd die te maken hebben met de kwaliteit en volledigheid van de onderliggende toetsdata of motieven die in de toekomst niet meer van toepassing zijn als scholen de rapportages zelf zouden kunnen maken met hun eigen leerlingvolgsysteem. Voorbeelden daarvan zijn 'de gegevens kloppen niet', 'aantal leerlingen is te klein om conclusies te trekken' en 'leerlingen staan niet per jaargroep maar door elkaar heen'. Opmerkelijk is verder dat er twee scholen zijn die de meerwaarde van de leerwinstrapportages niet inzien en daarom een onvoldoende geven voor de bruikbaarheid.

Opvattingen over toekomstig gebruik

Tabel 9 Meningen van directeuren ib-ers van reguliere pilotscholen over het nut en het toekomstig gebruik van rapportages over leerwinst (LW) en toegevoegde waarde (TW)

		% mee oneens	% neutraal	% mee eens
Leerwinst-schoolrapport	Ik vind het LW-rapport is een waardevolle aanvulling op wat er nu in LVS systemen zit	16%	24%	60%
	Ik ga het LW-rapport in toekomst gebruiken als het in ons LVS systeem beschikbaar komt	14%	24%	62%
	Ik denk dat het LW-rapport goed ingezet kan worden bij onderwijstoezicht door de Inspectie.	17%	22%	61%
Toegevoegde waarde schoolrapport	Het TW-rapport is een waardevolle aanvulling op groepsrapportages die nu in LVS systemen zitten.	14%	33%	53%
	Ik zal het TW-rapport in toekomst gebruiken voor opbrengstgericht werken als het voor onze school gemaakt wordt	14%	31%	56%
	Ik vind het Vaardigheidsverschil- beter dan Vaardigheidsgroei-model als het gaat om opbrengstgericht werken	19%	64%	17%
	Ik vind dat het TW-rapport een eerlijke vergelijking laat zien van de TW van mijn school met andere pilotscholen	28%	25%	47%
	Ik vind dat het TW-rapport wel door Inspectie gemaakt mag worden	31%	39%	31%
Ik ben meer opbrengstgericht gaan werken sinds de start van de pilot LTW-PO		31%	19%	50%

Aan de hand van stellingen is nagegaan wat de mening van directeuren en ib-ers is over het nut en gebruik van toekomstige rapportages over leerwinst en toegevoegde waarde. In tabel 9 zijn de resultaten gepresenteerd voor de reguliere pilotscholen.
Tabel 9 laat zien dat rapportages over leerwinst door 60% van de reguliere pilotscholen wordt beschouwd als een waardevolle aanvulling op de huidige leerling- en groepsrapportages in de LVS-systemen, dat 62% deze zal gaan gebruiken als het beschikbaar komt in hun eigen leerlingvolgsysteem en dat 61% dergelijke rapportages ook aan de inspectie zou geven bij een toekomstig schoolbezoek.

De rapportages over toegevoegde waarde vinden de reguliere pilotscholen in het algemeen iets minder vaak een waardevolle aanvulling op wat er al is (53%). Men denkt ze ook iets minder vaak te gaan gebruiken (56%). Het verschil met vergelijkbare stellingen over de leerwinstrapportages is gering (5% - 6%); er zijn wel meer scholen die geen duidelijke mening hebben. Slechts een kleine groep scholen (14% -16%) is van mening dat de ontwikkelde schoolrapportages niet nuttig en bruikbaar zijn. Daarbij maakt het niet uit of het om leerwinst of om toegevoegde waarde gaat.

Om een beter beeld te krijgen van de mening van de scholen over de toekomstige ontwikkeling van rapportages over toegevoegde waarde, zijn hierover nog drie aanvullende stellingen voorgelegd (zie tabel 9). De reguliere scholen uit de pilot blijken geen duidelijke voorkeur te hebben voor een van de twee toegevoegde waarde modellen: 64% van de scholen antwoordt neutraal, 17% vindt het vaardigheidsverschil-model beter dan het vaardigheidsgroei-model en 19% vindt precies het omgekeerde.

Een andere stelling betreft de ‘fairness’ van de vergelijking die met de toegevoegde waarde berekeningen beoogd wordt. Bijna de helft van de scholen vindt dat de toegevoegde waarde rapportages daarin slagen. Door rekening te houden met verschillen in fairness-kenmerken wordt de eigen school op een eerlijker manier vergeleken met andere scholen.
Tot slot is de stelling voorgelegd of de toegevoegde waarde in de toekomst door de Inspectie van het Onderwijs berekend mag worden. Bijna een derde deel van de reguliere scholen is hier duidelijk geen voorstander van. Daar staat tegenover dat een even groot deel van de scholen het geen probleem vindt als de inspectie dergelijke rapportages maakt. Bijna 40% van de scholen heeft hierover geen duidelijke mening.

Verbetering opbrengstgericht werken

Door deelname aan de pilot mag aangenomen worden dat de scholen meer of op een andere manier naar de leerprestaties van hun leerlingen zijn gaan kijken. De rapportages kunnen een positieve stimulans voor het opbrengstgericht werken teweeg hebben gebracht. Uit de internet-enquête blijkt dat veel pilotscholen dit inderdaad onderschrijven. De helft van de reguliere scholen is – naar eigen zeggen – sinds de start van de pilot meer opbrengstgericht gaan werken (zie tabel 9). Bijna een derde van de reguliere scholen zegt dat dit niet het geval is en bijna 20% is neutraal. Voor de zeven sbo-scholen zijn de verhoudingen ongeveer gelijk: drie sbo-scholen

zijn meer opbrengstgericht gaan werken, twee sbo-scholen zijn hierover neutraal en twee sbo-scholen zijn van mening dat dit niet zo is.

Gewenste verbeteringen van de schoolrapportages

Aan alle pilotscholen is gevraagd welke verbeteringen in de toekomst gewenst zijn om de schoolrapportages inzichtelijker en beter inzetbaar te maken. Voor leerwinst worden de volgende punten genoemd:

- andere vorm van rapporteren voor sbo-scholen;
- snellere beschikbaarheid van rapporten;
- beschikbaar maken in een leerlingvolgsysteem;
- verwijzingen naar toepassingen in de praktijk;
- betere toelichting op de resultaten;
- duidelijke handleiding met betrekking tot de interpretatie van de resultaten;
- uitgaan van (basis)groep in plaats van cohort;
- alle leerlingen meenemen.

Voor de rapportages over toegevoegde waarde zijn ook verbeterpunten genoemd. Ze komen deels overeen met de hierboven genoemde punten. Het gaat om:

- sbo-scholen vergelijken met andere sbo-scholen;
- snellere beschikbaarheid van rapporten;
- beschikbaar maken in een leerlingvolgsysteem;
- handvatten voor gebruik in de praktijk;
- betere toelichting;
- duidelijke handleiding met betrekking tot de interpretatie van de resultaten;
- meer 'fairness'- kenmerken erbij betrekken;
- liever vergelijken met landelijk gemiddelde dan met pilotscholen.

5.7. Conclusies

5.7.1. Leerwinst

Leerwinstberekeningen zijn in het algemeen gebaseerd op de groei in leerprestaties van leerlingen in een bepaald leergebied gedurende een langere onderwijsperiode. Van leerlingen uit bijvoorbeeld groep 4 kan de leerwinst worden berekend door hun prestaties te vergelijken met die van dezelfde leerlingen toen ze nog in groep 3 zaten. Uit de geconstateerde leerwinst kan blijken of die leerlingen gegroeid zijn in hun leerprestaties, ongeacht of zij aanvankelijk hoog of laag presteerden op een toets, of een zwakke of begaafde leerling zijn. Bij leerwinstberekeningen wordt echter niet gecontroleerd voor achtergrondkenmerken van leerlingen. Daardoor is uit de leerwinst niet af te leiden welke schoolse factoren tot de leerwinst hebben geleid. Dit gebeurt wel bij het schatten van toegevoegde waarde (Harris, 2011).

Omdat de vaardigheidsschalen die ten grondslag liggen aan de LVS-toetsen van Cito voor de verschillende leerstofgebieden niet hetzelfde zijn, kan leerwinst alleen per leerstofgebied worden berekend. De leerwinst in verschillende leerstofgebieden kan dus niet worden getotaliseerd. Om de leerwinst per leerstofgebied te kunnen beoordelen is er in de pilot voor

gekozen de groei van een leerling te vergelijken met de groei van leerlingen met hetzelfde startniveau. De norm voor voldoende of onvoldoende groei wordt ontleend aan de gemiddelde groei van deze groep leerlingen met dezelfde vaardigheidsscore. Door vervolgens de feitelijke groei van een leerling te vergelijken met het gemiddelde van de referentiegroep blijkt of een leerling in een bepaalde periode meer, gelijk, of minder gegroeid is dan andere leerlingen die aanvankelijk dezelfde vaardigheidsscore hadden. Er is dus gekozen voor een relatieve norm omdat de leerwinst van een individuele leerling wordt afgemeten aan de groei van een referentiegroep.

Het leerwinstmodel ontwikkeld door Cito en de twee modellen van de Universiteit Twente vertonen veel overeenkomsten. Voor beide modellen geldt dat de leerwinst wordt bepaald en beoordeeld door leerlingen met eenzelfde uitgangspositie met elkaar te vergelijken. Het Z-score model vergelijkt leerlingen op basis van vaardigheidsscores (zie figuur 6) en het Groeitempo-model doet dat op basis van vaardigheidsniveaus (zie figuur 4). Omdat het Z-score model uitgaat van vaardigheidsscores is daarmee de leerwinst preciezer te bepalen en duidelijker weer te geven dan met het Groeitempo-model. Daarom is door zowel de deelnemende scholen als de projectgroep de voorkeur uitgesproken voor het Z-score model. De technische verantwoording van dit model is te vinden in bijlage 5.

5.7.2. Toegevoegde waarde

In de pilot zijn twee methoden uitgeprobeerd om de toegevoegde waarde van een basisschool te bepalen: het vaardigheidsverschil-model en het vaardigheidsgroei-model. Beide modellen gaan uit van LVS-toetsen over hetzelfde leerstofgebied van dezelfde groep leerlingen. Het ene model is gebaseerd op de vaardigheidsscores op twee LVS-toetsen. Deze twee toetsen zijn op te vatten als een begin- en eindmeting van hetzelfde leerstofgebied. Daartussen is het verschil berekend en vandaar de naam vaardigheidsverschil-model. Met dit model kan op schoolniveau de toegevoegde waarde geschat worden over een langere onderwijsperiode waarbij slechts gebruik wordt gemaakt van gegevens van twee toetsmomenten die qua tijd zelfs ver uit elkaar kunnen liggen.

Het andere model is gebaseerd op vaardigheidsscores van meer toetsmomenten uit hetzelfde leerstofgebied van dezelfde groep leerlingen. Daarmee is op schoolniveau de groei van de leerlingen in een bepaald leerstofgebied in kaart te brengen. Vandaar dat aan dit model de naam vaardigheidsgroei-model is gegeven. Alle beschikbare toetsmomenten kunnen worden meegenomen in dit model. In het geval een school voor alle leerstofgebieden halfjaarlijks alle LVS-toetsen afneemt, kan op deze wijze voor elk leerstofgebied de toegevoegde waarde over nagenoeg alle tussenliggende leerjaren uit de hele basisschoolperiode worden bepaald.

Met het vaardigheidsgroei-model kan de leerwinst in een breder perspectief worden geplaatst. Bij het verschil-model, waarbij sprake is van twee meetmomenten, wordt de toegevoegde waarde uitgedrukt in een getal. Bij het groei-model wordt de ontwikkeling zichtbaar in de vorm van een groeilijn die bijvoorbeeld afbuigt of stijgt. De relatie met eerdere prestaties uit voorgaande schooljaren is dan direct te zien. Naast de mogelijkheid (half)jaarlijks terug te kijken op de toegevoegde waarde, biedt het groei-model ook de mogelijkheid (half)jaarlijks

vooruit te kijken naar het volgende moment waarop de toegevoegde waarde bepaald wordt. De groeilijn kan immers worden geëxtrapoleerd. Mocht daaruit bijvoorbeeld blijken dat de toegevoegde waarde lijkt af te nemen dan kunnen daar tijdig eventueel consequenties voor de inrichting en inhoud van het onderwijs aan verbonden worden.

Ofschoon de scholen de uitkomsten van beide modellen op hun waarde weten te schatten, gaat de voorkeur uit naar het vaardigheidsgroei-model. Dat model biedt zowel de mogelijkheid om terug te kijken op de geleverde toegevoegde waarde, als vooruit te kijken naar de mogelijke ontwikkeling van de toegevoegde waarde. Dit model kan dus in dienst staan van zowel het *schoolverbeterings-* als het *accountabilityperspectief.*

De pilot heeft laten zien dat het mogelijk is om de toegevoegde waarde van scholen te schatten op basis van leerwinstgegevens die gecorrigeerd worden voor relevante niet-schoolse factoren. De bepaling van toegevoegde waarde is echter nog te complex om deze in bestaande toets- en schoolinformatiesystemen toe te passen. Ook zitten er de nodige haken en ogen aan de nauwkeurigheid van toegevoegde waarde schattingen. Bekende knelpunten zijn: de invloed van meetfouten, het aantal en de soort correctiefactoren, scholen met kleine aantallen leerlingen, de stabiliteit van toegevoegde waarde-schattingen en eventueel strategisch gedrag van scholen (zie ook par. 3.5).

Deze knelpunten hebben zich niet allemaal in de pilot voorgedaan. Dat heeft onder andere te maken met de kleinschaligheid waarbinnen deze maten zijn ontwikkeld en uitgeprobeerd. Zo hebben bijvoorbeeld de stabiliteit van de schattingen over een langere termijn en de invloed van strategisch gedrag van scholen geen rol van betekenis gespeeld. In de pilot is niet geëxperimenteerd met de rol die de modellen kunnen spelen bij de verantwoording door de scholen over hun opbrengsten of bij de beoordeling van de opbrengsten door de onderwijsinspectie. Daardoor is er ook geen ervaring op gedaan met de effecten en eventuele neveneffecten van het gebruik dat leerwinst en toegevoegde waarde in *accountabilityperspectief*. Het zijn echter wel factoren waarmee rekening gehouden moet worden bij de toepassing van leerwinst en toegevoegde waarde in het primair onderwijs. Voor sommige van de genoemde knelpunten zijn oplossingen te bedenken, maar voor andere knelpunten heeft ook de wetenschap nog geen afdoende antwoord. Dat geldt bijvoorbeeld voor de vraag naar hoeveel en welke niet-schoolse factoren het best gecorrigeerd kan worden om tot een nauwkeurige schatting van de toegevoegde waarde te komen. Om die reden is nader onderzoek en ontwikkelingswerk noodzakelijk.

De projectgroep en de deelnemende scholen willen niet ontkennen dat er allerlei haken en ogen zitten aan de schatting van de toegevoegde waarde of aan het openbaar maken van deze gegevens, maar vinden dit geen reden om af te zien van de verdere ontwikkeling van deze vorm van opbrengstbeoordeling. De ontwikkelde maten zijn bruikbaar op schoolniveau en ze leveren een bijdrage aan een eerlijkere verantwoording over de leerresultaten aan de buitenwereld. De projectgroep pleit er daarom voor de verdere ontwikkeling van deze toegevoegde waarde modellen ter hand te nemen waarbij het initiatief bij de scholen ligt en blijft.

Door de verdere ontwikkeling en toepasbaarheid van - in het bijzonder - de toegevoegde waarde in een 'beschermde omgeving' en onder beheer van de scholen te laten plaatsvinden, is de kans op strategisch gedrag mogelijk bijzonder klein. Vandaar dat overwogen kan worden een plan op te stellen om de toegevoegde waarde en de toepassing ervan in de verantwoording aan derden over de leerprestaties van scholen, verder te ontwikkelen. De combinatie leerwinst en toegevoegde waarde biedt immers mogelijkheden om de leerprestaties van scholen op een eerlijkere manier te beoordelen dan op basis van de huidige systematiek het geval is.

5.7.3. Evaluatie schoolrapportages

In deze pilot is getracht door middel van rapportages over leerwinst en toegevoegde waarde basisscholen nieuwe instrumenten in handen te geven om de leerprestaties van hun leerlingen te meten en te evalueren. Via deze schoolrapporten kregen de deelnemende scholen de leerwinst van individuele en van groepen leerlingen in beeld. Bovendien konden de scholen op basis van de rapportages over de toegevoegde waarde een oordeel vellen over de leerwinst op hun school. Bestaande rapportages bieden deze mogelijkheden niet of minder zorgvuldig. De ontwikkelde modellen en de daarop gebaseerde rapportages beogen daarom in de eerste plaats basisscholen hulpmiddelen te bieden om enerzijds het onderwijs nog beter op hun leerlingen af te stemmen en anderzijds de leeropbrengsten van de school beter te beoordelen (schoolverbeteringsperspectief).

Bruikbaarheid en inzichtelijkheid

Uit de internet-enquête onder ib-ers en directeuren kan worden afgeleid dat de meeste deelnemende scholen vinden dat de pilot erin is geslaagd een nieuwe en bruikbare bijdrage te leveren aan de mogelijkheden die de bestaande toets- en schoolinformatiesystemen bieden voor schoolverbetering. Ruim de helft van de respondenten is van mening dat de rapportages inzichtelijk en goed bruikbaar zijn voor dit doel, een kwart tot een derde is hierover neutraal en slechts vijftien procent van de scholen is van mening dat dit niet het geval is.

Enige nuancering is wel op zijn plaats. De ontwikkelde manier van rapporteren wordt geschikter geacht voor reguliere basisscholen dan voor sbo-scholen. De leerwinstrapportages vinden de scholen inzichtelijker en bruikbaarder dan de rapportages over de toegevoegde waarde. De scholen vinden het essentieel dat de rapportages vergezeld gaan van:

- een goede uitleg van de modellen,
- een duidelijke toelichting op de rapportage en
- een handleiding voor de interpretatie van de resultaten.

Deze verbeterpunten gelden sterker voor rapportages over toegevoegde waarde dan voor rapportages over leerwinst. Vooral de toegevoegde waarde rapportages vinden scholen lastig om te begrijpen en om goed te interpreteren. Hier zal in de toekomst nog het nodige aan verbeterd moeten worden, voordat basisscholen het op grote schaal kunnen en zullen gaan gebruiken.

Accountabilityperspectief

In de tweede plaats beogen de rapportages ook een hulpmiddel te zijn voor accountabilitydoeleinden. Uit de internet-enquête onder ib-ers en directeuren is gebleken dat de meeste scholen de leerwinstrapportages daarvoor wel geschikt vinden. Zestig procent van de reguliere pilotscholen is van mening dat deze rapportages goed ingezet kunnen worden voor het toezicht op de kwaliteit van het onderwijs door de onderwijsinspectie. Twintig procent is hierover neutraal en slechts zeventien procent is het hier mee oneens.
Over de inzetbaarheid van toegevoegde waarde voor accountability bestaat minder eensgezindheid dan over leerwinst. Bijna de helft van de scholen is van mening dat de ontwikkelde rapportages in hun huidige vorm een eerlijke vergelijking van de toegevoegde waarde van hun school met die van andere pilotscholen laat zien. De achterliggende redenen zijn divers. Hier is dus nog het nodige ontwikkelwerk te verzetten.

Toch zijn er redenen genoeg om met toegevoegde waarde rapportages door te gaan op de ingeslagen weg. De helft van de scholen vindt dat ze door de rapportages meer opbrengstgericht zijn gaan werken sinds de start van de pilot. Dit is in overeenstemming met de resultaten uit recent onderzoek van Deunk en Doolaard (2013). Verder geeft bijna zestig procent van de reguliere scholen uit de pilot aan dat ze in de toekomst rapportages over de toegevoegde waarde zullen gaan inzetten voor opbrengstgericht werken als ze voor hun school beschikbaar zijn. Waar deze rapportages dan gemaakt moeten worden en door wie, is een kwestie die weliswaar buiten het bestek van de pilot valt, maar de scholen is wel naar hun mening hieromtrent gevraagd. Bijna een derde deel van de reguliere scholen vindt – blijkens de internet-enquête – dat de onderwijsinspectie rapportages over toegevoegde waarde rapportages mag maken, maar een even groot deel van scholen vindt precies het tegenovergestelde. Bijna veertig procent van de reguliere scholen heeft hier geen uitgesproken mening over. Deze resultaten zijn verrassend gezien de maatschappelijke discussie rond het gebruik van toetsen in het basisonderwijs, het gebruik van toetsen door de onderwijsinspectie en het eventuele strategisch gedrag dat dit bij scholen oproept.

5.7.4. Voorwaarden voor de invoering van leerwinst en toegevoegde waarde

In de pilot deden zich specifieke problemen voor bij de berekening van leerwinst en toegevoegde waarde. Deze werden vooral veroorzaakt doordat de leerwinst en de toegevoegde waarde niet in de toets- en schoolinformatiesystemen van de scholen berekend konden worden. Per deelnemende school moesten bestanden met toetsgegevens uit hun systemen worden geëxporteerd om de onderzoeksinstituten de berekeningen te laten uitvoeren. Daaruit is gebleken dat over de inrichting van bestanden afspraken met de scholen en de aanbieders van toets- en schoolinformatiesystemen gemaakt moeten worden, omdat anders erg veel data-cleaning nodig is om deze bestanden geschikt te maken voor analyses. Dat heeft enerzijds te maken met verschillen tussen toets- en schoolinformatiesystemen, maar anderzijds ook met de invoerdiscipline van de scholen en met het feit dat er nu eenmaal ook gegevens in die systemen ontbreken.

Voorts is het van belang met de scholen afspraken te maken over de gegevens die nodig zijn om de leerwinst en toegevoegde waarde te corrigeren. In de pilot is ervoor gekozen om in ieder geval gegevens over leerlingen te gebruiken waarover de scholen in principe beschikken, namelijk opleidingsniveau van ouders en het type zorgleerling. Overwogen kan worden de correctiefactoren te beperken tot kenmerken die ontleend kunnen worden aan externe databestanden, zoals de gemeentelijke registratie, het CBS of het onderwijsnummer. In dat verband is in de pilot nagegaan of de standaardmeetfout van de toetsen als correctiefactor is te gebruiken. Dit is namelijk een gegeven dat door de toetsleverancier kan worden aangeleverd en niet door de scholen behoeft te worden verzameld. Een eerste verkenning naar het effect van de standaardmeetfout in een toegevoegde waarde model leverde op dat dit een correctiefactor van enige betekenis is. De schatting van de toegevoegde waarde wordt er beter door en het gegeven kan op eenvoudig wijze beschikbaar gesteld worden (Rekers-Mombarg & Janssens, 2012).

Op basis van de uitkomsten van de pilot en de mening van de scholen daarover, zijn de volgende voorwaarden gedefinieerd waaraan moet worden voldaan om leerwinst en toegevoegde waarde met succes in de onderwijspraktijk te kunnen implementeren.

1. *De school maakt per leerstofgebied gebruik van kwalitatief goede toetsen gebaseerd op een vaardigheidsschaal*

De bepaling van leerwinst of toegevoegde waarde is afhankelijk van de mate waarop deze is gebaseerd op betrouwbare en valide gegevens. Dat betekent dat scholen kwalitatief goede toetsen volgens de voorschriften op de juiste momenten afnemen. De toetsen die in aanmerking komen voor de bepaling van de leerwinst en de toegevoegde waarde zijn de LVS-toetsen van Cito, inclusief de toetsen voor speciale leerlingen. Deze zijn, afhankelijk van het leerstofgebied, voor meerdere leerjaren beschikbaar, waaronder voor rekenen-wiskunde, technisch lezen, begrijpend lezen en woordenschat. Deze toetsen zijn gebaseerd op een zogenaamde vaardigheidsschaal. Zo'n schaal staat toe dat de leerwinst over meerdere leerjaren berekend kan worden omdat alle toetsen van een bepaald leerstofgebied aan die schaal kunnen worden gerelateerd. En aangezien 95% van de po-scholen deze toetsen gebruikt, hoeft omwille van het berekenen van de leerwinst of toegevoegde waarde de toetsdruk op de scholen niet toe te nemen.

Om tot een betekenisvolle interpretatie van de leerwinst of toegevoegde waarde te komen dienen deze maten gebaseerd te zijn op gegevens van toetsen uit één leerstofgebied. De verschillende schalen kunnen dus niet worden getotaliseerd. Door voor de leerstofgebieden afzonderlijk de leerwinst of toegevoegde waarde te bepalen wordt niet alleen recht gedaan aan de inhoudelijke verschillen tussen deze leerstofgebieden, maar wordt de scholen ook meer gelegenheid geboden om op de leerwinst en toegevoegde waarde te sturen. De mogelijkheden hiertoe en de betrouwbaarheid van de bepaling van de leerwinst en toegevoegde waarde worden vergroot naarmate deze is gebaseerd op meer dan twee metingen per leerling.

2. *Toets- en schoolinformatiesystemen bevatten een module om leerwinstberekeningen uit te voeren en normgegevens om de leerwinst te beoordelen*

De leerwinstberekening zoals uitgeprobeerd in de pilot is gebaseerd op relatief eenvoudige statistische berekeningen. De basis voor deze berekeningen wordt immers gevormd door de toetsgegevens (vaardigheidsscores) van de leerlingen van een school. De statistische analyses kunnen via een speciale module in de verschillende toets- en schoolinformatiesystemen worden ingebouwd.
Toetsaanbieders dienen tabellen met normgegevens beschikbaar te stellen om de leerwinst op individueel-, groeps- en schoolniveau te kunnen vaststellen. Dergelijke normgegevens dienen in de module te worden geïntegreerd.
Tot slot dienen de toets- en schoolinformatiesystemen te worden ingericht om overzichten te kunnen produceren van de leerwinst op leerling-, groeps- en schoolniveau.

3. *De toetspraktijk van scholen en de invoer van toetsgegevens door de school voldoen aan een aantal eisen.*

Door de pilot hebben de deelnemende scholen meer inzicht gekregen in hun eigen toetssystematiek en waar er eventueel hiaten zitten in de afnamesystematiek en de registratie. Voor een betrouwbare bepaling van de leerwinst van een leerstofgebied is het van belang dat toetsen op vaste momenten worden afgenomen en dat de uitkomsten op een systematische wijze worden geregistreerd. Daarom is het belangrijk dat scholen gebruik maken van een toetskalender die gebaseerd is op de toetsmomenten zoals in de handleidingen van de LVS-toetsen van Cito staan beschreven. Ook dienen de toetsen te worden afgenomen onder de condities zoals beschreven in die handleidingen.

Na de toetsafnames dienen de resultaten in een digitaal toets- of schoolinformatiesysteem te worden ingevoerd. Om een groep leerlingen te definiëren waarvoor de leerwinst moet worden bepaald is het noodzakelijk dat de gegevens voor alle leerlingen op een correcte en eenzelfde manier worden ingevoerd, bij voorkeur de toetsscores. Van belang is verder dat de gegevens van de leerling kloppen en worden bijgehouden. Daarbij gaat het niet alleen om de naam van de leerling, de geboortedatum, maar ook om het moment waarop de leerling op de school is gestart of ingestroomd en in welke leerjaar de leerling jaarlijks zit. Verder is het van belang dat per toets en per leerling de afnamedatum van de toets wordt geregistreerd en niet (alleen) de invoerdatum.

4. *Toets- en schoolinformatiesystemen bieden mogelijkheden om leerlingcohorten te definiëren waarover de leerwinst kan worden berekend.*

Een belangrijke voorwaarde voor de berekening van leerwinst of toegevoegde waarde is dat de groep leerlingen waarvoor deze worden berekend, duidelijk is gedefinieerd. Dat kan een jaargroep zijn, bijvoorbeeld alle leerlingen uit groep 3 waarvan de leerwinst in opeenvolgende leerjaren wordt gevolgd. De leerwinst wordt dan vastgesteld per groep waarbij genegeerd wordt of een leerling uit die groep is blijven zitten of een groep heeft overgeslagen sinds bijvoorbeeld groep 3. Dit levert in ieder geval overzichten op die betekenisvol zijn voor een groepsleerkracht, maar bij de beoordeling van de leerwinst op groepsniveau dient wel met de verschillen in genoten onderwijstijd rekening te worden gehouden. Voorts wisselt de

samenstelling van jaargroepen nogal eens en niet alle scholen zijn volgens het jaargroepprincipe georganiseerd, zoals sbo-scholen.

In de pilot is gekozen voor een benadering waarbij de leerwinst en toegevoegde waarde is bepaald op basis van de tijd die leerlingen op school zitten. Een cohort bestond uit leerlingen die allemaal tegelijk in groep 3 zijn gestart en waarvan de ontwikkeling in de tijd is gevolgd. Een beperking van deze aanpak is dat het cohort ook leerlingen bevat die zijn blijven zitten of een leerjaar hebben overgeslagen sinds groep 3. De leerlingen uit dat cohort hebben niet allemaal noodzakelijkerwijs hetzelfde onderwijsaanbod gehad en zitten mogelijk in verschillende leerjaren Als de leerwinst halfjaarlijks wordt berekend ontstaan er dus overzichten van leerlingen die in verschillende groepen of klassen kunnen zitten. Daarmee dient rekening te worden gehouden bij het produceren van leerwinstoverzichten. Een leerkracht moet de leerwinst van al zijn of haar leerlingen terug kunnen vinden in één overzicht.

Een specifiek probleem bij het samenstellen van cohorten voor de berekening van leerwinst en toegevoegde waarde is de tussentijdse in- en uitstroom van leerlingen. Niet alle leerlingen zijn op het zelfde moment met het onderwijs op een bepaalde school gestart en ook niet alle leerlingen ronden hun schoolloopbaan op dezelfde school af. Voor de beoordeling van de vraag of een leerling op een school voldoende leerwinst heeft geboekt, maakt het echter niet uit wanneer deze is ingestroomd of dat de leerling tussentijds naar een andere school gaat. Voor de tijd dat de leerling op dezelfde school zit kan de leerwinst worden bijgehouden en beoordeeld, omdat er geen rekening hoeft te worden gehouden met het instroomniveau of de bijdrage van een vorige school aan dat niveau. De nieuwe school zal moeten aansluiten bij de situatie waarmee een leerling instroomt.

Tussentijdse in- en uitstroom speelt vooral een rol bij de bepaling van de toegevoegde waarde omdat daar de vraag gesteld wordt naar de bijdrage van de school aan de leerwinst van leerlingen. Voor leerlingen die tussentijds zijn ingestroomd geldt dat alleen de leerwinst die bereikt is op de school waar ze feitelijk zitten, in de schatting van de toegevoegde waarde kan worden meegenomen. Overigens wordt het steeds meer praktijk dat bij tussentijdse in- en uitstroom via het onderwijskundige rapport de ontvangende of de verwijzende school ook de verschillende behaalde vaardigheidsscores rapporteert die vervolgens in de schoolinformatiesystemen kunnen worden ingevoerd. Daarmee is ook de geschiedenis van de leerwinst van deze leerlingen te reconstrueren en bij de beoordeling van de aanpassing van het onderwijs op de leerling te betrekken. Maar ook in dit geval dient bij de bepaling van de toegevoegde waarde rekening te worden gehouden met het vraagstuk welke school verantwoordelijk is voor welk deel van de leerwinst.

Overigens is het vaardigheidsverschil-model gevoeliger voor het ontbreken van leerlinggegevens dan het vaardigheidsgroei-model. Dat heeft vooral te maken met het feit dat in het eerste model slechts twee toetsafnames een rol spelen. Omdat in het groeimodel alle beschikbare toetsafnames worden betrokken, is het model beter bestand tegen incomplete gegevens dan het verschilmodel (zie ook par. 5.4.3).

6. Slotbeschouwing - conclusies en aanbevelingen

6.1. Inleiding

Het eerste doel van de pilot is na te gaan of het mogelijk is op basis van de bestaande toetspraktijk in het primair onderwijs en de daar in gebruik zijnde toets- en schoolinformatiesystemen, leerwinst te berekenen en toegevoegde waarde te schatten. Het tweede doel richt zich op de vraag welke bijdrage deze maten kunnen leveren aan opbrengstgericht werken en de beoordeling van de opbrengsten. Dit tweede doel is uiteengerafeld in twee subdoelen.
Leerwinst en toegevoegde waarde in het perspectief van:

1. Schoolverbetering
2. Accountability

Onder het *schoolverbeteringsperspectief* wordt verstaan dat leerwinst en toegevoegde waarde gebruikt kunnen worden voor zowel de afstemming van het onderwijs op de leerling, als voor het beoordelen van de leerprestaties door de school zelf, bijvoorbeeld in het kader van de kwaliteitszorg.
Leerwinst en toegevoegde waarde kunnen ook betrokken worden in de verantwoording door de scholen over de leerprestaties aan de ouders en de Inspectie van het Onderwijs. De inspectie op haar beurt kan leerwinst en toegevoegde waarde betrekken in de beoordeling van de leerprestaties van scholen. Omdat hier rekenschap en openbaarmaking een belangrijke rol spelen, wordt dit doel in het rapport omschreven als leerwinst en toegevoegde waarde in *accountabilityperspectief*.

Aan beide doelen ligt ook een aantal vragen ten grondslag, namelijk:

1. Zijn toetsen voor leerlingen uit groep 3 geschikt voor een beginmeting?
2. Welke toetsen van welke toetsaanbieders zijn bruikbaar voor de bepaling van leerwinst en toegevoegde waarde?
3. Kan voor verschillende groepen leerlingen binnen de school (zwak, gemiddeld, excellent) de leerwinst en toegevoegde waarde afzonderlijk worden bepaald?
4. Kan leerwinst en toegevoegde waarde een bijdrage leveren aan het gebruik van het ontwikkelingsperspectief voor individuele leerlingen in het speciaal (basis)onderwijs?

Ofschoon beide doelstellingen en de bovengenoemde vragen ruimschoots in de pilot aan de orde zijn gesteld, is er vooral aandacht besteed aan de ontwikkeling van verschillende maten voor leerwinst en toegevoegde waarde en aan de optimalisering van de rapportage over deze maten aan de scholen. Enerzijds hadden die rapportages tot doel de school zo duidelijk mogelijk te informeren over de leerwinst en toegevoegde waarde met het oog op consequenties voor het opbrengstgericht werken. Anderzijds stellen de rapportages de school ook in staat de kwaliteit van de leerwinst en de toegevoegde waarde te beoordelen. Het accent in de pilot lag dus op het gebruik van leerwinst en toegevoegde waarde binnen het schoolverbeteringsperspectief.

Om bovenstaande redenen kunnen op basis van bevindingen uit de pilot alleen conclusies getrokken worden over de toepasbaarheid en de bruikbaarheid van leerwinst en toegevoegde waarde in schoolverband. Om de bevindingen uit de pilot ook toepasbaar en bruikbaar te maken voor andere scholen uit het primair onderwijs wordt een aantal aanbevelingen gedaan gericht op de overheid, de scholen en op de aanbieders van toets- en schoolinformatiesystemen.

Naar de rol, de effecten en de eventuele neveneffecten van leerwinst en toegevoegde waarde in een accountabilityperspectief is in de pilot geen onderzoek gedaan. Er is bijvoorbeeld niet nagegaan op welke wijze de maten voor leerwinst en toegevoegde waarde door scholen uit de pilot gebruikt kunnen worden om zich over de leerprestaties te verantwoorden aan bijvoorbeeld ouders of de Inspectie van het Onderwijs. Ook is geen onderzoek gedaan naar de eventuele consequenties die de maten voor leerwinst en toegevoegde waarde hebben voor de beoordeling van opbrengsten door de Inspectie van het Onderwijs. Om die reden kunnen geen conclusies getrokken worden over de toepasbaarheid van leerwinst en toegevoegde waarde vanuit een accountabilityperspectief. Maar omdat er wel literatuuronderzoek is gedaan en omdat het accountabilityperspectief wel met de deelnemende scholen is bediscussieerd, worden in dit rapport toch enkele aanbevelingen gedaan voor een nadere verkenning van het gebruik van leerwinst en toegevoegde waarde voor dit doel.

Leeswijzer

Op basis van de uitkomsten van de pilot zijn in totaal vijf conclusies en 13 aanbevelingen geformuleerd over de mogelijkheid om leerwinst te berekenen en toegevoegde waarde te schatten. Daarbij is ook aandacht voor de vraag of deze berekeningen en schattingen in de bestaande toets- en schoolinformatiesystemen kunnen worden ingebouwd.

Hoewel er overlap tussen leerwinst en toegevoegde waarde bestaat, wordt er wat betreft de conclusies en aanbevelingen nadrukkelijk onderscheid tussen beide gemaakt. Deze keuze is een gevolg van enerzijds een verschil tussen beide maten in complexiteit en toepasbaarheid en anderzijds het perspectief van waaruit ze gebruikt worden. Leerwinst is voor scholen makkelijker te doorgronden dan toegevoegde waarde, zo blijkt uit de pilot. Verder bieden volgens de deelnemende scholen leerwinstgegevens meer mogelijkheden om er consequenties voor de inrichting van het onderwijs aan te verbinden, dan gegevens over de toegevoegde waarde van de school. Uit de evaluatie van de pilot onder de deelnemende scholen blijkt dat het gebruik van leerwinst en toegevoegde waarde in een accountabilityperspectief gevoelig ligt.

In paragraaf 6.2 wordt eerst ingegaan op de voorwaarden voor het kunnen berekenen van leerwinst en toegevoegde waarde, in het bijzonder op de toetsen die daarvoor gebruikt kunnen worden.

In paragraaf 6.3 worden conclusies en aanbevelingen geformuleerd over de berekening van *leerwinst*. Aan de orde komt de vraag of het mogelijk is maten voor leerwinst te ontwikkelen en of deze in toets- en schoolinformatiesystemen kunnen worden ingebouwd.

Conclusies en aanbevelingen over *toegevoegde waarde* worden geformuleerd in paragraaf 6.4. Daarbij gaat het om de vraag of de toegevoegde waarde van een school berekend kan worden en of dit een onderdeel zou kunnen zijn van de bestaande toets- en informatiesystemen. Er worden

aanbevelingen gedaan voor een nadere verkenning van leerwinst en toegevoegde waarde in een *accountabilityperspectief*.

Paragraaf 6.5 geeft enkele overwegingen bij de toepasbaarheid van leerwinst en toegevoegde waarde door de Inspectie van het Onderwijs.

6.2. Toetsinstrumentarium

Basisscholen maken volop gebruik van toetsen van verschillende toetsaanbieders om de vorderingen van hun leerlingen te volgen. Deze toetsen worden gebruikt om het onderwijs af te stemmen op de leerlingen en om de resultaten van het onderwijs te beoordelen. De bestaande toets- en schoolinformatiesystemen bieden verschillende mogelijkheden om rapportages te produceren op leerling-, groeps- en schoolniveau. Tot op heden boden deze systemen de scholen echter niet de mogelijkheid om een oordeel te vormen over de vraag of iedere leerling voldoende vorderingen maakt en deze informatie te gebruiken voor het opbrengstgericht werken. Ook bieden de bestaande systemen de scholen niet de mogelijkheid zich een oordeel te vormen over de vraag of er sprake is van toegevoegde waarde.

Met het oog op een brede implementatie in het primair onderwijs zijn twee voorwaarden geformuleerd waaraan de uitkomsten van de pilot moeten voldoen. Een van die voorwaarden is dat zoveel mogelijk aangesloten diende te worden op de bestaande toetspraktijk in het basisonderwijs. Afgesproken is dat de scholen die aan de pilot mee konden doen een serie gestandaardiseerde toetsen gebruikten, waarmee de vorderingen van leerlingen gedurende een langere periode konden worden geïnterpreteerd. Omdat alle participerende scholen LVS-toetsen van Cito gebruikten, zijn de modellen die in de pilot zijn ontwikkeld en uitgeprobeerd op deze toetsen gebaseerd. In de pilot is daarom geen ervaring opgedaan met de bepaling van leerwinst en toegevoegde waarde op basis van toetsen van andere aanbieders.

Een belangrijke voorwaarde voor het berekenen van leerwinst en het schatten van toegevoegde waarde is dat toetsen onderling vergelijkbaar zijn, waardoor de prestatiegroei van leerlingen in een bepaald leergebied zichtbaar wordt. Dat betekent dat er een serie toetsen per leerstofgebied beschikbaar moet zijn. Er zijn in Nederland verschillende aanbieders van toetsen voor het primair onderwijs die dergelijke series beschikbaar stellen, zoals Boom Test Uitgevers en Cito.

Toetsen die gebaseerd zijn op een vaardigheidsschaal die een lange onderwijsperiode beslaat zijn geschikt om de leerwinst en de toegevoegde waarde van een school te bepalen. Dit is het geval bij bijvoorbeeld de LVS-toetsen van Cito. Deze zijn voor verschillende leerstofgebieden beschikbaar. Voor andere aanbieders van toetsen voor het primair onderwijs geldt dat zij kunnen nagaan in welke mate hun toetsseries geschikt zijn of geschikt gemaakt kunnen worden om leerwinst en toegevoegde waarde te berekenen. Uit de pilot is gebleken aan welke voorwaarden voldaan moet worden om dit praktisch toepasbaar te maken.

6.3. Leerwinst

Hieronder worden twee conclusies getrokken die specifiek betrekking hebben op het in kaart brengen en gebruiken van de leerwinst door scholen.

CONCLUSIE 1
Het is mogelijk om leerwinst te berekenen. Deze berekeningen zijn een zinvolle uitbreiding van evaluatiegegevens die basisscholen kunnen gebruiken bij het opbrengstgericht werken.

Omdat alle scholen uit de pilot gebruik maken van LVS-toetsen van Cito is voldaan aan twee belangrijke voorwaarden om leerwinst te kunnen berekenen. In de eerste plaats staan deze toetsen interpretaties toe in termen van groei op een vaardigheidsschaal. Deze vaardigheidsschaal maakt het mogelijk om enerzijds de resultaten van een leerling op verschillende toetsmomenten met elkaar te vergelijken. Anderzijds kunnen met deze schaal ook de resultaten van leerlingen in dezelfde groep worden vergeleken die verschillende toetsen hebben gemaakt. Ieder leergebied waarvoor toetsen beschikbaar zijn, kent zijn eigen vaardigheidsschaal. Daarom kan de leerwinst alleen per leerstofgebied en niet over de verschillende leerstofgebieden heen worden uitgerekend.

In de tweede plaats dekken de LVS-toetsen belangrijke leerstofgebieden over meerdere leerjaren van het basisonderwijs, zoals rekenen-wiskunde en technisch- en begrijpend lezen. Daarmee zijn deze toetsen 'gevoelig' voor de kwaliteit van de geboden instructie en is er een relatie te leggen tussen de hoogte van de scores en de kwaliteit van het gegeven onderwijs. Dit is ook een belangrijke voorwaarde voor de bepaling van toegevoegde waarde. Via een toegevoegde waarde model wordt immers geprobeerd een schatting te maken van de bijdrage van de school aan de leerprestaties.

Periode waarover de leerwinst kan worden berekend

Technisch gesproken is het geen probleem om de leerwinst te berekenen over de hele onderwijsperiode waarvoor een vaardigheidsschaal beschikbaar is, zoals de periode van groep 3 tot en met groep 8. Daar zijn echter twee kanttekeningen bij te plaatsen. In de eerste plaats is het de vraag of basisvaardigheden over de jaren heen vergelijkbaar zijn: lezen of rekenen-wiskunde in groep 8 is niet hetzelfde als in groep 3. Naarmate het tijdsinterval tussen twee meetmomenten groter wordt, neemt de kans op een verandering in de inhoud van de gemeten vaardigheid toe.
In de tweede plaats beperkt de lengte van de periode waarover de leerwinst wordt berekend, de gebruiksmogelijkheden van deze maten. Voor scholen heeft een berekening van de leerwinst over bijvoorbeeld de periode vanaf groep 3 tot en met 8 met het oog op schoolverbetering weinig zin. Dit tijdsinterval biedt geen mogelijkheid om tussentijds maatregelen te treffen omdat er alleen op het bereikte resultaat kan worden teruggekeken. Een leerwinstberekening op basis van een kleiner tijdsinterval heeft de voorkeur, omdat deze betekenisvol is en omdat daarmee ook vooruit gekeken kan worden.

Norm om leerwinst te kunnen beoordelen
Voor de bepaling van de leerwinst van individuele of van een groep leerlingen was tot op heden geen goede norm voorhanden. Om de leerwinst in het LVS van Cito te duiden kan gebruik worden gemaakt van de vaardigheidsniveaus (A tot en met E of I tot en met V), waarbij de vaardigheid van leerlingen vergeleken kan worden met een landelijke steekproef. Het centrale referentiepunt hierbij is het landelijke gemiddelde. Hierdoor wordt zichtbaar hoe de vaardigheid van individuele leerlingen zich verhoudt tot de landelijke verdeling in de verschillende vaardigheidsniveaus.

Ofschoon een landelijke steekproef een betekenisvol referentiepunt is om na te gaan of de leerprestaties van leerlingen in de pas lopen, kan hiermee niet de vraag beantwoord worden of dit voor alle leerlingen geldt, gezien hun uitgangssituatie en het gegeven onderwijs. Niet voor alle leerlingen is de groei hetzelfde. De ene leerling groeit schoksgewijs, terwijl een andere leerling een meer vloeiende vooruitgang boekt. Sommige leerlingen groeien in een bepaalde periode meer dan andere leerlingen uit dezelfde groep. En ondanks dat leerlingen op eenzelfde niveau beginnen wil dat niet zeggen dat ze dezelfde groei doormaken. Daarom is het landelijk gemiddelde niet voor iedere leerling het meest geschikte referentiepunt om de leerwinst te beoordelen. Bovendien werkt de indeling in vaardigheidsniveaus in de hand dat de aandacht van leerkrachten en ib-ers vooral wordt gericht op leerlingen die in de laagste niveaus zitten. Leerlingen uit de hogere niveaus, die zich niet voldoende ontwikkelen of langzaam achteruitgaan, springen niet altijd in het oog.

In de pilot is daarom geprobeerd de scholen een specifieker referentiepunt dan het landelijk gemiddelde te bieden om de leerwinst voor alle leerlingen, ongeacht hun vaardigheidsniveau, te kunnen bepalen: de vergelijking met leerlingen die in aanvang dezelfde vaardigheidsscore hebben. Hierdoor worden leerlingen niet vergeleken met een landelijk gemiddelde in vaardigheidsniveaus, maar met leerlingen die een gelijke startsituatie hebben. De norm voor voldoende of onvoldoende groei wordt dan ontleend aan de gemiddelde groei van de groep leerlingen met dezelfde vaardigheidsscore. Door vervolgens de feitelijke groei van een leerling te vergelijken met het gemiddelde van de referentiegroep, blijkt of een leerling in een bepaalde periode meer, gelijk, of minder gegroeid is dan andere leerlingen die aanvankelijk dezelfde vaardigheidsscore hadden. Er is dus sprake van een relatieve norm, omdat de leerwinst van een individuele leerling wordt afgemeten aan de groei van een qua startniveau vergelijkbare referentiegroep.

Leerwinst in de zomervakantie
Binnen de pilot hebben enkele scholen twee weken na de zomervakantie opnieuw de E-versie van de reguliere eindmeting van het vorige leerjaar afgenomen. Dit werd gedaan voor de vakken rekenen-wiskunde, spelling en technisch lezen. De vaardigheidsscores van voor de vakantie werden afgetrokken van de vaardigheidsscores van na de vakantie. Dit verschil, de groei tijdens zomervakantie, werd vergeleken met de groei die de leerlingen hebben geboekt tijdens het schooljaar. We noemden dit het *seizoensgebonden leerwinstmodel*.

Dit model plaatst informatie over leerwinst en toegevoegde waarde in een nieuw perspectief. Ook buiten schooltijd doen leerlingen kennis en vaardigheden op. Voor sommige leerlingen geldt dat tijdens de zomervakantie hun prestaties achteruit gaan. Om die reden geeft een E-meting afgenomen aan het einde van het schooljaar geen nauwkeurige en actuele informatie over de startsituatie van leerlingen in het nieuwe schooljaar. Bovendien wordt de volgende serie LVS-toetsen pas een half jaar later weer afgenomen. Er zit dus een flinke periode tussen een E- en M-afname. Daarom is voor scholen informatie over de ontwikkeling van leerlingen tijdens de zomervakantie uiterst relevant.

In zekere zin levert het seizoensgebonden leerwinstmodel op een bijzondere manier een indicatie van de toegevoegde waarde van een school. Tijdens de zomervakantie heeft de school geen directe invloed op de vaardigheidsgroei van leerlingen. Mocht er tijdens de zomervakantie toch sprake zijn van groei, dan kan deze dus niet worden toegeschreven aan de school. En in het geval er geen groei of zelfs terugval is, dan is dat een indicatie dat onderwijs effect heeft.

Het seizoensgebonden model kan ook de inspanningen van de thuissituatie op de leerprestaties zichtbaar maken. Een van de deelnemende scholen is in 2012 gestart met een pilot om ouders meer te betrekken in het leesonderwijs. Een groep ouders is gericht uitgelegd waarom thuis lezen zo belangrijk is. Ze kregen ook instructie over de wijze waarop ze dit thuis kunnen aanpakken. Uit de rapportages over het seizoensgebonden leerwinstmodel bleek dat de kinderen van deze ouders tijdens de zomervakantie duidelijk vooruit waren gegaan in het lezen.

Sbo-scholen

Omdat via gegevens over de leerwinst de ontwikkeling van leerlingen nauwkeuriger gevolgd en bijgestuurd kan worden, kunnen deze gegevens van grote betekenis zijn voor de bepaling van het ontwikkelingsperspectief van individuele leerlingen. Via de leerwinstberekening kan immers worden bepaald hoe een individuele leerling zich ontwikkelt ten opzichte van vergelijkbare leerlingen. Dit is niet alleen van belang voor de aanpassing van de instructie en verwerking voor deze leerling, maar ook voor het te verwachten uitstroomniveau. Het ontwikkelingsprofiel kan dus mede op basis van leerwinstgegevens in de tijd worden bijgesteld. Wel dient vermeld te worden dat de toetspraktijk in het sbo een bijzonder karakter heeft, waardoor zich specifieke problemen kunnen voordoen die het in kaart brengen van de leerwinst van sommige leerlingen bemoeilijkt.

Leerwinstberekeningen maken een nauwkeuriger bepaling van de prestatiegroei van individuele en groepen leerlingen mogelijk, dan het geval is met de huidige toets- en schoolinformatiesystemen. Daardoor leveren ze vanuit het schoolverbeteringsperspectief een zinvolle bijdrage aan de gegevens die scholen kunnen gebruiken voor het opbrengstgericht werken en voor de interne kwaliteitszorg.

Aanbevelingen berekenen van leerwinst

De projectgroep beveelt met het oog op implementatie van leerwinst in het primair onderwijs het volgende aan:

Aanbieders van toets- en schoolinformatiesystemen

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

Scholen

Aanbeveling 2
Indien toets- en schoolinformatiesystemen de mogelijkheden bevatten om leerwinst te berekenen, ontwikkel een beleid voor de school om deze leerwinstberekeningen in te zetten voor schoolverbetering en in het bijzonder voor opbrengstgericht werken en voor de interne kwaliteitszorg.

Een belangrijk element van zo'n beleid richt zich op afspraken over een toetskalender, de nauwkeurige registratie van toetsgegevens, de leerlinggroepen waarvoor en de periode waarover de leerwinst wordt berekend. Door de leerwinst bijvoorbeeld halfjaarlijks te berekenen kan optimaal van dit gegeven gebruik worden gemaakt voor de aanpassing van de instructie en de leerstof aan de leerlingen. De berekening van de leerwinst gedurende de zomervakantie voegt informatie toe over de invloed van niet-schoolse factoren op de leerwinst en verschaft de leraren aan het begin van het schooljaar actuele informatie over het beginniveau van hun nieuwe groep leerlingen.

CONCLUSIE 2
Leerwinstberekeningen kunnen door scholen zelf worden uitgevoerd mits aan de onderstaande voorwaarden is voldaan[12]:

1. Er worden toetsen gebruikt die gebaseerd zijn op een vaardigheidsschaal.
2. De verschillende schoolinformatiesystemen die gebruikt worden om toetsscores te registreren en te verwerken, beschikken over de technische mogelijkheid om leerwinst te kunnen berekenen. Er zijn normtabellen beschikbaar om de leerwinst op leerling-, groeps- en op schoolniveau te beoordelen.
3. De toetspraktijk van scholen en de invoer van toetsgegevens door de school voldoen aan een aantal eisen.
4. Toets- en schoolinformatiesystemen bieden mogelijkheden om leerlingcohorten te definiëren waarover de leerwinst kan worden berekend.

[12] Deze zijn in paragraaf 5.6.3 nader uitgewerkt.

De kern van de leerwinstberekening bestaat uit de vergelijking van de groei in vaardigheidsscores van een leerling of een groep leerlingen met die van leerlingen met een gelijk startniveau. Via een module waarmee deze berekening geautomatiseerd kan worden uitgevoerd, kan deze mogelijkheid in bestaande toets- en schoolinformatiesystemen worden ingebouwd. De bepaling van de leerwinst kan echter niet op een betrouwbare manier plaatsvinden als deze uitsluitend wordt berekend op basis van de onderlinge vergelijking van leerlingen van dezelfde school. Daarom dient die vergelijking te geschieden op basis van een landelijk representatieve steekproef van leerlingen. Dat betekent dat de module ook normtabellen met landelijke referentiegegevens dient te bevatten om op leerling-, groeps- en schoolniveau leerwinst te kunnen berekenen. Deze tabellen zouden door toetsaanbieders beschikbaar gesteld moet worden.

Aanbevelingen om scholen zelf leerwinst te laten bepalen

Om scholen in staat te stellen zelf hun eigen leerwinst te berekenen en te beoordelen beveelt de projectgroep het volgende aan:

Toetsaanbieders

Aanbeveling 3
Stel normtabellen samen op basis waarvan enerzijds de leerwinst berekend kan worden voor leerlingen met een gelijk startniveau en anderzijds de schoolgemiddelde leerwinst van een school vergeleken kan worden met die van een representatieve steekproef scholen.

Aanbeveling 4
Maak mogelijk dat scholen rond de zomervakantie toetsen kunnen afnemen om de leerwinst tijdens de zomervakantie te kunnen berekenen, bijvoorbeeld door daarvoor speciale toetsversies te ontwikkelen.

Aanbieders van toets- en schoolinformatiesystemen

Aanbeveling 5
Bouw modules in de systemen in waarmee scholen op leerling-, groeps- en schoolniveau de leerwinst kunnen berekenen op basis van een vergelijking van gelijke startniveaus gebruikmakend van normtabellen van de toetsaanbieders.

Aanbeveling 6
Zorg in de systemen voor mogelijkheden om op flexibele wijze leerlingcohorten te definiëren waarover de leerwinst kan worden bepaald.

Aanbeveling 7
Maak leerwinstrapportages per vaardigheid op leerling-, groeps- en schoolniveau mogelijk.

Scholen

Aanbeveling 8
Voor een betrouwbare bepaling van de leerwinst van een leerstofgebied is het van belang dat:
1. toetsen worden gebruikt waarmee leerwinstberekeningen mogelijk zijn,
2. deze toetsen op vaste momenten worden afgenomen en
3. de uitkomsten op een systematische wijze in een toets- of schoolinformatiesysteem worden geregistreerd.

6.4. Toegevoegde waarde

Hieronder worden drie conclusies getrokken die specifiek betrekking hebben op het in kaart brengen en gebruiken van toegevoegde waarde.

CONCLUSIE 3
Het is mogelijk om per leerstofgebied de toegevoegde waarde van een school te schatten.

De bepaling van toegevoegde waarde van scholen is tot op heden in Nederland beperkt gebleven tot een schatting op basis van eindtoetsresultaten van leerlingen uit groep 8, gecorrigeerd voor relevante achtergrondkenmerken van leerlingen. Door in de schatting van de toegevoegde waarde ook rekening te houden met de beginsituatie van leerlingen en zelfs met de leerwinst op een leerstofgebied over een langere onderwijsperiode, is in de pilot geprobeerd de toegevoegde waarde van scholen beter te benaderen. Nagegaan is in welke mate scholen zelf profiteren van gegevens over hun toegevoegde waarde bij de beoordeling van hun leerprestaties.

Naast informatie over de mate waarin sprake is van leerwinst in een bepaald leerstofgebied - de vergelijking van leerlingen met eenzelfde startniveau - voegt de toegevoegde waarde nog een dimensie aan de duiding van de leerwinst toe, namelijk de vergelijking van de toegevoegde waarde van de school met die van vergelijkbare andere scholen. Of – om het preciezer te zeggen – de vergelijking van dat deel van de leerwinst dat overblijft na correctie voor fairness-kenmerken van een school met dat van andere scholen.

In de pilot zijn twee modellen voor de bepaling van toegevoegde waarde ontwikkeld: het vaardigheidsverschil- en het vaardigheidsgroei-model. De overeenkomst met de leerwinstmodellen is dat ook de groei in vaardigheidsscores op LVS-toetsen de basis vormt voor de bepaling van toegevoegde waarde. Dat betekent dat de toegevoegde waarde uitsluitend per leerstofgebied kan worden bepaald.

Het verschil met de leerwinstmodellen is dat de toegevoegde waarde uitsluitend op schoolniveau wordt vastgesteld, omdat de leerwinst van een leerlingcohort per leerstofgebied wordt gemiddeld. Bij toegevoegde waarde zijn we immers op zoek naar de bijdrage van de school aan prestatiegroei van alle leerlingen binnen een leerstofgebied. Om die bijdrage in beeld te brengen zijn correcties nodig om de invloed van niet-schoolse factoren op de prestatiegroei uit te zuiveren.

Periode waarover de toegevoegde waarde kan worden berekend

Net zoals dat het geval is bij een leerwinstberekening is het technisch gesproken geen probleem om de toegevoegde waarde te berekenen over de hele onderwijsperiode waarvoor een vaardigheidsschaal beschikbaar is, zoals de periode van groep 3 tot en met groep 8. Ook hier geldt de vraag of basisvaardigheden over de jaren heen vergelijkbaar zijn: lezen of rekenen-wiskunde in groep 8 is niet hetzelfde als in groep 3. Naarmate het tijdsinterval tussen twee meetmomenten groter wordt, neemt de kans op een verandering in de inhoud van de gemeten vaardigheid toe. Bovendien is het de vraag welke betekenis gegeven kan worden aan de toegevoegde waarde die over zo'n lange periode is berekend.
In de tweede plaats beperkt de lengte van de periode waarover de toegevoegde waarde wordt berekend, de gebruiksmogelijkheden van deze maten. Voor scholen heeft een berekening van de toegevoegde waarde over bijvoorbeeld de periode vanaf groep 3 tot en met 8 met het oog op schoolverbetering weinig zin. Dit tijdsinterval biedt geen mogelijkheid om tussentijds maatregelen te treffen omdat er alleen op het bereikte resultaat kan worden teruggekeken. Een toegevoegde waardeberekening op basis van een kleiner tijdsinterval heeft de voorkeur, omdat deze betekenisvol is en daarmee ook vooruit gekeken kan worden, bijvoorbeeld apart voor de onder-, midden-, of bouwbouw.

Norm om toegevoegde waarde te kunnen beoordelen

Ook bij een toegevoegde waarde-schatting is een norm nodig om de mate van de toegevoegde waarde te kunnen beoordelen. De kern van de toegevoegde waarde-schatting is dat de gecorrigeerde leerwinst van een school vergeleken kan worden met die van andere scholen. Dat is een waardevolle toevoeging op de leerwinstberekening van de school, omdat door de vergelijking duidelijk wordt hoe de netto leerwinst van een school zich verhoudt tot die van een groep vergelijkbare scholen. Ook bij een toegevoegde waarde-schatting is dus sprake van een relatieve norm waarbij in de pilot het gemiddelde van de deelnemende scholen als referentiepunt dient, maar in feite zou dat een landelijk gemiddelde moeten zijn.

Aanbevelingen voor de schatting van toegevoegde waarde

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

Aanbeveling 10
Neem als periodes voor de schatting van toegevoegde waarde betekenisvolle tijdseenheden, zoals een leerjaar of de onder- of bovenbouw van een basisschool. Hoe langer de periode waarover de toegevoegde waarde wordt geschat, hoe lastiger het is een verband te leggen met de eventuele bijdrage van de school aan de leerwinst en daaraan consequenties te verbinden.

CONCLUSIE 4
Toegevoegde waarde-schattingen op basis van correctie van niet-schoolse factoren kunnen niet door de scholen zelf worden uitgevoerd. Wel kunnen scholen hun schoolgemiddelde leerwinst vergelijken met die van andere scholen.

Voor de schatting van toegevoegde waarde zijn naast leerwinstgegevens ook gegevens nodig die als correctiefactor gebruikt worden. In de pilot betrof dit onder andere niet-schoolse factoren als het opleidingsniveau van de ouders en het type zorgleerlingen. Ofschoon dit soort gegevens in principe voor scholen beschikbaar zijn en soms al in schoolinformatiesystemen zijn opgeslagen, acht de projectgroep het uitgesloten dat de bestaande toets- en schoolinformatiesystemen geschikt gemaakt kunnen worden om de toegevoegde waarde te kunnen corrigeren voor relevante correctiefactoren. In de eerste plaats omdat dergelijke gegevens niet alleen van de betrokken school beschikbaar moeten zijn, maar ook van de landelijke representatieve steekproef waarmee de scholen vergeleken worden. Deze landelijke gegevens dienen dan onderdeel te zijn van de in gebruik zijnde toets- en schoolinformatiesystemen. Het is de vraag of de aanbieders van deze systemen daartoe geëquipeerd zijn. Afgezien van kwesties rond de privacybescherming van leerlinggegevens, spelen ook kosten voor het beheer en het onderhoud van deze landelijke gegevens een rol.

In de tweede plaats zijn ingewikkelde meerniveau statistische analyses nodig om de toegevoegde waarde van een school te bepalen. De toepassing van dergelijke analyses vergt wetenschappelijke expertise die niet standaard in dergelijke systemen kan worden ingebouwd.

En in de derde plaats zijn modellen voor toegevoegde waarde nog onderwerp van wetenschappelijk onderzoek. Daarbij gaat het bijvoorbeeld om vragen als: voor welke en hoeveel factoren dient gecorrigeerd te worden en hoe betrouwbaar en valide zijn de schattingen? Veel van deze modellen zijn thans nog in ontwikkeling en leveren niet steeds dezelfde uitkomsten.

Bovenstaande overwegingen leiden ertoe dat de projectgroep het niet realistisch vindt ervan uit te gaan dat op korte termijn gecorrigeerde schattingen van toegevoegde waarde in bestaande toets- en schoolinformatiesystemen kunnen worden ingebouwd.

Schoolgemiddelde leerwinst vergelijken

Ofschoon de opname van gecorrigeerde toegevoegde waardeschattingen in bestaande toets- en schoolinformatiesystemen momenteel niet realistisch is, betekent dit niet dat scholen hun leerwinst onderling niet met elkaar zouden kunnen vergelijken. De gemiddelde leerwinst van een school op een bepaald vaardigheidsgebied kan immers wel vergeleken worden met die van andere scholen.

Bij schoolgemiddelde leerwinst vindt geen correctie plaats voor niet-schoolse factoren. Daardoor kunnen geen uitspraken gedaan kunnen worden over de bijdrage van de school aan

de leerwinst. Omdat de vergelijking van schoolgemiddelde leerwinst is gebaseerd op de prestatiegroei van schoolpopulaties, wordt hiermee een belangrijk bezwaar ondervangen dat kleeft aan de vergelijking op basis van statusmetingen. Zoals in paragraaf 5.4.2 is toegelicht is de vergelijking van ongecorrigeerde schoolgemiddelden een eerste stap in de toegevoegde waardebepaling. Om de schoolgemiddelde leerwinst van scholen onderling te kunnen vergelijken, dienen de toets- en schoolinformatiesystemen representatieve normgegevens te bevatten om het schoolgemiddelde met die van een landelijke steekproef te kunnen vergelijken. Ook dergelijke referentiegegevens zouden verstrekt moeten worden door de toetsaanbieders (zie daarvoor conclusie 2).

CONCLUSIE 5
Om toegevoegde waarde goed te kunnen schatten is verder onderzoek naar en nadere ontwikkeling van de modellen noodzakelijk.

De pilot heeft laten zien dat het mogelijk is om de toegevoegde waarde van scholen te schatten op basis van leerwinstgegevens die gecorrigeerd worden voor relevante niet-schoolse factoren. De bepaling van toegevoegde waarde is niet alleen nog te complex om deze in bestaande toets- en schoolinformatiesystemen toe te passen, er zijn ook nog de nodige problemen met de nauwkeurigheid van toegevoegde waarde schattingen. Ook kennen toegevoegde waarde modellen beperkingen.

Net zoals dat het geval is bij de huidige opbrengstbeoordeling zijn er vraagtekens te plaatsen bij de toepassing van correcties voor niet-schoolse factoren. Door die correcties wordt de indruk bevestigd dat voor scholen met veel zorgleerlingen en veel leerlingen met taal- en cognitieve achterstanden, lagere verwachtingen gelden over hun leerresultaten dan voor andere scholen. Een ander punt is dat zowel de schoolgemiddelde leerwinst als de toegevoegde waarde gebaseerd is op een relatieve norm: de leerwinst van leerlingen wordt vergeleken met die van vergelijkbare leerlingen. Ook de schoolgemiddelde leerwinst en de toegevoegde waarde van scholen wordt afgemeten aan de prestaties van andere scholen. Er is geen sprake van een absolute maat zoals het geval is bij de referentieniveaus, waarbij de prestaties worden afgemeten aan een duidelijk criterium. Bij een relatieve norm blijven er dus altijd scholen die onder het gemiddelde van de vergelijkingsgroep presteren en in dit geval - relatief gezien - niet voldoende toegevoegde waarde leveren.

Tot slot, de kwaliteit van de schattingen van de toegevoegde waarde van scholen wordt in hoge mate bepaald door factoren die van invloed zijn op de nauwkeurigheid ervan, zoals de invloed van meetfouten, het aantal en de soort correctiefactoren, de stabiliteit van toegevoegde waarde-schattingen en strategisch gedrag van scholen (zie ook par. 3.5). Naar de mogelijkheden om de kwaliteit van de ontwikkelde toegevoegde waarde-schattingen te verbeteren, is in de pilot geen verder onderzoek gedaan. De kwaliteit van toegevoegde waarde-schattingen is, zo bleek uit de pilot, wel van groot belang voor het vertrouwen dat scholen zullen hebben in het op de juiste wijze functioneren van toegevoegde waarde in een *accountabilityperspectief*.

De projectgroep en de deelnemende scholen willen niet ontkennen dat er allerlei haken en ogen zitten aan de schatting van de toegevoegde waarde, maar vinden dit geen reden om af te zien van de verdere ontwikkeling van deze vorm van opbrengstbeoordeling. Niet alleen de projectgroep maar ook de deelnemende scholen vinden de ontwikkelde maten bruikbaar op schoolniveau. Ze leveren ook een bijdrage aan een eerlijker verantwoording over de leerresultaten aan de buitenwereld. De projectgroep pleit er daarom voor de verdere ontwikkeling van deze toegevoegde waarde modellen ter hand te nemen in nauw overleg met de scholen.

Aanvullend onderzoek naar de toepassing van toegevoegde waarde in een Nederlandse context is nodig op twee terreinen:

1. Methodisch-technische kwesties rond de verdere ontwikkeling van de modellen.
2. Openbaarmaking van de toegevoegde waarde van scholen.

Methodisch-technische kwesties

Nader onderzoek naar toegevoegde waarde zou zich kunnen richten op de vraag voor hoeveel en welke niet-schoolse factoren het best gecorrigeerd kan worden om tot een nauwkeurige schatting van de toegevoegde waarde te komen. In de pilot zijn deze correctiefactoren beperkt gebleven tot het opleidingsniveau van ouders en de zorgproblematiek van leerlingen. Overwogen kan worden de correctiefactoren te beperken tot kenmerken die ontleend kunnen worden aan externe databestanden, zoals de gemeentelijke registratie, het CBS of het onderwijsnummer. In de pilot is wel nagegaan of de zogenoemde standaardmeetfout van de toetsen als correctiefactor is te gebruiken. Dit is namelijk een gegeven dat niet door de scholen maar door de toetsleverancier kan worden geleverd. Een eerste verkenning leverde op dat deze correctiefactor van enige betekenis voor de betrouwbaarheid van de schattingen is en in ieder geval geen belasting voor de scholen meebrengt.

Aanbevelingen voor nader onderzoek naar toegevoegde waarde modellen

Aanbeveling 11
Doe nader onderzoek naar de - voor de Nederlandse context - relevante niet-schoolse factoren die in de modellering betrokken kunnen worden om de betrouwbaarheid en stabiliteit van de schattingen van toegevoegde waarde te bevorderen.

Aanbeveling 12
Niet-schoolse factoren dienen bij voorkeur gebaseerd te zijn op gegevens die relatief eenvoudig beschikbaar zijn, bijvoorbeeld doordat ze onderdeel zijn van schoolinformatiesystemen of omdat ze opgenomen zijn in toegankelijke externe databestanden, zoals de gemeentelijke registratie, het CBS of via het onderwijsnummer.

Openbaarmaking

Het gaat bij het verdere onderzoek naar de bruikbaarheid van toegevoegde waarde in het primair onderwijs niet alleen om verbeteringen van louter methodisch-technische aard. Er zijn ook kwesties die in de wetenschappelijke literatuur wel een rol spelen, maar niet in de pilot aan de orde zijn gesteld. Zo is in de pilot van strategisch gedrag van scholen niets gebleken, omdat het vooral ging om de vraag wat de scholen zelf aan rapportages over de leerwinst en toegevoegde waarde hadden. Er is ook geen onderzoek gedaan naar de effecten en neveneffecten van het gebruik dat de ontwikkelde maten in het perspectief van accountability. Dit zijn echter wel onderwerpen die binnen de Nederlandse context nader onderzoek vergen.

Uit de pilot is duidelijk naar voren gekomen dat binnen het *accountabilityperspectief* ook ethische en communicatieve kwesties een rol spelen. Die hebben vooral te maken met de openbaarmaking van gegevens over leerwinst en toegevoegde waarde. Bij openbaarmaking spelen eerlijkheid, rechtsgelijkheid en zorgvuldigheid van de beoordeling een belangrijke rol, maar ook de bestrijding van ongewenste effecten, zoals strategisch gedrag. Ook de communicatieve kant is, met het oog op openbaarmaking, een punt van nader onderzoek. Leerwinst en vooral toegevoegde waarde zijn geen makkelijk toegankelijke begrippen en roepen al snel misverstanden op zodra deze hun intrede doen in het publieke domein.

Of het nu gaat om een verantwoording door de scholen over leerwinst en toegevoegde waarde aan de buitenwereld of om de wijze waarop de inspectie ervan gebruik maakt, in alle gevallen is het van belang naar manieren te zoeken waarop deze zo eerlijk en zorgvuldig mogelijk openbaar gemaakt kunnen worden. Omdat in de pilot het accent lag op het intern gebruik door de scholen, zijn kwesties over de wijze van openbaarmaking, de wijze waarop de inspectie de leerwinst en toegevoegde waarde kan beoordelen en de voor- en nadelen daarvan, maar zijdelings aan de orde geweest. Daarom stelt de projectgroep voor hier nader onderzoek naar te doen en daarbij ook de scholen te betrekken. Naar de mening van de projectgroep is bijvoorbeeld de PO-Raad een instantie die de verdere ontwikkeling van toegevoegde waarde onder haar hoede kan nemen en zich ook kan buigen over het gebruik leerwinst en toegevoegde waarde voor accountability-doeleinden.

Aanbeveling voor de verdere ontwikkeling van leerwinst en toegevoegde waarde in accountabilityperspectief

Aanbeveling 13
De projectgroep adviseert de staatssecretaris van OCW om op basis van de uitkomsten van de pilot met de PO-Raad, de Inspectie van het Onderwijs en met wetenschappers op korte termijn een plan te ontwikkelen:

1. voor de wijze waarop de toegevoegde waarde van scholen zo nauwkeurig mogelijk voor niet-schoolse factoren kan worden gecorrigeerd;
2. welke instantie de noodzakelijke gegevens over toegevoegde waarde van scholen het beste zou kunnen verzamelen en in beeld kan brengen;
3. hoe deze gegevens teruggekoppeld kunnen worden aan de scholen;
4. op welke wijze leerwinst en toegevoegde waarde schattingen gebruikt kunnen worden door scholen om extern verantwoording af te leggen over hun opbrengsten;
5. op welke wijze de Inspectie van het Onderwijs leerwinst en toegevoegde waarde kan betrekken in het toezicht.

6.5. Toepasbaarheid van leerwinst en toegevoegde waarde in accountability-perspectief

Gegevens over leerwinst en toegevoegde waarde kunnen ook een rol spelen in de horizontale en verticale verantwoording door scholen over de leerprestaties en in de beoordeling van de opbrengsten door de inspectie. Naar de rol, de effecten en de eventuele neveneffecten van leerwinst en toegevoegde waarde in een *accountabilityperspectief* is in de pilot alleen literatuuronderzoek gedaan. Wel is de deelnemende scholen om hun mening hieromtrent gevraagd.

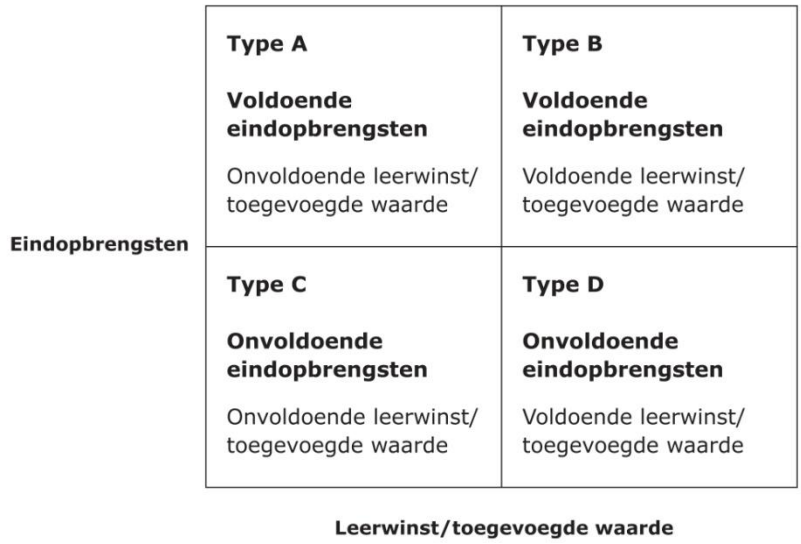
Uit de pilot is gebleken dat de meerderheid van de deelnemende scholen geen zwaarwegende bezwaren heeft tegen het gebruik van leerwinst of toegevoegde waarde door de onderwijsinspectie om de leerprestaties van scholen te beoordelen. Het belangrijkste is, zo bleek uit de pilot, dat deze beoordeling fair en transparant is. Het gebruik van leerwinst of toegevoegde waarde, zoals ontwikkeld in de pilot, maakt de beoordeling van de leerprestaties eerlijker, zo luidt de mening van de meerderheid van de deelnemende scholen. Leerwinst en toegevoegde waarde maken immers de bijdrage van de scholen aan de leerprestaties transparanter dan thans het geval is, waarbij de beoordeling van de leerprestaties overwegend op eindtoetsgegevens is gebaseerd.

In ons land geschiedt de beoordeling van de leerprestaties van basisscholen door de inspectie op basis van de gemiddelde scores op een eindtoets[13]. Het is echter algemeen bekend dat, ondanks correcties, deze eindtoetsscores in hoge mate correleren met de sociaal-economische status van de leerlingpopulatie. De huidige manier van opbrengstbeoordeling wordt door

[13] De beoordeling van de tussenopbrengsten in groep 3, 4 en 6 (bij kleine scholen ook groep 5) vindt plaats op basis van een waardering van gemiddelde vaardigheidsscores.

scholen vaak als oneerlijk beschouwd, omdat de prestaties worden beïnvloed door factoren waarop de school geen invloed heeft en omdat deze factoren niet gelijk verdeeld zijn over scholen en tussen groepen leerlingen binnen een school.

Als alleen naar de eindopbrengsten van een school wordt gekeken, is het mogelijk dat scholen die op de eindopbrengsten gemiddeld of daarboven presteren als effectief worden gezien, terwijl er relatief weinig leerwinst is geboekt. Over de hoogte van de eindprestaties van een school kan men dan tevreden zijn, terwijl de school dit vooral heeft te danken aan het hoge instroomniveau van hun leerlingen. Ook het omgekeerde komt voor, een school heeft een gemiddelde lage score op een eindtoets, terwijl - als naar de leerwinst gekeken zou worden - ook duidelijk is of er voldoende inspanning is gepleegd om bij alle leerlingen voldoende vorderingen te boeken. Het betrekken van leerwinst of toegevoegde waarde bij de beoordeling van de leerprestaties van scholen maakt deze beoordeling eerlijker omdat er rekening wordt gehouden met zowel de beginsituatie van de leerlingen als met de bijdrage die de school levert aan de prestatiegroei van de leerlingen. Dit is te illustreren aan de hand van figuur 11.



Figuur 11 Typen leeropbrengsten naar de combinatie van eindopbrengsten met leerwinst/toegevoegde waarde

In figuur 11 worden de eindopbrengsten van scholen gerelateerd aan de leerwinst of toegevoegde waarde berekend over tussenresultaten. Daarbij ontstaat voor scholen met lage eindopbrengsten de mogelijkheid om - bijvoorbeeld in een dialoog met de inspectie - zichtbaar te maken welke inspanningen zijn gepleegd om de leerlingen tot zo hoog mogelijke prestaties te brengen. Het verandert de kijk op welke scholen hoge en welke scholen lage opbrengsten hebben. Het zijn de scholen met de opbrengsten van het type A die ondanks hun hoge eindopbrengsten weinig groei of toegevoegde waarde laten zien. Type B-scholen vormen het ideaal: je leert er veel en de school biedt veel kansen om zo hoog mogelijk uit te stromen. Type C-scholen hebben lage eindopbrengsten, maar blijken ook niet in staat te zijn veel bij te dragen aan de prestaties van hun leerlingen. Scholen met opbrengsten van het type D hebben, mogelijk

vanwege de beperkingen van de schoolbevolking, relatief lage eindopbrengsten maar boeken wel leerwinst of laten toegevoegde waarde zien.

Inspectie van het Onderwijs

De Inspectie van het Onderwijs is van mening dat zij, door het betrekken van leerwinst en toegevoegde waarde in de beoordeling van de opbrengsten, een beter beeld krijgt van de bijdrage van de school aan de leerprestaties in het algemeen en aan de leerprestaties van specifieke groepen leerlingen binnen de school in het bijzonder zorgleerlingen of hoogbegaafden. Bij dit voornemen zijn wel enkele kanttekeningen te plaatsen.

1. Gezien het feit dat toegevoegde waarde-schattingen voor het primair onderwijs nog nader onderzoek vergen, is het niet realistisch ervan uit te gaan dat de inspectie deze op korte termijn kan toepassen. Daar komt nog bij dat de inspectie op basis van voor haar beschikbare data via BRON thans niet kan beschikken over informatie over het beginniveau van leerlingen uit het basisonderwijs.
2. Ook leerwinstmaten kan de inspectie niet berekenen omdat ze niet beschikt over datasets met gegevens uit de LVS-systemen van scholen. Het ideaal is dat scholen zelf over hun leerwinst aan de inspectie verantwoording afleggen. Echter, scholen kunnen op dit moment nog niet zelf hun leerwinst berekenen. Het is dus ook niet realistisch ervan uit te gaan dat de inspectie op korte termijn gebruik kan maken van leerwinstberekeningen die door de scholen zelf zijn uitgevoerd. Ze is daarbij afhankelijk van enerzijds de mogelijkheden die worden gecreëerd voor scholen om leerwinst te berekenen en anderzijds van de bereidwilligheid van scholen om deze informatie beschikbaar te stellen.
3. Daar komt nog bij dat we in Nederland nog geen ervaring hebben met het gebruik van leerwinst of toegevoegde waarde vanuit het accountabilityperspectief. Daarom heeft de projectgroep in de vorige paragraaf de staatssecretaris voorgesteld voor de verdere ontwikkeling en toepasbaarheid van leerwinst en toegevoegde waarde binnen een accountabilityperspectief een plan te ontwikkelen en daarbij de scholen, de inspectie en de wetenschap te betrekken.

Literatuur

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers*. (Briefing Paper #278). http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

Batenburg, T.A. , & Werf, M.P.C. van der (2004). *NSCCT niet schoolse cognitieve capaciteiten test voor groep 4,6 en 8 van het basisonderwijs, verantwoording, normering en handleiding*. Groningen: GION.

Batenburg, T. (2012). *Het signaleren en aanpakken van onderpresteren op de basisschool*. Groningen: GION.

Betebenner, D.W. (2007). *Estimation of student growth percentiles for the Colorado Student Assessment Program.* Dover, New Hampshire: National Centre for the Improvement of Educational Assessment (NCIEA).

Betebenner, D. W. (2009). *Growth, standards and accountability.* Denver: Colorado Department of Education: The centre for assessment.

Betebenner, D. W., & Linn, R. L. (2009). *Growth in student achievement: Issues of measurement, longitudinal data analysis, and accountability.* Paper presented at the Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda, Princeton, NJ. http://www.k12center.org/rsc/pdf/BetebennerandLinnPresenterSession1.pdf

Bosker, R.J. (2012). De toegevoegde waarde van een school: Begripsbepaling, meting en causale attributie. In A.B. Dijkstra & F.J.G. Janssens (Red.). *Om de kwaliteit van het onderwijs: Kwaliteitsbepaling en kwaliteitsbevordering (pp. 93-104)*. Den Haag: Boom/Lemma.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability. Washington, D.C.: The National Academies Press.

Briggs, D. (2008). *The goals and uses of value-added models.* Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. http://www.nap.edu/openbook.php?record_id=12820

Campbell, D. T. (1976). *Assessing the Impact of Planned Social Change*. Hanover New Hampshire: The Public Affairs Center, Dartmouth College.

Castelllano, K.E. & Ho, A.D. (2013). *A Practitioner's Guide to Growth Models*. The Council of Chief State School Officers. http://scholar.harvard.edu/files/andrewho/files/a_pracitioners_guide_to_growth_models.pdf

Center for Public Education. (2007). *Measuring student growth: A guide to informed decision making*. http://www.centerforpubliceducation.org/site/apps/nlnet/content3.aspx?c=lvIXIiN0JwE&b=5114813&ct=6857853.

De Loos Monitoring (z.j.). *De LeerWinst-methode: Een leerlingcapaciteitonafhankelijke berekening van de toegevoegde waarde van basisscholen*. z.u. http://deloos.net/inleiding-leerwinst-methode/

DePascale, C. A. (2006). *Measuring growth with the MCAS tests: A consideration of vertical scales and standards.* http://www.nciea.org/publications/MeasuringGrowthMCASTests_CD06.pdf

Deunk, M.I., & Doolaard S. (2013). *Attitude en handelen van basisschoolleerkrachten met betrekking tot het verbeteren en borgen van leerlingresultaten. Resultaten van vragenlijst Streef 2010 en 2012*. Groningen: Gion.

Dronkers, J. (2013). *Toelichting op de berekening van toegevoegde waarde van reguliere basisscholen op grond van hun gemiddelde scores op hun eindtoetsen 2010, 2011 en 2012*. Universiteit Maastricht. http://www.schoolcijferlijst.nl/basis/Toelichting.pdf

Fisher, T. H., & Twing, J. S. (2006). *Toward a growth-centric assessment model: A white paper from Pearson Educational Measurement*. http://www.pearsoned.com/RESRPTS_FOR_POSTING/ASSESSMENT_RESEARCH/AR1.%20PEMwp_GCA_y06n02.pdf

Goldschmidt, P. Roschewski, P., Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005). *Policymakers' guide to growth models for school accountability: How do accountability models differ?* http://beta.ccsso.org/Documents/2005/Policymakers_Guide_To_Growth_2005.pd f

Gong, B., Perie, M., & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under No Child Left Behind: An update to inform design decisions*. http://www.nciea.org/publications/GrowthModelUpdate_BGMAPJD07.pdf

Hamilton, L. S., McCaffrey, D. F., Koretz, D. M. (2006). Validating achievement gains in cohort-to-cohort and individual growth-based modeling contexts. In R. W. Lissitz (Ed.) *Longitudinal and value added models of student performance (pp. 407-435)*. Maple Grove, MN: JAM Press.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, 4, 319-350. doi: 10.1162/edfp.2009.4.4.319

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Hattie, J. & Anderman, E.M. (2013). *International guide to student achievement*. New York: Routledge.

Dijkstra, A.B. & Janssens, F.J.G. (Red.). *Om de kwaliteit van het onderwijs: kwaliteitsbepaling en kwaliteitsbevordering.* Den Haag: Boom/Lemma.

Inspectie van het Onderwijs. (2010). *Opbrengstgericht werken in het basisonderwijs*. Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2013). *De staat van het onderwijs. Onderwijsverslag 2011/2012.* Utrecht: Inspectie van het Onderwijs.

Kleij, F. M. van der, (2013). *Computer-based feedback in formative assessment*. Enschede: Universiteit Twente (diss.).

Koretz, D. (2008). A measured approach. Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*, fall 2008, p. 18-39.

Koedel, C., & Betts, J. (2009). *Value-added to what? How a ceiling in the testing instrument influences value-added estimation.* NBER Working Paper 14778. Cambridge, MA: National Bureau of Economic Research.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer Science+Business Media, Inc.

Kwaliteitswijzer basisonderwijs Amsterdam 2010-2011. http://www.obslaternamagica.nl/corp/downloads/836.pdf

Ledoux, G., Blok, H., & Boogaard, M. (2009). *Opbrengstgericht werken. Over de waarde van meetgestuurd onderwijs*. Amsterdam: SCO Kohnstamm Instituut.

Ledoux, G., Roeleveld, J., Langen, A. en Van, Paas, T. (2012). *COOLspeciaal Technisch rapport meting schooljaar 2010/2011*. Amsterdam: Kohnstamm Instituut.

Ledoux, G., Roeleveld, J., Langen, A. van, Smeets, E. (2012) *Cool Speciaal. Inhoudelijk rapport.* Amsterdam: Kohnstamm Instituut.

Ligon, G. D. (2008). *The optimal reference guide: Comparison of growth and value-add models, Growth models series—part II*. http://www.espsolutionsgroup.com/espweb/assets/files/ESP_Comparison_of_Growth_and_Value_Add_Models_ORG.pdf

Luyten, H. & Ten Bruggencate, G. (2011). The presence of Matthew effects in Dutch primary education, development of language skills over a six year period. *Journal of Learning Disabilities*, 44(5), 444-458.

Martin, D. J. (1985). *The measurement of growth in educational achievemen*t. Iowa City, IA.: University of Iowa. (Diss.)

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. & Lockwood, J.R. (2008). *Value-added models: Analytic issues*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. http://www.nationalacademies.org/bota/VAM_Workshop_Agenda.html.

McCaffrey, D., Sass, T.R., & Lockwood, J.R. (2008). *The inter temporal effect estimates.* Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin-Madison, April 22-24.

National Research Council & National Academy of Education (2010). *Getting Value Out of Value-Added: Report of a Workshop*. Washington, DC: The National Academies Press.

OECD (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools*. Paris: OECD.

Ofsted (2010). *The evaluation schedule for schools. Manchester*: The Office of Standards in Education.

Onderwijsraad (2011). *Een stevige basis voor iedere leerling*. Den Haag: Onderwijsraad.

Peschar, J.L. (2007). *Over leerwinst als stelselindicator.* Haren: z.u.

Popham, W.J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 89(2), 146-150.

Raudenbush, S.W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement?* The ninth annual William H. Angoff Memorial Lecture. Princeton, New Jersey. https://www.ets.org/Media/Education_Topics/pdf/angoff9.pdf

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 Department of Education Skills.

Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value added models for estimating school effects. *Educational Finance and Policy*, 4(4): 492-519.

Rekers-Mombarg, L. & Janssens, F. (2013). *Effect van correctie voor de meetfout van een LOVS-toets op de toegevoegde waarde van een basisschool*. Presentatie ten behoeve van de Onderwijsresearch Dagen te Brussel.

Roeleveld, J., Veen I. van der, & Ledoux, G. (2008). *Verkenning leerwinst als indicator voor onderwijskwaliteit*. Amsterdam: SCO-Kohnstamm Instituut van de Faculteit der Maatschappij- en Gedragswetenschappen, Universiteit van Amsterdam (SCO-rapport nr. 815, projectnummer 40330).

Rothstein, R. (2009). What's wrong with accountability by the number. *American Educator*, spring 2009, p. 20-33.

Rubin, D. B., Stuart, E. A. & Zanutto, E. L. (2004). A potential outcomes view of value added assessment in education. *Journal of Educational and Behavioral Statistics*, 29, 103-116.

Sanders, W. L. (2003). *Beyond No Child Left Behind*. In 2003 Annual Meeting American Educational Research Association.

Sanders, W. L. & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8, 299-311

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.

Schochet, P. Z., & Hanley S. C. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Smith, R. L., & Yen, W. M. (2006). Models for evaluating grade-to-grade growth. In R. W. Lissitz (Ed.) *Longitudinal and value added models of student performance* (pp. 82-99). Maple Grove, MN: JAM Press.

Stevens, J., & Zvoch, K. (2006). Issues in the implementation of longitudinal growth models for student achievement. In R. W. Lissitz (Ed.) *Longitudinal and value added models of student performance* (pp. 170-209). Maple Grove, MN: JAM Press.

Timmermans, A.C. (2012). *Value added in Educational Accountability: Possible Fair and Useful?* Groningen: University of Groningen. (Diss.)

Thomas, S., Sammons, P., Mortimore, P. & Smees, R. (1997). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School Effectiveness and School Improvement,* 8, 169-197.

Tong, Y., & Kolen, M.J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227-253.

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, 14(3), 321-349.

Visscher, A & Ehren, M. (2011). *De eenvoud en complexiteit van Opbrengstgericht Werken: Analyse in opdracht van de Kenniskamer van het Ministerie van Onderwijs, Cultuur en Wetenschap*. Enschede: Vakgroep Onderwijsorganisatie en -management, Universiteit Twente. http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/07/13/de-eenvoud-en-complexiteit-van-opbrengstgericht-werken.html

Willms, J.D. (2008). *Seven key issues for assessing "value-added" in education*. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. http://www.nap.edu/openbook.php?record_id=12820

Wolf, I.F. de & Janssens, F.J.G. (2007). Effects and side effects of school inspections and accountability in education: a review of empirical studies. *Oxford Review of Education*, 33, p. 379-396.

De Wolf, I. (2012). Opvattingen van scholen over het onderwijstoezicht. In. Dijkstra, A.B. & Janssens, F.J.G. (red.). *Om de kwaliteit van het onderwijs: kwaliteitsbepaling en kwaliteitsbevordering*. Den Haag: Boom | Lemma uitgevers.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.). *Linking and aligning scores and scales*. New York, NY: Springer.

Bijlage 1 Selectiecriteria scholen

naam school
brinnr
bestuur
e-mailadres school
postadres
bezoekadres
aantal leerlingen
percentage leerlingen dat langer dan 8 jaar over de basisschool heeft gedaan
percentage leerlingen dat later dan in groep 1 is ingestroomd
percentages leerlingmutaties 2010-2011
aantal leerkrachten
aantal OOP
aantal leden directie
aantal ib
denominatie
schoolconcept (Montessori, Dalton, etc.)
LVS systeem (ParnasSys, Esis, etc.)
leerlinggewicht 0 / 0,3 / 1,2
provincie
mate waarin wordt gewerkt met data* (cijfer en letter aangeven)
gebruikte reeksen toetsen (Cito, Kapinga, etc.)
inspectiebezoek ontvangen in de laatste twee jaar ja / nee (graag bij de mail voegen)
methode overstijgende toetsen in 1/2 (ja / nee en indien ja, welke)
deelname VVE-programma (ja / nee en voor alle leerlingen of een aantal)
type onderwijs (regulier, SBO of SO)

Scholen verschillen in de mate waarin ze opbrengstgericht werken:

Data worden verplicht ingevuld, maar niet tot nauwelijks bekeken door de leerkrachten

Data worden ingevuld en bekeken, maar er is weinig tot geen verband met plannen en acties(het zogenaamde weerbericht)

Data worden ingevuld en bekeken en vormen een vertrekpunt voor handelen.

Scholen verschillen in de niveaus waarop ze het opbrengstgericht werken toepassen:

Leerlingniveau, met onder andere het ontwikkelingsperspectief

Groepsniveau, met onder andere het leerrendement

Schoolniveau

Bestuursniveau

Bijlage 2 Deelnemende scholen

Michaëlschool, Amersfoort
Vrije basisschool de Regenboog, Eindhoven
Rehobothschool, Kootwijkerbroek
KBS De Zandberg, Breda
RKBS De Hoge Waai, Raamsdonk
SBO De Brug, Vianen
OBS De Carrousel, Dalfsen
De Zuidwester, Den Haag
CBS De Lindenborgh, Musselkanaal
Nieuwe Park Rozenburg School, Rotterdam
De Vrije School Almelo, Almelo
Prot. Chr. Jenaplanbasisschool De Peppels, Boxmeer
BS Kameleon, Weert
Holtkampschool, Goes
CBS De Evenaar, Krommenie
Basisschool Toermalijn, Den Haag
SBO De Rotonde, Gorinchem
De Klaver Carnisse, Rotterdam
Da Costa School, Rotterdam
SBO De Watergeus, Lelystad
SBO De Brug, Hulst
Emile Weslyschool – Suringarschool, Maastricht
Ariensschool, Utrecht
G. Th. Rietveldschool, Badhoevedorp
Koningin Wilhelminaschool, Leeuwarden
De Parkschool, Zwolle
OBS Overvecht, Utrecht
Pniëlschool, Rotterdam
ODS De Starter, Groningen
De Wegwijzer, Teteringen
Helder Camaraschool, Teteringen
Da Costa Kanaleneiland, Utrecht
OBS Sandeschool, Kloosterzande
Freinetschool OBS de Piramide, Heerlen
OBS Pantarijn (locatie Groeneweg en Oversteek), Rotterdam
Bos en Lommerschool, Amsterdam
De Schalm, Rotterdam
OBS ’t Prisma (locatie Tussenwater), Hoogvliet
St. Gerardusschool, Enschede
SBO De Dijk, Wageningen
Ericaschool SO/VSO, Vlaardingen
6e Montessorischool Anne Frank, Amsterdam
OBS De Duizendpoot, Almere
Shri Saraswatie School, Rotterdam
OBS De Albatros, Almere

Bijlage 3 Samenstelling projectgroep

Projectgroep:

Frans Janssens (projectleider), Universiteit Twente/Inspectie van het onderwijs
Boudewijn Spoorenberg (programmamanager), Ministerie van OCW
Job Cornelissen, Ministerie van OCW
Lyset Rekers-Mombarg, Rijksuniversiteit Groningen (GION)
Hans Luyten, Universiteit Twente
Renske de Leeuw, Universiteit Twente
Ilse Papenburg, Cito
Bernadette van Geest-Oosterman, CED-Groep
Carla Versteeg, CED-Groep
Ellen Lacor, CED-Groep

Namens de CED-Groep zijn de volgende adviseurs bij de pilot betrokken:

Afke Donker
Carla van Doornen
Hanke Geurts
Anne-Marie Gielen
Madeleine Hulsen
Margot Oomens
Christa Oudejans
Marijke Roetering
Elly van Seventer
Anna Veenstra
Vera Vergunst

Bijlage 4 Brochure deelnemende scholen

Modellen Leerwinst en Toegevoegde Waarde

Informatiebrochure voor scholen die deelnemen aan LTWPO
Januari 2012

Inleiding
Voor u ligt de folder met informatie met voorstellen van Cito, de Universiteit Twente en het GION voor modellen om Leerwinst en Toegevoegde Waarde te bepalen. Voordat we deze modellen bespreken, wordt eerst het project en de definities van leerwinst en van toegevoegde waarde nog eens beschreven, zoals deze ook binnen het project LTWPO gehanteerd worden.

Een van de doelen van het actieplan 'Basis voor Presteren' is 'leren van resultaten'. Om samen met scholen en hun besturen te beproeven langs welke weg leerwinst en toegevoegde van scholen beter in beeld zijn te brengen, is het project Leerwinst en Toegevoegde Waarde PO opgezet (zie http://www.ltwpo.nl).

In het project wordt gezocht naar een werkwijze waardoor leerwinst en toegevoegde waarde voor de school als hulpmiddel kan dienen om opbrengstgericht te werken. Verder wordt onderzocht of het instrument ook door de inspectie is te gebruiken bij het beoordelen van de leeropbrengsten en de kwaliteit van de school. Belangrijk is dat recht kan worden gedaan aan de inspanningen van de school en dat rekening kan worden gehouden met de samenstelling van de leerlingenpopulatie.

In het project wordt ook aandacht geschonken aan het moment en de manier waarop een begintoets op valide en betrouwbare wijze bij leerlingen kan worden afgenomen. Het gaat hier niet om de ontwikkeling van een toets voor jonge leerlingen, maar om de vraag hoe en wanneer de startsituatie van de leerlingen het best kan worden vastgesteld om de leerwinst en de toegevoegde waarde van de school zo zorgvuldig mogelijk te kunnen bepalen. Tijdens het project worden hiervoor de toetsen gebruikt die de school al afneemt in de onderbouw.

Vertrekpunt van het project is dat gezamenlijk wordt bekeken wat wel en wat niet werkt. Voor scholen is het bijvoorbeeld belangrijk, hoe gemakkelijk de verzamelde gegevens zijn te gebruiken, en voor onderzoekers of deze gegevens betrouwbaar zijn. Ook wordt goed bekeken of zich geen ongewenste neveneffecten voordoen, zoals 'teaching to the test' of een te eenzijdige aandacht voor taal en rekenen.

Leerwinst
Leerwinst is een maat voor de groei van leerprestaties van individuele leerlingen, van een groep leerlingen, of van een school tussen twee of meer toetsmomenten, gemeten met toetsen die met elkaar vergeleken kunnen worden. Omdat de prestaties van dezelfde leerlingen gevolgd worden, zegt leerwinst veel meer over de leervorderingen dan de afname van een enkele toets. Een leerwinstmaat geeft echter niet precies aan wat de school heeft bijgedragen aan de groei. Misschien spelen ook de ouders, een educatief buurthuis om de hoek of spelletjes op de computer een belangrijke rol. Om daarop meer zicht te krijgen kijken we ook naar de toegevoegde waarde.

Toegevoegde waarde
De toegevoegde waarde is een maat voor de bijdrage van de school aan de leerwinst. De toegevoegde waarde kan worden bepaald door de leerwinst in meer of mindere mate voor kenmerken van de leerling en/of omgevingsfactoren te corrigeren. Daarbij wordt de feitelijke leerwinst vergeleken met de leerwinst die, gezien de leerlingen en de schoolcontext, verwacht had mogen worden. Het verschil daartussen geeft aan of er sprake is van toegevoegde waarde.

Als school kunt u zelf een keuze maken uit de verschillende modellen. Tijdens het intakegesprek zal de adviseur van de CED-Groep u informeren over welke modellen het best bij u school passen. Daarbij wordt rekening gehouden met de gegevens die uw school beschikbaar heeft in uw administratie - en leerlingvolgsysteem (LVS).

Doel van het project
Het project LTWPO heeft tot doel het Opbrengstgericht Werken te bevorderen en de beoordeling van de kwaliteit van leerprestaties te verfijnen.

Opbrengstgericht Werken
In het Actieplan 'Beter presteren' wordt 'Opbrengstgericht Werken' (OGW) gedefinieerd als het gezamenlijk systematisch en doelgericht werken aan het maximaliseren van leerlingprestaties. Om maximale leerlingprestaties mogelijk te maken is het nodig om de onderwijsresultaten in beeld te brengen, deze te volgen en een plan op leerlingniveau te kunnen maken. Binnen het project LTWPO wordt dit gedaan door middel van de ontwikkeling van leerwinstmaten waarmee de volgende stap in de leerprocessen van leerlingen beter bepaald kan worden.

Eerlijkere beoordeling van de school

De meest gebruikelijke manier om naar de kwaliteit van de leerprestaties van een school te kijken, is een beoordeling van de prestaties op een eindtoets. De inspectie bijvoorbeeld, beoordeelt de eindtoetsprestaties van een school over drie schooljaren. Daarbij wordt gebruikt gemaakt van een correctie voor leerlinggewicht. Een faire beoordeling van de leerprestaties van de school is bijzonder lastig omdat steeds verschillende groepen leerlingen (van dezelfde school) met elkaar worden vergeleken die niet dezelfde startsituatie hebben. Omdat er geen rekening wordt gehouden in verschillen tussen de startsituaties blijft het erg lastig om te bepalen of de prestaties op die eindtoets toe te schrijven zijn aan het onderwijs of aan hetgeen de leerlingen van huis uit hebben meegekregen.
De beoordeling van de kwaliteit van de leerprestaties wordt een stuk eerlijker als we de prestatiegroei van leerlingen erbij kunnen betrekken en als we zouden kunnen aangeven welke bijdrage de school daaraan heeft geleverd. Met maten voor toegevoegde waarde wordt dat mogelijk.

Voorstellen

De drie instituten die in het project samenwerken stellen voor vier maten voor leerwinst en twee maten voor toegevoegde waarde samen met de scholen uit te proberen en verder te ontwikkelen.

Cito en de Universiteit Twente richten zich op de ontwikkeling van leerwinstmaten die voorlopig als werktitel hebben:

1. z-score model (Cito);
2. CuKuMu-model (UT);
3. Seizoensgebonden leerwinstmodel (UT);
4. Subdomeinen leerwinstmodel (UT).

Het GION ontwikkelt twee maten voor toegevoegde waarde die we voorlopig de volgende werktitel hebben gegeven:

1. Groeicurve-model;
2. Variantiecomponenten-model.

Bijlage 5 Z-score voor de analyse van leerresultaten en schooleffectiviteit

Inleiding

Het model van Keuning en Feskens (2013) voor de analyse van leerresultaten en schooleffectiviteit maakt gebruik van een vorm van risicostratificatie. Bij toepassing van deze strategie wordt de totale leerlingpopulatie verdeeld in subgroepen met een vergelijkbare predispositie (Deeks et al., 2003). De analyse wordt voor elk van de subgroepen apart uitgevoerd. Een belangrijk voordeel hiervan is dat subgroepen met leerlingen direct op basis van de uitkomstvariabele met elkaar vergeleken kunnen worden. Het is niet nodig om bij de interpretatie van de uitkomstvariabele rekening te houden met achtergrondvariabelen die al meegenomen zijn bij het indelen van de leerlingpopulatie in subgroepen. Keuning en Feskens (2013) maken de indeling in subgroepen op basis van het startniveau van de leerlingen. Dit betekent dat voor elke mogelijke pretestscore een verwachte posttestscore en een standaarddeviatie wordt berekend. De geobserveerde posttestscore wordt vervolgens vergeleken met de verwachte posttestscore voor een leerling en gedeeld door de standaarddeviatie. We verkrijgen een groeiscore die gestandaardiseerd is naar startniveau. Op het niveau van de leerling voorzien de groeiscores in een interpreteerbare maat voor leerwinst. De maat houdt rekening met individuele verschillen in groei. Op het niveau van de school maken de groeiscores het mogelijk om de effectiviteit van een school op een eerlijke manier te bepalen, zonder dat daarbij gecorrigeerd hoeft te worden voor leerling- en schoolkenmerken.

Berekeningsstappen in het risicostratificatiemodel

Het risicostratificatiemodel veronderstelt dat de vaardigheid van leerlingen op een bepaald vakgebied ten minste tweemaal getoetst wordt. Bij deze twee metingen wordt bij voorkeur gebruikgemaakt van toetsen die qua moeilijkheidsgraad aansluiten bij het vaardigheidsniveau van de leerlingen die de toetsen maken (Wilson, 2005). Dit betekent dat leerkrachten voor de verschillende metingen in de tijd veelal verschillende toetsen gebruiken. De ruwe scores die leerlingen behalen op deze toetsen zijn niet direct met elkaar te vergelijken. Het is bijvoorbeeld niet meteen duidelijk of een leerling die bij de eerste meting 25 opgaven correct maakt op de ene toets, en bij de tweede meting 27 opgaven op een andere toets, vooruit is gegaan. Bij het volgen van de vaardigheid van leerlingen is het dan ook niet betekenisvol om het leerresultaat in kaart te brengen met een set losse toetsen. Het is beter om gebruik te maken van meetschalen waarmee het leerresultaat onafhankelijk van de set toetsopgaven in kaart gebracht kan worden. Dergelijke meetschalen kunnen geconstrueerd worden met behulp van meetmodellen uit de item respons theorie (Hambleton, Swaminathan & Rogers, 1991; Embretson & Reise, 2000). Als het meetmodel geldt voor een verzameling toetsopgaven kunnen de ruwe scores van de leerlingen via het onderliggende meetmodel gecorrigeerd worden voor de moeilijkheidsgraad van de toets. De gecorrigeerde scores worden in de praktijk meestal schaalscores genoemd. Het risicostratificatiemodel gebruikt de ruwe- naar-schaalscore-tabel van de toets die wordt afgenomen bij de eerste meting als startpunt.

De groeiscore voor een leerling wordt in het risicostratificatiemodel in drie stappen berekend. Eerst wordt voor elke mogelijke ruwe score bij de eerste meting r_1, $r_1 = 0, ..., R_1$, de verwachte schaalscore $\mu_{\theta_2|r_1}$ en de bijbehorende standaarddeviatie $\sigma_{\theta_2|r_1}$ voor de tweede meting berekend. De verwachte schaalscore voor een leerling is gelijk aan:

$$\mu_{\theta_2|r_1} = \frac{1}{n_{r_1}} \sum_{i=1}^{n_{r_1}} \theta_{2i},$$

waarbij de sommatie loopt over alle $i \in \{1, 2, ..., n\}$ leerlingen met exact dezelfde ruwe score bij de eerste meting. De standaarddeviatie $\sigma_{\theta_2|r_1}$ wordt gegeven door:

$$\sigma_{\theta_2|r_1} = \sqrt{\frac{\sum_{i=1}^{n_{r_1}} \theta_{2i} - \mu_{\theta_2|r_1}}{n_{r_1} - 1}}.$$

De verwachte schaalscore wordt in de volgende stap van de geobserveerde schaalscore afgetrokken. Tot slot wordt deze verschilscore omgezet in een *z*-score door hem te delen door de standaarddeviatie die bij de verwachte schaalscore hoort. De groeiscore voor een leerling wordt dus als volgt berekend:

$$z_i = \frac{\theta_{2i} - \mu_{\theta_2|r_{1i}}}{\sigma_{\theta_2|r_{1i}}}.$$

We zien dat leerlingen vergeleken worden met andere leerlingen die exact hetzelfde startniveau hadden. Niet alle berekeningen hoeven na elke toetsafname verricht te worden. De verwachte scores voor de R_1 scoregroepen en de bijbehorende standaarddeviaties kunnen we eenmalig vooraf berekenen, zodat we in de praktijk gedeeltelijk gebruik kunnen maken van scoretabellen. Tabel A laat ter illustratie zien hoe de scoretabel vormgegeven kan worden. In de eerste twee kolommen staan de ruwe scores en de bijbehorende schaalscores die leerlingen kunnen behalen op de toets die bij de eerste meting wordt voorgelegd. In de derde kolom staat het aantal leerlingen per scoregroep. In de vijfde en zesde kolom staat voor elke scoregroep de verwachte schaalscore bij de tweede meting en de standaarddeviatie. In tabel A kan voor elke leerling dus de verwachte schaalscore bij de tweede meting opgezocht worden. De scoretabel kan gemaakt worden voor elk vakgebied en voor elke mogelijke combinatie van twee metingen. Het is dus mogelijk om metingen *medio 3* en *medio 5* met elkaar te vergelijken, maar we kunnen bijvoorbeeld ook metingen *einde 4* en *einde 7* met elkaar vergelijken. Een belangrijke voorwaarde voor het berekenen van de verwachte schaalscores is dat we via scholen kunnen beschikken over de longitudinale data die zij via het leerlingvolgsysteem verzamelen.

Tabel A Voorbeeld van een scoretabel met de verwachte schaalscores naar scoregroep

r_1	θ_1	n_{r_1}	$\mu_{\theta_2 \mid r_1}$	$\sigma_{\theta_2 \mid r_1}$
0	70.20	6	106.92	8.15
1	77.48	81	109.88	6.72
2	80.92	199	114.11	6.53
..	..	..	..	..
23	97.68	580	121.91	6.55
24	98.12	652	122.12	6.98
25	98.56	634	121.72	6.50
26	99.00	601	121.75	6.73
27	99.44	564	122.22	6.84
..	..	..	..	..
48	114.68	171	129.37	6.77
49	117.56	125	128.76	6.43
50	123.72	16	131.19	5.60

Bij de bespreking van het risicostratificatiemodel hebben we ons tot nog toe gericht op de analyse van de leerresultaten van individuele leerlingen. Het model kan uitgebreid worden naar het niveau van de school door een multilevel model te schatten met leerlingen *i* genest in scholen *j* en de eerder gedefinieerde maat voor leerwinst als afhankelijke variabele:

$z_{ij} = \beta_{0j} + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0,\ \sigma^2)$ Niveau 1 – leerlingniveau

$\beta_{0j} = \gamma_{00} + \mu_{0j},\ \mu_{0j} \sim N(0,\ \tau^2)$ Niveau 2 – schoolniveau

$z_{ij} = \gamma_{00} + \mu_{0j} + \varepsilon_{ij}$ Gecombineerd

We zien dat het model veronderstelt dat de scores van de leerlingen afhangen van de schoolgemiddelden β_{0j} en van de *random* individuele variatie ε_{ij} rond het schoolgemiddelde. Daarnaast veronderstelt het model dat het gemiddelde voor een school is opgebouwd uit een algemeen gemiddelde γ_{00} plus een afwijking van dat gemiddelde dat specifiek is voor de betreffende school μ_{0j}. Feitelijk scheidt het model de variantie van de afhankelijke variabele dus in een deel dat aan de scholen is toe te schrijven en in een deel dat aan de leerlingen is toe te schrijven. Op basis van het model kan de effectiviteit van een school als volgt gekwantificeerd worden:

$$\tilde{\beta}_{0j} = \frac{\beta_{0j} - \gamma_{00}}{\tau} \quad \text{met} \quad \beta_{0j} = \sum_{i=1}^{n_j} z_{ij} \div n_j .$$

De scores voor de scholen $\tilde{\beta}_{0j}$ kunnen op exact dezelfde manier geïnterpreteerd worden als de scores voor de leerlingen z_i. Zowel de scores op leerlingniveau als de scores op schoolniveau volgen namelijk een standaardnormale verdeling met een gemiddelde gelijk aan 0 en een standaarddeviatie gelijk aan 1. Een score van +1.14 betekent op beide niveaus dus dat we behoorlijk goed presteren in vergelijking met andere leerlingen met exact hetzelfde startniveau of in vergelijking met andere scholen.

In de hierboven beschreven maat voor schooleffectiviteit is geen rekening gehouden met de betrouwbaarheid van het schoolgemiddelde. Dit is onwenselijk, omdat scholen met weinig leerlingen hierdoor het risico lopen dat zij onterecht als zwak of excellent aangemerkt worden. Dit probleem kan opgelost worden door in de maat voor schooleffectiviteit ook informatie over de totale populatie mee te nemen. In het risicostratificatiemodel van Keuning en Feskens (2013) wordt de maat voor schooleffectiviteit $\tilde{\beta}_{0j}$ standaard gecorrigeerd voor schoolgrootte. Op basis van de geschatte varianties op school- en leerlingniveau wordt eerst een wegingsfactor λ_j voor een school uitgerekend (zie Snijders & Bosker, 1999):

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma^2 / n_j}.$$

De wegingsfactor drukt de betrouwbaarheid van het gemiddelde van school *j* uit in een getal. Vervolgens wordt het schooleffect als volgt gecorrigeerd voor schoolgrootte:

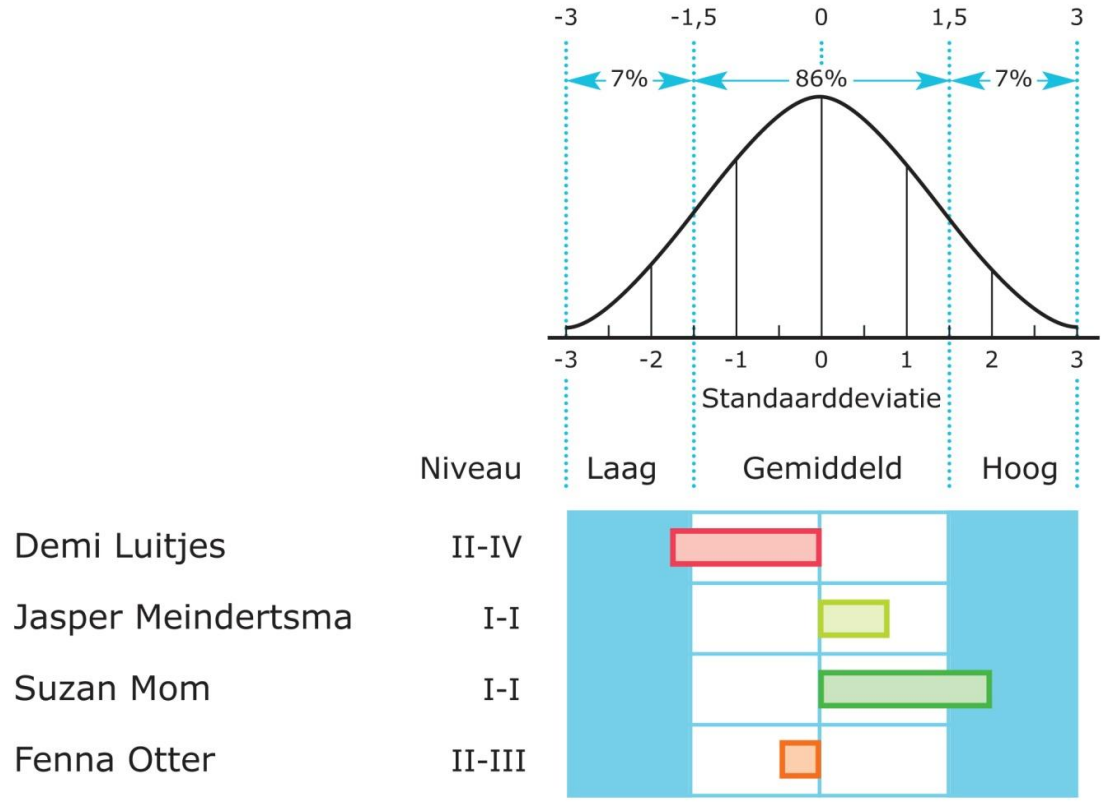
$$\tilde{\beta}_{0j}^{EB} = \gamma_{00}(1 - \lambda_j) + \tilde{\beta}_{0j}\lambda_j.$$

Uit de formule voor de berekening van de wegingsfactor valt af te leiden dat de grootte van σ^2 / n_j afhankelijk is van het aantal leerlingen binnen een school. Hoe groter een aantal leerlingen des te meer zal λ_j naderen tot 1. In dat geval wordt de gecorrigeerde maat voor schooleffectiviteit $\tilde{\beta}_{0j}^{EB}$ meer bepaald door de toetsresultaten van de leerlingen op school en minder door het algemene gemiddelde. Bij een kleiner leerlingaantal neemt λ_j verhoudingsgewijs af en wordt $\tilde{\beta}_{0j}^{EB}$ meer bepaald door het algemene gemiddelde. Door een dergelijke correctie toe te passen zorgen we ervoor dat we scholen alleen als zwak of excellent classificeren als we zeker weten dat de prestaties van de leerlingen daar aanleiding toe geven. Net als eerder bij de leerlingen kunnen we ook bij toepassing van het risicostratificatiemodel op schoolniveau grotendeels terugvallen op scoretabellen. De varianties op leerling- en schoolniveau worden eenmalig geschat op basis van de longitudinale data die we via scholen verkrijgen en de wegingsfactor kan voorafgaand aan toepassing van het model berekend worden voor verschillende schoolgroottes.

Het risicostratificatiemodel in de onderwijspraktijk

Hoewel het risicostratificatiemodel wat opzet betreft vrij eenvoudig is in vergelijking met sommige alternatieve modellen voor leerwinst en schooleffectiviteit (zie bijvoorbeeld, Timmermans, Doolaard, & De Wolf, 2011; Tekwe et al., 2004; Raudenbush, 2004; Kelly & Downey, 2010) mogen we niet verwachten dat een leerkracht of intern begeleider zich verdiept in de technische details van een model. Daarom is met het oog op de uitvoering van het project *LTW-PO* een conceptrapportage ontwikkeld. De rapportage is stapsgewijs in samenspraak met het onderwijsveld tot stand gekomen. Figuur A laat zien hoe het risicostratificatiemodel de resultaten op leerlingniveau rapporteert. Aan de linkerkant van Figuur A zien we de namen van de leerlingen in een klas. Direct achter de naam wordt het vaardigheidsniveau bij de eerste en de tweede meting uitgedrukt in de Romeinse cijfers I tot en met V. Er wordt uitgegaan van een kwintielschaal met vijf gelijke groepen: I = ver boven het gemiddelde (20 procent), II = boven het gemiddelde (20 procent), III = de gemiddelde groep leerlingen (20 procent), IV = onder het

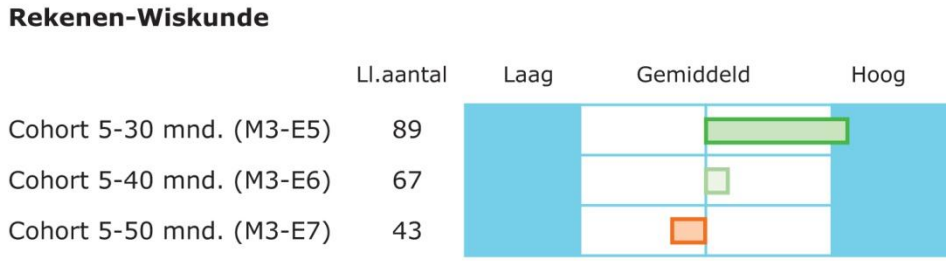
gemiddelde (20 procent) en V = ver onder het gemiddelde (20 procent). Ten slotte wordt de leerwinst per vakgebied gevisualiseerd in een afwijkingsgrafiek. Er zijn vier kleurcoderingen. De leerwinst die een leerling op een vakgebied laat zien tussen twee metingen is: opvallend groot (donkergroen), iets groter dan verwacht (lichtgroen), iets kleiner dan verwacht (oranje) of opvallend klein (rood). De lengte van het balkje geeft weer hoeveel standaarddeviaties de leerwinst van een leerling afwijkt van de leerwinst die andere leerlingen met exact hetzelfde startniveau gemiddeld laten zien.



Figuur A Voorbeeldrapportage op leerlingniveau

In de toelichting bij de rapportage wordt aangegeven hoe de leerwinst voor een leerling berekend wordt en hoe de uitkomst geïnterpreteerd moet worden. Er wordt onder meer aangegeven dat: (1) een *z*-score van 0 wijst op een gemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau, (2) een *z*-score > 0 wijst op een bovengemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau, en dat (3) een *z*-score < 0 wijst op een benedengemiddelde groei ten opzichte van andere leerlingen met hetzelfde startniveau. Daarnaast wordt uitgelegd dat een z-score die exact gelijk is aan 0 in de praktijk niet vaak zal worden waargenomen en dat kleine afwijkingen naar boven en beneden heel normaal zijn. Om dit extra te benadrukken, onderscheidt het leerlingrapport drie categorieën, namelijk *laag*, *gemiddeld* en *hoog*. De z-scores tussen –1.5 en +1.5 vallen in het wit gearceerde gebied en worden aangemerkt als gemiddeld. Naar verwachting behaalt 86 procent van de leerlingen een *z*-score tussen deze grenzen. De overige scores vallen in het grijs gearceerde gebied en worden aangemerkt als (opvallend) laag of hoog. Naar verwachting behaalt zeven procent van de leerlingen een *z*-score <1.5 en eveneens zeven procent een *z*-score >1.5. Als de *z*-score van een

leerling in het grijs gearceerde gebied valt, kan er reden zijn om het onderwijsaanbod iets bij te stellen. Hoewel de indeling in drie categorieën de interpretatie vereenvoudigt, komen de grenzen enigszins arbitrair tot stand. Er is uitgegaan van anderhalve standaarddeviatie, omdat deze grenswaarde vaker gebruikt wordt in het onderwijsveld (Resing et al., 2008; Verhoeven, Keuning, Horsels & Van Boxtel, 2013). De grens kan desgewenst ook bij ±1.65 of ±1.96 standaarddeviatie gelegd worden, zodat de prestaties van respectievelijk tien of vijf procent van de leerlingen als opvallend aangemerkt worden.
De resultaten op schoolniveau worden op dezelfde wijze gepresenteerd als de resultaten op leerlingniveau. Figuur B laat zien hoe de rapportage op schoolniveau eruit ziet. Zoals we kunnen zien, worden de resultaten geordend naar vakgebied. In het voorbeeld worden de resultaten voor het vakgebied rekenen-wiskunde gepresenteerd. Aan de linkerkant van figuur B zien we welke combinaties van twee metingen geanalyseerd zijn. De eerste regel heeft betrekking op metingen *medio 3* en *einde 5.* In het project LTW-PO is ervoor gekozen om de metingen te definiëren in termen van het aantal onderwijsmaanden. Bij meting *medio 3* hebben leerlingen normaliter vijf maanden onderwijs achter de rug en bij meting *einde 5* dertig maanden onderwijs. Vertraagde en versnelde leerlingen worden op basis het aantal onderwijsmaanden geclassificeerd in een bepaalde groep. Dit betekent dat in de vergelijking tussen metingen *medio 3* en *einde 5* niet alleen de leerlingen met een normale onderwijsloopbaan meegenomen zijn, maar ook een klein aantal vertraagde en versnelde leerlingen. Deze leerlingen hebben feitelijk niet aan meting *medio 5* meegedaan, maar aan meting *medio 4* (vertraagde leerlingen) of aan meting *medio 6* (versnelde leerlingen). Daarom wordt in de schoolrapportage in eerste instantie over cohorten met een bepaald aantal onderwijsmaanden gesproken. Over jaargroepen wordt pas in tweede instantie gesproken, omdat de jaargroepen voor de vertraagde en versnelde leerlingen geen correcte weergave geven van de werkelijkheid. Direct achter de cohortaanduiding staat het aantal leerlingen dat is meegenomen in de analyse. Daarna wordt de schooleffectiviteit gevisualiseerd in een afwijkingsgrafiek. Net als in de rapportage op leerlingniveau wordt gebruikgemaakt van vier kleurcoderingen en de categorieën *laag*, *gemiddeld* en *hoog*. De kleurcoderingen en de categorieën hebben in beide rapportages exact dezelfde betekenis.



Figuur B Voorbeeldrapportage op schoolniveau

Conclusies en discussie

Het gebruik van het risicostratificatiemodel van Keuning en Feskens (2013) brengt enkele voordelen met zich mee in vergelijking met andere modellen voor leerwinst en schooleffectiviteit. Op het niveau van de leerling voorziet het model in een transparante maat voor leerwinst. De maat houdt rekening met individuele verschillen in groei. Op het niveau van

de school maakt het model het mogelijk om de effectiviteit van een school in kaart te brengen zonder dat daarbij gecorrigeerd hoeft te worden voor een groot aantal leerling- en schoolkenmerken (zie Feskens & Keuning, in voorbereiding). De kenmerken komen via de voorgestelde risicostratificatie namelijk in de afhankelijke variabele terecht. Ze hoeven niet in het deterministische deel van het model meegenomen te worden. Dit maakt de implementatie eenvoudiger. Het is namelijk niet nodig om op grote schaal privacygevoelige informatie te verzamelen. Ook kan geen discussie ontstaan over de vraag of er wel of niet gecorrigeerd moet worden voor een bepaald kenmerk. Niettemin is onduidelijk voor welke kenmerken precies gecorrigeerd wordt via de voorgestelde risicostratificatie. Het model houdt rekening met individuele verschillen in startniveau, omdat de voorspellingen conditioneel verricht worden op basis van de toetsscores bij de eerste meting. Daarnaast is er sprake van *fairness* correctie, omdat de groeisnelheid per subgroep kan variëren. Het risicostratificatiemodel verschilt dan ook van alternatieve modellen voor schooleffectiviteit waarin géén *fairness* correctie wordt toegepast. Dat komt doordat dergelijke modellen veronderstellen dat het effect van de variabele *startniveau* voor alle leerlingen hetzelfde is. Variatie in de voorspelde groeisnelheid ontstaat pas als relevante leerling- en schoolkenmerken aan het model toegevoegd worden en we feitelijk dus een model mét *fairness* correctie gebruiken. We weten op dit moment niet precies hoe het risicostratificatiemodel zich verhoudt tot alternatieve modellen voor schooleffectiviteit. Vermoedelijk resulteren de verschillende modellen onder bepaalde omstandigheden in vergelijkbare uitkomsten.

Het risicostratificatiemodel is ontwikkeld als hulpmiddel voor scholen en leerkrachten. De informatie die volgt uit het model kan de leerkracht houvast geven in de planningscyclus. Het onderwijsaanbod kan op basis van de voorspelling afgestemd worden op het niveau van de leerling en leerresultaten kunnen geëvalueerd worden door de voorspelling te vergelijken met daadwerkelijk behaalde toetsresultaten. Het is geenszins de bedoeling dat het ontwikkelde risicostratificatiemodel bij toepassing in een leerlingvolgsysteem ook in een beoordelingskader terechtkomt. Dat zou er namelijk toe kunnen leiden dat scholen het leerlingvolgsysteem met tegenzin gaan gebruiken en/of manipulatief gaan inzetten, omdat zij afgerekend worden op de resultaten die uit het leerlingvolgsysteem volgen. De ondersteunende functie die een leerlingvolgsysteem heeft bij de vastlegging, analyse en interpretatie van leerresultaten van individuele leerlingen komt dan in het geding. Voordat het risicostratificatiemodel als hulpmiddel aangeboden kan worden aan scholen en leerkrachten is nader onderzoek noodzakelijk. Ten eerste is niet bekend hoe de onbetrouwbaarheid van een meetresultaat de uitkomsten op leerling- en schoolniveau beïnvloedt. Het is mogelijk dat de *z*-score informatie geeft over het leerresultaat van een individuele leerling die moeilijk te legitimeren is in het licht van de standaardmeetfout die we waarnemen bij de eerste en de tweede meting. Als we geen rekening houden met de standaardmeetfout zetten we scholen en leerkrachten mogelijk op het verkeerde spoor. Ten tweede weten we niet hoe de voorgestelde risicostratificatie exact functioneert in de praktijk. Een indeling in subgroepen op basis van de score bij de eerste meting is inhoudelijk fraai, maar vanuit statistisch oogpunt gezien misschien niet altijd te bewerkstelligen, omdat we bepaalde scores in de praktijk nauwelijks waarnemen. De subgroepen moeten in dat geval ingedikt worden. Het is de vraag wat we aan nauwkeurigheid winnen als de subgroepen ingedikt worden en welke informatie verloren gaat. Ten slotte,

brengt een leerlingvolgsysteem de prestaties van leerlingen herhaald in kaart. In de huidige vorm sluit het risicostratificatiemodel niet bij dit principe aan, omdat slechts twee metingen meegenomen worden in de analyse. Misschien is een uitbreiding naar een model waarin metingen genest zijn binnen leerlingen, en leerlingen genest zijn binnen scholen mogelijk.

Referenties

Deeks, J.J., Dinnes J, D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M., & Altman, D.G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*, 27.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Feskens, R., & Keuning, J. (in voorbereiding). *Analysis of Student Growth and School Effectiveness: Validation of a Risk Stratification Model*.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kelly, A. & C. Downey (2010). Value-added measures for schools in England: looking inside the 'black box' of complex metrics. *Educational Assessment: Evaluation and Accountability*, *22*, 181 – 198.

Keuning, J. & Feskens, R. (2013). *Meten van leerwinst en toegevoegde waarde op basis van niveau-gestandaardiseerde groeiscores*. Paper gepresenteerd tijdens Onderwijs Research Dagen, Brussel.

Raudenbusch, S.W. (2004). *Schooling, statistics and poverty: can we measure school improvement?* Educational Testing Service. Policy Evaluation and Research Center Princeton, NJ.

Resing, W.C.M., Evers, A., Koomen, H.M.Y., Pameijer, N.K. & Bleichrodt, N. (2008). I*ndicatiestelling speciaal onderwijs en leerlinggebonden financiering. Condities en instrumentarium*. Amsterdam: Boom Test Uitgevers.

Tekwe, C.D., Carter, R.L., Ma, C., Algina, J., Lucas, M.E., Roth, J., Ariet, M., Fisher T. & M. B. Resnick (2004). An empirical comparison of statistical models for value-added assessment of school. *Journal of Educational and Behavioral Statistics,* 29, 11 - 36.

Snijders, T.A.B., Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications, London.

Timmermans, A. C., Doolaard, S., & De Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement, 22*, 393 - 413.

Verhoeven, L, Keuning, J., Horsels, L. & Van Boxtel, H. (2013). *Testinstrumentarium taalontwikkelingsstoornissen (T-TOS)*. Arnhem: Cito.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Bijlage 6 Achtergrond van Toegevoegde waarde modellen

Op basis van een internationale wetenschappelijke literatuurstudie zijn twee methoden geselecteerd waarmee de toevoegde waarde van een school bepaald zou kunnen worden in de pilot. Het gaat om het Vaardigheidsverschil-model en Vaardigheidsgroei-model. In deze bijlage zal eerst toegelicht worden welke type modellen potentieel geschikt zijn om de toegevoegde waarde te bepalen (1) en hoe we tot de keuze van twee type modellen zijn gekomen (2). Daarna volgt een nadere technische toelichting op de twee gebruikte toegevoegde waarde modellen (3 en 4).

1. Beschikbare toegevoegde waarde modellen

Uitgangspunt voor de literatuurstudie is een rapport van de OECD (2008) over dit onderwerp geweest. Hierin worden vier verschillende categorieën modellen beschreven (A t/m D). Op basis van een aanvullende literatuurstudie is dit aantal uitgebreid met drie (E, F, G). De bespreking van de acht categorieën hieronder richt zich vooral op verschillen in de manier waarop de schatting van de score op de eindtoets (de afhankelijke variabele) wordt gedaan: welke aannames worden gemaakt over de variabiliteit in de eindtoetsscores van leerlingen en in de leerlingachtergrond- en schoolkenmerken waarvoor eventueel gecorrigeerd zou moeten worden. Ook worden nationale en internationale voorbeelden van toepassing in de genoemd.

A. Lineaire regressiemodellen

Dit is het meest basale toegevoegde waarde model; de score op de eindtoets van een leerling wordt voorspeld op basis van de score op zijn begintoets. Hiermee lijkt dit model sterk op de berekening van de leerwinst; de eindtoetsscore minus begintoetsscore. Het lineaire regressiemodel is eenvoudig uit te breiden met nog meer toetsscores(s) van de leerling en met leerlingachtergrond- en schoolcontextkenmerken die als correctiefactoren kunnen dienen (fairness-kenmerken). Deze extra kenmerken in het model worden covariaten genoemd. Zo wordt een multiple lineaire regressie model verkregen waarin het intercept de geschatte gemiddelde eindtoetsscore van alle leerlingen is als voor de invloed van de covariaten die in het model zijn meegenomen is gecorrigeerd.

Belangrijk nadeel van dergelijke lineaire regressie modellen is dat geen rekening gehouden wordt met hiërarchische structuur van een school: leerlingen zijn gegroepeerd binnen klassen en klassen zijn gegroepeerd binnen scholen. Deze structuur zorgt voor een bepaalde afhankelijkheid tussen de toetsscores; leerlingen binnen een klas lijken meer op elkaar dan willekeurig leerlingen. Als geen rekening wordt gehouden met deze samenhang tussen de gegevens kan dit interpretatiefouten te gevolg hebben (te kleine standaardfouten en ecologische valkuil; zie Snijders en Bosker, 2012). Een tweede beperking is dat lineaire regressie modellen uit alleen ‘fixed’ effecten bestaan. Er wordt verondersteld dat de leerlingen (of scholen – afhankelijk van het analyseniveau -) niet van elkaar verschillen in de manier waarop een kenmerk bijvoorbeeld sociaal milieu samenhangt met de score op de eindtoets; het is steeds een vaststaand algemeen effect.

Nationaal is dit model toegepast door Roeleveld (2003) op de PRIMA cohort data, internationaal door onder meer Ladd en Walsh (2002), Jakubowski (2008), U.S. Department of education (2011), Webster en Mendro (1997), Webster (2005), Klein, Freedman, Shavelson en Bolus (2008) en Isenberg en Hock (2011).

B. Random intercept-modellen

In deze varianten op het lineaire regressiemodel (A) wordt wél de hiërarchische structuur van de data verdisconteerd: de totale variantie in de eindtoetsscores wordt gesplitst in variantie op leerlingniveau (variantie tussen leerlingen) en variantie op schoolniveau (variantie tussen scholen). Het intercept – de gemiddelde score op de eindtoets – wordt daarbij verondersteld random te zijn op leerling- en schoolniveau. Dit betekent dat ervan uitgegaan wordt dat niet alleen leerlingen kunnen verschillen in eindtoetsscore maar dat ook scholen onderling, en dat deze verschillen kunnen samenhangen met eerdere toetsscore(s) of andere covariaten die in het model als onafhankelijke variabelen zijn meegenomen (bijvoorbeeld de begintoetsscore en sociaal milieu of grootte van een school). De covariaten kunnen in principe zowel 'random' als 'fixed' variabelen zijn. Voorbeelden van nationaal gebruik van dit model zijn Bosker, Béquin en Rekers-Mombarg (2001) en Wijnstra, Ouwens en Béguin (2003). Internationaal worden deze modellen ook veelvuldig gebruikt (Ray, 2006; Webster & Mendro, 1997; Webster, 2005; Antelius, 2006; Thomas, Peng & Gray, 2007).

C. Multivariate random-effectmodellen

Het belangrijkste verschil met de voorafgaande categorieën is dat met multivariate random-effectmodellen meerdere eindtoetsscores voorspeld worden met één model (multivariaat). Bijvoorbeeld de score op de Cito Eindtoets Basisonderwijs én de score op de rekentoets M8 van een leerling wordt tegelijkertijd gerelateerd aan eerdere toetsscores van deze leerling, met - eventueel – correctie voor relevante kenmerken (covariaten). Het voordeel van deze categorie meerniveau modellen is dat als er meerdere eindtoetsscore beschikbaar zijn op de scholen de samenhang tussen deze scores ook meegenomen kan worden in de berekening van de toegevoegde waarde. Hierdoor kunnen verschillen tussen scholen efficiënter en nauwkeuriger geschat worden dan met random intercept-modellen (categorie B). Het nadeel is de grotere complexiteit van de modellen en de moeilijk te onderbouwen aannames over de samenhang tussen variabelen in de modellen. Internationale voorbeelden van gebruik zijn Sanders en Horn (1984), Goldstein (1997) en Lauder, Kounali, Robinson en Goldstein (2010).

D. Groeicurve-modellen

De ontwikkeling in leerprestaties van een leerling wordt inzichtelijk door een meerniveau-model dat een schatting maakt van de vloeiend verlopende groeilijn (veelal een 1^e of 2^e graads polynoom) die achter de herhaaldelijk gemeten leerprestaties schuilgaat (latente groeicurve). Het model schat voor iedere leerling afzonderlijk het intercept (het aanvangsniveau) en de hellingshoek (ontwikkelingssnelheid) van zijn groeilijn. Groeicurve-modellen kunnen vrij eenvoudig worden uitgebreid met relevante correctiefactoren en met een derde niveau (schoolniveau). Het is niet nodig om de ontwikkeling gedurende de gehele basisschoolperiode in een keer te modelleren; opdelen in groeilijnen voor de onderbouw, middenbouw en bovenbouw behoort ook tot de mogelijkheden. Dit heeft belangrijke voordelen: vertekening

door instroom en uitstroom van leerlingen wordt beperkt, inhoudelijk zijn de toetsen voldoende op elkaar afgestemd en leerkrachten kunnen tijdig inzicht krijgen in de ontwikkeling van hun leerlingen. Voor elk (deel)model moeten echter wel minimaal twee toetsscores per leerling verspreid in de tijd beschikbaar zijn. Nationaal zijn groeicurve-modellen beschreven door Guldemond en Bosker (2009), internationaal door bijvoorbeeld Poniscial en Byrk (2005) en Choi en Seltzer (2005).

E. Cross classificatie random-effect groeicurvemodellen

Deze categorie modellen zijn een uitbreiding op bovenstaande groeicurve-modellen (D). De modellen verdisconteren ook dat leerlingen bij doubleren dezelfde toets meerdere keren doen en bij instroom en uitstroom tot meerdere basisscholen behoren (cross classificatie). Evenals bij de vorige categorie modellen wordt het intercept en de hellingshoek van de groeilijn voor iedere leerling geschat, gaat het om meerniveau modellen en zijn tenminste twee meetmomenten verspreid in de tijd voor iedere groeicurve nodig. Toepassing van dit model is door Palardy (2010) beschreven.

In de pilot moet een afweging gemaakt worden of de toegenomen complexiteit het rechtvaardigt om op deze manier voor doubleren en schoolwisseling te controleren. Door het maken van groeilijnen per bouw van de basisschool – zoals voorgesteld bij D – kunnen problemen met ontbrekende gegevens door doubleren en schoolwisselingen waarschijnlijk voldoende ondervangen worden. Hierdoor verdient deze categorie E in dit project niet de voorkeur.

F. Kwartiel-regressiemodellen

Kwartiel-regressiemodellen onderscheiden zich van alle voorafgaande modellen (A t/m E) doordat het hier niet gaat om het schatten van een gemiddelde score op de eindtoets, maar om het schatten van een bepaalde percentiel van de verdeling van de eindtoetsscores (quantile regression models). Voor iedere percentiel (bijvoorbeeld P10, P25, P50, P75, P90) wordt een aparte vergelijking gemaakt (growth percentiles). De P10-curve geeft bijvoorbeeld aan hoe de ontwikkeling verloopt van leerlingen die tot de slechtste 10% van de leerlingen in de normgroep behoren. Zo kan de positie van scholen die relatief ver van het gemiddeld presteren – bijvoorbeeld sbo-scholen - nauwkeurig worden geschat. Deze categorie is een variant op het lineaire regressie model (categorie A): met de hiërarchische structuur van de data wordt geen rekening gehouden. Verder is correctie voor fairness-kenmerken maar in beperkte mate mogelijk. Het model is gebruikt door Betebenner (2007) en door Haile en Nguyen (2008).

G. Kwartiel random-effect regressiemodellen

Deze modellen zijn de meerniveau variant van de kwartiel regressie modellen (F). Een bepaald percentiel in de verdeling van de eindtoetsscore wordt voorspeld uit eerdere toetsscores en correctiefactoren. Tzavidis, Salvati, Geraci en Bottai (2010) ontwikkelden het ‘M-quantile and expectile random effect regression models’. Omdat het een erg complexe methode is waarvan nog vrijwel geen toepassingen in de praktijk bekend zijn en waarvoor nauwelijks software beschikbaar is, valt deze categorie modellen af voor de pilot LTW-PO.

2. Toepasbare toegevoegde waarde modellen

Uit het bovenstaande overzicht van de beschikbare toegevoegde waarde modellen volgt dat cross classificatie random-effect groeicurvemodellen (E) en kwartiel random-effect regressiemodellen (G) *niet* geschikt zijn om te gaan uitproberen in de Nederlandse praktijk. Voor de vijf overgebleven categorieën toegevoegde waarde modellen is een globale inschatting gemaakt van de toepasbaarheid van het model voor de scholen en voor de Inspectie van het Onderwijs. Hierbij is een duidelijk verschil in doelstelling te constateren, hetgeen waarschijnlijk in een andere voorkeur voor maten zal resulteren.

Het gebruik van gegevens over de leervorderingen van leerlingen vormt een essentieel onderdeel van opbrengstgericht werken door scholen. Hierbij stellen scholen doelen en wordt aan de hand van leerprestaties en de ontwikkeling daarin gecheckt of deze doelen gehaald worden. Als scholen tijdig zicht hebben op de leervorderingen van hun leerlingen kan men hier gericht op sturen. Belangrijk hierbij is dat de gekozen maat of maten inzichtelijk zijn voor de diverse betrokkenen op de scholen. Lineaire regressiemodellen (A) en kwartiel-regressiemodellen (F) komen dan in principe in aanmerking.

De Inspectie van het Onderwijs heeft vooral belang bij een inzichtelijke en maatschappelijk geaccepteerde maat voor de toegevoegde waarde van een school. Belangrijk is ook de beschikbaarheid van een landelijk representatieve normgroep, zodat de inspectie kan beoordelen hoe goed de betreffende school het doet in vergelijk met andere (vergelijkbare) basisscholen in Nederland. De vergelijking moet wel ‘fair’ zijn; scholen moeten niet afgerekend worden op factoren waar ze geen invloed op hebben zoals een laag instroomniveau, een hoog percentage gewichtenleerlingen of een grote mobiliteit in de wijk waar de school staat. Dit maakt dat Kwartiel-regressiemodellen (F) voor de Inspectie minder geschikt zijn omdat deze slechts in beperkte mate rekening kunnen houden met de ‘fairness’-variabelen: er worden aparte modellen voor alleen de belangrijkste subgroepen van leerlingen of scholen berekend. Ook lineaire regressie modellen (A) zijn in het kader van het onderwijstoezicht minder geschikt; ze doen geen recht aan de structuur van de data en waardoor er foutieve conclusies getrokken kunnen worden over verschillen tussen scholen.

Random intercept-modellen (B) zijn methodologisch beter dan de (multiple) lineaire regressie modellen (A) omdat deze wel rekening houden met de hiërarchische structuur in de data en correctie voor fairness-kenmerken op het juiste niveau kan plaatsvinden: leerlingkenmerken worden op leerlingniveau (L1) meegenomen en schoolkenmerken op schoolniveau (L2). Verder zijn deze modellen flexibel: de invloed van fairness-kenmerken op de eindtoetsscore kan zowel ‘fixed’ als ‘random’ zijn. Daarbij komt dat een random intercept-model met leerwinst als afhankelijke variabele direct aansluit bij de leerwinstmodellen in de pilot; ze liggen in elkaars verlengde. Dit maakt het extra aantrekkelijk om deze categorie modellen te gaan uitproberen in de praktijk. Overigens is in de UK met dit type meerniveau-modellen ruime praktijkervaring opgedaan bij scholen en onderwijsinspectie. Daar wordt sinds 2005 de Contextual Value Added (CVA) van scholen berekend met behulp van random intercept-modellen.

Groeicurve-modellen (E) komen in principe ook in aanmerking om toegepast te worden in de pilot. Het zijn meerniveau-modellen waarbij correctie voor fairness-kenmerken eenvoudig uitvoerbaar is. Ook deze categorie modellen sluit nauw aan bij de leerwinstmodellen in de pilot. Belangrijk voordeel ten opzichte van random intercept-modellen (categorie B) is dat niet alleen de begin- en eindtoetsscore van een leerling wordt meegenomen, maar ook alle tussenliggende toetsscores. Zo wordt een gedetailleerder beeld van de ontwikkeling verkregen. De toepasbaarheid van deze categorie modellen is echter wel direct afhankelijk van het aantal beschikbare meetmomenten per leerling: minimaal 2 toetsscores verspreid in de tijd per groeicurve. Er zal daarom eerste geïnventariseerd moeten of aan deze voorwaarde in het algemeen voldaan wordt. Scholen die de nieuwste versie van de LVS-toetsen van het Cito al meerdere jaren gebruiken voldoen ruimschoots aan deze vereiste.

Resteert nog één categorie modellen: de multivariate random-effectmodellen (C). Dit zijn complexe meerniveau modellen waarbij meerdere eindtoetsscores, bijvoorbeeld Technisch lezen en Begrijpend lezen in groep 6, tegelijkertijd voorspeld worden uit eerdere toetsscores en waarbij eventueel ook nog gecorrigeerd kan worden voor fairness-kenmerken. In de pilot wordt de toepasbaarheid van een model in de eerste plaats bepaald door inzichtelijkheid en bruikbaarheid voor de scholen. Aan beide voorwaarden wordt hier niet voldaan. Voor de scholen leidt het samenvoegen van meerdere de eindtoetsscores ook tot interpretatieproblemen: ze geven de leerkracht, ib-er en directeur geen specifieke aanknopingspunten voor verbetering van het onderwijs aan hun leerlingen.

Samenvattend kan geconcludeerd worden dat twee categorieën toegevoegde waarde modellen het meest geschikt zijn voor de pilot: random intercept-modellen (B) en groeicurve-modellen (D). Lineaire regressiemodellen (A) en Kwartiel-regressiemodellen (F) zijn methodologisch gezien minder geschikt. Uitbreiding van kwartiel-regressiemodellen naar een meerniveau variant (G) is in principe mogelijk, maar hier is nu nog (te) weinig ervaring mee opgedaan (Tzavidis, Salvati, Geraci & Bottai, 2010). Tot slot, Multivariate random-effectmodellen (C) en Cross classificatie random-effect groeicurvemodellen (E) en zijn te complex en leveren waarschijnlijk ernstige interpretatieproblemen op voor de scholen.

3. Technische toelichting Vaardigheidsverschil-model

Het vaardigheidsverschil-model is een meerniveau random intercept-model (categorie B), ook wel variantie-componentenmodel of random-effectmodel genoemd. Voor uitvoerige bespreking van dit model wordt verwezen naar Snijders en Bosker (2012), pag. 49-56. Voor de analyses is het softwarepakket ML-win versie 2.27 gebruikt.

Het vaardigheidsverschil-model is een twee-niveau model; de leerwinstmetingen van een leerling (L1) zijn gegroepeerd binnen een school (L2). De bijbehorende wiskundige vergelijking ziet er als volgt uit:

$$Y_{ij} = \beta_0 + u_{0j} + e_{0ij} \qquad \text{(I)}$$

Met daarin:

Y_{ij} : Vaardigheidscore $_{\text{eindmeting}}$ – Vaardigheidsscore $_{\text{beginmeting}}$ van leerling *i* op school *j*. Oftewel, de leerwinst op een bepaald leerstofgebied – bijvoorbeeld spelling - van een individuele leerling *i* op school *j* gedurende een bepaalde rapportageperiode – bijvoorbeeld tussen 5 en 30 maanden onderwijs (M3 tot E5 voor nominaal doorstromende leerlingen).

β_0 : het algemeen gemiddelde (intercept). Om verder te gaan met het voorbeeld; het is de gemiddelde groei in vaardigheid voor spelling tussen 5 en 30 maanden onderwijs van alle leerlingen in pilot.

u_{0j} : het residu op schoolniveau; de afwijking van een school *j* van het algemeen gemiddelde. De aanname is dat de schoolresiduen u_{0j} normaal verdeeld zijn, met een gemiddelde waarde van 0 en een variantie van τ_0^2 .

e_{0ij} : het residu op leerlingniveau; de afwijking van een leerling *i* van het algemeen gemiddelde. De aanname is dat de leerlingresiduen e_{0j} normaal verdeeld zijn, met een gemiddelde waarde van 0 en een variantie van σ_0^2.

In de schoolrapportage Toegevoegde waarde wordt het algemeen gemiddelde β_0 gepresenteerd in een tabel als de gemiddelde bruto leerwinst van pilot-scholen. Het is de gemiddelde ongecorrigeerde) leerwinst van alle leerlingen op alle pilotscholen. Een voorbeeld van een dergelijke tabel staat hieronder (tabel A). De gemiddelde groei in spellingsvaardigheid tussen 5 en 30 maanden onderwijs voor alle pilot-scholen samen is 22,4.

De som van β_0 en u_{0j} is groei in gemiddelde vaardigheid tussen begin en eindmeting op een school. Uit de berekening volgt dat school *j* gemiddeld genomen 1,9 punt hogere leerwinst behaalt dan alle Pilot scholen samen. Op school *j* is de schoolgemiddelde leerwinst voor spelling dan gelijk aan 22,4 + 1,9 = 24,3. Dit is in tabel A weergegeven als de gemiddelde 'totale leerwinst' van uw school *j*.

Tabel A Toename in vaardigheidsscore spelling gedurende 25 maanden onderwijs op basis van het vaardigheidsverschil-model

	Gemiddelde Pilot-scholen	Gemiddelde van uw school *j* [Onzekerheidsgrenzen]	Uw school *j* afwijkend van Pilot-scholen?
Onderwijsmaanden 5 tot 30 (M3-E5):			
• Bruto leerwinst • Netto leerwinst	22,4 19,9	24,3 [22,2; 26,4] 21,0 [18,9; 23,2]	gelijk gelijk

De tweede regel in tabel A geeft weer wat de gemiddelde netto leerwinst van alle pilot-scholen samen en van school *j* is. Dit is bepaald door het basismodel voor de totale leerwinst (vergelijking I) uit te breiden met een zestal fairness-kenmerken: hoogste opleidingsniveau van

de ouders, etniciteit van het kind, dyslexie met indicatie, dyscalculie met indicatie, adhd of add met indicatie, autisme/ass/pdd-nos met indicatie. Ze zijn op volgende manier in de berekeningen zijn meegenomen:

- hoogste opleidingsniveau van beide ouders/verzorgers is een ordinale variabele (x_{1ij}):
- 1 = geen onderwijs gevolgd;
- 2 = 1-3 jaar basisonderwijs;
- 3 = 4-6 jaar basisonderwijs/svo *(=categorie 1 gewichtenregeling)*;
- 4 = 1-2 jaar lbo/vmbo bbl-kbl/(i(vbo);
- 5 = 3-4 jaar lbo/vmbo bbl-kbl/i(vbo) *(=categorie 2 gewichtenregeling)*;
- 6 = 1-2 jaar mavo/vmbo tl- gl;
- 7= 3-4 jaar mavo/vmbo tl- gl *(=categorie 3 gewichtenregeling)*;
- 8 = 1-3 jaar havo/vwo;
- 9 = 4-6 jaar havo/vwo;
- 10 = Mbo/leerlingwezen;
- 11 = Hbo;
- 12 = Universiteit.
- etniciteit van een leerling is bepaald op basis van het geboorteland van de ouders volgens de CBS definitie. Het kenmerk is meegenomen in de modellen als twee dummy variabelen Westers allochtoon (x_{2ij}) en niet-Westers allochtoon (x_{3ij}). Autochtoon fungeert als referentiecategorie.
- dyslexie is een dichotome variabele met géén dyslexie als referentiecategorie (x_{4ij}).
- dyscalculie is een dichotome variabele met géén dyscalculie als referentiecategorie (x_{5ij}).
- adhd of add is een dichotome variabele met géén adhd/add als referentiecategorie (x_{6ij}).
- autisme/ass/pdd-nos is een dichotome variabele met géén autisme, ass of pdd-nos als referentiecategorie (x_{7ij}).

De fairness-kenmerken die aan het model zijn toegevoegd, worden ook wel covariaten genoemd. Dit zijn variabelen die in de berekeningen worden betrokken omdat correctie ervoor wenselijk is. Doordat ze aan de modellen zijn toegevoegd wordt de invloed van deze fairness-kenmerken op de leerwinst geneutraliseerd. Hoe dit in zijn werk gaat is globaal als volgt. Stel dat niet-westers allochtone leerlingen gemiddeld genomen een leerwinst voor spelling behalen die 4 (vaardigheids)punten lager is dan van autochtone leerlingen. Westers allochtone leerlingen behalen bijvoorbeeld gemiddeld een 2 punten lagere score. In de berekeningen wordt hiervoor gecorrigeerd door bij niet-westers allochtone leerlingen 4 punten op te tellen bij zijn of haar werkelijk behaalde leerwinst voor spelling. Bij westers allochtone leerlingen komen er 2 punten bij. Met andere woorden, we voegen als het ware bij iedere allochtone leerling een stukje leerwinst toe die bepaald wordt door zijn etnische herkomst; als hij een autochtone leerling zou zijn geweest dan had hij naar verwachting respectievelijk 4 en 2 punten hoger gescoord. Zo wordt per leerling een corrigeerde leerwinstscore berekend die vrij is van de invloed van etniciteit. Er zijn als het ware alleen maar autochtone leerlingen op de pilotscholen.

Dit doen we op een vergelijkbare wijze ook voor de overige vijf fairness-kenmerken. Als naast etniciteit ook de andere 'fairness-kenmerken aan het model zijn toegevoegd, is de behaalde leerwinst niet alleen gezuiverd van de invloed van etniciteit, maar tegelijkertijd ook van het hoogste opleidingsniveau van de ouders en de invloed van zorgleerlingen (dyslexie met

indicatie, dyscalculie met indicatie, adhd of add met indicatie en autisme/ass/pdd-nos met indicatie). Zo verkrijgt ieder leerling een voor fairness-kenmerken gecorrigeerd leerwinst bepaling. Door vervolgens de gecorrigeerde leerwinsten van de leerlingen te middelen per school, wordt de gemiddelde netto leerwinst van een school verkregen. Het is dat deel van de leerwinst voor spelling dat met enige zekerheid aan de school is toe te schrijven. De niet-schoolse invloeden op de totale leerwinst voor spelling zijn er zo goed mogelijk uitgezuiverd.

De uitbreiding van het brutomodel (I) voor de leerwinst met de zes fairness-kenmerken ziet er in een wiskundige vergelijking als volgt uit:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + u_{0j} + e_{0ij} \quad \text{(II)}$$

Vergelijking II toont het model waarmee de netto leerwinst van een school kan wordt geschat. Dit is te beschouwen als een indicator van de toegevoegde waarde van een school. In de vergelijking zijn β_1 tot en met β_6 zogenaamde 'fixed' regressiecoëfficiënten. Daarmee veronderstellen we dat de manier waarop de zes fairness-kenmerken hun invloed uitoefenen op de leerwinst voor alle leerlingen en alle scholen hetzelfde is (fixed). Het intercept β_0 uit vergelijking II is de *gecorrigeerde* algemeen gemiddelde leerwinst van alle leerlingen in de pilot. In voorbeeldtabel A uit het schoolrapport Toegevoegde waarde wordt dit netto leerwinst van de pilotscholen genoemd (19,9). Uit de berekeningen volgt dat op school *j* - mét correctie voor de fairness-kenmerken – de leerlingen een leerwinst voor spelling behalen die gemiddeld genomen 1,1 punt hoger is dan voor alle leerlingen op alle pilotscholen samen. De som van β_0 en u_{0j}(21,0) geeft een indicatie van de gemiddelde toegevoegde waarde van school *j.*

Als een school van minder dan 10 leerlingen bruikbare data heeft, wordt er geen schoolgemiddelde netto of bruto leerwinst gepresenteerd in de schoolrapportage Toegevoegde waarde.

4. Technische toelichting Vaardigheidsgroei-model

Aan het vaardigheidsgroei-model ligt een meerniveau lineair groeicurve-model ten grondslag (categorie D). Er is hier sprake van een drie-niveau model; de metingen *t* (L1) zijn gegroepeerd binnen een leerling *i* (L2), en leerlingen zijn weer gegroepeerd binnen een school *j* (L3). Een gedetailleerde bespreking van dergelijke modellen is te vinden in Snijders en Bosker (2012, pag. 247-280). Voor het analyseren is gebruik gemaakt van het softwarepakket ML-win versie 2.27.

De kern van het vaardigheidsgroei-model is dat per rapportageperiode (bijvoorbeeld 5 tot 30 maanden onderwijs) van iedere leerling zoveel mogelijk meetmomenten – twee of meer - worden meegenomen in de berekeningen. Deze herhaalde metingen van de leerlingen kunnen gezien worden als een populatie van individuele groeilijnen. Elke groeilijn toont de ontwikkeling in vaardigheid van een leerling op een bepaald leerstofgebied waarbij de vaardigheidsscore is uitgezet tegen de (kalender)leeftijd op het moment van toetsafname. Verder is het zo dat een meetmoment – bijvoorbeeld de medio groep 3 toets – niet voor alle leerlingen in de pilot op exact dezelfde dag plaatsvindt; er is een marge toegestaan van

ongeveer twee maanden. Dit geldt ook voor alle andere meetmomenten. Een dergelijke opzet wordt ook 'variable occasion design' genoemd.

De vorm van de groeilijnen hoeft niet voor alle leerlingen gelijk te zijn, maar we veronderstellen dat er wel een bepaalde algemene wiskundige vergelijking achter schuilt gaat: de functie $F_{ij}(t)$. De aanname is dat een rechte lijn (1e graads polynoom) of gebogen lijn (2e graads polynoom) de samenhang tussen leeftijd en vaardigheidsscore het beste beschrijft. Welke vorm het beste past, volgt uit de berekeningen van de fit van het model. De wiskundige vergelijking van het vaardigheidsgroei-model ziet er als volgt uit:

$$Y_{tij} = F_{ij}(t) + u_{0j} + e_{0ij} + r_{0tij} \qquad \text{(III)}$$

Met daarin:

Y_{tij} : Vaardigheidsscore op meetmoment *t* van leerling *i* op school *j.*
Uitgaande van half jaarlijkse toetsafname zijn tussen bijvoorbeeld 5 en 30 maanden onderwijs van iedere leerling maximaal zes vaardigheidsscores beschikbaar (M3, E3, M4, E4, M5, E5). De waarde van *t* loopt daarmee uiteen van 1 tot 6. Het minimaal aantal meetmomenten per leerling is gesteld op twee. Het sporadisch ontbreken van vaardigheidsscores door bijvoorbeeld ziekte van de leerling is zijn algemeenheid niet problematisch. Dat wordt het wel als het gaat om veel leerlingen van één school. Daarom wordt voor een school met minder dan 10 leerlingen met bruikbare data geen schoolgemiddelde groeicurve gepresenteerd.

u_{0j} : het residu op schoolniveau; de afwijking van een school *j* van het algemeen gemiddelde. De aanname is dat de schoolresiduen u_{0j} normaal verdeeld zijn, met een gemiddelde waarde van 0 en een variantie van τ_0^2 .

e_{0ij} : het residu op leerlingniveau; de afwijking van een leerling *i* op school *j* van het algemeen gemiddelde. De aanname is dat de leerlingresiduen e_{0ij} normaal verdeeld zijn, met een gemiddelde waarde van 0 en een variantie van σ_0^2.

r_{0ij} : het residu op meetmoment-niveau; de afwijking van een meting *t* van leerling *i* op school *j* van het algemeen gemiddelde. De aanname is dat de meetmoment-residuen r_{0tij} normaal verdeeld zijn, met een gemiddelde waarde van 0 en een variantie van ρ_0^2.

$F_{ij}(t)$: De functie die de samenhang tussen leeftijd en vaardigheidsscore voor bijvoorbeeld spelling tussen 5 en 30 maanden onderwijs van alle leerlingen in de pilot beschrijft. Er wordt in de pilot uitgegaan van een 2e graads polynoom als passende functie; een afbuigende groeilijn met maximaal één top.

We veronderstellen verder dat zowel leerlingen als scholen van elkaar kunnen verschillen als het gaat om het vaardigheidsniveau bij de startmeting op 5 maanden onderwijs (random intercept; u_{0j}, e_{0ij}) en de hellingshoek van de groeilijn (random slope; u_{1j}, e_{1ij}). De afbuiging

van de lijn (leeftijd²) is daarentegen voor iedereen en alle scholen gelijk (fixed). De bijbehorende functie ziet er als volgt uit:

$$F_{ij}\,(t) = \beta_0 + \beta_1 leeftijd + \beta_2 leeftijd^2 + u_{1j} + e_{1ij} \qquad \text{(IV)}$$

Het samenvoegen van vergelijking III en IV resulteert in de volgende vergelijking:

$$Y_{tij} = \beta_0 + \beta_1 leeftijd + \beta_2 leeftijd^2 + u_{0j} + e_{0ij} + v_{0tij} + u_{1j} + e_{1ij} \qquad \text{(V)}$$

Het fixed deel van de vergelijking wordt gevormd door het intercept β_0, en de regressiecoëfficiënten β_1 en β_2. Ze vormen samen de wiskundige vergelijking voor de gemiddelde groeilijn van alle pilot-leerlingen. Op basis van alle beschikbare vaardigheidsscores van bijvoorbeeld spelling tussen 5 en 30 maanden onderwijs wordt een gemiddelde waarde voor het intercept β_0, de hellingshoek β_1 en de afbuiging β_2 bepaald. Vervolgens kan hiermee een pilotgemiddelde groeilijn berekend worden door steeds een andere waarde voor de leeftijd te kiezen. Bijvoorbeeld, de gemiddelde vaardigheidsscore van alle 6 jarige leerlingen op de pilot-scholen is:

$$Y_{tij} = \beta_0 + \beta_1 * 6 + \beta_2 * 6^2$$

Voor 7 jarigen wordt in plaats van 6 de waarde 7 ingevuld, enzovoort. Zo berekenen we als het ware de groeilijn van de 'gemiddelde' pilot-leerling voor spelling tussen 5 en 30 maanden onderwijs. In het Schoolrapport Toegevoegde waarde dit weergeven als de blauwe lijn in het linker panel (Bruto leerwinst). Figuur A geeft hiervan een voorbeeld.

De residuen $u_{0j}, e_{0ij}, v_{0tij}, u_{1j}$ en e_{1ij} vormen samen het random deel van de vaardigheidsgroei-model. Ze geven aan in welke mate een individuele meting, leerling of school afwijkend is van de pilot-gemiddelde groeilijn. Omdat we vooral geïnteresseerd zijn in de groeilijnen van scholen, beperken we ons tot de residuen die daarvoor van belang zijn. Het gaat dan om de residuen u_{0j} en u_{1j} die voor ieder school apart berekend worden (gemiddelde afwijking). Voor alle duidelijkheid, de waarden voor alle regressie coëfficiënten en residuen worden in één keer geschat; de gemiddelde waarden voor de coëfficiënten β_0, β_1 en β_2 en de residuen $u_{0j}, e_{0ij}, v_{0tij}, u_{1j}$ en e_{1ij} volgen uit dezelfde model doorrekening.

Door de schoolspecifieke residuen (u_{0j} en u_{1j}) op te tellen bij de waarden voor de regressiecoëfficiënten van de pilot-gemiddelde groeilijn (β_0 , β_1 en β_2) wordt duidelijk wat de gemiddelde vaardigheid voor een school is bij een bepaalde leeftijd van haar leerlingen. Bijvoorbeeld, op de leeftijd van 6 jaar is de vaardigheidsscore voor school *j* gelijk aan:

$$Y_{tij} = \beta_0 + \beta_1 * 6 + \beta_2 * 6^2 + u_{0j} + u_{1j}$$

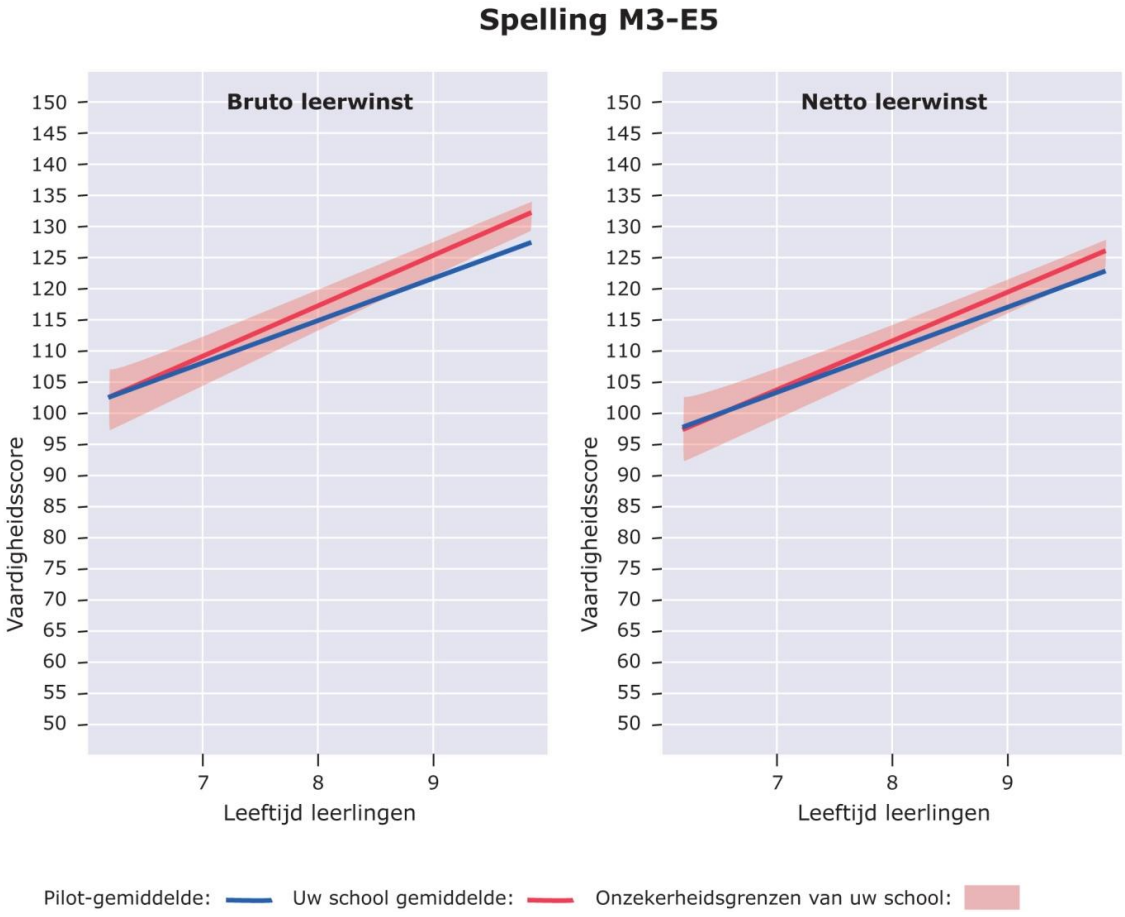
Voor zeven jarigen op school *j* kan in plaats van 6 weer de waarde 7 ingevuld worden, enzovoort. Dit levert de gemiddelde groeilijn van leerlingen op school *j*. In het schoolrapport

Toegevoegde waarde is deze schoolgemiddelde groeilijn weergegeven als een rode lijn in het linker panel van de figuur (zie Figuur A; Bruto leerwinst).

Vervolgens is het vaardigheidsgroei-model V is uitgebreid met fairness-kenmerken om zo de netto leerwinst, oftewel de toegevoegde waarde van een school zo goed mogelijk te bepalen. Het gaat om dezelfde zes fairness-kenmerken die op dezelfde manier zijn geoperationaliseerd als bij het vaardigheidsverschil-model. De bijbehorende wiskundige vergelijking is:

$$Y_{tij} = \beta_0 + \beta_1 leeftijd + \beta_2 leeftijd^2 + \beta_3 x_{1ij} + \beta_4 x_{2ij} + \beta_5 x_{3ij} + \beta_6 x_{4ij} + \beta_7 x_{5ij} + \beta_8 x_{6ij} + u_{0j} + e_{0ij} + v_{0tij} + u_{1j} + e_{1ij} \quad \text{(VI)}$$

Op basis van alle beschikbare vaardigheidsscores voor bijvoorbeeld spelling tussen 5 en 30 maanden onderwijs wordt vervolgens niet alleen het intercept, de hellingshoek en de afbuiging van de individuele groeilijnen berekend, maar tegelijkertijd ook het gemiddelde effect dat elk fairness-kenmerk afzonderlijk en onafhankelijk van elkaar heeft op de vaardigheidsscores voor spelling. Door de fairness-kenmerken aan het basale vaardigheidsgroei-model (V) toe te voegen wordt de invloed van deze kenmerken geneutraliseerd. De benodigde mate van neutralisatie wordt namelijk direct bepaald door het gemiddelde effect die elk afzonderlijk kenmerk heeft op de groeilijnen van de pilot-leerlingen. Dit is wordt weergegeven door de bijbehorend regressiecoëfficiënten (β_3 tot en met β_8 in vergelijking VI). Op deze manier wordt per leerling een corrigeerde groeilijn berekend die vrij is van de invloed van hoogste opleidingsniveau van de ouders, etniciteit en type zorgleerling. Dit is de toegevoegde waarde groeilijn; de groeilijn die weergeeft hoe de groei in vaardigheidsscore verloopt als zo goed mogelijk voor de invloed van niet-schoolse kenmerken is gecorrigeerd.



Figuur A Ontwikkeling in vaardigheidsscore spelling in 25 maanden onderwijs op basis van het vaardigheidsgroei-model

Op basis van de regressiecoëfficiënten β_0, β_1 en β_2 kan de gemiddelde gecorrigeerde groeilijn van alle pilotleerlingen worden berekend. Dit gaat op dezelfde manier als bij de bruto groeilijn voor alle pilotscholen samen. Zo wordt het pilot-gemiddelde van de netto groeilijnen verkregen. In het schoolrapport Toegevoegde waarde is dit de blauwe lijn in het rechter panel (Figuur A, Netto leerwinst).

De netto groeilijn voor school j wordt berekend door de schoolspecifieke residuen u_{0j} en u_{1j} op te tellen bij de waarden voor de regressiecoëfficiënten van de pilot-gemiddelde groeilijn (β_0 , β_1 en β_2). De werkwijze is gelijk aan die voor de bruto groeilijn van school *j.* In het schoolrapport Toegevoegde waarde is de gecorrigeerde schoolgemiddelde groeilijn weergegeven als een rode lijn in het rechter panel van figuur A: de netto groeilijn van school *j.* Hiermee wordt een indicatie gegeven van de ontwikkeling in de gemiddelde toegevoegde waarde van deze school.

Referenties

Antelius, J. (2006). *Value-Added Modelling in Sweden: A Background Report for the OECD Project on the Development of Value-added Models in Education Systems*. Zweden: Skolverket.

Betebenner, D.W. (2007) *Estimation of student growth percentiles for the Colorado student assessment program*. http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf

Bosker, R.J., Béquin, A., & Rekers-Mombarg, L.T.M. (2001). Hoe meten we de prestatie van een school? In: Dijkstra, A.B., Karsten, S., Veenstra, R., & Visscher, A. (Eds) *Het oog der natie: scholen op rapport, standaarden voor de publicatie van schoolprestaties*.(pp 121-135). Assen, Van Gorcum.

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*. March. Los Angeles, California: National Center for Research on Evaluation, Standards and Student Testing/UCLA.

Guldemond, H., & Bosker, R. J. (2009). School effects on students' progress – a dynamic perspective. *School Effectiveness and School Improvement*, 20 (2), 255-268.

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.

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.

Jakubowski, M. (2008). *Implementing value-added models of school assessment*. RSCAS Working Papers 2008/06, Florence: European University Institute.

Klein, S., Freedman, D., Shavelson, R., & Bolus, R. (2008). Assessing school effectiveness. *Evaluation Review*, 32 (6), 511-525.

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

OECD (2008). *Measuring Improvements in learning outcomes; best practices to assess the value-added of schools*. Paris: OECD Publications.

Palardy, G.J. (2010). The multilevel crossed random effects growth model for estimating teacher en school effects: issues and extensions. *Educational and Psychological Measurement*, 70 (3), 401-419.

Ponisciak, P. M., & Bryk, A.S. (2005). Value-added analysis of the Chicago public schools: an application of hierarchical models. In R.L.(Ed.), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.

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*. https://consumption.education.gov.uk/publications/eOrderingDownload/RW85.pdf

Roeleveld. J. (2003) *Herkomstkenmerken en begintoets. Secundaire analyses op het Prima-cohortonderzoek*. Nijmegen, ITS.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. Londen: Sage publications.

Thomas, S., Peng, W.J., & Gray, J. (2007). Value added trends in English secondary school performance over ten years. *Oxford Review of Education*, 33 (3), 261-295.

Tzavidis, N., Salvati, N., Geraci, M., & Bottai, M. (2010*). M-quantile and expectile random effects regression for multilevel data.* Working Paper M10/07. Southampton: University of Southampton, Southamptons Statistical Sciences Research Institute.

U.S. Department of education, office of planning, evaluation and policy development, policy and program Studies Service. (2011). *Final report on the evaluation of the growth model pilot project*. Washington, D.C. ED Pubs.

Webster, W.J., Mendro, R.L., Orsak, T.H., & Weerasinghe, D. (1996). *The applicability of selected regression and hierarchical linear models to the estimation of school and teacher effects*. Paper gepresenteerd op de Annual Meeting of the National Council in Measurement in Education, New York, NY. 8-12 april.

Webster, W. & Mendro. R. (1997). The Dallas value-added accountability system. In J. Millman (ed.), *Grading teachers, grading schools: is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.

Webster, W. J. (2005). The Dallas School-Level Accountability Model: The Marriage of Status and Value-Added Approaches. In R.W. Lissitz (ed.), *Value added models in education: Theory and Applications*. Maple Grove, MN: JAM Press.

Wijnstra, J., Ouwens, M. & Béguin, A. (2003). *De toegevoegde waarde van de basisschool*. Arnhem: Citogroep.