Centraal Planbureau

CPB Achtergronddocument

Centraal Planbureau
Van Stolkweg 14
Postbus 80510
2508 GM Den Haag

T (070)3383 380
I www.cpb.nl

Contactpersoon
Marc van der Steeg (06-21974089)
Sander Gerritsen (070-3383381)

Aan: Ministerie van OCW

Datum: 29 januari 2013
Betreft: Evaluatie pilot investeren in kwaliteit leraren

Samenvatting

De leraar is een sterk bepalende factor voor de kwaliteit van het onderwijs en voor latere levens van kinderen. Chetty e.a. (2011) laten bijvoorbeeld zien dat kinderen die op jonge leeftijd leraren hebben gehad met een hogere 'value added' later vaker college bezoeken, meer verdienen en in betere buurten wonen. Minder duidelijk is welke factoren nu echt bepalend zijn voor de kwaliteit van leraren. Klassieke factoren als het opleidingsniveau en ervaring blijken in de meeste studies niet tot nauwelijks verband te houden met de gerealiseerde leerwinst van de leerlingen. Recente studies in de Verenigde Staten vinden wel significante verbanden tussen subjectieve beoordelingen van bekwaamheid van leraren door schoolleiders en de leerwinst van hun leerlingen. Ook zijn positieve verbanden gevonden tussen evaluaties van bekwaamheid van leraren op basis van klassenobservaties en leerresultaten van de leerlingen. Het betreft klassenobservaties door getrainde observanten met behulp van gedetailleerde scoresystemen met heldere standaarden voor gedrag of vaardigheden van leerkrachten.

Dit document onderzoekt een pilot die is uitgevoerd bij 125 leerkrachten van zeven basisscholen in Amsterdam. De pilot vond plaats in het schooljaar 2011-12 en besloeg een periode van een half jaar.

De pilot is gericht is op het verbeteren van bekwaamheid van leraren en in het verlengde daarvan de prestaties van leerlingen. De pilot berust op de volgende drie pijlers:

1) Meten van bekwaamheid van leerkrachten door middel van lesobservaties door getrainde observanten en een gedetailleerd scoresysteem van verschillende gedragsaspecten die gerelateerd zijn aan opbrengstgericht werken. Het betreft een meetinstrument met 75 gedragsaspecten op het vlak van pedagogische

competentie, vakinhoudelijk-didactische competentie en organisatorische competentie. De deelnemende leerkrachten zijn gemeten voorafgaand en aan het eind van de pilot. In beide gevallen betrof het aangekondigde observaties.

2) Intensieve coaching en training op maat gericht op verbetering van leerkrachtvaardigheden. Dit is gebeurd door externe coaches met veelal lesgevende ervaring. De coaching heeft plaatsgevonden gedurende een periode van ongeveer een half jaar. Coaches hielden zich bezig met lesobservaties en feedback, video-interactiebegeleiding, individuele voortgangsgesprekken en teambijeenkomsten en -trainingen.

3) Prestatiebeloning, waarbij teams van leraren in aanmerking komen voor een teambudget en een individuele uitkering als minimaal 85 procent van de teamleden een minimaal niveau en een bepaalde groei in de score op het observatie-instrument wist te behalen. De doelstellingen voor groei waren hoger voor de zwakkere leerkrachten dan voor de betere leerkrachten.

De kosten van de coaching bedroegen 3000 euro per leerkracht, die van de prestatiebeloning 1000 euro per leerkracht als individuele uitkering en een teambudget van 8000 euro. In totaal hebben de leerkrachten gedurende een half jaar gemiddeld 35 uur besteed aan de pilot. Vijf van de zeven teams hebben voldaan aan de gestelde criteria en de prestatiebeloning behaald. De twee teams die het niet hebben gehaald zijn gestruikeld over het criterium dat de leerkrachten een bepaald minimum niveau moesten halen op het observatie-instrument.

Een belangrijke aanleiding voor de pilot is geweest het voornemen van het kabinet Rutte 1 om ervaring op te gaan doen met prestatiebeloning voor leraren. Het betreffende bestuur waaronder de zeven deelnemende scholen vallen bleek geïnteresseerd om een pilot op te starten met prestatiebeloning, mits dit gepaard zou gaan met gerichte interventies om leerkrachtvaardigheden te verbeteren en mits de prestatiecriteria ook aan leerkrachtvaardigheden zouden worden verbonden. De Dienst Maatschappelijke Ontwikkeling (DMO) van Amsterdam heeft de middelen voor de coachingsinterventies ter beschikking gesteld, na goedkeuring van de diverse projectplannen van de betrokken scholen. De voornaamste rol van het CPB is de verantwoordelijkheid voor het onderzoek naar de pilot. Ook heeft het CPB in de beginfase meegedacht over de opzet van de pilot voor wat betreft de prestatiebeloning. De uiteindelijke besluitvorming over de vormgeving van de pilot lag bij DMO en het schoolbestuur, na consultatie en goedkeuring van de personeelsgeleding van de medezeggenschapsraad.

Dit document gaat achtereenvolgens in op de volgende deelvragen:

1. Welke veranderingen in gedrag en vaardigheden zijn waargenomen bij deelnemende leerkrachten?
2. In hoeverre vormt het gehanteerde meetinstrument van bekwaamheid van leerkrachten een goede voorspelling voor de leerwinst die leerlingen boeken?
3. Hoe is de houding van de leerkrachten ten aanzien van de pilot, en dan voornamelijk ten aanzien van prestatiebeloning?
4. Welke meer algemene lessen kunnen we trekken uit deze pilot?

Veranderingen in vaardigheden en gedrag

De groei in waargenomen leerkrachtvaardigheid in termen van de score op het observatie-instrument tussen begin en eind van de pilot is substantieel. De gemiddelde groei bedroeg 21 gedragsaspecten op een te behalen totaal van 75 gedragsaspecten. Dit komt overeen met ruim een standaarddeviatie. De waargenomen groei is het sterkst bij de bij aanvang zwakste leerkrachten en neemt af naarmate de score op de nulmeting hoger is. De groei is het grootst bij pedagogische vaardigheden en vakinhoudelijk-didactische vaardigheden, en minder groot bij organisatorische vaardigheden. Er zijn drie nuanceringen te maken bij de omvang van de waargenomen groei in leerkrachtvaardigheden. Ten eerste, we kunnen hieraan geen uitspraken ontlenen over de effecten van de pilot op bekwaamheid. Dat zou een experiment vergen met overtuigende controlegroepen waar de betreffende interventies niet worden uitgevoerd.[1] Ten tweede, coaches en directeuren hebben gemiddeld gesproken een wat negatiever beeld over de daadwerkelijk doorgemaakte groei dan het beeld dat volgt uit de metingen. Een combinatie met onaangekondigde lesobservaties zal naar verwachting een betrouwbaarder beeld opleveren. Ten derde, zowel coaches als directeuren verwachten dat gemiddeld ongeveer de helft van de doorgemaakte groei na twee jaar weggeëbd zal zijn als geen verder onderhoud en investering wordt gepleegd in de bekwaamheid van betreffende leerkrachten.

Leraren rapporteren meer interactie met collega's op allerlei terreinen ten opzichte van het jaar voor de pilot. De grootste veranderingen hierin hebben zich voorgedaan op het vlak van het bespreken van de planning van de lessen, het analyseren van resultaten van de eigen leerlingen met collega's en het advies vragen en geven aan collega's ter verbetering van de eigen leerkrachtvaardigheid respectievelijk die van een collega. Het observeren van lessen van collega's of het mentor zijn voor een collega komt relatief weinig voor en hier is ook relatief de minste verandering in

[1] Bij deze pilot was een experimentele setting met overtuigende controlegroepen niet haalbaar in het schooljaar 2011-12. Dat zou (veel) meer scholen vergen en een loting tussen welke scholen wel en niet de interventies zouden krijgen. Het doel was nadrukkelijk eerst een kleinschaliger pilot op te zetten en lessen te trekken uit de eerste ervaringen met prestatiebeloning, coaching en meten van leerkrachtvaardigheden, om vervolgens op te schalen naar een echt experiment in het schooljaar 2012-13. Vanwege het besluit tot stopzetting van de middelen voor prestatiebeloning heeft opschaling niet meer plaats kunnen vinden.

opgetreden. Zowel leraren als coaches als directeuren geven aan dat er veel van elkaar geleerd kan worden als leraren vaker bij elkaar in de les zouden kijken.

Relatie score op observatie-instrument en leerprestaties

Het in de pilot gehanteerde meetinstrument op basis van lesobservaties door getrainde observanten is goed in staat om verschillen in kwaliteit van leraren zichtbaar te maken. Er bestaat een positief significant verband tussen de score op het meetinstrument en de leerprestaties van leerlingen. Dit verband is het sterkst bij rekenen en spelling, gevolgd door lezen. Ter illustratie, het vervangen van een leraar uit het slechtste kwart door een leraar uit het beste kwart op basis van de score op het meetinstrument gaat gepaard met een gemiddelde winst in leerprestaties van circa 0.4 standaarddeviatie bij rekenen en spelling (en 0.24 bij lezen). Dit is een niet triviale winst die overeenkomt met ongeveer het verschil in gemiddelde scores tussen leerlingen met een havo/vwo-advies en leerlingen met een vwo-advies. Leerlingen die twee jaar op rij een slechte leraar hebben in plaats van een goede leraar zouden daarmee een heel niveau lager uit kunnen komen in het vervolgonderwijs. Deze resultaten suggereren dat het meetinstrument daadwerkelijk vaardigheden meet die ertoe doen voor de prestaties van leerlingen. Dit is een belangrijke bevinding, omdat veel traditionele kenmerken van leraren als vooropleiding en ervaring niet of nauwelijks voorspellend blijken voor de kwaliteit van leraren. Zelfs zwakke signalen van verschillen in kwaliteit kunnen informatie verschaffen die zeer waardevol is, onder andere voor ontwikkelings- en beloningsbeleid op scholen. Het observatie-instrument geeft specifieke informatie op welke competenties leraren zich nog kunnen verbeteren.

Houding ten aanzien van meten en meetinstrument

Leraren, coaches en directeuren zijn overwegend positief over het meten van leerkrachtvaardigheden door middel van lesobservaties. Hetzelfde geldt voor het in de pilot gehanteerde meetinstrument de Amsterdamse Kijkwijzer. Wel zijn alle actoren overwegend van mening dat een combinatie van aangekondigde en onaangekondigde lesobservaties tot een betrouwbaarder beeld zal leiden van de bekwaamheid van de leerkracht. Schoolleiders verwachten dat een dergelijk meetinstrument prima gebruikt kan worden voor beoordelings- en ontwikkelgesprekken en als startpunt om gericht te gaan werken aan verbetering van leerkrachtvaardigheden.

Houding ten aanzien van prestatiebeloning

De houding van leraren ten aanzien van prestatiebeloning en de mate waarin men zich daardoor extra geprikkeld voelt blijkt behoorlijk gemengd. Meer ervaren leraren en leraren die meer risicomijdend zijn staan negatiever tegenover prestatiebeloning. Hetzelfde geldt voor leraren die de criteria te zwaar vinden en leraren die onbekend zijn met de hoogte van de te behalen beloning. De mate waarin men zich extra geprikkeld voelt door de prestatiebeloning hangt sterk samen met de hoogte van de te behalen prestatiebeloning. Leraren staan ongeveer even positief tegenover een

uitkering in de vorm van een individuele uitkering als in de vorm van een teambudget.

Houding ten opzichte van coaching

De houding van leerkrachten ten aanzien van het niveau van de coaches en het type uitgevoerde interventies is positief tot zeer positief. Zowel coaches als schoolleiders als leraren zijn het minst positief over de groepsbrede trainingsbijeenkomsten. Dit suggereert dat van een individuele maatwerkaanpak om leerkrachtvaardigheid te verbeteren meer kan worden verwacht dan van een groepsaanpak. De bijdrage van de coaching aan de geboekte resultaten van de pilot wordt door zowel leraren als schoolleiders relatief het hoogst ingeschat, gevolgd door het stellen van doelen voor groei in leerkrachtvaardigheid, en op enige afstand door de prestatiebeloning.

Lessen uit pilot

De pilot biedt geen harde inzichten in de effecten op de bekwaamheid van leraren en leerprestaties. Daarvoor zijn experimenten nodig met overtuigende controlegroepen. De pilot biedt wel enige lessen en aanbevelingen.

Evaluatiesysteem

- Evaluatiesystemen zoals in deze pilot gehanteerd bieden de mogelijkheid om leraren onmiddellijke en specifieke feedback te geven op hun kwaliteit van lesgeven. Ook kunnen ze nadere inzichten bieden aan schoolleiders over de bekwaamheid van hun leerkrachten, die meegenomen kan worden bij het HRM- en professionaliseringsbeleid.
- Een evaluatie op basis van meerdere observaties per leerkracht vergroot de betrouwbaarheid. Hetzelfde geldt voor uitbreiding met onaangekondigde observaties.
- Uniformiteit in het beoordelen is van belang voor de betrouwbaarheid. Dat vergt training en afstemming tussen observatoren.
- Het is van belang om voldoende onderscheidend vermogen te hebben bij een observatie-instrument of een classificatie van de bekwaamheid van leraren. Onderzoek naar de relatie tussen het meetinstrument en leerlingprestaties is aan te bevelen, alvorens een bepaald instrument breed op te schalen.
- De objectiviteit van de beoordelaar is belangrijk, vooral als er beloningen aan de beoordelingen gekoppeld worden of anderszins belangen aan de uitkomsten verbonden zijn. Externe en onafhankelijke observanten verdienen de voorkeur.
- Voldoende draagvlak onder leerkrachten voor het meetinstrument en het meten lijkt van belang. Dit vereist volgens betrokkenen bij het project het creëren van een gevoel van urgentie en eigenaarschap, duidelijke communicatie vooraf over het observatie-instrument en het meetproces, reductie van (mogelijke) angst onder leerkrachten, en nabesprekingen van uitkomsten vanuit de observator met de leerkracht.

Prestatiebeloning

- Of prestatiebeloning voor leraren werkt of niet, kunnen we niet concluderen op basis van deze pilot. Het betreft een specifieke vorm van prestatiebeloning, en een experimentele setting met controlegroepen ontbreekt.
- Het vergt een stevige voorbereiding en inspanning om een prestatiebeloningsysteem te ontwikkelen en goed uit te voeren. Het gaat dan onder andere om de keuze voor de prestatiecriteria, de strengheid van de criteria, de hoogte van de beloning, het type beloning (teambudget en/of individuele uitkering), en spelregels over bijzondere situaties die zich voor kunnen doen.
- Het is van belang dat de gekozen prestatiemaat niet gemakkelijk te manipuleren is door degenen die de beloning kunnen krijgen. Bij een evaluatiesysteem zoals in de pilot gehanteerd met een aangekondigde lesobservatie bestaat het risico dat leraren zich heel specifiek gaan voorbereiden op die ene les en daar precies laten zien wat er van hen verwacht wordt, terwijl ze dit dan niet elke les laten zien. Een combinatie met een onaangekondigde lesobservatie kan dit risico verkleinen.
- Bij teambeloning is het van belang om duidelijke afspraken te maken over hoe wordt omgegaan met uitval van leraren gedurende het traject, die om allerlei redenen kan plaatsvinden.
- Het is van belang dat uitkomstmaten als prestatiecriteria worden gekozen die een bewezen verband hebben met onderwijskwaliteit. Het blijkt uit deze pilot dat evaluaties van leraren op basis van de Amsterdamse Kijkwijzer positief verband houden met de prestaties van leerlingen op rekenen en taal.[2]
- Een absolute lat (zoals in deze pilot gehanteerd) heeft als nadeel dat je van tevoren niet weet of de lat op het juiste niveau gelegd wordt. Het risico bestaat dat de lat te laag of juist te hoog gelegd wordt. Vooral bij gebrek aan historische gegevens over de uitkomstcriteria is dat een risico. Een model waarbij de beste zoveel procent scholen of leraren de prestatiebeloning ontvangen heeft dit risico niet. Het voordeel van een absolute lat kan daarentegen zijn dat er een gewenste standaard gezet wordt, zodat (teams van) leraren van tevoren precies helder hebben hoe hoog ze moeten scoren.
- De hoogte van de beloning lijkt sterk uit te maken voor de mate van extra stimulans die de leraren voelen. Een lage beloning zal weinig extra stimulans opleveren.
- Duidelijke communicatie over de hoogte van te behalen beloning richting de leraren verhoogt de extra stimulans die uitgaat van prestatiebeloning.
- Het ervaren van een (te) zware lat vermindert de extra stimulans die uitgaat van prestatiebeloning.

[2] De vraag is of dat bij allerlei andere maten die soms worden voorgesteld om mee te nemen (zoals oudertevredenheid of leerlingtevredenheid) ook het geval is.

- Doelen die uitgaan van groei in plaats van (alleen) een absoluut te bereiken niveau kunnen ervoor zorgen dat meer scholen (leraren) zich geprikkeld voelen, zowel de zwakkere als de betere.

Coaching

- Een individuele maatwerkaanpak is volgens betrokkenen effectiever dan een groepsaanpak (met groepstrainingen of groepsbijeenkomsten) om leerkrachtvaardigheid van leraren te verbeteren.
- Voor duurzame groei in leerkrachtvaardigheid wordt door directeuren en coaches een langer coachingstraject zinvol geacht dan in deze pilot heeft plaatsgevonden. In deze pilot betrof het een gemiddelde tijdsinzet van 35 uur per leerkracht over een periode van een half jaar.
- Het lijkt erop - afgaande op de nulmetingen in deze pilot, maar ook uit eerder onderzoek van de Inspectie - dat een deel van de leraren onvoldoende is toegerust om hun werk goed te doen. De intensieve coaching grijpt hier direct op aan. Een belangrijke vraag is in hoeverre financiële prikkels en het stellen van concrete doelen daarnaast bij kunnen dragen aan de motivatie van leerkrachten om te werken aan verbetering van hun leerkrachtvaardigheden. Volgens zowel directeuren als leraren hebben beide elementen voor een niet onbelangrijk deel positief bijgedragen aan de behaalde resultaten van de pilot.

Vervolgstappen en evaluatie

- Er zijn enkele voorbeelden van interventies met meten, feedback geven en coachen van leerkrachtvaardigheden die effectief gebleken zijn (zie o.a. Taylor en Tyler, 2011 en Allen, 2011). Het verdient aanbeveling bij de ontwikkeling en implementatie van dergelijke interventies deze voorbeelden goed te bestuderen en te leren van ervaringen die men daarmee in de loop der tijd mee heeft opgedaan.
- Een experiment met aselecte toewijzing van meten, coachen en belonen van leerkrachtvaardigheid aan bepaalde scholen en niet aan andere scholen is nodig om de effecten van dergelijke interventies overtuigend vast te kunnen stellen. Het is daarbij van belang dat voldoende scholen deelnemen om de effecten vast te kunnen stellen.
- Aandacht voor het meten van de langere termijn effecten van dergelijke interventies is belangrijk. Dit vanwege de mogelijkheid van het uitdoven, of juist pas vertraagd optreden van effecten.
- In de loop der tijd zijn verschillende observatie-instrumenten van leerkrachtvaardigheden ontwikkeld. Het verdient aanbeveling deze instrumenten tegen elkaar af te zetten op een aantal aspecten. Het belangrijkste aspect is de mate waarin de score op het instrument samenhangt met leerlingprestaties. De observatie-instrumenten waarbij dit in sterke mate het geval is, verdienen de voorkeur.

1 Inleiding

De leraar is een sterk bepalende factor voor de kwaliteit van het onderwijs. Onderzoek in vooral het afgelopen decennium laat duidelijk zien dat er grote en persistente variatie is in leerwinst tussen leerlingen die aan verschillende leraren zijn toegewezen (zie bv. Rockoff, 2004; Rivkin e.a., 2005; Aaronson e.a., 2007, Kane & Staiger, 2008; Hanushek en Rivkin, 2010). De schattingen van het effect van een standaarddeviatie betere leraar liggen in de orde van grootte van 0.10 - 0.20 standaarddeviatie hogere prestaties van leerlingen.[3] Chetty e.a. (2011) laten zien dat kinderen die op jonge leeftijd leraren hebben gehad met een hogere toegevoegde waarde ('value added') later vaker aan vervolgonderwijs deelnemen na de middelbare school, meer verdienen en in betere buurten wonen. Het vervangen van een leraar in de onderste vijf procent van de verdeling door een gemiddelde leraar levert 250 duizend dollar meer inkomen op over de hele levensloop voor een gemiddelde klas.[4]

De leerkracht kan dus een groot verschil maken. Minder duidelijk is welke factoren nu echt bepalend zijn voor de kwaliteit van leraren. Veel studies zijn verricht naar het effect van klassieke kenmerken als het (initiële) opleidingsniveau en ervaring van de leerkracht. Verreweg de meeste studies vinden geen verband tussen het initiële opleidingsniveau van de leraar en de leerwinst van leerlingen (Hanushek en Rivkin, 2006). Bij ervaring wijzen de meeste studies erop dat ervaring veel toevoegt in de eerste twee a drie jaar van de loopbaan, maar dat de effecten van ervaring daarna afvlakken (Staiger en Rockoff, 2010).

In de zoektocht naar observeerbare factoren die wel verschillen in kwaliteit tussen leraren bloot kunnen leggen hebben enkele studies zich gericht op het ontdekken van verbanden tussen beoordelingen van bekwaamheid van leraren door schoolleiders (o.a. Armour, 1976; Jacob en Lefgren, 2008) en leerwinst van leerlingen. Deze studies vinden dat schoolleiders redelijk goed in staat zijn verschillen in kwaliteit tussen leraren te identificeren.[5]

Een andere serie studies kijkt naar de relatie tussen leerwinst van leerlingen en beoordelingen van allerlei competenties van leraren op basis van klassenobservaties

[3] De maatstaf van een standaarddeviatie wordt gewoonlijk gebruikt om het effect van interventies in het onderwijs uit te drukken. Ter indicatie, een standaarddeviatie hogere toetsscores komt overeen met het verschil in gemiddelde score op de CITO-eindtoets tussen leerlingen met een vmbo-TL/havo advies en leerlingen met een VWO-advies.

[4] Staiger en Rockoff (2010) komen door een ruwe berekening uit op een totale waarde van 330 tot 760 duizend dollar veroorzaakt door een 1 standaarddeviatie betere leerkracht. Ze hebben hierbij gekeken naar literatuur die het verband tussen toetsscores en later looninkomen onderzoekt (Murnane et al. 1995; Neal & Johnson, 1996). Ze nemen verder aan dat 1 leerkracht impact heeft op 20 leerlingen.

[5] Jacob en Lefgren (2008) laten zien dat schoolleiders vooral in staat zijn om de zwakke en de goede leerkrachten te identificeren, en wat minder goed in het identificeren van verschillen in het midden van de kwaliteitsverdeling.

(o.a. Milanowski e.a., 2004; Holtzapple, 2003a en b; Rockoff en Speroni, 2010; Kane e.a., 2011, Journal of Human Resources) en de leerwinst van hun leerlingen. Deze studies vinden significant positieve verbanden. Kane e.a. (2011) vinden bijvoorbeeld dat eenderde tot ruim 40 procent van de totale bijdrage van leraren aan de leerprestaties van hun leerlingen kan worden verklaard uit de beoordelingen vanuit een evaluatiesysteem. Dit is een belangrijke bevinding, want betere informatie over de kwaliteit van leraren kan gebruikt worden bij beslissingen over het geven van vaste contracten, promoties, en om gericht te kunnen investeren in de bekwaamheid van leerkrachten. Zelfs zwakke signalen van kwaliteit kunnen uiteindelijk verschillen tussen leraren identificeren die hoge baten kunnen hebben (Staiger & Rockoff, 2010).

Het kunnen onderscheiden van kwaliteitsverschillen tussen leraren is belangrijk, maar hoe kan de kwaliteit van leraren verbeterd worden? Er is enige evidentie over hoe diverse interventies gericht op leraren kunnen leiden tot betere leerprestaties. Zo zijn er aanwijzingen dat scholing/training van leraren positieve effecten kan opleveren (Angrist & Lavy, 2001; Jacob & Lefgren, 2004). Ook zijn positieve effecten gevonden van (intensieve) mentoring door getrainde mentoren van beginnende leraren (Rockoff, 2008; Bressoux e.a., 2009).[6] Recent onderzoek laat zien dat het herhaaldelijk meten van bekwaamheid van leerkrachten op basis van een gedetailleerd evaluatiesysteem in combinatie met gerichte feedback door getrainde experts een positieve invloed heeft op leerprestaties (Taylor en Tyler, 2011). Het gaat om 0.06 standaarddeviatie hogere scores in het jaar van deelname, en 0.11 standaarddeviatie hogere scores in de twee jaar na deelname. De betrokken leerkrachten waren leerkrachten in het midden van hun carrière. De kosten van de interventie bedroegen circa 7000 dollar per leerkracht. Taylor en Tyler schatten in dat de baten de kosten overtreffen. Onderzoek van Allen e.a. (2011) op basis van een gerandomiseerd experiment suggereert dat individuele coaching voor voortgezet onderwijs docenten op basis van een meetinstrument van bekwaamheid positieve effecten heeft op leerlingprestaties. Het gaat om een interventie die 3700 dollar per docent kost en circa 20 uur training vergt die zich uitstrekt over een periode van een jaar. De coaching richtte zich op het verbeteren van interacties tussen leerkrachten en leerlingen en beoogt de motivatie en inzet van de leerlingen te verhogen. Leerlingen van leerkrachten die gecoacht waren scoorden 0.22 standaarddeviatie hoger in het jaar na de coaching dan leerlingen van leerkrachten die niet gecoacht waren.

Prestatiebeloning van leraren is een andere interventie waar het afgelopen decennium in diverse landen ervaring mee is opgedaan. De evaluaties leveren een gemengd beeld op;[7] enkele studies laten positieve effecten zien van zowel teambeloning (Ladd, 1999; Lavy, 2002; Glewwe e.a., 2010) als individuele

[6] Zie Webbink e.a. (2009) voor een nadere discussie van de uitkomsten van deze literatuur.
[7] Zie Webbink e.a. (2009) voor een nadere bespreking van de literatuur over prestatiebeloning.

prestatiebeloning (Elberts e.a., 2002; Lavy, 2003; Atkinson e.a., 2004; Muralidharan en Sundararamen, 2008). Recentelijk is echter ook een aantal studies verschenen die geen effecten vinden (Goodman and Turner (nog te verschijnen); Fryer, 2011; Springer et al., 2010; Glazerman and Seifullah, 2010). Neal (2011) benadrukt dat de specifieke vormgeving van prestatiebeloning van belang is en geeft enkele aanbevelingen hiervoor.

Deze notitie beschrijft de resultaten van een pilot in het Nederlandse basisonderwijs die gericht is op het verbeteren van bekwaamheid van leraren en in het verlengde daarvan de prestaties van leerlingen. De pilot berust op drie pijlers, die alle drie raakvlakken hebben met eerder in het buitenland onderzochte en hierboven genoemde interventies: 1) meten van bekwaamheid van leerkrachten door middel van lesobservaties en een scoresysteem van verschillende competenties 2) coaching en training op maat gericht op verbetering van leerkrachtvaardigheden; en 3) prestatiebeloning.

Het onderzoek spitst zich toe op de volgende hoofdvragen:

1. Welke veranderingen in gedrag en vaardigheden zijn waargenomen bij deelnemende leerkrachten?
2. In hoeverre hangen de scores van leerkrachten op het gehanteerde meetinstrument van leerkrachtvaardigheden samen met de leerwinst die leerlingen boeken op taal en rekenen?
3. Hoe is de houding van de leerkrachten ten aanzien van de pilot, en dan voornamelijk ten aanzien van prestatiebeloning?
4. Welke meer algemene lessen kunnen we trekken uit deze pilot?

De pilot is uitgevoerd in het schooljaar 2011-12 bij 125 leerkrachten van zeven basisscholen in Amsterdam. De looptijd van de pilot was een half jaar. Leerkrachten zijn gemeten door getrainde observatoren voor aanvang van de pilot en direct na afloop van de pilot.

Het unieke aan de pilot is dat de combinatie van deze interventies bij ons weten nog nergens is uitgevoerd. Dit onderzoek legt de belangrijkste ervaringen met de pilot vast. Benadrukt dient te worden dat vanwege het ontbreken van een experimentele setting met controlegroepen die deze interventies niet hebben gekregen de uitkomsten niet causaal mogen worden geïnterpreteerd. Deze notitie kan dan ook geen uitsluitsel geven over de effecten van de pilot op leerprestaties. Dat zou in een vervolgexperiment moeten worden onderzocht. De pilot verschaft wel lessen en aanbevelingen waar rekening mee gehouden kan worden in dergelijke vervolgexperimenten. Een andere toegevoegde waarde van dit onderzoek is dat we in staat zijn een verband te leggen tussen evaluatiescores op een gedetailleerd meetinstrument van bekwaamheid en leerprestaties van leerlingen.

De opzet van deze notitie is als volgt. Paragraaf 2 gaat in op de opzet, resultaten en houding van de betrokken actoren ten aanzien van het meten, coachen en de prestatiebeloning. Paragraaf 3 bespreek de relatie tussen het meetinstrument en de leerprestaties van leerlingen. Paragraaf 4 trekt lessen uit de pilot, en paragraaf 5 sluit af met conclusies en discussie.

2 Meten, coachen en prestatiebeloning

De pilot kent drie pijlers. De eerste pijler is het meten van vaardigheden van leerkrachten op basis van een observatie-instrument waarbij getrainde observanten de leerkrachten scoren op pedagogische, didactische en organisatorische gedragsaspecten. De tweede pijler is het intensief coachen van leerkrachten door externe coaches, waarbij de nadruk ligt op een individuele maatwerkaanpak. De derde pijler is prestatiebeloning, waarbij teams van leraren in aanmerking komen voor een teambudget en een individuele uitkering als de teamleden een minimaal niveau en een minimale groei in de score op het observatie-instrument weten te behalen.

2.1 Aanleiding en opzet pilot

De pilot die we in deze notitie onderzoeken is uitgevoerd op zeven Amsterdamse basisscholen in het schooljaar 2011-12 bij 125 leerkrachten. De pilot is genaamd (H)erkennen van Ontwikkeling en Kwaliteit en was gericht op het verbeteren van de kwaliteit van leerkrachten van de deelnemende scholen.[8] Deze paragraaf beschrijft de drie pijlers van de aanpak. De pilot is gestoeld op de volgende drie pijlers:

1. Meten van vaardigheden van leerkrachten
2. Coachen en trainen van leerkrachten op maat
3. Prestatiebeloning

Meten van vaardigheden van leerkrachten

Voorafgaande aan de pilot (oktober 2011) en aan het einde (juni 2012) zijn deelnemende leerkrachten geobserveerd in de les en is hun bekwaamheid beoordeeld aan de hand van een observatie-instrument, de Amsterdamse Kijkwijzer.

[8] De pilot sluit aan bij twee losstaande trajecten. Het eerste traject is de Kwaliteitsaanpak Basisonderwijs Amsterdam (KBA) die wordt uitgevoerd door de gemeente Amsterdam in samenwerking met de schoolbesturen. Een van de pijlers daarin is de professionalisering van medewerkers. Binnen deze pijler biedt KBA schoolbesturen de mogelijkheid om te participeren in professionaliseringstrajecten, en ontwikkelt men beroepsstandaarden voor onder andere leerkrachten en schoolleiders. Het tweede traject waar de pilot bij aansluit is het plan van het vorige kabinet om te experimenteren met prestatiebeloning voor leraren in het onderwijs.

De Kijkwijzer is ontwikkeld door de KPC Groep in samenwerking met De Kwaliteitsaanpak Basisonderwijs Amsterdam (KBA) en de schoolbesturen verenigd in het Breed Bestuurlijk Overleg (BBO). In de Kijkwijzer zijn de competenties uit de beroepsstandaard (de zogenoemde SBL-competenties) en de belangrijkste onderdelen van het Inspectiekader vertaald naar concreet observeerbaar gedrag. De Kijkwijzer wordt binnen de KBA verbeteraanpak gebruikt om een beeld te vormen van de onderwijskwaliteit op scholen. De observaties in het kader van de Verbeteraanpak worden uitgevoerd door oud-Inspecteurs van de Inspectie van het Onderwijs.

De Kijkwijzer kent 75 verschillende gedragsaspecten die te maken hebben met opbrengstgericht werken. De gedragsaspecten meten vaardigheden binnen drie competentiegebieden: pedagogische competentie, didactisch-vakinhoudelijke competentie en organisatorische competentie. De Kijkwijzer kent 45 (basis) gedragsaspecten die een 'startbekwame' leerkracht zou moeten bezitten (opleiding afgerond, tot twee jaar ervaring), en 30 (over het algemeen complexere) gedragsaspecten die behoren bij een 'vakbekwame' leerkracht (opleiding afgerond, meer dan twee jaar werkervaring). Appendix D bevat een lijst met alle betreffende gedragsaspecten.

De observaties en metingen zijn verricht door speciaal daarvoor getrainde observanten, externe onderwijsexperts en schoolleiders. Bij de nulmetingen betrof het bij 5 van de 7 scholen schoolleiders en 2 van de 7 externe experts. Bij de eindmetingen waren nog bij 2 van de 7 scholen schoolleiders betrokken en bij de overige volledig externe experts. Daar waar schoolleiders betrokken waren bij de eindmeting betrof het schoolleiders van een andere school om te voorkomen dat de slager zijn eigen vlees zou keuren. Als extra waarborg voor de onafhankelijkheid van de metingen bij deze scholen is circa de helft van de observaties uitgevoerd door externe experts. De leerkrachten is gevraagd een rekenles voor te bereiden waarin ze alle 75 gedragsaspecten konden laten zien. Zowel de nul- als eindmetingen hebben aangekondigd plaatsgevonden.

Coachen en trainen van leerkrachten op maat

Op basis van de nulmetingen hebben gedurende een half schooljaar gerichte coaching en training interventies plaatsgevonden om de leerkrachtvaardigheden van de leerkrachten te verbeteren.

De schoolleiders speelden een belangrijke rol bij het opstellen van een plan van aanpak voor zijn of haar school voor de coachingstrajecten. De plannen van aanpak moesten worden geaccordeerd door het bestuur van de stichting waaronder de zeven scholen vallen en door de projectleider namens de gemeente, die de middelen voor de interventies ter beschikking stelde. Na de keuze van de competenties, waarin het schoolteam in het algemeen en de individuele leerkracht in het bijzonder zich wil ontwikkelen, hebben de directeuren een 'doelengesprek' gevoerd met hun

leerkrachten (oktober 2011). Tijdens dit gesprek, dat in veel gevallen ook diende als functioneringsgesprek, zijn de voor de pilot te behalen doelen voor de leerkracht concreet in kaart gebracht. Met behulp van de gegevens van de eindmeting voerden de directeuren een eindgesprek ten behoeve van de pilot. Afhankelijk van de behaalde resultaten zijn (in overleg met HRM) vervolgacties georganiseerd. Deze kunnen uiteenlopen van verdere ontwikkeling naar excellent leraarschap (bevordering naar schaal LB) tot, in het meest extreme geval, een outplacementtraject.

De interventies zijn in alle gevallen uitgevoerd door externe coaches. De gemiddelde ervaring van de coaches als coach bedroeg 13 jaar (variërend van 2 tot 35 jaar), waarvan 11 jaar lesgevend.

Gemiddeld geven leerkrachten aan dat ze 35 uur in de pilot gestoken hebben. Tweederde heeft 10 uur of meer in de pilot gestoken, eenderde meer dan 30 uur en 10 procent meer dan 50 uur. De ondernomen interventies besloegen lesobservaties, teambijeenkomsten, individuele voortgangsgesprekken, teamtrainingen en video-interactiebegeleiding. Het gemiddelde aantal contacten tussen coach en leerkracht bedroeg 6,4 (variërend van 2 tot 20), het gemiddelde aantal lesobservaties 3,8 (variërend van 0 tot 10), het gemiddelde aantal individuele voortgangsgesprekken 2,4 (0 tot 12), het gemiddelde aantal groepsbrede trainingen 1,6 (van 0 tot 10), aantal keren dat deelgenomen is aan video-interactiebegeleiding 1,2 (van 0 tot 10), en overige gesprekken 2,8 (0 tot 12).

De intensiteit/frequentie van de coaching verschilde per leerkracht afhankelijk van de mate waarin men begeleiding nodig achtte. De zwakkere leerkrachten hebben volgens de coaches gemiddeld bijna twee keer zo veel coaching/aandacht gekregen dan de betere leerkrachten.[9]

Prestatiebeloning

Teams van leraren konden in aanmerking komen voor prestatiebeloning. Deze prestatiebeloning omvatte een combinatie van teambudget en een individuele uitkering op de rekening van de leerkracht. De besteding van het teambudget is door het team te bepalen.

De maatstaf voor het behalen van de prestatiebeloning is een absolute norm. Anders gezegd, het is niet zo dat scholen met elkaar concurreren om de prestatiebeloning en alleen de beste zoveel procent scholen de prestatiebeloning kon binnen halen. Dit betekent dat bij het gegeven vooraf bepaalde totale budget de hoogte van de te behalen teambudgetten en individuele uitkeringen nog niet vast stond, maar af zou

[9] 44 procent van de inzet van de coaches is besteed aan de zwakkere leerkrachten, 34 procent aan de gemiddelde en 23 procent aan de betere leerkrachten (bron: enquête onder coaches)

hangen van het aantal teams dat aan de criteria zou voldoen. Er is wel gecommuniceerd naar de leerkrachten en schoolleiders dat de minimale individuele uitkering 1000 euro (bruto) zou bedragen.

De prestatiebeloning wordt uitgekeerd als minstens 85 procent van de deelnemende leerkrachten hun persoonlijke doelen hebben behaald. De gedachte hierachter is dat dan een leerling in ten minste 7 van de 8 jaar onderwijs krijgt van voldoende niveau. Het persoonlijke doel is tweeledig: de leerkrachten moeten minimaal 40 van de 45 zogenoemde startbekwame aspecten scoren (ofwel een minimaal basisniveau halen) en de leerkrachten moeten een bepaalde groei laten zien in aantal gescoorde aspecten ten opzichte van de nulmeting. Bij een score van 0 tot en met 45 is het doel een groei van 20 gedragsaspecten. Bij een score van 46 tot en met 60 is het doel een minimale score van 66 (een groei variërend van 5 tot 20 aspecten). Bij een score van 61 tot en met 65 is het doel een groei van 5 aspecten. Bij een score van 66 of meer is het doel een groei naar 70 gedragsaspecten (groei variërend van 0 tot 5).

De bij aanvang zwakker scorende leerkrachten moesten dus meer groei bewerkstelligen dan de bij aanvang beter scorende leerkrachten.

Alle leerkrachten die bij het bestuur in dienst zijn zouden in principe deelnemen met uitzondering van degenen die onder de uitsluitingscriteria vielen.[10] In totaal zijn 125 leerkrachten gestart met de pilot. Bij 106 daarvan heeft een eindmeting plaatsgevonden. De resterende 19 hebben de eindmeting gemist door uiteenlopende redenen, bijvoorbeeld zwangerschapsverlof, ontslag, vertrek of (langdurige) ziekte. Deze leerkrachten telden niet mee voor het bepalen of aan de criteria was voldaan, maar maakten ook geen aanspraak meer op de prestatiebeloning.

Opgemerkt dient te worden dat dit een specifieke vorm van prestatiebeloning betreft. Eerder in het buitenland beproefde prestatiebeloningssystemen hanteren vrijwel altijd een rechtstreekse koppeling aan (vooruitgang in) leerprestaties, eventueel aangevuld met andere criteria. In deze pilot is ervoor gekozen om de prestatiebeloning niet te koppelen aan leerlingprestaties, maar aan vaardigheidsmetingen.

De kosten van de maatwerkinterventies bedroegen gemiddeld 3000 euro per deelnemende leerkracht. Op 125 leerkrachten komt dat neer op 375 duizend euro. De kosten van de prestatiebeloning bedroegen in totaal 180 duizend euro.

[10] Het ging om personeelsleden bij wie bij aanvang of gedurende de pilot een juridische ontslagprocedure in gang is gezet, leerkrachten bij wie voor eind 2011 geen beginmeting heeft kunnen plaatsvinden, en leerkrachten die incidenteel voor de klas staan of minder dan 1 dag per week structureel in dezelfde groep lesgeeft.

2.2 Veranderingen in bekwaamheid en gedrag

De groei in gemeten leerkrachtvaardigheid in termen van de score op het observatie-instrument tussen begin en eind van de pilot is substantieel, van 45 naar 66 gedragsaspecten gemiddeld (op een maximumscore van 75). De waargenomen groei is het sterkst bij de bij aanvang zwakste leerkrachten en neemt af naarmate de score op de nulmeting hoger is. Dit kan te maken hebben met hogere groeidoelen en een grotere tijdsinzet van de coaches op de zwakkere leerkrachten. Een experiment met controlegroepen van leraren die de betreffende interventies niet krijgen is nodig om te onderzoeken in hoeverre het pakket interventies uit de pilot tot betere leerprestaties leidt. Er zijn twee nuanceringen te maken bij de forse gemeten groei. Ten eerste, coaches en directeuren hebben gemiddeld gesproken een wat negatiever beeld over de daadwerkelijk doorgemaakte groei dan het beeld dat volgt uit de metingen. Een combinatie met onaangekondigde lesobservaties zal naar verwachting een betrouwbaarder beeld opleveren. Ten tweede, zowel coaches als directeuren verwachten dat gemiddeld ongeveer de helft van de doorgemaakte groei na twee jaar weggeëbd zal zijn als geen verder onderhoud en investering wordt gepleegd in de bekwaamheid van betreffende leerkrachten. Mogelijk is een langer traject noodzakelijk voor (grotere) duurzame groei.

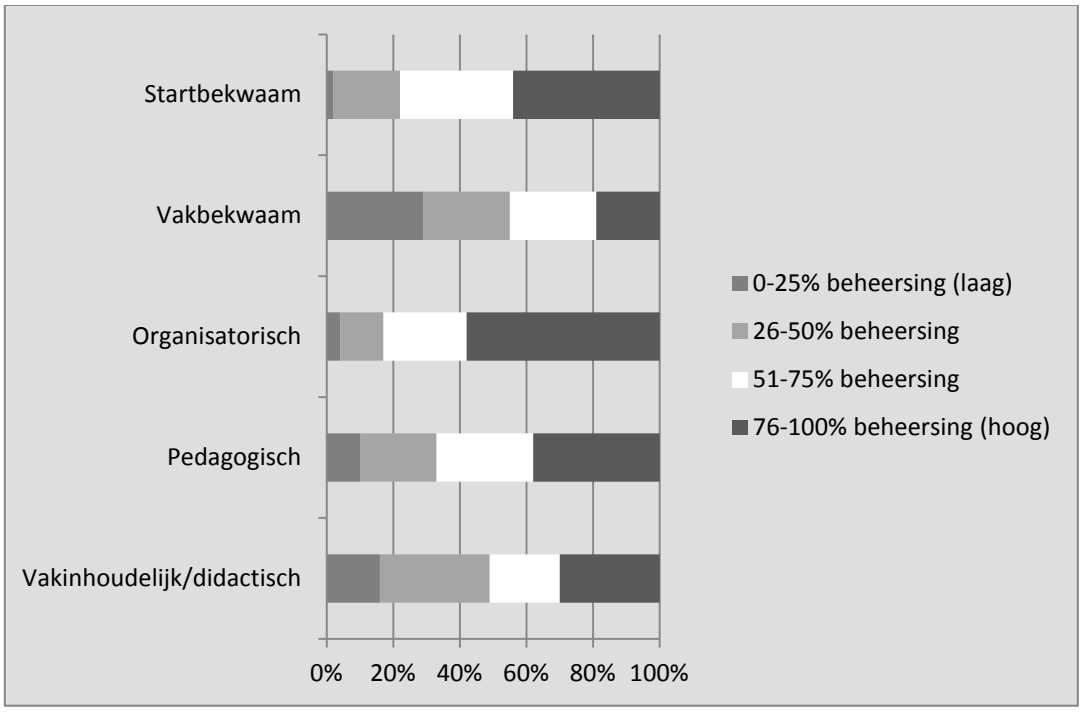
De verdeling van waargenomen leerkrachtvaardigheden voor de start van de pilot

Figuur 2.1 geeft een beeld van de verdeling van de waargenomen vaardigheden van de betrokken leerkrachten bij de pilot. Deze is gebaseerd op de beginmetingen, die voorafgaande aan de pilot hebben plaatsgevonden. De figuur laat zien dat leerkrachten voor aanvang van de pilot de meeste moeite hadden met vakinhoudelijke/didactische vaardigheden, en het minst met organisatorische vaardigheden. Bij de vakinhoudelijke/didactische vaardigheden liet ongeveer de helft van de leerkrachten minder dan de helft van de onderliggende gedragsaspecten zien, en ruim 1 op de 6 zelfs minder dan een kwart. Bij organisatorische bekwaamheid is het percentage dat driekwart of meer van de onderliggende aspecten laat zien met 58 procent bijna twee keer zo hoog als bij vakinhoudelijke/didactische bekwaamheid. Gemiddeld beheersten de leerkrachten 55 procent van de onderliggende vakinhoudelijke/didactische gedragsaspecten tegenover 73 procent van de organisatorische gedragsaspecten.

Een tweede bevinding uit figuur 2.1 is dat leraren de startbekwame gedragsaspecten beter beheersen dan de vakbekwame aspecten. Minder dan de helft van de vakbekwame aspecten werd beheerst voor aanvang van de pilot, tegenover ruim tweederde van de startbekwame aspecten. Minder dan een op de vijf leraren laat meer dan driekwart van de vakbekwame aspecten zien. Dit beeld komt overeen met het beeld van de Onderwijsinspectie (2011) die constateert dat leraren veel vaker de complexere vaardigheden niet beheersen dan dat ze de basisvaardigheden niet beheersen.

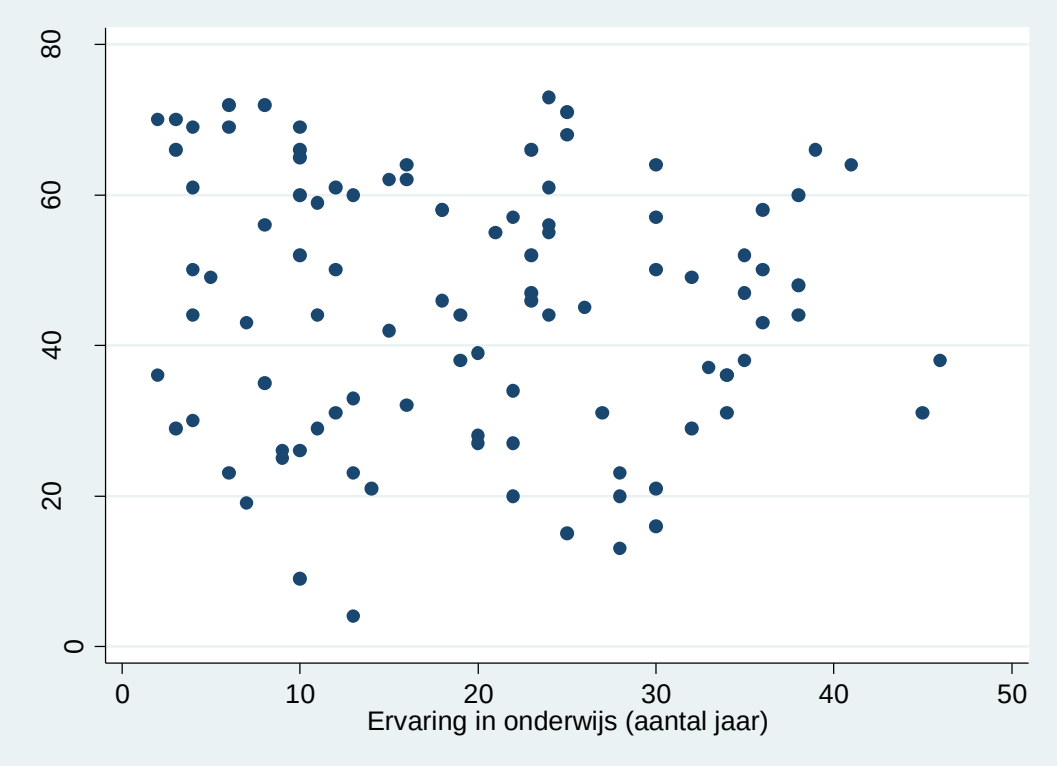
Appendix A laat zien hoe de leraren bij aanvang scoren op de 18 gedragsindicatoren van de Kijkwijzer. De gemiddelde mate van beheersing van deze indicatoren loopt behoorlijk uiteen: van 30 tot 80 procent van de onderliggende gedragsaspecten. Leraren scoorden bij aanvang het slechtste op “Laat leerlingen reflecteren op (diverse) oplossingsstrategieën” en het beste op “Besteedt de tijd daadwerkelijk aan het geplande lesdoel”.

Figuur 2.1 Verdeling van leerkrachtvaardigheden voorafgaande aan pilot



Figuur 2.2 laat zien dat er geen systematisch verband lijkt te bestaan tussen ervaring van leerkrachten en de score op de Kijkwijzer bij de nulmeting. Dit sluit aan bij eerder genoemde bevindingen in de internationale literatuur.

Figuur 2.2 Relatie score op meetinstrument voor aanvang pilot en ervaring leerkracht



Resultaten pilot op hoofdlijnen

Vijf scholen hebben aan de criteria voldaan voor prestatiebeloning en twee niet. De twee scholen die de prestatiebeloning niet behaald hebben zijn gestruikeld op basis van het doel ten aanzien van startbekwaamheid. Bij aanvang van de pilot beheerste geen enkele deelnemende leerkracht op deze twee scholen minimaal 40 van de 45 startbekwame aspecten, aan het eind was dit opgelopen tot de helft. Dit is een duidelijk verschil met de scholen die de doelen wel gerealiseerd hebben, waar 91 procent van de leerkrachten het minimaal te behalen niveau op de startbekwame gedragsaspecten liet zien.

Gemiddeld over alle deelnemende scholen bedraagt de waargenomen groei 21 gedragsaspecten (van 45 naar 66). De standaarddeviatie op de nulmeting bedroeg 17 gedragsaspecten. De gemiddelde waargenomen groei in leerkrachtvaardigheid van

21 gedragsaspecten komt derhalve overeen met een toename van 1.2 standaarddeviatie.[11]

Tabel 2.1 Resultaten pilot Herkennen en erkennen van ontwikkeling en kwaliteit

Variabele	Alle deelnemende scholen	Scholen prestatiebeloning behaald	Scholen prestatiebeloning niet behaald
Aantal scholen	7	5	2
Aantal deelnemende leerkrachten	106	80	26
Gemiddeld aantal gedragaspecten nulmeting (0-75)	45	48	35
Gemiddeld aantal gedragsaspecten eindmeting (0-75)	66	68	60
Gemiddelde groei in aantal gedragsaspecten	21	20	25
% leerkrachten met minstens 40 startbekwame aspecten nulmeting	25	34	0
% leerkrachten met minstens 40 startbekwame aspecten eindmeting	82	91	50

De beloning bedraagt 1000 euro (bruto) voor alle leerkrachten die hun individuele doelen hadden behaald en waar het teamdoel voor de school ook behaald is. Dit waren er 72. De teambudgetten voor de vijf scholen die aan het teamdoel hebben voldaan bedraagt 8000 euro per school. Twaalf leerkrachten van de twee scholen die het teamdoel niet behaald hebben, hebben wel hun individuele doelen behaald. Zij krijgen een geringere beloning van 350 euro. Ook directeuren (750 euro) en onderwijsassistenten (350 euro) van de scholen het teamdoel behaald hadden kregen een beloning.

Als we wat verder inzoomen op de behaalde resultaten, dan vallen de volgende zaken op. We beginnen met een uitsplitsing van belangrijkste bevindingen naar type vaardigheden, gevolgd door een korte analyse van samenhangende leraarkenmerken factoren met behaalde groei. Daarna bespreken we een tweetal nuanceringen bij de waargenomen groei.

Waargenomen groei naar type vaardigheden

De groei is zichtbaar bij alle typen vaardigheden en significant. Er bestaan wel verschillen in de mate van waargenomen groei (zie tabel 2.2). Zo is de groei groter bij de vakbekwame gedragsaspecten dan bij de startbekwame gedragsaspecten. Naar de drie typen gedragsaspecten bezien is de waargenomen groei het laagst bij de

[11] Ter vergelijking, de gemiddelde toename op een observatie-instrument gehanteerd in de Verenigde Staten (basisscholen in Cincinatti) bedroeg tweederde standaarddeviatie tussen twee opeenvolgende metingen, waarbij het meest voorkomende verschil in tijdsbestek tussen beide metingen drie jaar bedroeg (Kane e.a., 2011). Bij metingen binnen hetzelfde schooljaar liep de toename uiteen van eenderde tot tweederde standaarddeviatie, afhankelijk van de ervaring van de leerkracht. Voor alle duidelijkheid, bij deze leraren heeft geen intensieve coaching en prestatiebeloning plaats gevonden, zoals in de onderhavige pilot wel het geval was.

organisatorische gedragsaspecten. Bij de pedagogische bekwaamheid en de vakinhoudelijke/didactische bekwaamheid is de waargenomen groei in procenten van het totale te behalen aantal gedragsaspecten ruim twee keer zo hoog.

Tabel 2.2 Beheersing naar type vaardigheden, nulmeting, eindmeting en ontwikkeling

Type vaardigheid	Beginmeting (% van gedragsaspecten)	Eindmeting (% van gedragsaspecten)	Toename (% van gedragsaspecten)
Vakbekwaam	47 (27)	82 (20)	35*** (22)
Startbekwaam	69 (20)	91 (13)	23*** (17)
Pedagogisch	63 (29)	93 (14)	30*** (25)
Vakinhoudelijk/didactisch	55 (25)	86 (17)	31*** (21)
Organisatorisch	73 (20)	87 (17)	14*** (17)
Totaalscore	60 (22)	87 (15)	28***(18)

Standaarddeviaties gerapporteerd tussen haakjes. *** = toename is significant op 1 procent significantieniveau.

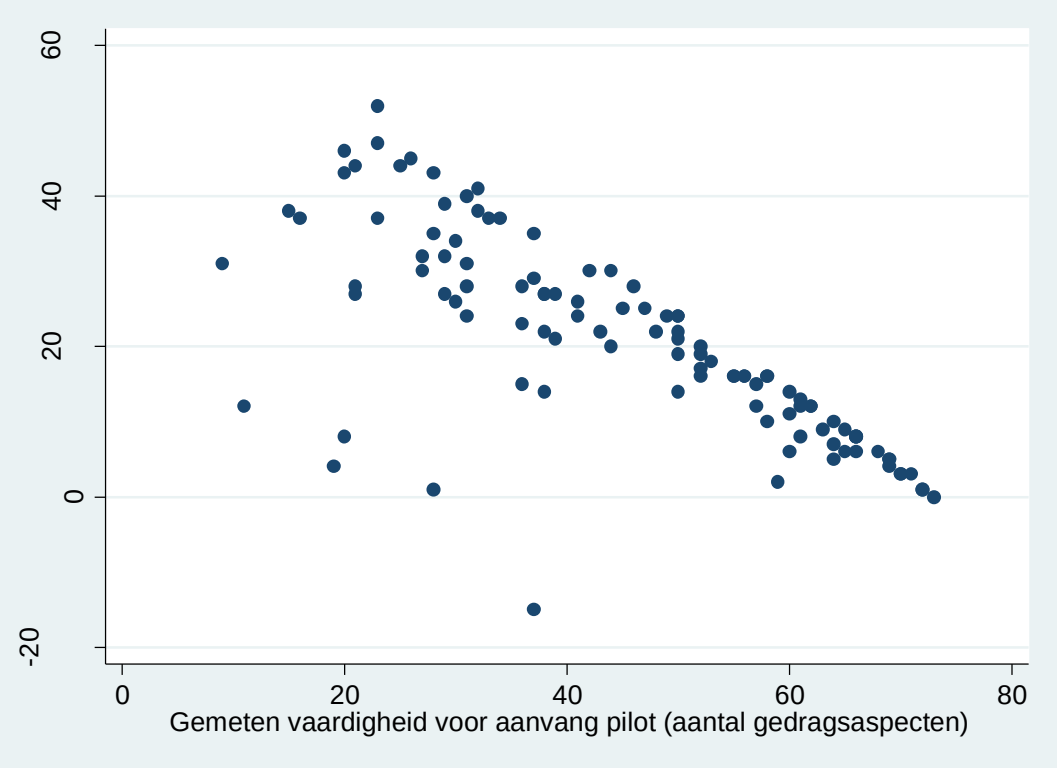
De gedragsindicatoren met de grootste winst op de Kijkwijzer tussen begin- en eindmeting zijn: “Gaat na of de lesdoelen bereikt zijn” (+40 procentpunt), “Verduidelijkt bij aanvang van de les de lesdoelen” (+37 procentpunt), “Maakt voor leerlingen de opbouw van de les inzichtelijk” (+34 procentpunt), en “Bevordert het toepassen van het geleerde” (+33 procentpunt), “Stemt de verwerking van de lesstof af op relevante verschillen tussen leerlingen” (+33 procentpunt stijging), en “Stemt instructie af op relevante verschillen tussen leerlingen” (+30 procentpunt).[12]

Samenhangende factoren met waargenomen groei in bekwaamheid

De waargenomen winst in gedragsaspecten hangt het sterkst samen met het startniveau: hoe lager het startniveau, des te hoger de waargenomen groei. Dat is duidelijk te zien in onderstaande figuur. Opvallend is dat Taylor en Tyler (2011) vinden dat de effecten van de introductie van evaluatie en feedback op leerprestaties groter zijn bij zwakkere leraren en afnemen naarmate het gemeten startniveau van bekwaamheid hoger ligt. De resultaten in deze pilot laten hetzelfde beeld zien voor wat betreft de groei in gemeten leerkrachtvaardigheid.

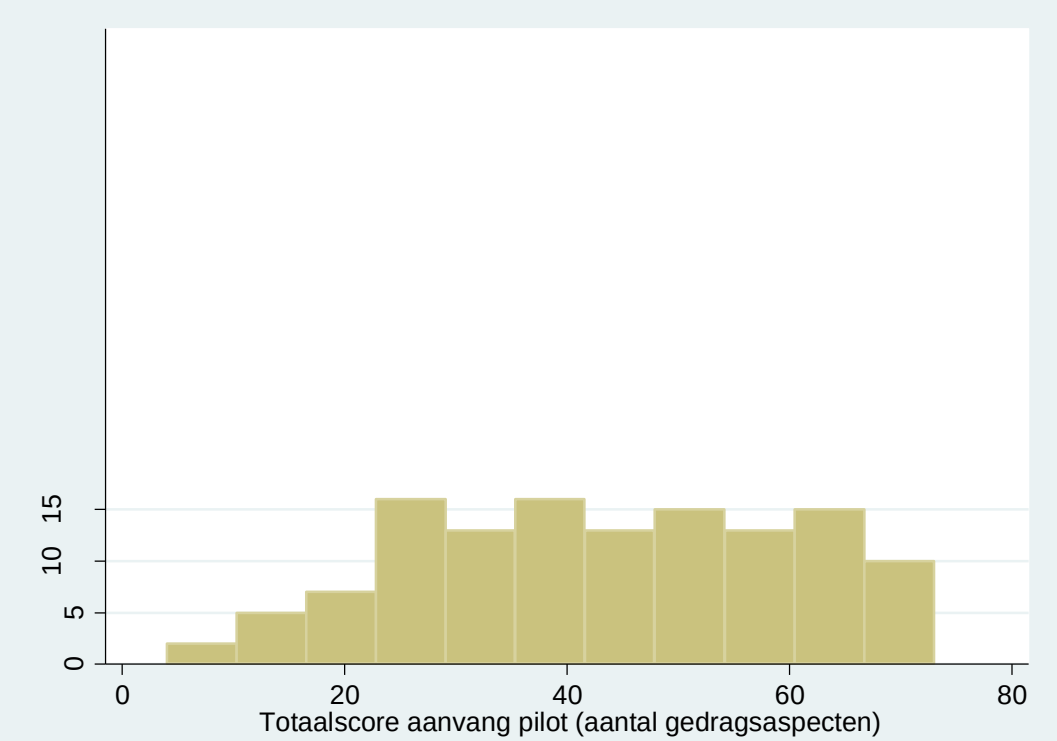
[12] Deze laatste twee vaardigheden vallen bij de indeling die de Inspectie hanteert onder de complexere vaardigheden. Op basis van ruim 1700 door de Inspectie bekeken lessen in het basisonderwijs in 2011 stemt 60 procent van de leerkrachten de instructie, en 75 procent de verwerkingsopdrachten, af op verschillen tussen leerlingen. Deze vaardigheden behoren daarmee tot de vaardigheden waar leraren in het basisonderwijs het slechtste op scoren. Ter vergelijking, de percentages leraren die de verschillende *basis*vaardigheden beheersen lopen uiteen van 91 tot 97 procent (zie Inspectie van het Onderwijs, 2011).

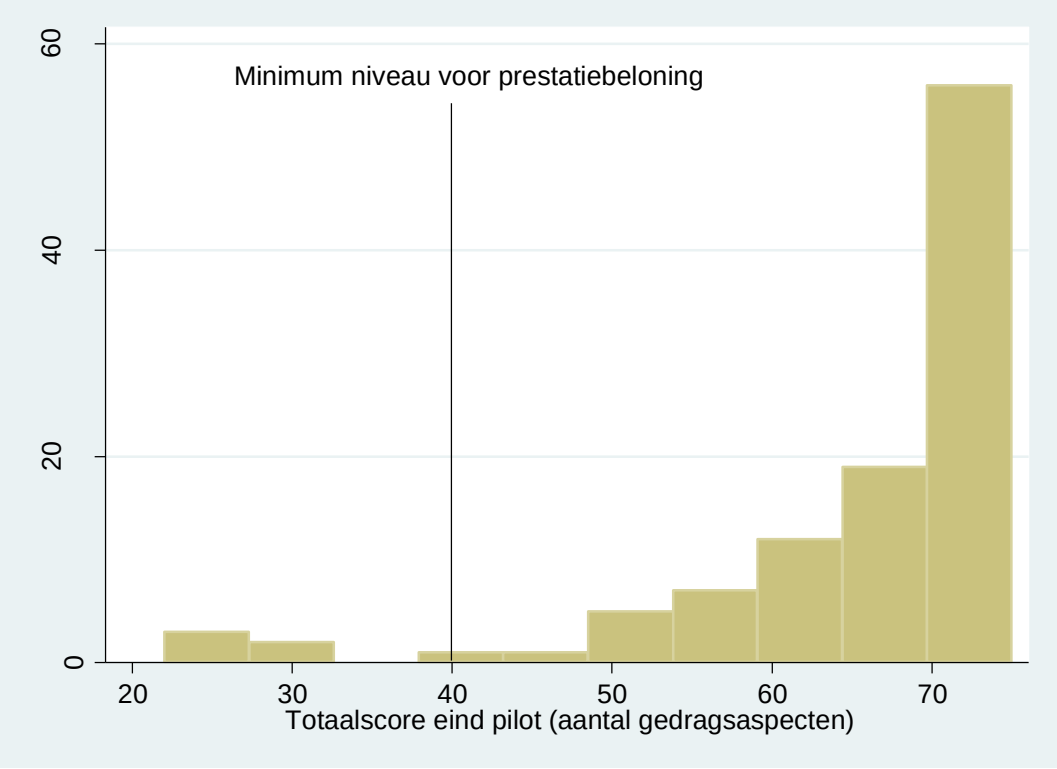
Figuur 2.3 Leraren met de laagste score aan het begin laten de grootste progressie zien



Dit patroon betekent dat de spreiding in scores op de eindmeting lager is geworden dan die op de nulmeting. De standaarddeviatie van de scores op de eindmeting ligt ongeveer een derde lager dan die op de nulmeting. Onderstaande figuur laat ook zien dat veel leraren aan het eind van de pilot zijn opgeschoven van de linkerkant naar de rechterkant van de gemeten bekwaamheidsverdeling.

Figuur 2.4 Verdeling vaardigheidsscores nulmeting (boven) en eindmeting (onder)





Een voor de hand liggende verklaring voor de sterkere groei onder de zwakkere leerkrachten is dat de coaches meer tijd en aandacht hebben besteed aan deze groep. Een andere mogelijke verklaring is dat er voor de zwakkere leerkrachten meer gedragsaspecten waren die nog relatief eenvoudig konden worden bijgeleerd (laaghangend fruit), terwijl de beste leerkrachten met name de complexere vaardigheden nog niet lieten zien die moeilijker zijn aan te leren. Ook kan hebben meegespeeld dat de gestelde doelen qua groei hoger waren naarmate men lager scoorde op de nulmeting. Dit kan een grotere prikkel hebben gegeven aan de zwakkere leerkrachten om meer te groeien. Een laatste factor is dat er zeker bij het hoogste kwartiel (61-75 gedragsaspecten) natuurlijke grenzen aan de nog te behalen groei zitten, gegeven de maximaal te behalen score van 75 gedragsaspecten.

Conditioneel op het startniveau lijkt de waargenomen groei in aantal gedragsaspecten iets sterker naarmate de leerkracht aangeeft zich sterker geprikkeld te voelen door de te behalen individuele beloning, naarmate de leerkracht hoger scoort op het persoonlijkheidskenmerk zorgvuldigheid en naarmate de leerkracht minder extravert is.[13] Deze relaties zijn echter minder sterk en minder significant dan het verband van het startniveau met de waargenomen groei.[14] Ervaring, een route via het MBO naar de lerarenopleiding, en de overige drie gemeten persoonlijkheidskenmerken van de zogenoemde 'big five' lijken geen verband te houden met de gerealiseerde groei. Tabel 2.2 toont de geschatte coëfficiënten.

Tabel 2.3 Relatie behaalde winst in leerkrachtvaardigheid met kenmerken leerkracht

Leraarkenmerken	Geschatte Coëfficiënt	P-waarde
Score nulmeting	**-0,65*** **	0,00
Aantal jaar ervaring in onderwijs (1-46)	0,04	0,54
MBO-route naar lerarenopleiding	-1,50	0,46
Sterk (of zeer sterk) extra gestimuleerd door prestatiebeloning (0 of 1)	3,10	0,11
Extraversie (1-5)	**-0,83***	0,08
Zorgvuldigheid (1-5)	**1,15***	0,10
Vriendelijkheid/inschikkelijkheid (1-5)	0,48	0,23
Emotionele stabiliteit (1-5)	-0,48	0,36
Openheid voor ervaring / ideeën (1-5)	0,02	0,98
Aantal waarnemingen	82	

* / ** / *** betekent effectschatting significant op 10 / 5 / 1 procent significantieniveau. De p-waarden zijn bepaald op basis van voor clustering op schoolniveau gecorrigeerde standaardfouten.

[13] Er zijn aanwijzingen dat het effect van feedback afhankelijk is van factoren buiten de feedback, zoals persoonlijkheidskenmerken (Ilgen, Fisher & Taylor, 1979).

[14] Ter indicatie, het effect van 1 standaarddeviatie lagere score op de nulmeting op de behaalde groei is een 11 aspecten hogere groei, terwijl het geschatte effect van 1 standaarddeviatie lagere score op extraversie anderhalf aspect meer groei is en het geschatte effect van een 1 standaarddeviatie hogere score op nauwkeurigheid twee gedragsaspecten meer groei. Het zich (zeer) sterk extra gestimuleerd voelen door de prestatiebeloning houdt volgens de puntschatting verband met drie gedragsaspecten hogere groei.

Veranderingen in (zelfgerapporteerd) gedrag

We hebben leerkrachten ook gevraagd naar veranderingen in gedrag na de pilot ten opzichte van het jaar voor de pilot. De resultaten staan vermeld in onderstaande tabel, telkens gerangschikt van hoog naar laag naar de mate van verandering ten opzichte van het jaar voor de pilot.

Op het gebied van interactie met collega's (zie panel C) hebben veranderingen zich met name voorgedaan op het vlak van het bespreken van de planning van de lessen, het analyseren van resultaten van de eigen leerlingen en het advies vragen en geven aan collega's ter verbetering van de eigen leerkrachtvaardigheid respectievelijk die van een collega. Dit laatste kan te maken hebben met de prestatiebeloning, waarbij de vooruitgang van de bekwaamheid van het team maatgevend was voor het behalen van de teambeloning.

Het observeren van lessen van een collega of het laten observeren van eigen lessen door een collega vindt wel iets vaker plaats, maar komt in zijn algemeenheid nog relatief weinig voor. Veertig procent van de leraren geeft aan dit nooit te doen. Opvallend is dat vrijwel alle leraren (93%) het eens zijn met de stelling dat leerkrachten veel meer van elkaar zouden kunnen leren door vaker bij elkaar in de klas te gaan kijken. Ook alle schoolleiders zijn het eens met deze stelling, en 91 procent van de coaches. Mentoring of coaching geven aan of ontvangen van een collega komt relatief het minste voor en daar heeft de pilot ook weinig verandering in gebracht.

Qua focus van instructie (zie panel A) valt op dat leraren zich vooral meer zijn gaan concentreren op de leerlingen die net niet voldoende presteren en de beste leerlingen.[15] Voor wat betreft het gebruik van toetsresultaten (zie panel B) rapporteren leraren de grootste gedragsverandering bij het gebruik hiervan om instructies af te stemmen op individuele behoeften van leerlingen. Beide bovengenoemde gedragsveranderingen sluiten aan bij de eerdere observatie dat leraren bij de metingen relatief veel winst hebben geboekt op de twee gedragsindicatoren die gaan over het afstemmen van instructie en verwerking van lesstof op relevante verschillen tussen leerlingen.

[15] De focus op de beste leerlingen blijft relatief het laagst van alle typen leerlingen. Ongeveer een derde geeft aan zich na de pilot bijna altijd of altijd te focussen op de beste leerlingen. Dit is anderhalf keer zo weinig als op de gemiddelde leerlingen en bijna twee keer zo weinig als op de leerlingen die net niet voldoende presteren.

Tabel 2.4 Zelf gerapporteerde gedragsveranderingen leerkrachten na pilot

	Meer of veel meer dan jaar voor pilot (% respondenten)	Bijna altijd of altijd (% respondenten)
A. Focus instructie		
- Ik concentreer me op leerlingen die net niet voldoende presteren	58	59
- Ik concentreer me op de beste leerlingen	55	32
- Ik concentreer me op de zwakste leerlingen	45	53
- Ik concentreer me op de gemiddelde leerlingen	36	49
B. Gebruik van toetsresultaten		
- Instructies afstemmen op individuele behoeftes van leerlingen	53	56
- Identificatie van individuele leerlingen die speciale hulp nodig hebben	46	56
- Opstellen van leerdoelen voor individuele leerlingen	45	52
C. Activiteiten samen met collega's		Nooit / meer dan eens per maand
- Bespreken planning van lessen	42	1 / 90
- Analyseren van resultaten van mijn leerlingen	37	3 / 54
- Advies gevraagd aan collega('s) om mijn leerkrachtvaardigheid te verbeteren	34	12 / 51
- Advies gegeven aan collega('s) om hun leerkrachtvaardigheden te verbeteren	31	19 / 48
- Laten observeren van een of meerdere van mijn lessen	20	39 / 18
- Observeren een of meerdere lessen van collega's	19	40 / 14
- Mentor of coach voor een of meerdere van mijn collega's geweest	9	70 / 11
- Mentoring of coaching gekregen van een of meerdere van collega's	7	61 / 16

Betrouwbaarheid en persistentie van waargenomen groei in leerkrachtvaardigheden

De metingen suggereren een substantiële groei in leerkrachtvaardigheid tussen begin en eind van de pilot. Omdat de pilot geen experiment betrof met zorgvuldig gekozen controlegroepen die niet zijn blootgesteld aan deze interventies kunnen we de gemeten groei niet als het causale effect van de pilot interpreteren. Anders gezegd, we weten niet wat de gerealiseerde groei in leerkrachtvaardigheid zou zijn geweest in de afwezigheid van de pilot.

De vraag is in hoeverre de gemeten groei een betrouwbaar beeld geeft van de daadwerkelijk doorgemaakte groei in leerkrachtvaardigheid en in hoeverre deze groei veroorzaakt is door de pilot. Een deel van de groei kan komen door het natuurlijke effect van meer ervaring, een deel kan ook komen doordat leraren door de nulmeting meer bekend zijn geraakt met de Kijkwijzer en het gedrag dat daarin wordt gescoord. Dit toegenomen bewustzijn kan leiden tot een toename in de scores doordat leraren hebben geïnvesteerd in het leren en permanent adopteren van gedragsaspecten uit de Kijkwijzer, of doordat leraren strategisch de gedragsaspecten laten zien op het moment dat ze geobserveerd worden (zie Kane e.a., 2010). Het in het vooruitzicht stellen van een beloning kan dit laatste effect versterkt hebben. De data kunnen hierover geen uitsluitsel geven.

De visie van coaches en directeuren op de behaalde resultaten leiden tot een tweetal nuanceringen bij de waargenomen groei in leerkrachtvaardigheden.

Ten eerste, zowel het eigen beeld van coaches als dat van directeuren over de doorgemaakte groei is gemiddeld minder positief dan de gemeten groei in leerkrachtvaardigheid (zie tabel 2.4). Met andere woorden, de gemeten groei zou wel eens een iets te rooskleurig beeld kunnen geven van de daadwerkelijk doorgemaakte groei. Bij de coaches geeft ruim 60 procent aan dat de gemeten groei groter is dan het eigen beeld van de doorgemaakte groei, bij de directeuren gaat het om dertig procent. Geen enkele coach of directeur is van mening dat de gemeten groei een te negatief beeld geeft van de doorgemaakte groei. Opvallend is dat het eigen beeld van leerkrachten over de doorgemaakte groei bij twintig procent van de leerkrachten positiever is dan de gemeten groei, en dat slechts drie procent aangeeft dat de gemeten groei hoger is dan het eigen beeld over de doorgemaakte groei.

Tabel 2.5 Eigen beeld van leraren, coaches en directeuren ten opzichte van gemeten bekwaamheid bij eindmeting

Eigen beeld ten opzichte van gemeten score	Leraren	Directeuren	Coaches
Lager	3	29	63
Gelijk	77	71	37
Hoger	20	0	0

De verwachting is dat een combinatie van aangekondigde met onaangekondigde lesobservaties een betrouwbaarder of completer beeld zou geven van de doorgemaakte groei. Meer dan driekwart van de coaches en meer dan 70 procent van de schoolleiders is het eens met de stelling dat een combinatie van aangekondigde en onaangekondigde lesobservaties gedurende een jaar tot een completer en betrouwbaarder beeld zou leiden. De helft van de leraren is het eens met de stelling dat onaangekondigde observaties een realistischer beeld zouden geven, tegenover 18 procent oneens. Een combinatie van aangekondigde en onaangekondigde lesbezoeken wordt ook gehanteerd bij een van de meest onderzochte evaluatiesystemen in de Verenigde Staten (zie Taylor & Tyler, 2011). Een argument voor meerdere lesbezoeken per leerkracht is dat Kane en Staiger (2012) laten zien dat de betrouwbaarheid van evaluaties van leerkrachten significant toeneemt met het aantal lesobservaties per leerkracht.

Een tweede nuancering bij de waargenomen groei in leerkrachtvaardigheid is dat zowel coaches als directeuren verwachten dat er sprake zal zijn van een wegebeffect. Vrijwel alle coaches verwachten dat na twee jaar nog minder dan de helft van de gerealiseerde groei zichtbaar is als geen verdere investeringen worden gepleegd.[16]

[16] 80 procent hiervan denkt dat tussen een kwart en de helft van de groei nog zichtbaar is, en 20 procent hiervan denkt dat minder dan een kwart nog zichtbaar is, als er geen vervolg wordt gegeven aan de pilot.

Directeuren zijn iets positiever hierover en verwachten dat na twee jaar gemiddeld nog zo'n 60 procent van de waargenomen groei zichtbaar is.

De verwachting van het wegebben van effecten kan te maken hebben met de duur van de coachingstrajecten (een half jaar). Zowel directeuren (85%) als coaches (80%) geven in overgrote meerderheid aan deze duur kort of te kort te vinden. Opvallend is dat het overgrote deel van de leraren (80 procent) aangeeft de duur van de coaching precies goed te vinden. Slechts 11 procent van hen geeft aan deze duur kort of te kort te vinden.

Dit suggereert dat schoolleiders en coaches in veel grotere mate dan leraren zelf vinden dat er meer moet gebeuren om de bekwaamheid blijvend op een hoger niveau te tillen. Deze suggestie wordt verder ondersteund door de eerdere bevinding uit tabel 2.5 dat directeuren en coaches vaker een negatiever beeld hebben dan het gemeten beeld (en nooit een positiever beeld), terwijl leraren vaker een positiever beeld hebben dan het gemeten beeld (en vrijwel nooit een negatiever beeld).

3 Relatie observatiescore en leerprestaties

Het meten van leerkrachtvaardigheden met een gedetailleerd observatie-instrument en getrainde observatoren maakt verschillen in de kwaliteit van leraren zichtbaar. De score op het gehanteerde observatie-instrument blijkt significant samen te hangen met de leerwinst die leerlingen boeken. Als een leerkracht uit de 25 procent slechtste leerkrachten op het observatie-instrument wordt vervangen door een leerkracht uit de 25 procent leerkrachten met de hoogste observatiescore gaan de leerlingen er gemiddeld 0,37 standaarddeviatie meer op vooruit bij rekenen, 0,44 standaarddeviatie bij spelling en 0,24 standaarddeviaties bij lezen. Deze verschillen zijn niet gering. Door twee jaar lang een (gemeten) slechte leraar te hebben in plaats van een goede leraar kan een basisschool leerling een heel niveau lager uitkomen in het vervolgonderwijs, dus bijvoorbeeld van een in potentie vwo-niveau op havo-niveau.

In deze paragraaf kijken we naar het verband tussen de score op de Kijkwijzer en de leerprestaties van de leerlingen. Laat een leerkracht die goed scoort op de Kijkwijzer zijn of haar leerlingen ook beter presteren op belangrijke domeinen als rekenen, spelling en lezen? Dit is een belangrijke vraag, want het zegt iets over in hoeverre de gedragsaspecten die gemeten worden ertoe doen voor leerlingprestaties. Als de score op het instrument voorspellend is voor leerlingprestaties biedt dit waardevolle informatie, aangezien andere observeerbare kenmerken als vooropleiding en ervaring volgens de meeste onderzoeken niet of nauwelijks voorspellend zijn voor de kwaliteit van leraren (gemeten door de prestaties van leerlingen).

Om het verband tussen de score op het observatie-instrument en leerlingprestaties te onderzoeken maken we gebruik van regressieanalyse, waarbij we controleren voor

belangrijke verschillen in leerling- en klassenkenmerken tussen klassen. Deze techniek zorgt ervoor dat de uitkomsten niet vertekend worden. Als we hier niet voor zouden controleren zouden we onterecht verschillen in leerling- en klassenkenmerken kunnen toewijzen aan de leerkracht. Door hier wel voor te controleren is een goede vergelijking tussen de leerprestaties van de kinderen in verschillende klassen met verschillende leraren mogelijk.

Tabel 3.1 Het verband tussen de Kijkwijzer en de leerprestaties van leerlingen

	Afhankelijke variabele, score op:		
Onafhankelijke variabele:	rekenen	spelling	lezen
Kijkwijzer (gemiddelde nul- en eindmeting)	0,154***	0,178***	0,107**
	(0,051)	(0,046)	(0,050)
Observaties	2084	2110	2135
Aantal klassen	99	99	99
Controles	ja	ja	ja
R-kwadraat	0,449	0,375	0,398
***p<0,01, **p<0,05, *p<0,1			

De geschatte coëfficiënten geven het verband weer van een 1 punt hogere score op Kijkwijzer met de toetsscore van leerlingen op het betreffende domein, uitgedrukt in standaarddeviaties. De score op de Kijkwijzer kan uiteenlopen van 0 tot 75 aspecten.

Tabel 3.1 presenteert de resultaten van de regressieanalyse. Zij presenteert het verband tussen de totaalscore op de Kijkwijzer van de leerkracht en de toetscore van de leerlingen aan het einde van het schooljaar 2011/2012, voor de vakken rekenen, spelling en lezen. We controleren voor verschillen in aanvangsniveau van leerlingen in klassen, door het opnemen van toetscores aan het eind van het vorige schooljaar (schooljaar 2010/2011), zittenblijven, nationaliteit, afkomstig uit een ouder gezin, leerling-gewicht[17], leeftijd en geslacht. Naast individuele kenmerken controleren we ook voor verschillen in klassenkenmerken: het gemiddelde aanvangsniveau, de klassengrootte, het percentage meiden in de klas, het percentage gewichtenleerlingen, het percentage kinderen uit een eenoudergezin, of de klas een combinatieklas is en het aantal leerkrachten dat op de klas staat. In het geval van meerdere leraren op een klas wegen we de Kijkwijzer in met het aantal dagen dat de leraren op de klas staan. Verder gebruiken we het gemiddelde van de begin- en eindmeting op de Kijkwijzer in plaats van alleen 1 van de metingen, omdat meerdere observaties een betrouwbaarder beeld geven van de kwaliteit van de leerkracht (zie Kane en Staiger, 2012). Het model dat we schatten is gelijk aan een veelgebruikt model dat in de buitenlandse literatuur wordt gehanteerd, zoals in o.a. Kane e.a. (2011), zie Appendix B.

[17] Leerling-gewicht wordt bepaald op basis van het opleidingsniveau van de ouders. Er bestaan drie categorieën, 0, 0.3 en 1.2. Hoe hoger het gewicht, hoe meer financiering de school krijgt voor een leerling.

De cijfers in tabel 3.1 laten zien hoeveel beter de leerlingen op een toets scoren, gemeten in standaarddeviaties, als zij een leerkracht toegewezen krijgen die 1 standaarddeviatie hoger scoort op de Kijkwijzer. 1 standaarddeviatie hogere score op de Kijkwijzer komt overeen met 13.5 meer gedragsaspecten. Voor rekenen vinden we dat leerlingen gemiddeld 0.15 standaarddeviaties beter scoren voor rekenen als een leerkracht 1 standaarddeviatie hogere evaluatiescore heeft. Voor spelling gaat het om een 0.18 standaarddeviatie hogere score, en voor lezen om een 0.11 standaarddeviatie hogere score.

We kunnen, na een kleine omrekening, ook een andere interpretatie geven aan deze cijfers: als een leerkracht uit de slechtste 25 procent slechtste leerkrachten wordt vervangen door een leerkracht uit de 25 procent beste leerkrachten van de verdeling op de Kijkwijzer gaan de leerlingen er gemiddeld 0,37 standaarddeviatie meer op vooruit bij rekenen, 0,44 standaarddeviatie bij spelling en 0,24 standaarddeviatie bij lezen.[18] Dit zijn substantiële verschillen. Een basisschoolleerling die in potentie een vwo-niveau heeft kan een heel niveau lager uitkomen in het vervolgonderwijs als hij of zij twee jaar lang een (gemeten) slechte leraar heeft in plaats van een (gemeten) goede leraar.[19]

De Kijkwijzer lijkt hiermee in staat grotere verschillen in leraarkwaliteit te identificeren dan bijvoorbeeld het Teacher Evaluation System (TES) dat gehanteerd is in Cincinatti in de Verenigde Staten.[20] Dit komt mogelijk door een grotere mate van detail bij de Kijkwijzer in die zin dat er meer gedragsaspecten worden gemeten (75 gedragsaspecten bij de Kijkwijzer tegenover 29 gedragsaspecten bij TES). Maar het kan ook komen door verschillen in de specifieke gedragsaspecten die gemeten worden. Een ander verschil is dat de gehanteerde Kijkwijzer in de pilot meet of een leraar een bepaald gedragsaspect laat zien of niet (score 0 of 1), terwijl bij het TES in Cincinatti op elk gedragsaspect een score op een vierpuntsschaal gegeven dient te worden.

De analyse laat zien dat leerlingen die zijn toegewezen aan een leerkracht die goed scoort op de Kijkwijzer gemiddeld beter presteren dan vergelijkbare leerlingen die zijn toegewezen aan een leerkracht die laag scoort op de Kijkwijzer. Met andere woorden: het meetinstrument de Kijkwijzer lijkt leraargedrag te meten dat er toe doet en maakt daarmee behoorlijke verschillen in kwaliteit tussen leerkrachten

[18] Dit verschil is een gemiddeld verschil. Dat betekent dus niet dat elke leerling deze winst zal halen, sommigen zullen meer dan gemiddeld profijt hebben, anderen minder dan gemiddeld.
[19] Het verschil in gemiddelde scores op de CITO eindtoets tussen leerlingen met een havo-advies en leerlingen met een vwo-advies bedraagt circa 0.7 standaarddeviatie (bron: eigen berekening op basis van PRIMA cohortonderzoeken).
[20] Bij dit evaluatiesysteem zijn de schattingen van de vooruitgang door het hebben van een leraar uit het beste kwartiel ten opzichte van een leraar uit het slechtste kwartiel 0.09 bij rekenen en 0.13 bij lezen (zie Kane et al., 2011). Onze schattingen voor de Kijkwijzer zijn dus ruwweg twee (lezen) tot vier (rekenen) keer zo groot.

zichtbaar. Dat is een belangrijke bevinding, want zelfs zwakke signalen van (verschillen in) leraarkwaliteit kunnen informatie verschaffen die zeer waardevol is, bijvoorbeeld met het oog op ontwikkel-, belonings- en personeelsbeleid. De Kijkwijzer laat aan leerkrachten en schoolleiders specifiek zien op welke gedragsaspecten en competenties leerkrachten zich kunnen verbeteren en maakt daarmee gerichte ontwikkelactiviteiten mogelijk.

4 Houding ten aanzien van pilot

4.1 Houding ten aanzien van prestatiebeloning

De houding ten aanzien van prestatiebeloning en de mate waarin men zich daardoor extra geprikkeld voelt blijken behoorlijk gemengd. Meer ervaren leraren en leraren die meer risicomijdend zijn staan negatiever tegenover prestatiebeloning. Hetzelfde geldt voor leraren die de criteria te zwaar vinden en leraren die onbekend zijn met de hoogte van de te behalen beloning. De mate waarin men zich extra geprikkeld voelt door de prestatiebeloning hangt sterk samen met de hoogte van de te behalen prestatiebeloning. Leraren staan ongeveer even positief tegenover een uitkering in de vorm van een individuele uitkering als in de vorm van een teambudget.

Tabel 2.4 laat de resultaten zien van vragen die zijn voorgelegd aan de leerkrachten voorafgaande aan de pilot. Het blijkt dat er een behoorlijk gemengd beeld bestaat onder de deelnemende leerkrachten over prestatiebeloning. Dit beeld is het meest positief over prestatiebeloning op basis van vooruitgang in leerkrachtvaardigheden, zoals dat in deze pilot plaatsvindt. Het oordeel van de leerkrachten over een beloning in de vorm van een individuele uitkering is ongeveer gelijk aan dat over een beloning in de vorm van een teambudget. Als leraren wordt gevraagd een voorkeur uit te spreken, dan kiest de helft voor een individuele uitkering en de andere helft voor een teambeloning. Het hangt wel af van het doel waaraan het teambudget besteed mag worden. Als het teambudget moet worden besteed aan scholing of training, dan kiest 70 procent liever voor een individuele uitkering. Als het teambudget een vrije invulling heeft die kan worden bepaald door de leerkrachten zelf, dan zijn de leraren indifferent tussen een individuele uitkering en een teambudget.

Tabel 4.1 Houding van leraren ten aanzien van prestatiebeloning

Houding	Negatief	Neutraal	Positief	Gemiddeld (1-5)
Prestatiebeloning in zijn algemeenheid	29	37	34	3,1
Het extra belonen van (teams van) leerkrachten op basis van vooruitgang in leerkrachtvaardigheden	22	24	54	3,4
Prestatiebeloning in de vorm van teambudget voor innovatie, teamontwikkeling en/of professionalisering	23	30	47	3,3
Prestatiebeloning in de vorm van een individuele uitkering	27	29	44	3,2

Gesteld voor een keuze hebben de leraren een lichte voorkeur voor prestatiebeloning op basis van de bekwaamheid van het team (62 procent) ten opzichte van op basis van hun eigen bekwaamheid (38 procent). Eenzelfde mate van voorkeur geldt voor beloning op basis van een meting van leerkrachtvaardigheden (63 procent) ten opzichte van op basis van gemeten vooruitgang van leerprestaties van leerlingen (rekening houdend met de achtergrondkenmerken van de leerlingen). Dat is een interessante bevinding, omdat bij de meeste beproefde prestatiebeloningsystemen in het buitenland tot op heden de leerprestaties van leerlingen als belangrijkste criterium golden voor de beloning.

De leerkrachten geven aan duidelijk sterker geprikkeld te worden door een hogere beloning. Dit geldt in vergelijkbare mate voor de hoogte van de teambeloning als voor de hoogte van de individuele uitkering. Ter illustratie, bij een individuele beloning van 250 euro zou 10 procent zich sterk tot zeer sterk gestimuleerd voelen (en 52 procent totaal niet of niet). Bij een beloning van 10.000 euro gaat het om 63 procent die zich sterk tot zeer sterk gestimuleerd zou voelen, tegenover slechts 15 procent totaal niet of niet. De gecommuniceerde minimaal te behalen individuele uitkering was overigens 750 euro.

Verband houding ten aanzien van prestatiebeloning met leraarkenmerken

De houding ten aanzien van prestatiebeloning, en daarmee mogelijk de effecten daarvan, kunnen samenhangen met bepaalde persoonskenmerken en -voorkeuren. Dohmen en Falk (2010) laten met behulp van zowel veld- als experimentele labdata zien dat voorkeuren voor variabele beloning (en baanzekerheid) duidelijk verschillen en dat mensen zichzelf op basis van die voorkeuren selecteren in bepaalde beroepen met verschillende prikkelstructuren. Ze vinden bijvoorbeeld dat leraren meer risico-avers zijn dan werknemers in andere beroepen. Tegelijkertijd constateren ze dat het lerarenberoep zich kenmerkt door een vaste beloning en een hoge mate van baanzekerheid. Deze bevindingen sluiten goed aan bij de verwachtingen vanuit de theorie over voorkeuren voor vaste beloning en meer baanzekerheid al naar gelang men meer risico-avers is. Ook vinden ze (binnen het lerarenberoep) dat basisschoolleraren meer risico-avers zijn dan leraren in het voortgezet onderwijs, wat goed lijkt te sporen met de geringere variatie in beloning in het basisonderwijs.

We onderzoeken in deze paragraaf in welke mate de houding ten aanzien van prestatiebeloning en de mate waarin men zich extra geprikkeld voelt samenhangt met de houding ten aanzien van risico en een aantal andere kenmerken, zoals ervaring en bekwaamheid. De resultaten van de OLS regressies staan vermeld in tabel 2.5. De volgende factoren spelen een rol.

Leraren die meer risicozoekend zijn staan positiever tegenover prestatiebeloning in zijn algemeenheid en in de vorm van een individuele uitkering in het bijzonder (zie kolommen 2 en 3). Ter illustratie, van het kwartiel met de grootste risicoaversie is 37 procent positief of zeer positief over prestatiebeloning in de vorm van een

individuele uitkering, terwijl dit percentage bij het kwartiel met de laagste risicoaversie 62 procent bedraagt. Er is echter geen statistisch significant verband tussen risicozoekendheid en de mate waarin men zich gestimuleerd voelt door de te behalen individuele uitkering (zie kolom 1).

De ervaren zwaarte van de criteria hangt negatief samen met de stimulans en de houding ten aanzien van prestatiebeloning in de vorm van een individuele uitkering. Dit suggereert dat het van belang is voor de prikkelwerking van prestatiebeloning om een lat neer te leggen die leraren ervaren als haalbaar (niet te zwaar). Gemiddeld 66 procent van de leerkrachten beschouwde de lat in deze pilot als hoog (51 procent) of te hoog (15 procent).

Leraren die bekend zijn met de hoogte van de te behalen individuele uitkering voelen zich vaker gestimuleerd. Ter illustratie, van degenen die niet bekend zijn met de hoogte voelt 19 procent zich sterk tot zeer sterk gestimuleerd, van degenen die wel bekend zijn is het percentage ruim twee keer zo hoog (39 procent). Ook de houding ten opzichte van prestatiebeloning in de vorm van een individuele uitkering is positiever. Dit suggereert dat winst te behalen valt door een duidelijke communicatie over de hoogte van de te behalen prestatiebeloning.[21] Zestig procent van de leerkrachten was vlak voor de start de pilot niet op de hoogte van de minimaal te behalen individuele uitkering.

De mate van stimulans neemt af met ervaring. Ter illustratie, van het kwartiel meest ervaren leraren (meer dan 28 jaar) voelt 18 procent zich sterk of zeer sterk gestimuleerd door de te behalen individuele beloning. Bij het kwartiel minst ervaren leraren (minder dan 10 jaar) is dit ruim twee keer zo veel (40 procent). Ervaren leerkrachten staan ook negatiever ten opzichte van prestatiebeloning in zijn algemeenheid en in de vorm van een individuele uitkering in het bijzonder (zie kolommen 2 en 3). Ter illustratie, van het kwartiel meest ervaren leraren staat 1 op de 3 positief of zeer positief tegenover prestatiebeloning in de vorm van een individuele uitkering. Bij het kwartiel minst ervaren leraren is dit bijna twee keer zoveel (63 procent).

Een hogere score op de nulmeting lijkt in beperkte mate samen te hangen met de houding ten aanzien van prestatiebeloning in zijn algemeenheid en met de mate van prikkeling. Ter illustratie, van het kwartiel leraren met de laagste score staat 34 procent negatief of zeer negatief tegenover prestatiebeloning, tegenover 24 procent bij het kwartiel leraren met de hoogste score.

[21] Deze bevinding lijkt te sporen met een studie van Englmaier e.a. (2012), die aantoont dat de opvallendheid (salience) van beloningsprikkels van belang is voor het bereiken van effecten. Ze vinden dat alleen het veranderen van de communicatie over een gebruikte beloningsmethodiek voldoende is om een significante invloed op prestaties van werknemers te hebben, zonder aan de beloningssystematiek zelf iets te veranderen.

Tabel 4.2 Samenhangende factoren met houding ten aanzien van prestatiebeloning

Variabele	Mate van zich extra gestimuleerd voelen door de te behalen individuele uitkering (1-5)	Houding ten aanzien van prestatiebeloning in vorm van een individuele uitkering (1-5)	Houding ten aanzien van prestatiebeloning in het algemeen (1-5)
	(1)	(2)	(3)
1. (Ervaren) zwaarte criteria (1-5)	**-0,285**** (0,106)	**-0,351**** (0,125)	**-0,444***** (0,111)
2. Bekend met hoogte bedrag individuele uitkering (0 of 1)	**0,469**** (0,151)	**0,355**** (0,132)	-0,027 (0,148)
3. Risicozoekendheid (10-310; st.dev. = 82)	0,001 (0,001)	**0,003*** (0,001)	**0,003**** (0,001)
4. Ervaring in onderwijs (0-46 jaar)	**-0,019*** (0,009)	**-0,021**** (0,006)	**-0,017**** (0,002)
5. Nulmeting leerkrachtvaardigheid (4-73, st.dev. =17)	**0,008*** (0,004)	0,005 (0,004)	**0,008***** (0,001)
Aantal waarnemingen	97	97	97

Robuuste standaardfouten staan tussen haakjes, rekening houdend met clustering van leerkrachten op schoolniveau.

De hoogte van de beloning lijkt sterk uit te maken voor de mate waarin men zich gestimuleerd voelt door de prestatiebeloning. Ter illustratie, bij een beloning van 250 zou 10 procent zich sterk of zeer sterk gestimuleerd voelen, tegenover 63 procent bij een beloning van 10.000 euro. Bij oplopende teambudgetten neemt de mate van stimulans in vergelijkbare mate toe.

4.2 Houding ten aanzien van coaching

De houding van leerkrachten ten aanzien van het niveau van de coaches en het type uitgevoerde interventies is positief tot zeer positief. Zowel coaches als schoolleiders als leraren zijn het minst positief over de groepsbrede trainingsbijeenkomsten. Dit suggereert dat van een individuele maatwerkaanpak om leerkrachtvaardigheid te verbeteren meer kan worden verwacht dan van een groepsaanpak. De bijdrage van de coaching aan de geboekte resultaten van de pilot wordt relatief het hoogst ingeschat, gevolgd door het stellen van doelen voor groei in leerkrachtvaardigheid en op enige afstand de prestatiebeloning.

We hebben gevraagd aan coaches, leerkrachten en directeuren welke interventies van de coaches naar hun inzicht het meest hebben bijgedragen aan de gerealiseerde groei in leerkrachtvaardigheid. Alle directeuren geven aan dat fysieke lesobservaties het meeste hebben bijgedragen. Bij coaches is het beeld wat meer diffuus. Individuele voortgangsgesprekken wordt het vaakst genoemd als interventie met de hoogste bijdrage (36%), maar ook video-interactiebegeleiding (29%) en lesobservaties (21%). Als minst effectieve interventie noemen zowel de coaches als directeuren teamtrainingen en -bijeenkomsten.

Leraren zijn overwegend positief over alle typen ingezette interventies, met gemiddelden van 3.9 tot 4.3 op een vijfpuntsschaal van negatief naar positief. Leraren zijn het meest en in vergelijkbare mate positief over lesobservaties, voortgangsgesprekken met de coach en overige gesprekken/overleggen met de coach. Leraren zijn net als coaches en directeuren het minst positief over de groepsbrede trainings- of cursusbijeenkomsten. Deze resultaten suggereren een hogere effectiviteit van een individuele maatwerkaanpak dan van een collectieve aanpak.

Leraren zijn behoorlijk positief over de competentie/bekwaamheid van hun coach en over de bijdrage van hun coach aan hun bekwaamheid (gemiddeld 4.2 op een vijfpuntsschaal).

De leerkrachten en schoolleiders is gevraagd naar de relatieve bijdrage van de coaching ten opzichte van de andere twee componenten van de pilot (het stellen van doelen en prestatiebeloning) door 100 punten te verdelen. De grootste bijdrage wordt volgens de leerkrachten gevormd door de leerkrachtnabije ondersteuning (43 punten), op de voet gevolgd door het stellen van doelen voor teams in termen van groei en minimale bekwaamheid (37 punten).[22] Het in het vooruitzicht stellen van prestatiebeloning heeft volgens leerkrachten minder bijgedragen (gemiddeld 19 punten). Directeuren geven dezelfde volgorde aan, maar volgens hen heeft leerkrachtnabije ondersteuning verreweg de grootste bijdrage gehad van de drie componenten (respectievelijk 64, 24 en 11 punten). Directeuren zijn het allen eens met de stelling dat het stellen van concrete doelen duidelijke sturing gaf aan het traject en 60 procent van de directeuren denkt ook dat dit tot extra inzet en motivatie bij de leraren heeft geleid. Deze bevindingen onder directeuren en leraren suggereren dat het stellen van concrete doelen niet onbelangrijk is geweest voor de geboekte resultaten.

4.3 Houding ten aanzien van meten en meetinstrument

Leraren, coaches en directeuren zijn overwegend positief over het meten van leerkrachtvaardigheden door middel van lesobservaties. Hetzelfde geldt voor het in de pilot gehanteerde meetinstrument de Amsterdamse Kijkwijzer. Wel zijn alle actoren overwegend van mening dat een combinatie van aangekondigde en onaangekondigde lesobservaties tot een betrouwbaarder beeld zal leiden van de bekwaamheid van de leerkracht. Schoolleiders verwachten dat een dergelijk meetinstrument prima gebruikt kan worden voor beoordelings- en ontwikkelgesprekken en als startpunt om op maat te gaan werken aan verbetering van leerkrachtvaardigheid.

[22] Locke en Latham (2002) laten zien dat positieve effecten gevonden zijn van het stellen van specfieke, moeilijke doelen positief bijdraagt aan prestaties. Dat is gebleken in zowel laboratorium als veldonderzoek in experimentele en quasi-experimentele settings.

De betrokken leraren staan overwegend positief tegenover het meten van leerkrachtvaardigheden door middel van een lesbezoek. 60 procent is (zeer) positief, 7 procent is (zeer) negatief en 33 procent is neutraal. Een klein deel van de leraren (13 procent) denkt dat men door een lesobservatie geen behoorlijk beeld kan krijgen van hun bekwaamheid. De houding van leraren ten aanzien van de Kijkwijzer als specifiek gehanteerd meetinstrument is ook overwegend positief, maar wel iets meer gemengd dan tegenover het meten van bekwaamheid in zijn algemeenheid: 46 procent is (zeer) positief, 17 procent (zeer) negatief en 37 procent neutraal. Slechts 3 procent van de leraren is van mening dat de gemeten score op de eindmeting veel hoger of veel lager is dan het eigen beeld dat ze hebben van hun bekwaamheid. Bij de nulmeting betrof dit 8 procent.

Het verband tussen het eigen beeld van de bekwaamheid en de score op de Kijkwijzer (bij de nulmeting) is het sterkst met de inschatting van de eigen bekwaamheid op rekenen: een 1 punt hogere inschatting op een tienpuntsschaal van de eigen bekwaamheid op rekenen hangt samen met een 3.5 punten hogere score op de Kijkwijzer.[23] Het verband van de Kijkwijzerscore met de eigen inschatting van de algemene bekwaamheid of met de eigen inschatting van de bekwaamheid op het vlak van lezen en spelling is wat minder sterk en niet significant.

Directeuren zijn ook overwegend positief over een evaluatie-instrument als de Kijkwijzer. Dit blijkt uit de antwoorden op diverse stellingen hierover. Slechts 14 procent denkt dat de gehanteerde Kijkwijzer niet in staat is om minder goede van de goede leerkrachten te onderscheiden. 86 procent denkt dat de Kijkwijzer een goed startpunt is om gericht en op maat te werken aan verbetering van leerkrachtvaardigheden. Eenzelfde percentage is voorstander van periodieke metingen van de bekwaamheid van hun leerkrachten en ook 86 procent denkt dat een observatie-instrument als de Kijkwijzer prima gebruikt kan worden als input voor beoordelings- en persoonlijke ontwikkelgesprekken. Alle directeuren zijn van mening dat de Kijkwijzer prima mee zou kunnen wegen bij eventuele beslissingen omtrent ontslag of promotie naar een hogere schaal.

Ook coaches zijn overwegend positief over de Kijkwijzer. Tweederde van de coaches denkt dat de gehanteerde Kijkwijzer goed in staat is om de minder goede van de goede leerkrachten te onderscheiden, eenderde denkt van niet. Ook vindt tweederde de Kijkwijzer een goed startpunt om gericht en op maat te gaan werken aan

[23] Het geschatte verband is significant op 1-procent significantieniveau. 3.5 punten is gelijk aan ongeveer een vijfde standaarddeviatie van alle leraren die bij de nulmeting gemeten zijn. De standaarddeviatie van de inschatting van de eigen bekwaamheid bij rekenen bedraagt 1.04 punt. In de regressie waarin het verband geschat wordt is gecontroleerd voor de school waarop de leraren werkzaam zijn. Het is goed mogelijk dat leraren bij de inschatting van hun eigen bekwaamheid zichzelf hebben afgezet tegenover die van de andere leraren op hun school.

verbetering van leerkrachtvaardigheden, tegenover een kwart die het hier niet mee eens is.

Zoals eerder gemeld zijn coaches, leraren en directeuren overwegend van mening dat een combinatie van onaangekondigde en aangekondigde observaties tot een betrouwbaarder beeld zou leiden van de bekwaamheid van leerkrachten. Dit beeld spoort met bevindingen van Kane en Staiger (2012), die vinden dat de betrouwbaarheid van de evaluaties toeneemt met het aantal observaties per leerkracht.

5 Lessen en aanbevelingen

Omdat eerdere bevindingen geen definitieve evidentie verschaffen voor het effect van dit type interventies, kunnen we geen rechtstreekse aanbevelingen doen of andere schoolbesturen dit type beleid zouden moeten invoeren. De pilot biedt wel lessen en aanbevelingen voor schoolbesturen, gemeenten of beleidsmakers die (systemen van) lerarenevaluatie aan het ontwikkelen of herzien zijn, of die willen gaan werken met leerkrachtnabije ondersteuning of prestatiebeloning voor leraren. Onderstaand vatten we de belangrijkste lessen en aanbevelingen samen voor de verschillende onderdelen van de pilot. Deze lessen zijn mede gebaseerd op ervaringen in het buitenland met dergelijke interventies.

Lessen voor evaluatiesysteem van leraren

- Evaluatiesystemen gebaseerd op (geaccepteerde) standaarden met betrekking tot wat leraren moeten doen en kennen bieden de mogelijkheid om leraren onmiddellijke en specifieke feedback te geven op hun kwaliteit van lesgeven. Ook kunnen ze nader inzicht bieden aan schoolleiders over de bekwaamheid van hun leerkrachten, die meegenomen kan worden bij het HRM- en professionaliseringsbeleid.
- Een evaluatie op basis van meerdere observaties per leerkracht vergroot de betrouwbaarheid (zie Kane en Staiger, 2012).
- Uitbreiding van aangekondigde met onaangekondigde observaties geeft naar verwachting van betrokkenen een betrouwbaarder beeld van de bekwaamheid van de leraren.
- Het is van belang dat een voldoende mate van uniformiteit bereikt wordt bij het hanteren van het scoresysteem tussen de verschillende beoordelaars. Dat vergt training en afstemming tussen observatoren. In de onderhavige pilot heeft dergelijke training plaatsgevonden door een extern bureau. Dit was van groot belang, te meer daar het wel of niet ontvangen van teambeloning direct af hing van de scores op het observatie-instrument. Kane en Staiger (2012) laten zien dat bij zorgvuldig getrainde observatoren verschillen in oordelen beperkt blijven die

kunnen ontstaan doordat sommige mogelijk consistent hoog scoren en andere mogelijk consistent laag.[24] Hetzelfde beeld rijst op uit een studie van Van de Grift en Lam (1998). Zij laten zien dat de zogenoemde 'interbeoordelaarsbetrouwbaarheid' hoog was (gemiddeld 88 procent) bij een observatie-instrument voor didactisch handelen dat gehanteerd is door getrainde Inspecteurs van de Onderwijsinspectie in het basisonderwijs. Dit is bepaald door te kijken naar verschillen in evaluatiescores tussen twee Inspecteurs van dezelfde les (duo-observaties). In deze studie hebben we de interbeoordelaarsbetrouwbaarheid niet kunnen onderzoeken.[25]

- Het is van belang om voldoende onderscheidend vermogen te hebben bij een observatie-instrument of een classificatie van de bekwaamheid van leraren. Het blijkt dat regelmatig systemen zijn gehanteerd waarbij slechts heel kleine percentages in de laagste categorie werden ingedeeld (zie Weisberg et al, 2009). Dat spoort niet met wat bekend is over de variatie in kwaliteit van leraren.
- Onderzoek naar de relatie tussen het meetinstrument en leerlingprestaties van leerlingen is aan te bevelen, alvorens een bepaald observatie-instrument breed op te schalen. Het verdient aanbeveling te werken met observatie-instrumenten waarbij deze relatie overtuigend is aangetoond.
- De objectiviteit van de beoordelaar is belangrijk, te meer als er beloningen aan gekoppeld worden of andere belangen aan de uitkomsten verbonden zijn.[26] Externe onafhankelijke observanten verdienen de voorkeur.
- Voldoende draagvlak onder leerkrachten voor het meetinstrument en het meten lijkt van belang. Dit vereist volgens betrokkenen bij het project het creëren van een gevoel van urgentie en eigenaarschap, duidelijke communicatie vooraf over het observatie-instrument en het meetproces, reductie van (mogelijke) angst onder leerkrachten[27], en nabesprekingen van uitkomsten vanuit de observator met de leerkracht.

---

[24] Ten hoogste tien procent van de totale variantie in scores werd veroorzaakt door dergelijke 'main rater effects'. Kane en Staiger raden naast training ook certificering van beoordelaars aan, alsook audits op systeemniveau om de betrouwbaarheid te controleren.

[25] Opgemerkt dient te worden dat het risico op een gebrekkige betrouwbaarheid waarschijnlijk hoger ligt bij observatie-instrumenten waar per gedragsaspect of competentie op een schaal dient te worden gescoord hoe goed de leraar deze beheerst (in plaats van of een leraar een aspect wel of niet laat zien zoals in de hier onderzochte pilot). Beoordeling op een vierpuntsschaal per aspect is het geval bij bijvoorbeeld het Teacher Evaluation System in Cincinatti, of bij ICALT, een meetinstrument ontwikkeld ten behoeve van vergelijking van lerarenkwaliteit in een aantal Europese landen (van de Grift, 2007).

[26] Rothstein (2012) beschouwt het als een topprioriteit te onderzoeken of en in hoeverre de maatstaven voor kwaliteit van leraren gecorrumpeerd raken als de belangen worden vergroot. Wij hebben niet kunnen vaststellen of daar in deze pilot sprake van is geweest, maar het is een belangrijk aandachtspunt.

[27] In de pilot is getracht dit te doen door de focus sterk op ontwikkeling te leggen en niet op beoordelen of afrekenen.

Voorbeeld lerarenevaluatie in VS: Teacher evaluation system (Cincinatti)

Het Teacher Evaluation System (TES) is eind jaren negentig opgezet in publieke scholen in Cincinatti. Dit is een van de systemen waar de meeste ervaring mee is opgedaan. Het is een systeem dat gebruik maakt van getrainde beoordelaars en een gespecificeerd en op onderzoek gebaseerd rubriceringssysteem. Het systeem omvat meerdere lesobservaties per leraar gedurende een schooljaar. Het minimum is over het algemeen vier evaluaties die periodiek worden uitgevoerd gedurende het schooljaar. Om in aanmerking te komen als een 'peer' beoordelaar moet een senior leraar een intensieve training volgen waarin men een opgenomen les adequaat moet scoren om te mogen beginnen als een beoordelaar. Alle nieuwe leraren in het district moeten participeren in TES gedurende hun eerste jaar in het district en opnieuw als ze een vaste aanstelling hebben gekregen. Leraren met een vaste aanstelling moeten vervolgens elke vijf jaar opnieuw deelnemen aan TES.

TES bestaat uit vier domeinen, 15 standaarden en 32 elementen die het gedrag, de vaardigheden en de eigenschappen beschrijven die effectieve leraren zouden moeten bezitten and moeten gebruiken. De domeinen omvatten vier categorieën: voorbereiding, klassenmanagement, pedagogische en vakkennis en toepassing, en collegiale verantwoordelijkheden en betrokkenheid. De scores op de tweede en derde categorie worden afgeleid uit lesobservaties, die op de eerste en de vierde op basis van aan te leveren documentatie en bewijzen.

Binnen elk domein worden leraren beoordeeld op basis van een set standaarden, die zijn onderverdeeld in elementen. Elk element omvat beschrijvingen die de prestaties van een leerkracht beschrijft op elk van de vier niveaus waarop een leerkracht gerubriceerd kan worden (distinguished, proficient, basis en unsatisfactory).

Alleen de eerste lesobservatie in de evaluatiecyclus is aangekondigd, de andere lesobservaties kunnen onverwacht zijn. Na elk lesbezoek moeten beoordelaars een evaluatierapport verschaffen aan de leraar. Aan het eind van het jaar bekijkt de beoordelaar alle informatie voor een betreffende leraar uit de verschillende evaluaties en komt dan uit op een uiteindelijke score voor elk domein. De beoordelaar verschaft de leraar een eindejaarsrapport met de uiteindelijke scores.

Lessen voor prestatiebeloning

- Of prestatiebeloning voor leraren werkt of niet, kunnen we niet concluderen op basis van deze pilot. Het betreft een specifieke vorm van prestatiebeloning, en een experimentele setting met controlegroepen ontbreekt.
- Het vergt een stevige voorbereiding en inspanning om een prestatiebeloningsysteem te ontwikkelen en goed uit te voeren. Het gaat dan onder andere om de keuze voor de prestatiecriteria, de strengheid van de criteria, de hoogte van de beloning, het type beloning (teambudget en/of individuele uitkering), en spelregels over bijzondere situaties die zich voor kunnen doen.
- Het is van belang dat de gekozen prestatiemaat niet gemakkelijk te manipuleren is door degenen die de beloning kunnen krijgen.[28] Bij leerresultaten als prestatiemaat bestaat het risico dat leraren leerlingen precies gaan voorbereiden op de specifieke toets waar de beloning van afhankelijk is of door zwakke leerlingen de betreffende test niet te laten maken. Bij een evaluatiesysteem zoals in de pilot gehanteerd met een aangekondigde lesobservatie bestaat het risico dat leraren zich heel specifiek gaan voorbereiden op die ene les en daar precies laten

[28] Rothstein (2012) zegt hierover: "Teachers facing strong incentives may be able to raise their measured performance without improving their overall productivity, by redirecting effort from unmeasured to measured dimensions (Glewwe et al., 2010) or simply by distorting the performance measure directly, such as by cheating (Jacob and Levitt, 2003) or teaching to the test."

zien wat er van hen verwacht wordt, terwijl ze dit dan niet elke les laten zien. Een combinatie met een onaangekondigde lesobservatie kan dit risico verkleinen.

- Bij teambeloning is het van belang om duidelijke afspraken te maken over hoe wordt omgegaan met uitval van leraren gedurende het traject. Het blijkt in deze pilot dat degenen die niet hebben meegedaan aan de eindmeting een lager dan gemiddelde score hadden op de beginmeting. Een duidelijke prikkel tegen uitval was wel dat leraren die niet meededen aan de eindmeting ook geen aanspraak konden maken op de individuele uitkering.
- Het is van belang dat uitkomstmaten als prestatiecriteria worden gekozen die een bewezen verband hebben met onderwijskwaliteit. Het blijkt uit deze pilot dat evaluaties van leraren op basis van de Amsterdamse Kijkwijzer positief verband houden met de prestaties van leerlingen op rekenen en taal.
- Een absolute lat (zoals in deze pilot gehanteerd) heeft als nadeel dat je van tevoren niet weet of de lat op het juiste niveau gelegd wordt. Het risico bestaat dat de lat te laag of juist te hoog gelegd wordt. Vooral bij gebrek aan historische gegevens over de uitkomstcriteria is dat een risico. Een model waarbij de beste zoveel procent scholen of leraren de prestatiebeloning ontvangen heeft dit risico niet. Het bijkomende voordeel van een dergelijk model bij een gegeven budget voor prestatiebeloning is dat van tevoren al vast kan staan hoe hoog de beloning precies zal worden. Bij een absolute lat weet je nooit van te voren hoeveel scholen (leraren) de beloning zullen gaan krijgen. Het voordeel van een absolute lat kan zijn dat er een gewenste standaard gezet wordt, zodat (teams van) leraren van tevoren precies helder hebben hoe hoog ze moeten scoren. Bij een relatieve lat is dat op voorhand niet duidelijk.
- Duidelijke communicatie over de hoogte van te behalen beloning richting de leraren verhoogt de extra stimulans die uitgaat van prestatiebeloning.
- Het ervaren van een (te) zware lat vermindert de extra stimulans die uitgaat van prestatiebeloning. Doelen die uitgaan van groei in plaats van (alleen) een absoluut te bereiken niveau kunnen ervoor zorgen dat meer scholen (leraren) zich geprikkeld voelen, zowel de zwakkere als de betere. Bij een doel alleen gericht op het behalen van een bepaald niveau kan het zo zijn dat de zwakke scholen niet gemotiveerd raken omdat ze weten dat de lat te ver buiten hun bereik ligt, terwijl de beste scholen ook niet extra gemotiveerd raken omdat ze weten dat de toch wel boven de lat gaan uit komen.
- De hoogte van de beloning lijkt sterk uit te maken voor de mate van extra stimulans die de leraren voelen. Een lage beloning zal weinig extra stimulans opleveren.

Lessen voor coaching

- Een individuele maatwerkaanpak is volgens betrokkenen effectiever dan een groepsaanpak (met groepstrainingen) om leerkrachtvaardigheid van leraren te verbeteren.

- Voor duurzame groei in leerkrachtvaardigheid wordt door directeuren en coaches een langer coachingstraject zinvol geacht dan in deze pilot heeft plaatsgevonden. In deze pilot betrof het een gemiddelde tijdsinzet van 35 uur per leerkracht over een periode van een half jaar.
- Het lijkt erop - afgaande op de nulmetingen in deze pilot, maar ook uit eerder onderzoek van de Onderwijsinspectie - dat een deel van de leraren onvoldoende is toegerust om hun werk goed te doen. De intensieve coaching grijpt hier direct op aan. Een belangrijke vraag is in hoeverre financiële prikkels en het stellen van concrete doelen daarnaast bij kunnen dragen aan de motivatie van leerkrachten om te werken aan verbetering van hun leerkrachtvaardigheden en de resultaten daarvan. Volgens zowel directeuren als leraren hebben beide elementen voor een niet onbelangrijk deel positief bijgedragen aan de behaalde resultaten van de pilot.

Mogelijke vervolgstappen en evaluatie

- Er zijn diverse voorbeelden van interventies met meten, feedback geven en coachen op leerkrachtvaardigheid die effectief gebleken zijn. Het verdient aanbeveling bij de verdere ontwikkeling en implementatie van dergelijke interventies deze voorbeelden goed te bestuderen en te leren van ervaringen die men daarmee in de loop der tijd mee heeft opgedaan.
- Een experiment met aselecte toewijzing van meten, coachen en belonen van leerkrachtvaardigheid aan bepaalde scholen en niet aan andere scholen is nodig om de effecten van dergelijke interventies vast te kunnen stellen. Het is daarbij van belang dat voldoende scholen deelnemen om de effecten overtuigend vast te kunnen stellen.
- Aandacht voor het meten van de langere termijn effecten van dergelijke interventies is belangrijk. Uit een onderzoek van Taylor en Tyler (2011) naar de effecten van meten en feedback op leerkracht blijkt bijvoorbeeld dat de grootste effecten worden gerealiseerd in de twee jaar na afloop van het traject, en dat de effecten gedurende het traject minder sterk zijn. Maar omgekeerd weten we ook uit veel studies van onderwijsinterventies dat uitdoving van effecten plaats zou kunnen vinden. Dit suggereert dat het wenselijk is om ten minste effecten in het schooljaar na afloop van de interventies te meten.
- In de loop der tijd zijn verschillende observatie-instrumenten van leerkrachtvaardigheden ontwikkeld.[29] Het verdient aanbeveling deze

[29] Voorbeelden in Nederland zijn onder andere een instrument dat gehanteerd wordt door de Onderwijsinspectie om leraren te scoren op beheersing van een aantal eenvoudige en complexe vaardigheden (zie Inspectie van het Onderwijs, 2011), een observatie-instrument ontwikkeld ten behoeve van internationaal vergelijkend onderzoek in een viertal landen waaronder Nederland (zie van de Grift, 2007), en een instrument dat pedagogisch-didactische vaardigheid meet van leraren in het basisonderwijs (van de Grift et al., 2011). Merk op dat er naast verschillen in specifieke gedragsaspecten ook enige mate van overlap zit in de gedragsaspecten die gemeten worden in deze observatie-instrumenten.

instrumenten tegen elkaar af te zetten op een aantal aspecten. Het belangrijkste aspect is de mate waarin de score op het instrument samenhangt met leerlingprestaties. De observatie-instrumenten waarbij dit in sterke mate het geval is, verdienen de voorkeur. Andere aspecten kunnen zijn de interbeoordelaarsbetrouwbaarheid (komen verschillende observanten tot dezelfde scores), de acceptatie van het instrument door schoolleiders en leerkrachten en de mate waarin het hen inzicht en handvatten kan verschaffen om gericht aan verbetering te werken, en de kosten die het hanteren van het betreffende observatie-instrument met zich meebrengt (zoals trainingskosten voor observatoren en uitvoeringskosten).

6 Conclusies en discussie

De belangrijkste conclusie van dit onderzoek is dat metingen van leerkrachtvaardigheden door middel van lesobservaties door getrainde observatoren significante verschillen in kwaliteit van leraren zichtbaar weten te maken. Anders gezegd, leerlingen van leraren die goed scoren op dit meetinstrument halen gemiddeld betere toetsresultaten dan leerlingen van leraren die zwak scoren op dit meetinstrument. De verschillen zijn niet gering. Een leerling die twee jaar achter elkaar een zwakke leraar heeft in plaats van een goede kan een heel niveau lager uitkomen in het vervolgonderwijs, bijvoorbeeld in plaats van vwo-niveau op havo-niveau. Dit is een belangrijke bevinding, omdat veel traditionele kenmerken van leraren zoals vooropleiding en ervaring niet of nauwelijks voorspellend blijken voor de kwaliteit van leraren. Zelfs zwakke signalen van verschillen in kwaliteit kunnen informatie verschaffen die zeer waardevol is, onder andere voor ontwikkel- en personeelsbeleid op scholen.

De tweede hoofdconclusie is dat lesobservaties met een gedetailleerd scoresysteem door getrainde observatoren in combinatie met gerichte coaching en feedback door experts kansrijke interventies lijken. De pilot die we hier onderzocht hebben bevat dit type interventies, aangevuld met het stellen van concrete doelen en teambeloning. De resultaten wijzen op een significante toename in de gemeten leerkrachtvaardigheid van deelnemende leerkrachten na afloop van de pilot ten opzichte van de meting voor aanvang van de pilot. Meer inzicht in de effecten van meten, coachen en belonen van leerkrachtvaardigheid in de Nederlandse context vergt een experimentele aanpak met scholen die deze interventies wel krijgen en - aselect gekozen - vergelijkbare scholen die niet blootgesteld worden aan deze interventies. Ook is het belangrijk dat - naast effecten op leerkrachtvaardigheden - ook effecten op leerprestaties een of twee jaar na afloop van de interventies worden onderzocht om te zien in hoeverre effecten mogelijk uitdoven of juist enige tijd nodig hebben om zichtbaar te worden. In het buitenland zijn de afgelopen jaren enkele interventies op het vlak van meten van leerkrachtvaardigheden en gerichte coaching en feedback uitgevoerd die positieve effecten hebben gehad op leerlingprestaties (zie

o.a. Taylor en Tyler, 2011 en Allen, 2011). Ook in Nederland wijzen een aantal studies op positieve effecten van gerichte coaching en feedback op leerkrachtvaardigheden en leerlingprestaties (zie o.a. Houtveen en van de Grift, 2007 en 2012, en Houtveen et al., 2004).[30]

Uit eerder onderzoek blijkt dat er ruimte is voor verbetering van leerkrachtvaardigheden van de huidige lerarenpopulatie in Nederland. Dit maakt effectonderzoek naar de wijze waarop verbetering van leerkrachtvaardigheden effectief kan worden vormgegeven urgent. De Onderwijsinspectie laat bijvoorbeeld op basis van een grote hoeveelheid lesobservaties zien dat 1 op de 8 leraren in het Nederlandse basisonderwijs onvoldoende scoort op ten minste een van de basisvaardigheden. Op de complexere vaardigheden, zoals het afstemmen van instructie op verschillen tussen leerlingen, is de beheersing vaker onvoldoende. Op zwakke en zeer zwakke basisscholen is de beheersing van leerkrachtvaardigheden het minst. Er blijken hier tweeënhalf keer zo veel leraren te zijn die niet alle drie de basisvaardigheden beheersen en twee keer zo veel die niet alle complexere vaardigheden tijdens de les laten zien (Inspectie van het Onderwijs, 2011). Uit een onderzoek onder Rotterdamse schoolleiders in het basisonderwijs (Researchned, 2012) blijkt dat bijna 4 op de 10 schoolleiders problemen ervaart met de kwaliteit bij minstens een kwart van hun leerkrachten op het vlak van pedagogisch-didactische kennis en vaardigheden. Ook benoemen bestuurders dat de uitdaging op dit gebied vooral groot is bij de achterstandsscholen. Gezien deze zorgen omtrent de bekwaamheid van leraren is het opvallend dat een kwart van de leraren in het primair onderwijs aangeeft het afgelopen jaar geen functionerings- of beoordelingsgesprek te hebben gevoerd, en dat eenderde van de leraren aangeeft (nog) geen bekwaamheidsdossier te hebben (Research voor Beleid, 2011).

Een interessante vervolgvraag is daarnaast hoe de kosten en baten van coaching op leerkrachtvaardigheden zich verhouden tot die van formele op- en bijscholing van leraren. Vanuit het beleid lijkt de laatste jaren vooral ingezet te worden op het stimuleren van formele bij- en omscholing, getuige de middelen die zijn ingezet voor lerarenbeurzen.

Tot slot, er zijn diverse aanwijzingen dat de schoolleider een belangrijke speler is bij het verbeteren van de onderwijskwaliteit en meer specifiek de kwaliteit van leraren op hun scholen. Branch e.a. (2012) laten zien dat er significante verschillen bestaan in de toegevoegde waarde van schoolleiders in termen van leerwinst van leerlingen. Van der Grift (2001) constateert dat schoolleiders op (zeer) zwakke scholen doorgaans

[30] Het betreft allen studies met controlegroepen van scholen/leerkrachten waar dergelijke interventies niet hebben plaatsgevonden. Het betreft echter geen gerandomiseerde experimenten, dat wil zeggen met aselecte toewijzing aan een experimentele of een controleconditie.

onvoldoende ondersteuning geven aan het verbeterproces van zwakke leraren. De Inspectie (2011) constateert dat op zwakke scholen veel minder gedaan wordt aan scholing dan op overige scholen. Rand (2012) benadrukt het belang van training van schoolleiders in het observeren van lessen van hun leraren en het geven van feedback. Onderzoek van de Inspectie door Maarten Balvers laat zien dat leraren op scholen waar feedback plaatsvindt door de schoolleider 0.24 standaard deviatie hogere scores laten zien op metingen van didactische vaardigheden (presentatie op Onderwijsresearchdagen 2012). Woessmann e.a. (2009) vinden op basis van PISA-data dat leerlingen van scholen waar schoolleiders lesobservaties doen significant beter presteren dan op scholen waar schoolleiders dit niet doen. Een vervolgvraag voor onderzoek is dan ook hoe schoolleiders effectief geprikkeld en gefaciliteerd kunnen worden om de kwaliteit van hun lerarenkorps en het geboden onderwijs verder te verbeteren.

Literatuur

Aaronson, D., L. Barrow, en W. Sander, 2007, Teachers and Student Achievement in the Chicago Public High Schools, *Journal of Labor Economics*, vol. 25, nr. 1, pp. 95–135.

Allen, J., R. Pianta, G. Mikami en J. Lun, 2011, An interaction-based approach to enhancing secondary school instruction and student achievement, *Science*, vol. 333, pp 1034−37.

Armour, D., 1976, Analysis of the school preferred reading program in selected Los Angeles minority schools, Santa Monica, Ca: Rand Corporation.

Branch, G., E. Hanushek, en S. Rivkin, 2012, Estimating the effect of leaders on public sector productivity: the case of school principals, *NBER Working Paper*, nr. 17803.

Chetty, R., N. Friedman, en J. Rockoff, The long-term impact of teachers: teacher value-added and student outcomes in adulthood, *NBER Working Paper*, nr. 17699.

Fryer, R., 2011, Financial Incentives and Student Achievement: Evidence from Randomized Trials, *Quarterly Journal of Economics*, 126(4), 1755-1798.

Fryer, R., forthcoming, Teacher incentives and student achievement: Evidence from New York City Public Schools, *Journal of Labor Economics*, forthcoming.

Grift, W. van de, en J. Lam, 1998, Het didactisch handelen in het basisonderwijs, *Tijdschrift voor Onderwijsresearch*, vol. 23, nr. 3, pp. 224-241.

Glazerman, S. en A. Seifullah, 2010, An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report, Mathematica Policy Research Inc., Washington D.C.

Grift, W. van de, 2007, Quality of teaching in four European countries: a review of the literature and application of an assessment instrument, *Educational Research*, vol. 49, nr. 2, pp. 127-152.

Grift, W. van de, M. van der Wal, en M. Torenbeek, 2011, Ontwikkeling in de pedagogisch didactische vaardigheid van leraren in het basisonderwijs, *Pedagogische Studiën*, vol. 88, pp. 416-432.

Hanushek, E., and S. Rivkin, 2010, Using Value-Added Measures of Teacher Quality, *American Economic Review*, vol. 100, nr. 2, pp. 267–71.

Holtzapple, E., 2003, Criterion-related validity evidence for a standards-based teacher evaluation system, *Journal of Personnel Evaluation in Education*, vol. 17, nr. 3, pp. 207-219.

Houtveen, A., W. van de Grift, en B. Creemers, 2004, Effective school improvement in mathematics, *School Effectiveness and School Improvement*, vol. 15, nr. 3-4, pp. 37-376.

Houtveen, A., en W. van de Grift, 2007, Effects of metacognitive strategy instruction and instruction time on reading comprehension, *School Effectiveness and School Improvement*, vol. 18, nr. 2, pp. 173-190.

Houtveen, A., en W. van de Grift, 2012, Improving reading achievements of struggling learners, *School Effectiveness and School Improvement*, vol. 23, nr. 1, pp. 71-93.

Ilgen, D. R., Fisher, C. D., en Taylor, M. S., 1979, Consequences of individual feedback on behavior in organizations, *Journal of Applied Psychology*, vol. 64, pp.349-371.

Jacob, B., en L. Lefgren, 2008, Principals as agents: Subjective performance measurement in education, *Journal of Labor Economics*, vol. 26, nr. 1, pp. 101-136.

Kane, T., and Douglas O. Staiger. 2008, Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, *National Bureau of Economic Research Working Paper*, nr. 14601.

Kane, T., E. Taylor, J. Tyler, en A. Wooten, 2011, Identifying effective classroom practices using student achievement data, *Journal of Human Resources*, vol 46, nr. 3, pp. 587-613.

Kane, T., and D. Staiger, 2012, Gathering feedback for teaching: combining high-quality observations with students surveys and achievement gains, *Measures of Effective Teaching Research Paper*.

Locke, E., en G. Latham, 2002, Building a practically useful theory of goal setting and task motivation: a 35-year odyssey, *American Psychologist*, vol. 57, nr. 9, pp. 705-717.

Milanowski, A., 2004, The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati, *Peabody Journal of Education*, vol. 79, nr. 4, pp. 33-53.

Rivkin, S., E. Hanushek, en J. Kain. 2005, Teachers, Schools and Academic Achievement, *Econometrica*, vol. 73, nr. 2, pp. 417–58.

Rockoff, J., 2004, The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data, *American Economic Review*, vol. 94, nr. 2:247-252.

Rockoff, J., D. Staiger, T. Kane, en E. Taylor, 2010, Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools, *National Bureau of Economic Research Working Paper,* nr. 16240.

Rockoff, J., en C. Speroni, 2010, Subjective and objective evaluations of teacher effectiveness, *American Economic Review, Papers & Proceedings*, vol. 100, pp 261-266.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., en Stecher, B., 2010, Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Staiger, D., en J. Rockoff, 2010, Searching for effective teachers with imperfect information, *Journal of Economic Perspectives*, vol. 23, nr. 3, pp 97-118.

Webbink, D., I. de Wolf, L. Woessmann, R. van Elk, B. Minne, en M. van der Steeg, 2009, Wat is bekend over de effecten van kenmerken van onderwijsstelsels? Een literatuurstudie, *CPB Document*, nr. 187.

Weisberg, D., Sexton, S., Mulhern, J. en Keeling, D., 2009, The widget effect: Our national failure to acknowledge and act on teacher effectiveness. New York City, N.Y.: The New Teacher Project.

Woessmann, L., E. Luedemann, G. Schuetz and M. R.West, 2009, School Accountability, Autonomy and Choice around the World,Edward Elgar, Cheltenham.

Appendix A Beheersing 18 gedragsindicatoren Kijkwijzer bij nulmeting

Gedragsindicator	Gemiddelde beheersing (% van onderliggende gedragsaspecten)	Standaarddeviatie
Laat leerlingen reflecteren op (diverse) oplossingsstrategieën	32	39
Stimuleert het hanteren van controleactiviteiten	44	39
Gaat na of de lesdoelen bereikt zijn	46	35
Bevordert het toepassen van het geleerde	48	40
Stemt de verwerking van de leerstof af op relevante verschillen tussen leerlingen	54	36
Stimuleert reflectie door middel van interactieve instructie- en werkvormen	54	40
Geeft feedback aan leerlingen	55	26
Leert leerlingen strategieën voor denken en leren	57	34
Biedt zwakke leerlingen extra leer- en instructietijd	58	39
Maakt voor leerlingen de opbouw van de les inzichtelijk	58	34
Verduidelijkt bij aanvang van de les de lesdoelen	59	36
Maakt duidelijk hoe de les aansluit aan bij voorgaande lessen	63	35
Stemt instructie af op relevante verschillen tussen leerlingen	65	34
Laat de les verlopen volgens een adequate planning	69	18
Laat leerlingen hardop denken	73	41
Geeft expliciet blijk van hoge verwachtingen	74	35
Geeft duidelijke uitleg van de leerstof en opdrachten	77	25
Besteedt de geplande tijd daadwerkelijk aan het lesdoel	79	32
Aantal leerkrachten	125	

Appendix B Model relatie leerkrachtvaardigheden en leerprestaties

Model dat wordt geschat in tabel 3.1

We gebruiken de specificatie die o.a. wordt gebruikt door Rockoff en Sperroni (2010) om het effect van een teacher evaluation system (TES), zoals de Kijkwijzer, op toetsscores te schatten:

$$Y_t = \beta_0 + \beta_1 Kijkwijzer + \beta_2 Y_{t-1} + \beta_3 Y_{t-1}^2 + \beta_4 LK + \beta_5 KK + \chi_s + \varphi_l + \varepsilon$$

Hierin is Y de toetsscore (rekenen, spelling, lezen) aan het eind van het schooljaar 2011/2012, Y_{t-1} de toetsscore van vorig jaar (uit schooljaar 2010/2011), zijn de LK leerling-karakteristieken, de KK klaskararakteristieken, de χ_s school fixed effects (dummies voor de zeven scholen), de φ_l leerjaar fixed effects (dummies voor de leerjaren 1 t/m 8) en ε de storingsterm. We controleren voor school- en leerjaar fixed effects en toetsscore vorig jaar en toetsscore vorig jaar kwadraat. Wat betreft de leerlingkarakteristieken controleren we voor leeftijd, leeftijd kwadraat, dummy's voor geslacht, gewicht, Nederlandse nationaliteit, afkomstig uit eenoudergezin en zittenblijven. Verder controleren we voor de volgende klassenkarakteristieken: klassengrootte, percentage meisjes, percentage gewichtenleerlingen, aantal leraren op de klas en dummy voor combinatieklas. Bij meerdere leraren op een klas wegen we de kijkwijzer in met de bezetting van de leraren op de klas. We gebruiken een dummy voor als de toetsscore in het vorige jaar ontbreekt. De toetsscore wordt in dat geval op het gemiddelde gezet. We controleren ook voor ervaring en ervaring kwadraat.

Appendix C Beschrijvende statistiek

Tabel C.1 Beschrijvende statistiek leerlingen, klassen en leraren

Kenmerken leraren	Percentage
Vrouw	88
Hoogst afgeronde opleiding hbo	98
Vooropleiding MBO	49
Ervaring in onderwijs (jaar)	19
5 jaar of minder ervaring in onderwijs	13
Aanstelling (% FTE)	87
In schaal LB	7
Aantal leraren	125
Kenmerken klassen	
Gemiddelde observatie score Kijkwijzer	52
Klassengrootte (aantal leerlingen)	24
Meisjes	51
Gewichtenleerling	39
Combinatieklas (meerdere leerjaren)*	20
Meerdere leraren op 1 klas	33
Aantal klassen	99
Kenmerken leerlingen	
Meisje	50
Gewicht 0.3	16
Gewicht 1.2	22
Uit eenoudergezin	49
Nederlandse nationaliteit	90
Zittenblijven**	7
Aantal leerlingen	2110
* Dit betreft voornamelijk gecombineerde groep 1/2 klassen.	
** Dit betreft ook kinderen die een jaar langer ‘kleuteren’.	

Appendix D 75 gedragsaspecten Kijkwijzer

Deze appendix bevat een overzicht van de 75 in de pilot gebruikte gedragsaspecten van de Amsterdamse kijkwijzer.

		2. PEDAGOGISCH COMPETENT
2.1		Geeft expliciet blijk van hoge verwachtingen
	2.1a	Spreekt positieve verwachtingen uit naar leerlingen over wat zij aankunnen
	2.1b	Weet van iedere leerling wat hij van hen kan verwachten en communiceert dit naar hen
	2.1c	Bevestigt het als leerlingen verwachtingen waarmaken
	2.1d	Bemoedigt leerlingen als zij verwachtingen niet waarmaken
2.3		Stemt instructie af op relevante verschillen tussen leerlingen
	2.3a	Zet leerlingen die minder instructie nodig hebben (alvast) aan het werk
	2.3b	Geeft aanvullende instructie aan groepjes of individuele leerlingen
	2.3c	Richt zich niet alleen op de middenmoot
	2.3d	Laat zijn instructie aansluiten op de wijze waarop een leerling leert
2.4		Stemt de verwerking van de leerstof af op relevante verschillen tussen leerlingen
	2.4a	Maakt tussen leerlingen verschil in de omvang van opdrachten
	2.4b	Laat sommige leerlingen gebruik maken van hulpmaterialen
	2.4c	Geeft niet alle leerlingen dezelfde tijd voor de opdracht
	2.4d	Geeft leerlingen verwerkingsopdrachten die aansluiten op de wijze waarop zij leren
2.5		Biedt zwakke leerlingen extra leer- en instructietijd
	2.5a	Geeft zwakke leerlingen extra leertijd
	2.5b	Geeft zwakke leerlingen extra oefeningen
	2.5c	Geeft zwakke leerlingen ‘voor’-of ‘na’-instructie
		3. VAKINHOUDELIJK EN DIDACTISCH COMPETENT
3.1		Maakt duidelijk hoe de les aansluit aan bij voorgaande lessen
	3.1a	Bespreekt het voorgaande werk met betrekking tot hetzelfde onderwerp
	3.1b	Haalt relevante voorkennis op en vat deze samen
	3.1c	Vermeldt hoe de les aansluit bij wat voorafgegaan is
	3.1d	Noteert voorkennis op het bord (‘wat weten we al?’)
3.2		Verduidelijkt bij aanvang van de les de lesdoelen
	3.2a	Informeert de leerlingen bij de aanvang van de les over de lesdoelen
	3.2b	Noteert de lesdoelen op het bord (‘wat gaan we leren?’)
	3.2c	Maakt duidelijk wat het doel van de opdrachten is en wat de leerlingen ervan zullen leren
3.3		Maakt voor leerlingen de opbouw van de les inzichtelijk
	3.3a	Gebruikt duidelijk herkenbare componenten in de les (bv. uitleg, begeleid inoefenen, zelfstandig verwerken)
	3.3b	Maakt aan kinderen duidelijk volgens welke stappen de les gaat verlopen
	3.3c	Maakt bij iedere nieuwe stap in de les duidelijk hoe deze past in het totaal
3.4		Geeft duidelijke uitleg van de leerstof en opdrachten
	3.4a	Legt uit in opeenvolgende stappen
	3.4b	Stelt vragen die door leerlingen worden begrepen
	3.4c	Vat van tijd tot tijd de leerstof samen
	3.4d	Zet aanschouwelijke en ondersteunende middelen in die leerdoelen ondersteunen
3.8		Geeft feedback aan leerlingen
	3.8a	Gaat tijdens de instructie na of leerlingen de leerstof hebben begrepen
	3.8b	Gaat tijdens de verwerking na of leerlingen de opdrachten op een juiste manier uitvoeren
	3.8c	Refereert bij feedback expliciet aan de doelen
	3.8d	Refereert bij feedback expliciet aan de fasering van de les of van de opdracht
	3.8e	Geeft feedback op de wijze waarop leerlingen tot hun antwoord komen
	3.8g	Geeft feedback op het sociaal functioneren bij de uitgevoerde taak
3.9		Gaat na of de lesdoelen bereikt zijn
	3.9a	Laat de leerlingen vertellen wat ze geleerd hebben
	3.9b	Vat samen (op het bord) wat kinderen hebben geleerd

	3.9c	Grijpt expliciet terug op doelen
	3.9d	Laat leerlingen vertellen wat goed ging, wat niet goed ging en wat ze de volgende keer anders gaan doen
	3.9e	Gaat na wat de prestaties van de leerlingen zijn
3.10		Stimuleert reflectie door middel van interactieve instructie- en werkvormen
	3.10a	Is niet alleen zelf aan het woord maar stimuleert reflectie door middel van interactie met leerlingen
	3.10b	Gebruikt werkvormen waarbij interactie tussen leerlingen leidt tot reflectie
3.11		Laat leerlingen hardop denken
	3.11a	Geeft leerlingen de gelegenheid hardop oplossingen te bedenken
	3.11b	Vraagt leerlingen oplossingen te verbaliseren
3.12		Leert leerlingen strategieën voor denken en leren
	3.12a	Leert leerlingen oplossingsmethodieken (algoritme, analogie, regeltoepassing)
	3.12b	Leert leerlingen het gebruik van ordeningsmiddelen aan
	3.12c	Geeft leerlingen aanwijzingen voor het oplossen van problemen
	3.12d	Biedt leerlingen checklisten voor het oplossen van problemen
	3.12e	Demonstreert denkstrategieën door mondeling of hardop denken
	3.12f	Vereenvoudigt problemen door ze in stukken te hakken
3.13		Laat leerlingen reflecteren op (diverse) oplossingsstrategieën
	3.13a	Laat leerlingen verschillende oplossingsstrategieën met elkaar vergelijken
	3.13b	Brengt structuur aan in de verschillende oplossingsstrategieën
	3.13c	Laat leerlingen de handigste oplossingsstrategie bepalen
	3.13d	Evalueert de bruikbaarheid van oplossingsstrategieën
	3.13e	Geeft leerlingen niet alleen feedback op het resultaat maar ook op het proces
3.14		Stimuleert het hanteren van controleactiviteiten
	3.14a	Schenkt aandacht aan schattend rekenen en voorspellend lezen
	3.14b	Laat oplossingen relateren aan de context
	3.14c	Stimuleert het gebruik van alternatieve oplossingen en strategieën
3.15		Bevordert het toepassen van het geleerde
	3.15a	Plaatst de leerstof in een betekenisvolle context
	3.15b	Vraagt leerlingen waarvoor het geleerde (ook) gebruikt kan worden
	3.15c	Daagt leerlingen uit het geleerde in andere leergebieden toe te passen
		4. ORGANISATORISCH COMPONENT
4.2		Besteedt de geplande tijd daadwerkelijk aan het lesdoel
	4.2a	Laat geen tijd verloren gaan tijdens de les
	4.2b	Laat geen ‘dode’ momenten ontstaan
	4.2c	Laat de leerlingen niet wachten
	4.2d	Laat zich niet afleiden door irrelevante zaken of gebeurtenissen
	4.2e	Houdt zelf lestijd en lesdoel in de gaten
4.3		Laat de les verlopen volgens een adequate planning
	4.3a	Heeft de les gepland
	4.3b	Stemt zijn planning van de les af op de doelen
	4.3c	Stimuleert leerlingen door te werken
	4.3d	Voorkomt irrelevante uitweidingen
	4.3e	Wisselt de afwisseling in instructie en begeleiding in de groep(en) evenwichtig af
	4.3f	Verdeelt in combinatiegroepen de instructie evenwichtig over beide groepen
	4.3g	Vertelt leerlingen aan welke opdracht of taak zij moeten werken
	4.3h	Vertelt leerlingen hoeveel tijd ze hebben voor taken
	4.3i	Geeft leerlingen relevante opdrachten als zij minder dan de geplande tijd nodig hebben