Centraal Planbureau

CPB Notitie | 28 november 2014

Neveneffecten sturingsinstrumenten

Uitgevoerd op verzoek van de Ambtelijke Commissie Vernieuwing Publieke Belangen

Centraal Planbureau

CPB Notitie

Centraal Planbureau
Van Stolkweg 14
Postbus 80510
2508 GM Den Haag

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

Contactpersoon
Flora Felso
Robin Zoutenbier

Datum: 28 november 2014
Betreft: Neveneffecten sturingsinstrumenten

1 Inleiding

Op 11 september 2013 presenteerde de Commissie Behoorlijk Bestuur haar bevindingen waarin zij het bestaan van een aantal weeffouten constateerde in de politiek-bestuurlijke ordening van de semipublieke sector. Eén van de aanbevelingen van de Commissie is om sturingsinstrumenten tegen het licht te houden voor onbedoelde gedragseffecten:

"De Commissie adviseert het kabinet: [...] 2. Om perverse prikkels te voorkomen de meet- en sturingsinstrumenten altijd gepaard te laten gaan met vereisten van kwaliteit, alsook nieuwe regels en rapportageplichten steeds te toetsen op de te verwachten gedrags- en perverse effecten. "(p.24)

Als voorbeeld noemt de Commissie diplomabekostiging in het hoger onderwijs. Het zou een perverse prikkel geven om snel veel studenten op te leiden ten koste van de kwaliteit van de opleiding. Een tweede voorbeeld gaat in op een verpleeghuis met problemen rond de (papieren) verantwoording van contacten met de patiënten. Een betere verslaglegging van zorgcontacten gaat echter ten koste van de tijd die kan worden besteed aan contact met de bewoners. De gerapporteerde kwaliteit stijgt, maar de tevredenheid van bewoners en verplegers gaat achteruit. De bestuurskundige literatuur noemt dit de 'performance-paradox' (Van Thiel en Leeuw 2002).

Tegen deze achtergrond heeft de Ambtelijke Commissie Vernieuwing Publieke Belangen het CPB gevraagd om een notitie op te stellen over wat de literatuur zegt over averechtse effecten van verschillende meet- en sturingsinstrumenten. De focus in deze notitie ligt op de inzet van sturingsinstrumenten bij professionals en bij instellingen. Centraal in deze notitie staan de omstandigheden die sturingsinstrumenten minder effectief maken.

De theoretische literatuur schetst een aantal redenen waarom sturingsinstrumenten minder effectief kunnen zijn. Zo kunnen er problemen zijn bij de toepassing van sturingsinstrumenten door moeilijk meetbare output, onzekerheid bij productie, meerdere taken, conflicterende sturingsinstrumenten of als het werk in teamverband wordt uitgevoerd. Een andere mogelijke oorzaak voor ineffectiviteit is strategisch gedrag (*gaming*). Voorbeelden zijn *upcoding*, *downcoding* of andere manipulatie van de registratie, *cherry picking* en *cream skimming*. Ten slotte kan verdringing van intrinsieke motivatie ertoe leiden dat de resultaten anders uitpakken dan beoogd. Een (financiële) prikkel kan door de werknemer als een signaal worden gezien dat de werkgever hem niet als een intrinsiek gemotiveerd of als een kundige persoon ziet, wat de professional ontmoedigt.

De empirische literatuur over effectiviteit van meet- en sturingsinstrumenten is omvangrijk, maar niet alles is even relevant. Dat resultaatafhankelijke beloning bij een productiebedrijf werkt zegt nog niet zoveel over de inzetbaarheid van vergelijkbare beloningsstructuren in de zorg of in het onderwijs. Deze notitie richt zich daarom nadrukkelijk op empirische studies die relevant zijn in de context van de semipublieke sector en op instrumenten die door de overheid (kunnen) worden ingezet.

De meeste empirische studies bespreken of een specifiek instrument al dan niet heeft gewerkt in een bepaalde setting, maar slechts zelden wordt expliciet stilgestaan bij (onbedoelde) neveneffecten. Daarmee geeft de empirische literatuur slechts indirect een beeld van neveneffecten. Een studie die een positief netto effect laat zien, toont aan dat de averechtse effecten niet opwegen tegen de positieve effecten. Dit geldt strikt genomen voor de activiteit die gestimuleerd wordt. Soms is de gestimuleerde activiteit echter een afgeleide van een bredere doelstelling. Zelfs als er meer van de gestimuleerde activiteit teweeg wordt gebracht, is dat niet altijd in lijn met de bredere doelstelling. Het doel van het overzicht van de empirische literatuur is om een beeld te schetsen van de effectiviteit van sturingsinstrumenten en de bijwerkingen die (meestal in de marge van) effectiviteitstudies worden benoemd.

De opbouw van deze notitie is als volgt. In paragraaf 2 wordt de theoretische discussie over de effectiviteit van sturingsinstrumenten belicht. Zowel financiële als niet-financiële prikkels komen aan de orde. Paragraaf 3 stipt de factoren aan die averechtse effecten kunnen veroorzaken. Paragraaf 4 geeft een overzicht van de meest relevante empirische studies naar effectiviteit van sturingsinstrumenten. Eerst komen financiële prikkels aan de orde. Daarna worden de lessen van de kwalitatief goede empirische studies in de sfeer van de zorg en het onderwijs besproken. Hierbij staan we ook stil bij de gerapporteerde gedrags- en perverse effecten van zowel financiële- als niet-financiële prikkels. Paragraaf 5 bevat de conclusies en aanbevelingen.

2 Sturingsinstrumenten

Wanneer een opdrachtgever taken naar een uitvoerder delegeert, kunnen sturingsinstrumenten de verschillende belangen van de opdrachtgever en uitvoerder (meer) op één lijn brengen. Denk bijvoorbeeld aan de werkgever die de baten van inspanning verkrijgt, terwijl de werknemer de kosten van inspanning draagt. Een ander voorbeeld betreft een situatie waarin de overheid de kosten van het leveren van een publieke dienst wil verlagen, maar de instelling de geleverde hoeveelheid van de dienst wil verhogen. In beide gevallen heeft een sturingsinstrument als doel het gedrag van de uitvoerder in lijn te brengen met het belang van de opdrachtgever.

Opdrachtgevers hebben de keuze uit verschillende sturingsinstrumenten. Instrumenten kunnen zowel financieel als niet-financieel van aard zijn. Ook kan het delegeren van taken op verschillende niveaus plaatsvinden. Zo kan een manager een taak delegeren aan een medewerker, of kan de overheid taken delegeren aan een (semipublieke) instelling. De directie van die instelling vertaalt de prikkel van buitenaf vervolgens richting de professional die daadwerkelijk de taak uitvoert.

Tabel 2.1 geeft een (niet-uitputtend) overzicht van sturingsinstrumenten, ingedeeld naar aard van instrument (financieel of niet-financieel) en niveau waarop het instrument wordt ingezet (individueel of groepsniveau). Er zijn parallellen tussen instrumenten voor individuen en groepen. Zo vertoont resultaatafhankelijke beloning voor professionals veel gelijkenis met outputfinanciering voor instellingen en kunnen de prestaties van zowel professionals als instellingen gemonitord worden. Er kunnen ook verschillen zijn in de werking van instrumenten op individueel dan wel instellingsniveau. Denk bijvoorbeeld aan het meeliftgedrag dat teambeloning parten kan spelen, maar geen rol speelt bij sturingsinstrumenten op individueel niveau.

Tabel 2.1 Voorbeelden meet- en sturingsinstrumenten

	Individuele prikkels (bijv. professionals binnen instellingen)	**Teambeloning** (bijv. aansturing van instellingen of teams binnen instellingen)
Financiële instrumenten	Resultaatafhankelijke beloning (bijv. stuksbeloning, bonus, periodieken)	Resultaatafhankelijke beloning (bijv. outputfinanciering, prestatiesturing, budgetsturing, boete)
Niet-financiële instrumenten	Monitoren Targets Verantwoording Autonomie/*Empowerment* Feedback Waardering	Monitoren Targets Verantwoording/*Accountability* Autonomie/Taakopdracht Feedback Transparantie (zoals *naming and shaming*, benchmarking)

Hieronder beginnen we met de discussie over financiële prikkels, en meer specifiek over resultaatafhankelijke beloning. Het tweede deel van deze paragraaf geeft een overzicht van niet-financiële prikkels en bespreekt in welke mate deze prikkels vergelijkbaar zijn met financiële prikkels.

Resultaatafhankelijke beloning

Een belangrijk instrument om het gedrag van de uitvoerder van een taak te sturen is het belonen (of straffen) van de uitvoerder op basis van het behaalde resultaat. Deze vorm van sturen is uitgebreid bestudeerd in de economisch-wetenschappelijke literatuur met behulp van het principaal-agent raamwerk (zie Prendergast 1999 voor een overzicht). In dit raamwerk worden situaties geanalyseerd waarbij een principaal (werkgever) een taak delegeert aan een agent (werknemer). Een belangrijke veronderstelling bij deze modellen is dat de werknemer een informatievoorsprong heeft op de werkgever: de werkgever is minder goed in staat om de handelingen van de werknemer te observeren en verifiëren dan de werknemer zelf.[1] De werknemer kan meerwaarde creëren bij de uitvoering van deze taak door meer inspanning uit te oefenen. Het leveren van inspanning is echter kostbaar voor de werknemer. Wanneer een werkgever de inspanning van de werknemer niet perfect kan observeren, en deze beloont op basis van een vast bedrag, zal de werknemer minder inspanning uitoefenen (moreel gevaar). De principaal kan een agent ook laten delen in de opbrengst van zijn of haar handelen. Op deze manier brengt resultaatafhankelijke beloning de belangen van de principaal en agent meer op één lijn, het is in het belang van beiden een goed resultaat te behalen.

Het principaal-agent raamwerk voorspelt twee belangrijke effecten van het invoeren van resultaatafhankelijke beloning. Het eerste effect is het prikkeleffect: wanneer de werknemer niet alleen de kosten van inspanning draagt, maar ook (een deel) van de baten van inspanning krijgt, zal de werknemer een andere afweging maken over hoeveel in te spannen. Het selectie-effect is het tweede effect van resultaatafhankelijke beloning. De totale beloning zal voor productieve werknemers hoger zijn dan voor minder productieve werknemers; productievere werknemers zullen immers betere resultaten behalen en daarmee een hogere resultaatafhankelijke beloning ontvangen. Productieve werknemers zullen dus sneller geneigd zijn een baan met resultaatafhankelijke beloning te accepteren dan minder productieve werknemers.

Een empirische studie naar de invoering van stuksbeloning bij een bedrijf dat autoruiten zet, illustreert deze effecten (Lazear 2000). De studie toont aan dat de totale productiviteit na het invoeren van stuksbeloning met 44% stijgt. Ongeveer de helft van dit effect wordt veroorzaakt doordat minder productieve werknemers het bedrijf verlaten en productievere werknemers aangetrokken worden (selectie-effect).

[1] Wanneer deze informatievoorsprong afwezig is, kan de werkgever de werknemer simpelweg belonen voor het uitvoeren van de juiste acties, onafhankelijk van het uiteindelijke resultaat.

De overige helft van het effect kan verklaard worden doordat de huidige werknemers harder zijn gaan werken (prikkeleffect).

Niet-financiële instrumenten

Bij sturingsinstrumenten zullen veel mensen direct denken aan financiële sturingsinstrumenten, bijvoorbeeld een geldbonus voor goede prestaties, een loonsverhoging, outputbekostiging etc. Sturingsinstrumenten kunnen echter ook niet-financieel van aard zijn. Bij niet-financiële prikkels gaat het vaak om een combinatie van monitoren, verantwoorden, stellen van targets, transparantie en het geven van feedback, zonder een directe financiële beloning. Het voorbeeld uit het rapport van de Commissie Halsema over het verzorgingshuis dat de papierenverantwoording tracht te verbeteren kan ook worden gezien als een niet-financieel sturingsinstrument. Daar worden normtijden voor deeltaken opgelegd om zo het proces te sturen. Een andere (dwingende vorm) van een target is wetgeving of richtlijnen; de naleving van wetten en regels wordt immers ook intensief gemonitord (en kan achteraf tot veroordeling leiden).

Het monitoren van prestaties houdt in dat een principaal de voortgang van een agent volgt, dit kan de principaal doen in combinatie met andere niet-financiële instrumenten (maar ook in combinatie met andere financiële instrumenten). Bengstsson en Engström (2013) bestuderen bij non-profit organisaties welk effect het monitoren van projectrapportages heeft op de uiteindelijke uitkomsten van projecten. Bij een willekeurig gekozen groep projecten kondigt de Zweedse goededoelenautoriteit vooraf aan dat ze de rapportages zullen monitoren op de juistheid van de uitkomsten. De resultaten van het onderzoek tonen aan dat de projecten waar de controle vooraf is aangekondigd, betere resultaten laten zien, minder uitgeven en minder fouten maken dan de projecten waarbij controle niet van tevoren is aangekondigd.

De gemonitorde resultaten kunnen ook worden teruggekoppeld naar een professional of instelling in de vorm van feedback; als de resultaten openbaar worden gemaakt spreken we van transparantie. Er zijn echter ook vormen van transparantie waarbij de resultaten niet direct herleidbaar zijn naar de uitvoerder, althans niet voor het grote publiek. Dit is bijvoorbeeld het geval bij de geanonimiseerde rapporten van fraude in de zorg. Een groot aantal studies heeft aangetoond dat het geven van feedback een positief effect heeft op de uiteindelijke prestatie, zelfs als daar geen financiële prikkels tegenover staan.[2] De feedback in deze experimenten is vaak relatief, dat wil zeggen, professionals worden geïnformeerd over hun prestatie ten opzichte van een vooropgestelde target of prestaties van collega's. In een vergelijkbare studie tonen Bradler et al. (2013) aan dat het geven van publieke erkenning voor goede resultaten ook een positief effect kan hebben op prestaties.

[2] Azmat en Iriberri (2010a), (2010b), Blanes i Vidal en Nossol (2011), Kuhnen en Tymula (2012), Tran en Zeckhauser (2012), Delfgaauw et al. (2013) en Gerhards en Siemer (2014).

In tegenstelling tot het nauwlettend volgen van de prestaties van een agent kan ook autonomie (taakopdracht) geboden worden, of *empowerment* nagestreefd worden. Bij autonomie of empowerment krijgt de agent meer speelruimte en bevoegdheden om prioriteiten te stellen. Hanushek et al. (2013) onderzoeken het effect van autonomie van scholen op de prestaties van scholieren. Zij vinden een positieve relatie tussen de hoeveelheid speelruimte die een school wordt geboden en de prestatie van de school, gemeten via PISA-scores van studenten.

3 Beperkingen en averechtse effecten

Deze paragraaf schetst een aantal situaties waarin sturingsinstrumenten (financieel of niet-financieel) mogelijk minder effectief zijn. De oorzaak ligt deels aan problemen met de toepassing: soms is het lastig om prikkels te introduceren. Daarnaast kan strategisch gedrag of de verdringing van intrinsieke motivatie bij professionals ertoe leiden dat prikkels averechts werken.

Moeilijk meetbare output

Het standaard principaal-agent raamwerk veronderstelt dat de acties van een agent niet goed te meten zijn, maar het uiteindelijke resultaat wel. Er bestaan echter taken waarbij ook het meten van het resultaat lastig is. Denk bijvoorbeeld aan curatieve zorg, waar het beoogde resultaat vaak een verbetering in gezondheid is, maar, de gezondheidswinst als gevolg van een behandeling bij een patiënt lastig te meten en verifiëren is. Het gevolg van moeilijk meetbare output is dat er geen duidelijke relatie bestaat tussen de actie van de agent en het resultaat; sturingsinstrumenten zijn dan minder effectief. De agent zal immers veronderstellen dat het maken van de juiste keuzes maar weinig effect zal hebben op zijn beoordeling (of beloning). Soms kan, als het resultaat niet direct te meten is, gestuurd worden op afgeleiden van het resultaat (denk aan cito-scores in het onderwijs als afgeleide van de ontwikkeling van leerlingen). Het gevaar is dat belonen op cito-scores gedrag uitlokt wat de cito-scores maximaliseert maar verder niet ten goede komt van het onderwijs (dit is vooral een probleem als het resultaat meerdere dimensies heeft, zie ook de alinea over ‘meerdere taken’ hieronder).

Onzeker productieproces

Een gebrek aan een direct verband tussen de acties van een agent en het resultaat kan ook voortkomen uit onzekerheid in productie. Bijvoorbeeld wanneer de juiste acties van een agent de kans op een goede uitkomst verhogen maar het behalen van het resultaat voor een groot deel afhangt van willekeur of omgevingsfactoren. Ook hier geldt dat de agent weinig invloed kan uitoefenen op het resultaat met zijn acties, wat de inzet van sturingsinstrumenten mogelijk minder effectief maakt. Denk weer aan het voorbeeld van een behandeling in de gezondheidszorg; naast het handelen van de specialist bepalen ook de fysieke gesteldheid van de patiënt of andere

complicaties de kans van slagen. Als de kans op een succesvolle behandeling meer afhangt van de fysieke gesteldheid van de patiënt of complicaties dan van het handelen van de arts, dan heeft de inzet van sturingsinstrumenten minder effect.

Meerdere taken

Als de output meerdere dimensies heeft, zoals kenmerkend is voor veel (semi)publieke sectoren, kan het toepassen van sturingsinstrumenten ertoe leiden dat de agent te weinig aandacht schenkt aan de dimensies die relatief minder worden beloond of minder goed meetbaar zijn (Hölmstrom en Milgrom 1991 en Baker 1992). Bijvoorbeeld, de sociale en niet-cognitieve ontwikkeling van leerlingen is moeilijk te meten (en dus te belonen). Wanneer leraren beloond worden op basis van testscores zullen zij mogelijk geneigd zijn om meer (of zelfs te veel) aandacht te geven aan het verbeteren van het kennisniveau van leerlingen ten koste van de sociale en niet-cognitieve ontwikkeling (ook wel 'teaching to the test' genoemd). In hoeverre de prestaties op de verschillende taken beloond worden, hangt in de praktijk vaak af van het gemak waarmee de prestaties gemeten kunnen worden. Dit is dan ook de reden dat taken met meerdere dimensies minder vaak worden beloond op basis van resultaatafhankelijke beloning (Prendergast 1999: p.9).

Conflicterende sturingsinstrumenten

Interactie tussen meerdere sturingsinstrumenten kan ook de effectiviteit van individuele instrumenten belemmeren. Denk hierbij aan een sturingsinstrument gericht op kwaliteitsverbetering en een ander instrument gericht op kostenbeheersing. Een specifiek voorbeeld wordt beschreven door Aalbers et al. (2013). Hij onderzoekt het effect van milieubeleidsinstrumenten op het reduceren van CO2 uitstoot. De totale hoeveelheid emissierechten is vastgesteld via het Europese Emissiehandelssysteem, aanvullende milieubeleidsinstrumenten gericht op het reduceren van C02 uitstoot hebben geen additioneel effect op de totale hoeveelheid uitstoot. Wel kunnen deze instrumenten duurzame innovatie of het verbeteren van de energiezekerheid verbeteren.

De kans op conflicterende sturingsinstrumenten is groter als er meerdere principalen zijn. In de context van semipublieke organisaties valt te denken aan het politiekverantwoordelijke ministerie, toezichthouders zoals ACM, maar ook belangenorganisaties zoals patiëntenverenigingen. Als deze meerdere principalen het gebruik van sturingsinstrumenten niet onderling afstemmen, kan dit nadelige gevolgen hebben voor de uitwerking van deze instrumenten.

Werken in teamverband

In sommige werksituaties is samenwerken erg belangrijk, bijvoorbeeld omdat meerdere experts samen tot een resultaat moeten komen. Werken in teamverband gaat niet altijd goed samen met resultaatafhankelijke beloning (Alchian en Demsetz 1972). Als individuele prestaties beloond worden, kan dit een nadelig effect hebben op de prestatie van het team, omdat werknemers minder geneigd zijn om hun

collega's te helpen. Het geven van prikkels op teamniveau kan dan een uitkomst zijn (zie de paragraaf over beperkingen individuele- versus teamprikkels).

Eigenschappen semipublieke instellingen maakt gebruik van financiële prikkels lastig

Dixit (2002) stelt dat de standaardtheorie over resultaatafhankelijke beloning (zie paragraaf 2) mogelijk meer van toepassing is op organisaties in de private sector dan in de publieke sector. Dixit beschrijft een aantal belangrijke kenmerken van organisaties, kenmerken die veel voorkomen bij (semi)publieke organisaties, die resultaatafhankelijke beloning mogelijk minder effectief maken (zie ook Bijlsma en de Bijl 2013 voor een overzicht van de economische kenmerken van semipublieke instellingen). Ten eerste zijn prestaties in (semi)publieke organisaties vaak moeilijk te meten en kwantificeren. Ten tweede komt het resultaat vaak voort uit meerdere taken en heeft het resultaat meerdere dimensies. Ten derde wordt vaak in teamverband gewerkt om een resultaat te behalen. Als laatste geldt dat semipublieke organisaties vaak meerdere principalen hebben (Dixit 1997).

Strategisch gedrag

Naast de aard van het werk waarbij resultaatafhankelijke beloning minder effectief is als sturingsmiddel, zijn er ook situaties waarin het belonen van resultaten averechts kan werken. In de literatuur worden twee belangrijke oorzaken van averechtse effecten aangewezen, strategisch gedrag (ook wel *gaming of incentives* genoemd) en verdringing van intrinsieke motivatie door extrinsieke motivatie. We beginnen met de belangrijkste voorbeelden van strategisch gedrag.

Taken kunnen verschillen in de complexiteit van de doelstellingen. Bijvoorbeeld, omdat het resultaat meerdere dimensies heeft. Als doelstellingen complex zijn, is het niet altijd mogelijk om alle facetten van het beoogde resultaat nauwkeurig te meten. Het gevolg is dat de agent de meetinstrumenten zo kan manipuleren of verstoren dat de beloning (of beoordeling) hoger wordt, maar de waarde van de uitgevoerde handelingen voor de principaal niet. Dit fenomeen wordt aangeduid als strategisch gedrag of *gaming*. Zo kan de agent bijvoorbeeld proberen de resultaten beter voor te doen dan ze daadwerkelijk zijn (bijvoorbeeld via *upcoding*), zich richten op de taken of klanten die het meeste opleveren voor de agent (*cream skimming*), of de timing van prestaties beïnvloeden. Ook bij andere vormen van sturing kan strategisch gedrag een probleem zijn. Bij relatieve beloning (behaalde resultaat ten opzichte van directe collega's) is het bijvoorbeeld mogelijk om collega's te saboteren in plaats van harder te werken. Bij subjectieve beoordeling door de manager kunnen medewerkers veel tijd besteden aan het opbouwen van een goede band met de manager in plaats van meer tijd te besteden aan de daadwerkelijke taak. Hieronder volgt een aantal voorbeelden van situaties waarin sturingsinstrumenten leiden tot strategisch gedrag.

Upcoding komt bijvoorbeeld voor in de zorg, als artsen diagnoses registreren die beloond worden met een hogere prijs. Silverman en Skinner (2004) illustreren dit door te laten zien dat longontsteking vaker als diagnose wordt gesteld wanneer aan de behandeling van deze aandoening een hogere vergoeding wordt toegekend dan

aan andere aandoeningen gerelateerd aan ademhalingsproblemen. Daarnaast laat deze studie zien dat *upcoding* een groter probleem is als professionals strenger worden afgerekend op resultaten: *upcoding* komt vaker voor bij for-profit ziekenhuizen dan bij non-profit ziekenhuizen (23 procentpunt ten opzichte van 10 procentpunt). De hoogste stijging in het registreren van de duurste diagnose is echter gemeten bij ziekenhuizen in de overgangsfase van non-profit naar for-profit (37 procentpunt).

Cream skimming is het selecteren op eenvoudig uit te voeren taken. Een voorbeeld, Dranove et al. (2003) laten zien dat het openbaar maken van prestaties van zorgaanbieders in de Verenigde Staten ertoe heeft geleid dat aanbieders proberen te voorkomen dat ze moeilijk te genezen patiënten opnemen en juist eenvoudig te genezen patiënten proberen aan te trekken.

Een andere vorm van strategisch gedrag is het manipuleren van de timing van prestaties. Dit speelt in situaties waarbij de professional of instelling discretie heeft over de timing van een prestatie. Courty en Marschke (2004) bestuderen arbeidsbureaus (Job Participation Training Act) die als doel hebben het opleiden en klaarstomen van werklozen voor de arbeidsmarkt. Deze arbeidsbureaus worden beloond op basis van de prestaties van hun cliënten (werkzame status, salaris, werkzame uren) op het moment van afstuderen bij het arbeidsbureau. Courty en Marschke vinden dat arbeidsbureaus de timing van het melden van slagen gebruiken om een zo hoog mogelijke beoordeling te krijgen. Bij deelnemers met een goed salaris en veel werkzame uren wordt snel een datum van slagen vastgelegd, terwijl bij deelnemers met slechte resultaten zolang mogelijk gewacht wordt met het toekennen van een datum (in de hoop dat de resultaten nog verbeteren). Verder laten zij ook zien dat deze verandering in timing niet ten goede komt aan de deelnemers van de trainingsprogramma's.[3] Een ander belangrijk effect van de beloning was dat arbeidsbureaus probeerden te selecteren op de kandidaten met de meeste potentie (*cream skimming*). Het gevolg was dat juist de werklozen die de hulp het hardste nodig hadden de minste hulp kregen.

Strategisch gedrag is altijd een punt van aandacht bij de inzet van sturingsinstrumenten. Dit geldt ook voor niet-financiële sturingsinstrumenten, zoals in Dranove et al. (2003). Als goede prestaties niet direct met geld worden beloond, bestaat toch de neiging om zich zo goed mogelijk te presenteren. Dat strategisch gedrag aandacht verdient, betekent overigens niet dat sturingsinstrumenten per saldo ineffectief zijn. Ondanks het strategische gedrag kan het inzetten van een sturingsinstrument per saldo nog steeds beter zijn dan het niet gebruiken van sturingsinstrumenten.

[3] Vergelijkbare timing effecten zijn gevonden in situaties waarbij medewerkers per periode worden beoordeeld en bij de compensatie van managers en CEO's (zie Prendergast 1999).

Verdringing van intrinsieke motivatie

De motivatie om hard te werken of te handelen in het belang van de opdrachtgever kan uit verschillende bronnen voortkomen. Hierbij onderscheiden we extrinsieke bronnen en intrinsieke bronnen. Wanneer een uitvoerder gemotiveerd wordt door de inzet van sturingsinstrumenten wordt dit gedefinieerd als extrinsieke motivatie. Intrinsieke motivatie is gedefinieerd als motivatie die ontstaat uit het plezier of de voldoening van het uitvoeren van een taak of de resultaten die daarmee geboekt worden.

Intrinsieke motivatie in de psychologische en de bestuurskundige literatuur

De term intrinsieke motivatie komt uit de psychologische literatuur waar de discussie over de wisselwerking tussen intrinsieke en extrinsieke motivatie al decennia voortduurt. Een van de toonaangevende overzichtsstudies binnen dit veld is Deci et al. (1999). In hun meta-analyse bestuderen ze verdringing van intrinsieke motivatie door het inspanningsniveau van kinderen en studenten te vergelijken voor, tijdens en nadat ze blootgesteld zijn aan extrinsieke motivatie (gecontroleerd met een controle groep zonder extrinsieke prikkels). De bevindingen tonen aan dat het inspanningsniveau lager is na het afschaffen van de extrinsieke prikkel dan voor het invoeren van de extrinsieke prikkel. Zij concluderen hieruit dat deelnemers die zijn blootgesteld aan extrinsieke prikkels een lagere intrinsieke motivatie hebben dan werknemers die niet zijn blootgesteld aan extrinsieke prikkels. Toch zijn er kanttekeningen te plaatsen bij deze studie. Zo zijn de resultaten voor kinderen en studenten mogelijk niet direct te vertalen naar professionals. Bovendien is niet uit te sluiten dat hier twee effecten plaatsvinden: het niet krijgen van extrinsieke prikkels en het afpakken van extrinsieke prikkels. Indien het tweede effect inderdaad een rol speelt, dan zijn de resultaten niet door te trekken naar de vergelijking tussen wel of geen prikkels. Als laatst wordt ook vermoeidheid als mogelijke oorzaak genoemd, de kinderen en studenten hebben tijdens de periode met extrinsieke prikkels hard gewerkt en zijn daardoor mogelijk vermoeider dan de kinderen en studenten die in dezelfde periode geen extrinsieke prikkels kregen.

Een andere meta-studie (Eisenberger et al. 1999) komt tot een andere conclusie op basis van ongeveer dezelfde literatuur. Zij stellen dat slecht opgestelde of minimale prestatiedoelstellingen intrinsieke motivatie verdrijven, maar dat specifieke en hoge doelstellingen intrinsieke motivatie bevorderen. Hoge doelstellingen reflecteren mogelijk dat de uitgevoerde taak een groot persoonlijk en sociaal belang dient.

Jenkins et al. (1998) komen op basis van 47 studies eveneens tot de conclusie dat prikkels prestaties met 12 procent verhogen, terwijl het effect op kwaliteit verwaarloosbaar is. Condley et al. (2003) vinden op basis van 64 studies dat prikkels prestaties met gemiddeld 22% verhogen. Verder blijken financiële prikkels effectiever dan niet-financiële prikkels en prikkels voor een groep effectiever dan individuele prikkels.

Een voorbeeld van een meta-studie uit de bestuurskundige literatuur is Weibel et al. (2010) die op basis van 38 studies concludeert dat prikkels niet effectief zijn.

De literatuur onderscheidt drie mechanismen waardoor extrinsieke motivatie een negatieve invloed kan hebben op intrinsieke motivatie. Ten eerste kan een extrinsieke prikkel mensen in een andere gedachtesetting brengen. Het introduceren van een extrinsieke prikkel kan bijvoorbeeld de keuze van een persoon veranderen van een sociale keuze naar een monetaire keuze, waardoor de persoon geen rekening meer houdt met intrinsieke overwegingen (Gneezy en Rustichini 2000a en Heyman en Ariely 2004). Ten tweede kan de intrinsieke motivatie van een werknemer ook verdreven worden, omdat de werknemer mogelijk niet erkend wordt als een

intrinsiek gemotiveerd persoon wanneer hij of zij hard werkt en daarvoor een hoge beloning ontvangt (Bénabou en Tirole 2006). Tot slot kan het introduceren van een extrinsieke prikkel de agent ook meer informatie geven over het vertrouwen van de principaal in hem of haar, of over de plezierigheid van de taak (Bénabou en Tirole 2003).

De eerste reden waardoor extrinsieke motivatie intrinsieke motivatie kan verdrijven, ligt in de verandering in interpretatie van de taak. In de psychologie wordt beschreven dat een persoon een taak interpreteert aan de hand van de redenen en motivaties voor het uitvoeren van de taak (Bem 1965 en 1967). Wanneer een taak uitgevoerd moet worden zonder dat daar een beloning tegenover staat, interpreteert een persoon de taak als intrinsiek. Als een beloning wordt gegeven, dan verandert het interpretatiekader van de werknemer van intrinsiek naar extrinsiek. Gneezy en Rustichini (2000a) verklaren de uitkomsten van twee experimenten aan de hand van deze theorie. In één van deze twee experimenten zamelen middelbare scholieren geld in voor goede doelen. Een willekeurig gekozen groep ontvangt geen financiële vergoeding voor het verzamelde bedrag, een tweede groep wordt beloond met een bescheiden resultaatafhankelijk bedrag en de derde groep ontvangt een hogere financiële prikkel. Wanneer een monetaire beloning gegeven wordt, heeft dat een negatieve invloed door de verandering van het evaluatiekader, maar, als de beloning hoog genoeg is kan het totale effect van de beloning alsnog positief zijn, zo blijkt uit de resultaten.

Intrinsieke motivatie is van specifiek belang voor de (semi)publieke sector (zie volgend tekstkader). In een theoretische analyse onderzoeken Bénabou en Tirole (2006) een situatie waarbij een werknemer niet alleen gemotiveerd is, maar ook graag erkend wil worden als intrinsiek gemotiveerd. Bijvoorbeeld, omdat een werknemer in de semipublieke sector het belangrijk vindt dat de samenleving hem of haar ziet als iemand die passie heeft voor de publieke zaak. Bénabou en Tirole stellen een model op waarbij de agent een nutsfunctie heeft die rekening houdt met intrinsieke motivatie en imago. Wanneer geen extrinsieke beloning wordt toegekend, zal het gedrag een goede afspiegeling zijn van de intrinsieke motivatie van de werknemer. Werknemers die waarde hechten aan het gemotiveerd overkomen zullen dan extra hun best doen. Als er wel extrinsieke beloningen worden toegekend zou hard werken ook een signaal kunnen zijn dat de werknemer materialistisch is, dit betekend dat een intrinsiek gemotiveerde werknemer dus minder zijn of haar best zal doen. In deze situatie zou een extrinsieke prikkel dus een zwakker of averechts effect kunnen hebben op de prestatie van gemotiveerde werknemers.

Gerelateerde studies onderzoeken een situatie waarin de principaal betere informatie heeft dan de agent over de taak die uitgevoerd moet worden (Bénabou en Tirole 2003). Wanneer de principaal kiest voor het inzetten van een sturingsinstrument kan dit informatie vrijgeven aan de agent over de uit te voeren taak. Zo kan een sterke prikkel een teken zijn dat de principaal niet vertrouwt op de kunde en motivatie van

de agent, of kan het een signaal zijn dat de uit te voeren taak moeilijk of onprettig is. De achterliggende gedachte is de volgende. Als de taak leuk of eenvoudig is en de medewerker kundig, zal de principaal geen sturingsinstrumenten inzetten en vertrouwen op een goede afloop. Wanneer de principaal echter geen vertrouwen heeft in een goede afloop zullen sturingsinstrumenten wel ingezet worden om een nadelige uitkomst te voorkomen.

Motivatie van werknemers in de (semi)publieke sector en private sector

Het effect van de inzet van sturingsinstrumenten op intrinsieke motivatie is belangrijk voor de (semi)publieke sector. Uitgebreide literatuur in de bestuurskunde en economie heeft aangetoond dat intrinsieke motivaties een belangrijke rol spelen voor werknemers in deze sector (zie Perry et al. 2010 of Francois en Vlassopoulos 2008 voor een overzicht van de literatuur). Individuen met een sterke intrinsieke motivatie om bij te dragen aan het publieke belang werken vaker in de publieke sector en rapporteren meer tevreden te zijn in een publieke baan (zie bijvoorbeeld Steijn 2008, Buurman et al. 2012 en Dur en Zoutenbier 2014). De theorie van publieke service motivatie (Perry en Wise 1990) stelt dat drie soorten motivatie belangrijk zijn voor werknemers in een publieke functie: deelname aan het formuleren van beleid, de wil om bij te dragen aan een specifiek beleidsonderwerp of meer algemene gevoelens van onbaatzuchtigheid (zie Perry et al. 2010 voor een overzicht van de empirische studies).

Ondanks een aantal theoretische verklaringen is er nog weinig empirisch bewijs, zeker vanaf de werkvloer, voor de verdringing van intrinsieke motivatie door de inzet van sturingsinstrumenten. Wel zijn er enkele aansprekende voorbeelden te vinden van verdringing van intrinsieke motivatie.[4] Zo bestuderen Gneezy en Rustichini (2000b) het effect van een negatieve prikkel (een geldboete) bij een kinderopvang op het ophaalgedrag van ouders. Na het invoeren van een boete voor het te laat ophalen van een kind steeg het aantal te laat opgehaalde kinderen. Gneezy en Rustichini beargumenteren dat ouders in de uitgangsituatie hun kinderen vaak op tijd ophaalden uit intrinsieke overwegingen (zoals schuldgevoel naar de leidsters). Het introduceren van een boete gaf de mogelijkheid om het schuldgevoel af te kopen en zo de kinderen later op te halen.

Een voorbeeld van een (lab)experiment met niet-financiële prikkels is Falk en Kosfeld (2006), waarbij de principaal een deelnemer een minimaal inspanningsniveau kan opleggen. De resultaten tonen aan dat deelnemers aan wie een minimaal inspanningsniveau wordt opgelegd, minder inspanning uitoefenen dan deelnemers die niet gecontroleerd worden.

Individuele prikkels en teamprikkels

Het volgende tekstkader vat de mogelijke problemen bij de inzet van sturingsinstrumenten gericht op individuen samen. Zoals eerder aangegeven spelen dezelfde factoren een rol bij financiële en niet-financiële prikkels. Monitoren, verantwoorden, stellen van targets, transparantie en feedback werken niet

[4] Naast verdringing van intrinsieke motivatie kan het gebruik van sturingsinstrumenten ook nadelige effecten hebben op innovatief gedrag en creativiteit, zie bijvoorbeeld Azoulay et al. (2011) en Ederer en Manso (2013).

(optimaal) als (een deel van de) prestaties niet goed gemeten kunnen worden. De neiging tot strategisch gedrag is eveneens aanwezig, zelfs als daar geen financiële beloning tegenover staat. Ten slotte kan strenger toezicht een vergelijkbaar signaal afgeven als een directe financiële prikkel, met als gevaar dat de intrinsieke motivatie van werknemers verdrongen wordt.

Overzicht mogelijke problemen bij inzet

Instrumenten niet goed inzetbaar indien:

- Moeilijk meetbare output
- Onzeker productieproces
- Meerdere taken
- Conflicterende sturingsinstrumenten
- Werken in teamverband[(a)]

Mogelijk averechtse effecten:

- Strategisch gedrag:
  - Gaming
  - Upcoding
  - Cream skimming
- Verdringing intrinsieke motivatie[(b)]:
  - Gedachtesetting verandert van intrinsiek naar extrinsiek
  - Prikkel signaal dat agent niet gemotiveerd
  - Prikkel signaal dat agent niet kundig
  - Prikkel signaal dat taak vervelend is

[(a)] Er is een verschil tussen het effect van individuele prikkels en teamprikkels. Bij individuele prikkels bestaat het gevaar dat er te weinig wordt samengewerkt. Bij teamprikkels is meeliftergedrag mogelijk een probleem.

[(b)] Dit geldt voor individuele prikkels. Bij prikkels toegesneden op de organisatie hangt het effect af van hoe prikkels via interne aansturing worden vertaald naar de professionals.

Sturingsinstrumenten kunnen ook gericht zijn op een team. Prikkels op teamniveau (of instellingsniveau) kunnen via interne aansturing worden vertaald naar prikkels toegesneden op de organisatie.[5] Daarnaast kunnen leidinggevenden de prestaties verbeteren door belangrijke taken toe te wijzen aan de meest productieve medewerkers. Burgess et al. (2009) laten zien dat financiële prikkels bij de douane in het Verenigd Koninkrijk teamprestaties verbeteren. De verbetering in productiviteit is te verklaren door verbeterde individuele prestaties van teamleden en een aanpassing van de taakverdeling binnen een team.

De beschreven problemen bij de inzet van sturingsinstrumenten op teamniveau zijn in grote lijnen vergelijkbaar met de problemen bij de inzet op individueel niveau. Er zijn twee uitzonderingen. Ten eerste kan bij teambeloning een meeliftsituatie ontstaan waarbij individuele teamleden minder hard werken en meeliften op de

[5] Het geven van prikkels op teamniveau is vergelijkbaar met sturing van instellingen.

prestaties van andere teamleden. Ten tweede kunnen de mechanismen verschillen op het punt van verdringing van intrinsieke motivatie.

Zoals eerder aangegeven kunnen individuele prikkels een nadelig effect hebben op de prestatie van het team doordat werknemers minder geneigd zijn om elkaar te helpen. Dit probleem kan worden opgelost door de prestatie van het team als geheel te belonen. Dit is vooral een uitkomst wanneer de individuele bijdrage van teamleden niet te onderscheiden zijn. Helaas kan ook dit nadelige effecten met zich meebrengen. Er kan dan een zogenaamde meeliftsituatie ontstaan waarin individuele teamleden kunnen meeliften op de inspanningen van andere teamleden en meedelen in de baten van het team als geheel (Hölmstrom 1982).

Newhouse (1973) onderzoekt dit fenomeen in de gezondheidszorg. Hij laat zien dat artsen minder uren werken en meer kosten maken als zij kosten en opbrengsten delen in een groep van artsen. Vormen van groepsdruk of sterke groepsnormen kunnen het effect van meeliftgedrag afremmen (Kandel en Lazear 1992). Werken in teamverband kan ook op instellingsniveau geïnterpreteerd worden: soms moeten instellingen samenwerken om het beoogde resultaat te behalen. Ook hier kan het meelifteffect samenwerking parten spelen als de beloning afhangt van de prestatie van de groep, en andersom komt samenwerking moeilijk tot stand als instellingen afzonderlijk worden beloond.

Ook op het punt van intrinsieke motivatie zijn verschillen tussen individuele en teamprikkels. De verdringing van intrinsieke motivatie kan uitsluitend voorkomen bij personen. De theoretische studies in de economische literatuur stellen dat sturingsinstrumenten nieuwe informatie vrijgeven aan professionals over zichzelf. Zo is resultaatafhankelijke beloning een signaal dat de werknemer niet als een intrinsiek gemotiveerd of als een kundige persoon wordt gezien, wat de professional ontmoedigt. Deze argumentatie is vooral van toepassing bij aansturing binnen een organisatie. Een signaal van een meer op afstand staande centrale overheid heeft aanzienlijk minder informatiewaarde over de eigenschappen van het individu. De doorvertaling van centraal beleid in interne aansturing is cruciaal voor de mate waarin prikkels intrinsieke motivatie verdringen. Echter, overtuigend empirisch bewijs vanaf de werkvloer voor verdringing van intrinsieke motivatie door prikkels ontbreekt (nog).

4 Empirisch onderzoek

Er is veel empirisch onderzoek gedaan naar de effectiviteit van sturingsinstrumenten. Deze studies onderzoeken of een bepaald sturingsinstrument per saldo effectief is. Als een instrument niet werkt, wordt zelden besproken waarom het instrument niet werkt. Ook eventuele neveneffecten worden meestal slechts informeel

bediscussieerd. De meeste studies zijn dan ook slechts indirect informatief over averechtse effecten.

Een studie waarin een positief netto effect wordt gerapporteerd, toont aan dat de averechtse effecten niet opwegen tegen de positieve effecten. Dit geldt strikt genomen voor de activiteit die gestimuleerd wordt. Soms is de gestimuleerde activiteit echter een afgeleide van een bredere doelstelling. Zelfs als er meer van de gestimuleerde activiteit teweeg wordt gebracht, is dat niet altijd in lijn met de bredere doelstelling.

Hieronder geven we allereerst een overzicht van studies naar de effectiviteit van financiële prikkels. We gebruiken hierbij de indeling van Hasnain et al. (2014) en focussen op studies met een hoge interne- en externe validiteit. Vervolgens gaan we dieper in op de empirische studies naar sturingsinstrumenten ingezet in de zorg en het onderwijs. In deze sectorspecifieke overzichten komen zowel financiële- als niet-financiële instrumenten aan de orde alsook eventueel gerapporteerde averechtse- en neveneffecten.

Effectiviteit financiële prikkels

De meest recente overzichtsstudie naar de effectiviteit van resultaatafhankelijke beloning is Hasnain et al. (2014), waarin 153 empirische studies worden geordend naar kwaliteit van het empirisch werk. De auteurs concluderen dat financiële prikkels effectief *kunnen* zijn; de effectiviteit hangt echter af van de context waarin de prikkel gegeven wordt. De gepresenteerde classificatie van studies biedt de mogelijkheid om in te zoomen op het meest relevante deel van de literatuur, voor deze notitie zijn dat kwalitatief goede studies gericht op werkzaamheden die vergelijkbaar zijn met die in semipublieke sectoren en gericht op ervaringen uit OECD-landen.

Kwaliteit van empirisch werk is ingedeeld in vijf klassen en varieert van kwalitatieve (case) studies tot (veld) experimenten. Daartussen zitten kwantitatieve analyses op basis van data uit bestaande bronnen; de kwaliteitsindeling van deze studies hangt af van steekproefomvang en de gebruikte statistische methode. Experimenten hebben de hoogste mate van interne validiteit; de resultaten kunnen causaal worden geïnterpreteerd. Ook quasi-experimenten worden (in zowel Hasnain et al. als in deze notitie) tot de kwalitatief goede studies gerekend.[6]

Naast interne validiteit, zijn studies ook beoordeeld op externe validiteit, ofwel de mate waarin de resultaten doorgetrokken kunnen worden naar een andere context. Zo is de interne validiteit van labexperimenten hoog (causaliteit is niet discutabel), terwijl de externe validiteit beperkt is. Dat komt omdat de deelnemers aan

[6] Bij quasi-experimenten worden deelnemers niet willekeurig toegewezen aan de behandel- of controle groep. Er wordt meestal gebruik gemaakt van een criterium waardoor een deel van de populatie de behandeling krijgt en een ander deel dat niet krijgt. Bij de toepassing van de juiste econometrische methoden kunnen de resultaten causaal geïnterpreteerd worden.

labexperimenten (vaak studenten) mogelijk anders zijn dan de professionals die in de werkelijkheid beslissingen nemen. Bovendien is de hoogte van de prikkels in labexperimenten vaak kleiner dan gangbare prikkels in een professionele setting. Ook zijn de steekproeven van labexperimenten over het algemeen klein. Resultaten van labexperimenten zijn voor deze notitie dan ook minder relevant.

Zoals beschreven in paragraaf 3, hangt de effectiviteit van prikkels samen met de aard van de werkzaamheden. Zo kunnen prikkels voor eenvoudige en goed te meten taken effectiever zijn dan prikkels voor meer complexe opdrachten. Om hiervoor te corrigeren, gebruiken Hasnain et al. (2014) Wilson's classificatie van banen om de studies in te delen in vier categorieën. Wilson's (1989) classificatie stelt twee vragen, namelijk kan output gemeten worden en kunnen de stappen van het productieproces gemonitord worden. Tabel 4.1 schetst deze indeling. Banen met simpele repetitieve taken, waarvoor geen specifieke competenties nodig zijn, heten productiebanen. Voorbeelden zijn fabriekswerk of het ophalen van afval. Daarnaast zijn er banen waar output niet goed gemeten kan worden, maar de stappen (of input) goed te volgen zijn. Het leger wordt als voorbeeld voor dit type functie genoemd. Banen waar een brede set vaardigheden moet worden toegepast en waar het productieproces moeilijk door een buitenstaander gemonitord kan worden, zijn kenmerkend voor vele functies in de semipublieke sector. Als het eindproduct goed gemeten kan worden zijn dit *craft jobs*, wanneer het einde product niet goed gemeten kan worden zijn dit *coping jobs*. Hasnain et al. beschouwen onderwijs, de belastingdienst, re-integratiediensten en specifieke goed afgebakende taken in de zorg, als *craft jobs*. *Coping jobs* zijn functies in het hart van de publieke sector zoals beleidsfuncties en ook management functies in private bedrijven.

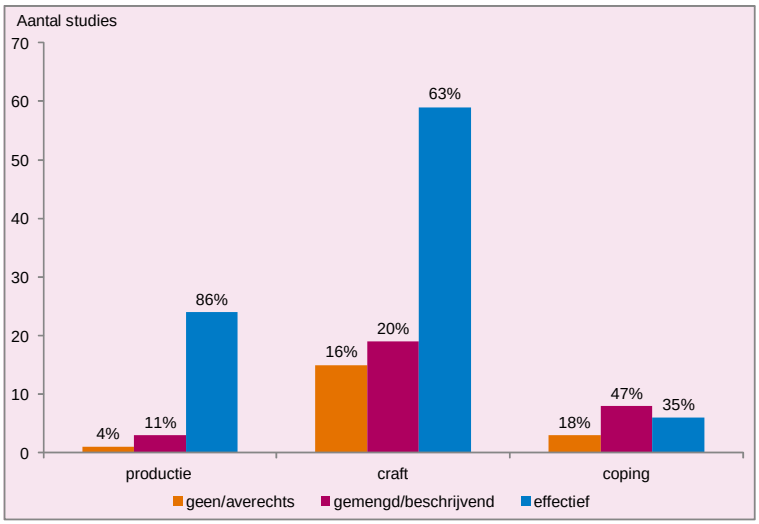
Tabel 4.1 Wilson's classificatie van functies

		Stappen of werkprocessen	
Productie		Waarneembaar	Niet-waarneembaar
	Meetbaar	Productie: specifieke competenties, simpele repetitieve handelingen (verkoop, afval ophalen, fabriekswerk)	Craft: toepassing brede set vaardigheden, eindproduct meetbaar (onderwijs, belastingdienst, re-integratie en zorg)
	Niet goed meetbaar	Procedureel: specifieke competenties, vaste taken, eindproduct uniek (leger)	Coping: toepassing brede set vaardigheden, eindproduct uniek. (beleidsfuncties, management)

Hasnain et al (2014) stelt dat alleen studies naar banen binnen een cel onderling goed vergelijkbaar zijn. Voor typische banen in de semipublieke sector zijn de studies van *craft* en *coping* jobs het meest relevant, aldus Hasnain et al. (2014). Figuur 4.1 illustreert de bevindingen uit de literatuur per type baan.[7] Studies die aangemerkt zijn als 'effectief' vinden empirisch bewijs voor de effectiviteit van financiële prikkels. 'Gemengd/beschrijvend' staat voor kwalitatief beschrijvende studies en voor studies met gemengde (tegenstrijdige) resultaten. Empirische studies die geen aantoonbaar effect vinden of die per saldo een effect vinden dat tegenovergesteld is aan het beoogde effect, krijgen de label 'Geen/averechts'.

Figuur 4.1 laat zien dat voor alle drie type banen meer studies zijn die resultaatafhankelijke beloning effectief vinden dan studies die geen effect of een averechts effect laten zien.[8] In lijn met de verwachting, lijkt resultaatafhankelijke beloning voor productiebanen effectiever dan voor *coping jobs*. Figuur 4.2 is vergelijkbaar met Figuur 4.1, maar toont alleen de 52 studies met een hoge interne en externe validiteit: veldexperimenten en quasi-experimenten. Ook met deze strengere selectie van studies lijkt de boodschap hetzelfde: resultaatafhankelijke beloning is per saldo vaker effectief dan niet.

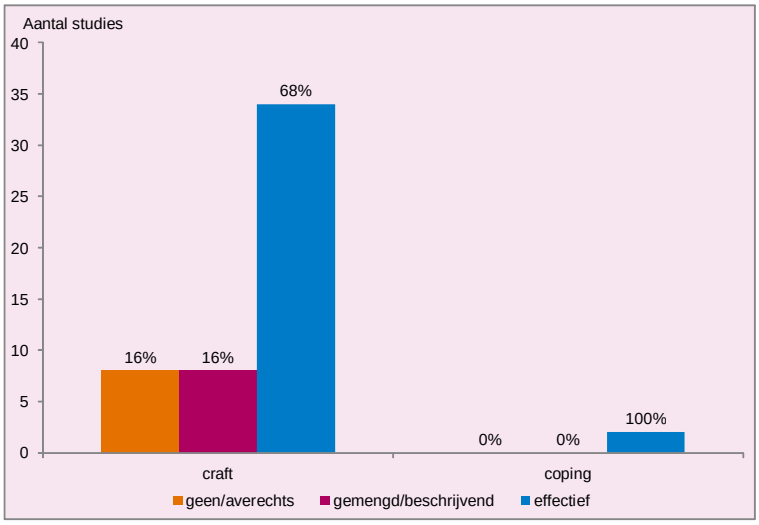
Figuur 4.1 Aantal studies naar resultaatafhankelijke beloning uitgesplitst naar gevonden effect (en type baan)



Bron: Hasnain et al. (2014), p.16.

[7] Merk op dat de figuren mogelijk een enigszins vertekenend beeld geven. Dit is het geval als de kans op publicatie hoger is voor studies die een positief effect vinden. Echter, het overzicht bevat zowel gepubliceerde als niet gepubliceerde (working) papers. Hierdoor is een eventuele *publication bias* redelijk beperkt.
[8] Er blijken geen studies te zijn naar resultaatafhankelijke beloning bij procedurele banen.

Figuur 4.2 Aantal veld- en quasi-experimentele studies naar resultaatafhankelijke beloning uitgesplitst naar gevonden effect (en type baan)



Bron: Hasnain et al. (2014), p.17.

Lessen uit studies in de gezondheidszorg

In de gezondheidszorg worden zowel op individueel niveau als op instellingsniveau sturingsinstrumenten ingezet. Doel van de ingezette instrumenten is het belonen van output of het stimuleren van de kwaliteit van de geleverde zorg. Specialisten en ziekenhuizen worden bijvoorbeeld beloond voor het uitvoeren van verrichtingen (DBC's) en de prestaties van zorginstellingen worden transparant gemaakt met behulp van kwaliteitsindicatoren.[9]

Verschillende kenmerken van de gezondheidszorg maken de toepassing van sturingsinstrumenten lastig (zie ook sectie 3). Ten eerste is de uitkomst van de verleende zorg is vaak lastig te meten en te waarderen. Wanneer een patiënt een ingewikkelde behandeling ondergaat in de curatieve zorg, is het bijvoorbeeld moeilijk om te meten wat het effect van de behandeling is, omdat er geen counterfactual beschikbaar is. Dat wil zeggen, we weten niet wat de gezondheid van de patiënt was geweest als er geen of een andere behandeling was uitgevoerd. Ten tweede is het lastig om het effect van het handelen van de professional op de gezondheidswinst te kwantificeren. Bij een behandeling werken vaak meerdere professionals samen (verpleegkundigen, specialisten, etc.) en het beoogde resultaat heeft vaak meerdere dimensies. In de langdurige zorg kan het eenvoudiger zijn om de uitkomst te meten (zorg voor de patiënt), maar, ook hier, werken meerdere professionals samen om een goed resultaat te behalen zonder dat de individuele bijdrage van elke professional te onderscheiden is.

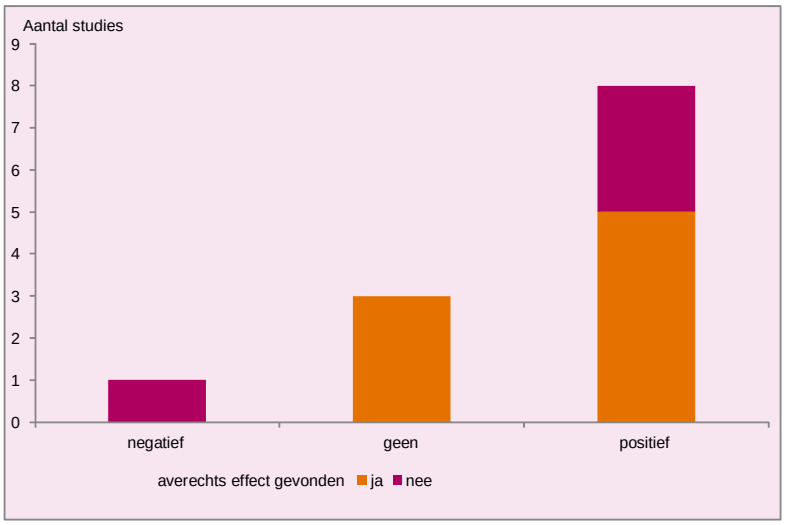
[9] Sturingsinstrumenten worden ook ingezet om het gedrag van de gebruikers van zorg te sturen, bijvoorbeeld via een eigen bijdrage of eigen risico om het gebruik van zorg te remmen, maar ook prikkels voor preventie. Deze notitie beperkt zich echter tot de inzet van sturingsinstrumenten bij professionals of instellingen.

Toch is er een groeiend aantal studies dat de effectiviteit van sturingsinstrumenten in de gezondheidszorg analyseert. De meeste studies gebruiken een geldbeloning als sturingsinstrument, maar er zijn ook een aantal voorbeelden waarbij wordt gekeken naar het monitoren en transparant maken van prestaties of het instellen van targets. Bijlage 1 geeft een overzicht van veldexperimenten en quasi-experimenten die de effectiviteit van sturing van zorginstellingen of professionals in de zorg onderzoeken.[10]

Figuur 4.3 geeft een overzicht van de resultaten van de studies over sturingsinstrumenten in de gezondheidszorg uit bijlage 1. In de figuur is een sturingsinstrument dat door de auteurs van de studie effectief wordt bevonden aangeduid als 'positief'. Studies die geen aantoonbaar effect vinden, vallen in de categorie 'geen'. Voorbeelden waar het invoeren van een sturingsinstrument het tegenovergestelde effect heeft bereikt dan oorspronkelijk bedoeld, zijn aangemerkt als 'negatief' (anders dan in Figuur 4.1 en 4.2 is hier 'geen effect' gesplitst van 'averechts effect').

Een deel van de studies vindt dat de invoering van een sturingsinstrument effectief is geweest, zelfs als er averechtse effecten aanwezig zijn. Bijvoorbeeld door een verbetering in prestatie op de beloonde uitkomst maar nadelige effecten op niet beloonde uitkomsten. Bij drie van de zeven studies met een positief effect worden aanwijzingen van averechtse effecten gevonden. Geobserveerde vormen van strategisch gedrag zijn het beïnvloeden van de resultaten op papier, manipulatie of selectie van minder ernstige patiënten waarbij eenvoudig betere resultaten te behalen zijn.

Figuur 4.3 Gevonden effecten bij studies naar sturingsinstrumenten in de gezondheidszorg



[10] Het literatuuroverzicht in de bijlage is gebaseerd op de literatuur overzichten door Hasnain et al. (2014) en Petersen et al. (2006). Alleen de studies met hoge interne en externe validiteit zijn meegenomen in de tabel.

Bij ziekenhuizen is vooral onderzoek gedaan naar sturing op kwaliteit door kwaliteitsindicatoren te belonen of transparant te maken. Lindenauer et al. (2007) vinden dat prestatiebeloning op kwaliteitsindicatoren een sterke verbetering op kwaliteitsindicatoren tot gevolg had. Er worden echter ook averechtse effecten gevonden op instellingsniveau, zoals cream skimming of manipulatie van meetinstrumenten. Propper et al. (2010) vinden aanwijzingen dat het monitoren van wachtlijsten bij ziekenhuizen ertoe leidt dat sommige patiënten tijdelijk of permanent van de lijst werden gehaald, wat een aanwijzing kan zijn voor manipulatie van de maatstaf in plaats van een verbetering in prestatie. Dranove et al. (2003) vinden dat transparantie over kwaliteit bij bypassoperaties in ziekenhuizen leidt tot selectie van eenvoudiger te behandelen patiënten.

Norton (1992) laat zien dat sturing ook toegepast kan worden bij verzorgingshuizen. Verzorgingshuizen werden een geldbonus geboden voor het opnemen, verbeteren van gezondheid en ontslaan van patiënten. Norton vindt een positief effect op kwaliteit en efficiency: verzorgingshuizen namen zwaardere patiënten aan en de behandeltijd was korter.

Ook op het niveau van de professional zijn sturingsinstrumenten onderzocht. Er is bijvoorbeeld onderzoek gedaan naar de effectiviteit van sturen op vaccinatiegraden, screening op ziekten, begeleiding bij stoppen met roken en het volgen van richtlijnen door professionals. Het bewijs voor effectiviteit is gemengd. Zo vinden Fairbrother et al. (1999, 2001), in twee sterk gerelateerde studies, dat de invoering van stuksbeloning bij vaccinatie vooral het aantal geregistreerde vaccinaties verhoogt maar het aantal daadwerkelijk uitgevoerde vaccinaties niet. De prikkel had dus vooral een effect op de administratieve nauwkeurigheid. Kouides et al. (1998) vinden wel een positief effect van een stuksbeloning voor het uitvoeren van vaccinaties, maar Hillman et al. (1998, 1999) en Grady et al. (1997) vinden geen effect van het belonen voor het volgen van richtlijnen of doorverwijzingen voor screening.

Belangrijke kanttekening is dat de experimenten beschreven in bijlage 1 maar een deel van de verrichtingen in de zorg vertegenwoordigen, wat de resultaten niet direct generaliseerbaar maken naar andere vormen van zorg. Deze studies richten zich vooral op eenvoudig te meten uitkomsten, zoals wachttijden, preventief screenen, vaccinatiegraden of het volgen van richtlijnen. Naast deze experimenten zijn ook een aantal evaluaties van bestaande data te vinden over het invoeren van prestatiebeloning voor artsen (Hillman 1991, Campbell et al. 2005, Campbell et al. 2007, Steel et al. 2007, Vaghela et al. 2009 en Chalkley et al. 2010). Het merendeel van deze studies rapporteert een positief effect van het invoeren van prestatie beloning. Shen (2003) en Doran et al. (2006) vinden echter bewijs voor strategisch gedrag.

Kort samengevat: de studies in de gezondheidszorg onderzoeken vooral de inzet van sturingsinstrumenten die worden ingezet om nauw gedefinieerde uitkomsten te

stimuleren. Gezien de diversiteit in zorg rijst de vraag of deze studies representatief zijn voor de gehele zorg of alleen voor de eenvoudig te meten uitkomsten. Het merendeel van de onderzochte studies rapporteert een positief effect van het inzetten van sturingsinstrumenten. Desondanks wordt bij een deel van deze studies aanwijzingen van strategisch gedrag ontdekt.

Lessen uit onderwijsstudies

In het onderwijs worden sturingsinstrumenten ingezet om leerlingen, ouders, leraren of scholen ertoe te bewegen om onderwijsprestaties te verbeteren. We beginnen met de prikkels gericht op leerlingen en ouders en gaan vervolgens in op het sturen van leraren en scholen.

Deelnemers

Gneezy et al. (2011) geven een overzicht van grootschalige veldexperimenten waarbij leerlingen of ouders financiële prikkels ontvangen. In deze experimenten is het doel ofwel om participatie te verhogen (bijvoorbeeld stimuleren van instroom en het voorkomen van spijbelen) of bepaalde onderwijsprestaties (zoals testresultaten) te behalen. De eerste categorie wordt ook wel 'sturen op input' en het tweede 'sturen op output' genoemd.

In het algemeen lijkt het sturen op input effectief in het behalen van het gestelde doel. Gneezy et al. (2011) stellen dat de eenduidigheid van de opdracht hierbij belangrijk is. Evaluaties van een experiment (in Mexico) met financiële prikkels voor gezinnen om onderwijsdeelname te stimuleren laten zien dat sturen op input effectief is (Behrman et al. 2005). Leerlingen stromen op jongere leeftijd in, blijven minder vaak zitten, de doorstroom is beter, uitval is lager en het aandeel van leerlingen dat terugkeert na uitval is hoger. Omdat de prikkels gericht zijn op gezinnen en niet specifiek op de kinderen is het risico dat de intrinsieke motivatie van kinderen wordt aangetast kleiner. Een opmerkelijk resultaat van dit experiment is dat doorstroom naar hogere groepen ook verbeterde voor jongere broers en zussen die deze prikkels niet hadden ontvangen.

De ervaringen met het verhogen van prestaties (sturen op output) zijn gemengd en lijken in verband te staan met de mate waarin de doelgroep de te nemen stappen kan overzien (Gneezy et al. 2011). Fryer (2010) onderzoekt financiële prikkels gericht op leerlingen en concludeert dat prikkels gericht op output (bijvoorbeeld hogere testresultaten) minder effectief zijn dan prikkels gericht op input zoals aanwezigheid of goed gedrag. Een mogelijke verklaring is dat een leerling input beter kan beïnvloeden dan output. Een andere studie laat zien dat financiële prikkels voor leerlingen gekoppeld aan begeleiding effectiever zijn dan zonder begeleiding (Angrist et al. 2006). Mogelijke verklaring is dat begeleiding helpt bij het concretiseren van doelen. Weer een ander experiment met financiële prikkels gericht op leerlingen laat zien dat resultaten behaald in rekenen verbeteren, maar de testresultaten van de

overige vakken niet zijn vooruitgegaan (Bettinger 2010). Deze bevinding is weer in lijn met de voorspelling dat extrinsieke motivatie effectief is voor concrete taken, maar niet voor meer algemene doelen.

Veel studies vinden heterogene effecten: sturingsinstrumenten hebben een groter effect voor de ene subgroep dan voor de andere subgroep. Zo bleek een financiële prikkel, gericht op eerstejaars economiestudenten op de Universiteit van Amsterdam om de propedeuse binnen een jaar te behalen, effectief voor de beste studenten, maar werkte averechts voor de minder vaardige studenten (Leuven et al. 2010). Een ander experiment met financiële prikkels, gericht op leerlingen in zwakke scholen in Chicago, bleek juist effectief zijn voor leerlingen op de 'drempel', dat zijn leerlingen die zonder de prikkel het diploma net niet hadden behaald (Levitt et al. 2012)

Over langetermijneffecten van dergelijke sturingsinstrumenten is nog niet veel bekend. Dat komt doordat de meeste veldexperimenten nog vrij recent zijn. Voor onderwijs is het echter zo dat kortetermijneffecten blijvende effecten kunnen zijn; het kan net het verschil maken tussen schooluitval en het behalen van een bepaald schoolniveau, of het aanleren van essentiële vaardigheden.

Een voorbeeld van een averechts effect bij een vergelijkbaar sturingsinstrument komt uit een evaluatie van studiebeurzen gekoppeld aan studieprestaties. Cornwell et al. (2006) laten zien dat de studiebeurzen tot betere cijfers leiden, maar dat studenten dan ook geneigd zijn makkelijkere vakken te kiezen. Dit lijkt op het fenomeen van *cherry picking* of *cream skimming,* zoals hierboven uitgelegd.

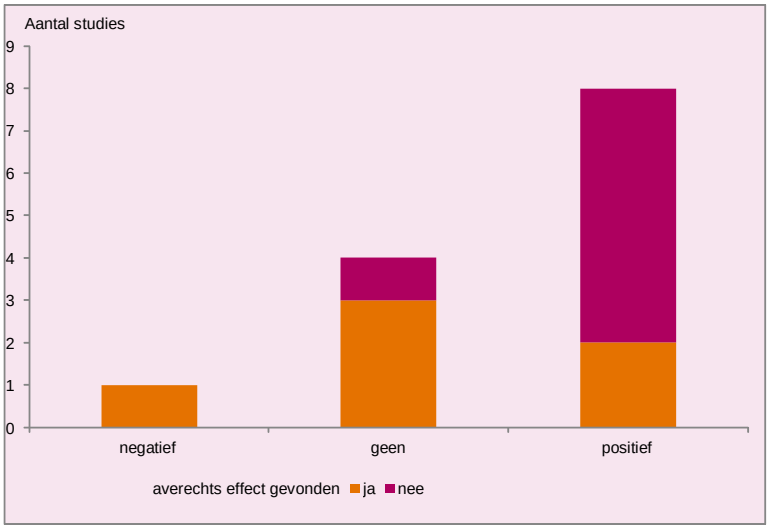
Gneezy et al. merken op dat de meeste grootschalige veldstudies in het onderwijs goed doordachte experimenten zijn waar veel aandacht is besteed aan het minimaliseren van averechtse- en neveneffecten. In het bijzonder is de omvang van de prikkel over het algemeen vrij groot, zodat het prikkeleffect naar verwachting groter uitvalt dan averechtse effecten. Het is mogelijk dat minder goed ontworpen sturingsinstrumenten meer averechtse- of neveneffecten vertonen.

Professionals en instellingen

Prikkels kunnen ook gericht zijn op scholen of leraren. De afgelopen jaren zijn bijvoorbeeld veel initiatieven geweest om de verantwoording van scholen te verbeteren, voornamelijk in de VS, maar ook in het VK en in Nederland. Dat kan door expliciete prikkels, zoals bonussen bij goede prestaties of door het dreigen met het sluiten of reorganiseren van een school bij slechte prestaties. Ook zijn er voorbeelden van meer impliciete, niet-financiële prikkels, zoals het transparant maken van toetsresultaten van scholen. Een belangrijk punt hierbij is of en hoe de prikkels voor de school door vertaald worden naar het onderwijzende personeel.

Bijlage 2 geeft een gedetailleerd overzicht van veldexperimenten en quasi-experimenten die de effectiviteit van sturing van scholen of leraren onderzoeken.[11] Figuur 4.2 schetst een beeld van de gevonden effecten en de gesignaleerde averechtse- of neveneffecten. Een sturingsinstrument wordt aangeduid als 'positief' als het effectief wordt bevonden. Studies die geen aantoonbaar effect vinden, vallen in de categorie 'geen' . Voorbeelden waar het invoeren van een sturingsinstrument per saldo het tegenovergestelde effect heeft bereikt dan oorspronkelijk bedoeld, zijn aangemerkt als 'negatief'. Merk op dat, anders dan in Figuur 4.1 en 4.2, omwille van de beeldvorming 'geen effect' gesplitst is van 'averechts effect'.

Figuur 4.4 Gevonden effecten bij studies naar sturingsinstrumenten gericht op leraren of scholen



Over het algemeen zijn sturingsinstrumenten effectief, als effectiviteit wordt afgemeten aan de gestimuleerde activiteit. Bij zes van de acht studies met een overall positief effect worden aanwijzingen van averechtse- of neveneffecten gevonden. Het gaat daarbij om te complexe regels, waardoor leraren niet weten wat zij moeten doen om het doel te behalen, meeliftergedrag en verschillende vormen van strategisch gedrag.

Een grootschalig veldexperiment met teamprikkels vond plaats in New York City, waar scholen een bonus konden ontvangen voor het behalen van een target uit de *school report card*. Scholen waren vrij om de schoolbonus naar eigen inzicht te verdelen onder de leraren. De meeste scholen kozen ervoor om de bonus gelijk te verdelen over leerkrachten. De omvang van de prikkel per leraar was ongeveer 4% van het gemiddelde salaris. Fryer (2011) analyseert de resultaten en komt tot de conclusie dat het instrument niet heeft geleid tot een verbetering van de prestaties.

[11] Studies naar prikkels ingezet in ontwikkelingslanden blijven buiten beschouwing. De ervaringen van deze experimenten zijn overwegend positiever dat die in OECD-landen. De context is echter grotendeels anders, deze studies zijn dus minder relevant voor deze notitie.

Een mogelijke verklaring is dat de regels te complex waren. De *progress report card* bleek een ingewikkelde maat, waar leraren moeilijk grip op kregen. Hierdoor was het onduidelijk wat een leraar kon doen om de prestaties te verbeteren. Goodman en Turner (2010) evalueren hetzelfde experiment en geven speciale aandacht aan meelifteffecten. Zij vinden dat leraren minder vaak absent zijn bij kleine scholen, maar niet bij grote scholen. Het effect is echter te klein om testresultaten te verbeteren. Dat er een verschil is tussen gedrag van leraren in kleine dan wel grote scholen suggereert dat meeliftergedrag een rol speelde bij de slechte resultaten.

Individuele prikkels zijn niet altijd effectief. Springer et al. (2010) bestuderen individuele prikkels aangeboden aan een willekeurig gekozen groep wiskundeleraren middelbare scholen in de VS. Een leraar kon een bonus ontvangen van ongeveer 22% van het jaarsalaris als de testresultaten van zijn klas tot de 95ste percentiel behoorde binnen de regio. Een bonus van circa 7% van het jaarsalaris was al bereikbaar bij een relatief kleine verbetering van de testresultaten (behorend tot de 80ste percentiel). Ondanks de sterke individuele financiële prikkel, bleek het instrument ineffectief. Ook deze auteurs schrijven de resultaten toe aan de complexiteit van de regels. De target is namelijk gespecificeerd ten opzichte van andere scholen, waardoor de leraar niet goed kan inschatten wat precies nodig is om het gewenste resultaat te behalen.

Andere studies naar financiële prikkels rapporteren betere resultaten, althans in het behalen van het gestelde doel. De introductie van resultaatafhankelijke beloning voor leraren in Engeland leidde tot een verbetering van testresultaten (Atkinson et al. 2009). Aanwijzingen voor *gaming* zijn niet gevonden in deze studie. Clotfelter (2008) laat zien dat financiële prikkels effectief waren in het behouden van leraren bij scholen in achterstandswijken of bij andere zwakke scholen in de VS. Eberts et al. (2002) evalueren het effect van een bonus als de leraar kan voorkomen dat leerlingen zijn vak laten vallen, gekoppeld aan een additionele bonus afhankelijk van een positieve evaluatie door de leerlingen. De prikkel bleek effectief in het reduceren van uitval onder leerlingen. Nadeel was dat testresultaten verslechterden. Het is onduidelijk wat daar de reden van was. Mogelijk werden de testresultaten omlaag gehaald door leerlingen die anders met het vak gestopt waren. Een andere mogelijke verklaring is dat leraren hun lesmethoden aanpasten om het vak aantrekkelijker te maken (meer studiereizen), maar dat het uiteindelijk ten koste van de kwaliteit van de lessen is gegaan.

Er zijn ook een aantal studies naar niet-financiële prikkels in het onderwijs. Vaak gaat het om zogenaamde school *accountability* programma's in de VS. Jacob (2005) evalueert het programma 'No Child Left Behind Act' dat in 1996-1997 geïmplementeerd is: alle staten in de VS werden verplicht om scholieren van groep 3, 6 en 8 jaarlijks te toetsen en de resultaten transparant te maken. Slecht presterende scholen liepen hierbij de kans om gesloten te worden. Jacob laat zien dat de toetsresultaten in wiskunde en lezen aanzienlijk verbeteren, maar een groot deel van de verbetering is toe te schrijven aan strategisch gedrag. Er zijn aanwijzingen dat

leraren proberen te beïnvloeden welke studenten de toets maken en welke resultaten meetellen voor de cijfers voor de school. Dat kan door leerlingen aan te merken als kandidaat voor bijzonder onderwijs of leerlingen een jaar te laten overdoen. Een ander aanwijzing voor strategisch gedrag is dat de prestaties op *low-stakes* toetsen niet veranderen, terwijl de resultaten van de zogenaamde *high-stakes* toetsen wel verbeteren.

Niet-financiële prikkels worden ook in Nederland ingezet. Van Elk en Kok (2014) evalueren het pakket aan maatregelen, geïntroduceerd door de Kwaliteitsaanpak Basisonderwijs Amsterdam (KBA), gericht op het verbeteren van zwakke scholen. De aanpak behelst verschillende maatregelen, waaronder evaluatie van leraren, feedback, scholing en coaching en het invoeren van nieuwe lesmethoden. Het beleid heeft, in vergelijking met de controlegroep, geleid tot een daling van de CITO-scores in de eerste vier jaar na invoering. Het per saldo negatieve effect heeft volgens schooldirecteuren te maken met het intensieve en veeleisende karakter van het programma dat tot weerstand heeft geleid onder leraren en tot het vertrek van leraren heeft geleid. Mogelijk gaat het om aanpassingskosten. Van Elk en Kok (2014) gaat niet in op intrinsieke motivatie, maar voor zover tevredenheid van leraren informatief is over intrinsieke motivatie van leraren, is dit een aanwijzing dat een dergelijk programma intrinsieke motivatie van leraren mogelijk aantast.

Aan de reeks voorbeelden van strategisch gedrag kan ook vals spelen worden toegevoegd. Jacob en Levitt (2003) laten zien dat school *accountability* programma's de kans op fraude verhoogt: leraren bewerken *high-stakes* tests van hun leerlingen. Dit komt vaker voor als de prikkels om goed te presteren sterker zijn.

Een minder voor de hand liggend voorbeeld van strategisch gedrag door scholen is het manipuleren van het caloriegehalte van de schoollunch (Figlio en Winicki 2005). Scholen in de gevarenzone voor mogelijke sancties blijken het menu aan te passen (een bekende stimulans voor cognitieve vaardigheden op de korte termijn). Deze opmerkelijke strategie blijkt overigens effectief: de slagingskans verbetert met 11% voor wiskunde en met 6% voor Engels, geschiedenis en maatschappijleer.

Samenvattend, sturingsinstrumenten zijn voor een groot deel effectief wanneer effectiviteit wordt afgemeten aan de gestimuleerde activiteit. Soms strookt dat echter niet helemaal met de bredere doelstelling van het onderwijs. Daarnaast is op basis van de literatuur te stellen dat prikkels in het onderwijs niet werken als het sturingsinstrument complex is, bijvoorbeeld door ingewikkelde wegingen of uitzonderingsmogelijkheden waardoor leraren niet weten hoe de doelstellingen te behalen. Meelifteffecten spelen mogelijk een rol bij teamprikkels. Financiële prikkels of meer monitoring kunnen bovendien ertoe leiden dat leraren een andere baan zoeken, mogelijk een uiting van aantasting van intrinsieke motivatie. Daarnaast wordt strategisch gedrag, zoals leerlingen niet toe te laten bij de toetsen (kwalificeren als kandidaat voor bijzonder onderwijs of als leerling met taalachterstand) als

belangrijke factoren genoemd. '*Teaching to the test*' wordt vaak als mogelijke oorzaak genoemd, maar is niet direct empirisch aangetoond in de studies besproken in dit overzicht. [12]

5 Conclusies en aanbevelingen

Deze notitie bespreekt de literatuur over de effectiviteit van meet- en sturingsinstrumenten ingezet bij professionals en bij sturing van instellingen. De omstandigheden die sturingsinstrumenten minder effectief maken staan hierbij centraal. Sturingsinstrumenten kunnen om verschillende redenen minder effectief zijn: zoals door problemen bij de toepassing, strategisch gedrag of verdringing van intrinsieke motivatie. Het onderstaande tekstkader vat de valkuilen van sturingsinstrumenten nog eens kort samen.

Overzicht mogelijke problemen bij inzet

Instrumenten niet goed inzetbaar indien:

- Moeilijk meetbare output
- Onzeker productieproces
- Meerdere taken
- Conflicterende sturingsinstrumenten
- Werken in teamverband[(a)]

Mogelijk averechtse effecten:

- Strategisch gedrag:
  - Gaming
  - Upcoding
  - Cream skimming
- Verdringing intrinsieke motivatie[(b)]:
  - Gedachtesetting verandert van intrinsiek naar extrinsiek
  - Prikkel signaal dat agent niet gemotiveerd
  - Prikkel signaal dat agent niet kundig
  - Prikkel signaal dat taak vervelend is

(a) Er is een verschil tussen het effect van individuele prikkels en teamprikkels. Bij individuele prikkels bestaat het gevaar dat er te weinig wordt samengewerkt. Bij teamprikkels is meeliftergedrag mogelijk een probleem.
(b) Dit geldt voor individuele prikkels. Bij prikkels toegesneden op de organisatie hangt het effect af van hoe prikkels via interne aansturing worden vertaald naar de professionals.

[12] Glewwe et al. (2003) vinden aanwijzingen voor teaching to the test in het kader van een veldexperiment in Kenia. Vanwege de keuze om alleen ervaringen uit OECD landen te behandelen, valt deze studie in dit overzicht buiten beschouwing.

De empirische literatuur over sturingsinstrumenten is omvangrijk. Consensus over de effectiviteit ervan ontbreekt. De discussie in de literatuur gaat dan ook eerder over de specifieke voorwaarden waaronder, en de context waarin, dergelijke instrumenten effectief kunnen zijn. Het grootste deel van de literatuur focust op de vraag of een specifiek instrument effectief is, waarbij slechts zijdelings wordt verwezen naar neveneffecten. Toch is bij verschillende veldexperimenten strategisch gedrag gesignaleerd of zijn aanwijzingen voor specifieke problemen bij de toepassing van het instrument. De aanwezigheid van neveneffecten betekent overigens niet automatisch dat een sturingsinstrument niet effectief is. De positieve effecten wegen soms zwaarder dan de negatieve.

De belangrijkste les uit de literatuur is dat een zorgvuldig design een belangrijke bijdrage levert aan de effectiviteit van een instrument, zeker gezien het belang van de context en de omgeving. Uit de literatuur komen de volgende aandachtspunten voor een zorgvuldig design.

Ten eerste is het belangrijk om vanuit het perspectief van de agent te kijken naar de prikkel. Welke acties leveren een goede beloning op en welke (belangrijke) acties dragen niet bij aan een goede beloning? Het is belangrijk om alle dimensies van het resultaat in kaart te brengen en, waar mogelijk, te belonen. Als niet alle dimensies te belonen zijn kunnen mogelijk averechtse effecten optreden.

Ten tweede is het belangrijk om goed na te denken over het sturingsinstrument op zich. Als een sturingsinstrument complex is, bijvoorbeeld door ingewikkelde wegingen of uitzonderingsmogelijkheden, weten agenten mogelijk niet goed wat ze moeten doen om de gestelde doelen te bereiken. Duidelijke en eenvoudige doelstellingen hebben de voorkeur. Het bepalen van de omvang van de prikkel is een ander punt van aandacht. Veelzeggend zijn de titels van verschillende artikelen over dit onderwerp, zoals “Pay enough or don't pay at all” (Gneezy en Rustichini 2000a) of “Large Stakes and Big Mistakes” (Ariely et al. 2009). De prikkel moet dus niet te hoog of te laag zijn. De exacte hoogte van de optimale prikkel is sterk afhankelijk van de context.

Als laatste is het belangrijk goed na te denken over de framing en communicatie van een prikkel. Om mogelijke verdringing van intrinsieke motivatie te vermijden is het aan te raden om na te gaan welk signaal uitgaat van het introduceren van een prikkel. Denk eerst goed na over wat de drijfveren zijn en hoe sturingsinstrumenten dat kunnen aanvullen of misschien verdrijven. Communiceer naar de uitvoerder ook duidelijk de doelstellingen van de prikkel.

Het zorgvuldig ontwerpen en invoeren van sturingsinstrumenten is niet eenvoudig. De empirische literatuur leert dat bij al deze stappen valkuilen zijn. Veldexperimenten (of pilots met een goede controlegroep) kunnen dan ook een belangrijke rol spelen in het zorgvuldig ontwerpen van sturingsinstrumenten (zie

Buurman et al. 2014 en Van Elk en Kok 2014 voor Nederlandse voorbeelden van veldexperimenten). Veldexperimenten leveren gefundeerde resultaten op, rekening houdend met de specifieke context waarin de instrumenten ingezet worden.

Literatuur

Aalbers, Rob, Victoria Shestalova en Gijsbert Zwart, 2013, Interactie milieubeleidsinstrumenten met het ETS, CPB notitie.

Alchian, Armen A. en Harold Demsetz, 1972, Production, Information Costs, and Economic Organization, *American Economic Review*, vol. 62(5): 777-95.

Angrist, Joshua, Daniel Lang en Philip Oreopoulos, 2006, Incentives and Services for College Achievement: Evidence from a Randomized Trial, *American Economic Journal: Applied Economics*, vol. 1(1): 136-63.

Ariely, Dan, Uri Gneezy, George Loewenstein en Nina Mazar, 2009, Large Stakes and Big Mistakes, *Review of Economic Studies*, vol. 76(2), 451-69.

Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater and D. Wilson, 2009, Evaluating the Impact of Performance-Related Pay for Teachers in England. *Labour Economics,* vol. 16(3): 251-61.

Azmat, Ghazala en Nagore Iriberri, 2010a, The Provision of Relative Performance Feedback Information: An Experimental Analysis of Performance and Happiness, Mimeo.

Azmat, Ghazala en Nagore Iriberri, 2010b, The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students, *Journal of Public Economics*, vol. 94(7): 435-52.

Azoulay, Pierre, Joshua S. Graff Zivin en Gustavo Manso, 2011, Incentives and Creativity: Evidence from the Academic Life Sciences, *RAND Journal of Economics*, vol. 42(3): 527-54.

Baker, George P., 1992, Incentive Contracts and Performance Measurement, *Journal of Political Economy,* vol. 100(3): 598-614.

Behrman, Jere R., Piyali Sengupta en Petra Todd, 2005, Progressing Through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico, *Economic Development and Cultural Change*, vol. 54(1): 237-75.

Bem, Daryl J., 1965, An Experimental Analysis of Self-Persuasion, *Journal of Experimental and Social Psychology*, vol. 1(3): 199-218.

Bem, Daryl J., 1967, Self-Perception: An Alternative Interpretation of Cognitive Dissonance Phenomena, *Psychological Review*, vol. 74(3): 183-200.
Bénabou, Roland en Jean Tirole, 2003, Intrinsic and Extrinsic Motivation, *Review of Economic Studies,* vol. 70(3): 489-520.

Bénabou, Roland en Jean Tirole, 2006, Incentives and Prosocial Behavior, *American Economic Review,* vol. 96(5): 1652-78.

Bengstsson, Niklas en Per Engström, 2013, Replacing Trust With Control: A Field Test of Motivation Crowd Out Theory, *Economic Journal,* vol. 124: 833-58.

Bettinger, Eric P., 2010, Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores, NBER Working Papers 16333.

Bijlsma, Michiel en Paul de Bijl, 2013, De governance van semi-publieke instellingen. CPB-notitie.

Blanes i Vidal, Jordi en Mareike Nossol, 2011, Tournaments Without Prizes: Evidence from Personnel Records, *Management Science*, vol. 57(10): 1721-36.

Bradler, Christiane, Robert Dur, Suzanne Neckermann en Arjan Non, 2013, Employee Recognition and Performance: A Field Experiment, ZEW Discussion Paper 13-017.

Burgess, Simon, Carol Propper, Marisa Ratto, Stepanie von Hinke Kessler Scholder en Emma Tominey, 2009, Smarter Task Assignment Or Greater Effort: The Impact of Incentives On Team Performance, *The Economic Journal*, vol. 120(september): 968-89.

Buurman, Margaretha, Josse Delfgaauw, Robert Dur en Seth van den Bossche, 2012, Public Sector Employees: Risk Averse and Altruistic?, *Journal of Economic Behavior and Organization*, vol. 83(3): 279-91.

Buurman, Margaretha, Josse Delfgaauw, Robert Dur en Robin Zoutenbier, 2014, The Effect of Students' Feedback on Teachers: Evidence from a Field Experiment, mimeo, Erasmus University Rotterdam.

Campbell, S.M., M. Roland, E. Middleton en D. Reeves, 2005, Improvements in the Quality of Clinical Care in English General Practice: Longitudinal Observational Study, *British Medical Journal*, vol. 331: 1121.

Campbell, S.M., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbad en M. Roland, 2007, Quality of Primary Care in England with the Introduction of Pay for Performance, *New England Journal of Medicine*, vol. 357: 181-90.

Chalkley, M., C. Tilley, L. Young, D. Bonetti en J. Clarkson, 2010, Incentives for Dentists in Public Service: Evidence from a Natural Experiment, *Journal of Public Administration Research and Theory*, vol. 20(2): 207-23.

Clotfelter, C., E. Glennie, H. Ladd en J. Vigdor, 2008, Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina, Journal of Public Economics, vol. 92, 1352-70.

Commissie Behoorlijk Bestuur, 2013, Een lastig gesprek- Advies Commissie Behoorlijk Bestuur.

Condley, S., R. Clark en H. Stolovitch, 2003, The Effects of Incentives on Workplace Performance: A Meta-Analytic Review of Research Studies, *Performance Improvement Quarterly*, vol. 16 (3): 46-63.

Cornwell, Christopher, David B. Mustard en Deepa J. Sridhar, 2006, The Enrollment Effects of Merrit-Based Financial Aid: Evidence from Georgia's HOPE Program. *Journal of Labor Economics*, vol. 24(4): 1-27.

Cooper, S.T. en E. Cohn, 1997, Estimation of a Frontier Production Function for the South Carolina Educational Process, *Economics of Education Review*, vol. 16(3): 313-27.

Courty, Pascal en Gerald Marschke, 2004, An Empirical Investigation of Gaming Responses to Explicit Performance Incentives, *Journal of Labor Economics*, vol. 22(1): 23-56.

Christensen, Dale B., N. Neil, W.E. Fassett, D.H. Smith, G. Holmes en A. Stergachis, 2000, Frequency and Characteristics of Cognitive Services Provided in Response to a Financial Incentive, *Journal of the American Pharmaceutical Association*, vol. 40(5): 609-17.

Deci, Edward L., Richard Koestner en Richard M. Ryan, 1999, A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation, *Psychological* Bulletin, vol. 125(6): 627-68.

Dixit, Avinash, 1997, Power of Incentives in Public versus Private Organization, *American Economic Review Papers and Proceedings,* vol. 87(2): 378-82.

Dixit, Avinash, 2002, Incentives and Organizations in the Public Sector: An Interpretative Review, *Journal of Human Resources*, vol. 37(4): 696-727.

Dranove, David, Daniel Kessler, Mark McClellan en Mark Satterthwaite, 2003, Is More Information Better? The Effects of "Report Cards" on Health Care Providers, *Journal of Political Economy,* vol. 111(3): 555-88.

Delfgaauw, Josse, Robert Dur, Joeri Sol en Willem Verbeke, 2013, Tournament Incentives in the Field: Gender Differences in the Workplace, *Journal of Labor Economics*, vol. 31(2): 305-26.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E.Kontopantelis, U. Hiroeh en M. Roland, 2006, Pay-for-Performance Programs in Family Practices in the United Kingdom, *New England Journal of Medicine*, vol. 355: 375-84.

Dur, Robert en Robin Zoutenbier, 2014, Working for a Good Cause, *Public Administration Review*, vol. 74(2): 144-55.

Eberts, R., K. Hollenbeck en J. Stone, 2002, Teacher Performance Incentives and Student Outcomes, *Journal of Human Resources*, vol. 37 (4): 913-27.

Ederer, Florian en Gustavo Manso, 2013, Is Pay for Performance Detrimental to Innovation?, *Management Science*, vol. 59(7): 1496-1513.

Eisenberger, Robert W., David Pierce en Judy Cameron, 1999, Effects of Reward on Intrinsic Motivation - Negative, Neutral, and Positive: Comment on Deci, Koestner, and Ryan (1999), *Psychological Bulletin*, vol. 125(6): 677-91.

Elk, R. van en S. Kok, 2014, The impact of a comprehensive school reform for failing schools on educational achievement; Results of the first four years. CPB Discussion Paper 264.

Fairbrother, Gerry, Karla Hanson, Stephen Friedman en Gary C. Butts, 1999, The Impact of Physician Bonuses, Enhanced Fees, and Feedback on Childhood Immunization Coverage Rates, *American Journal of Public Health*, vol. 89(2): 171-75.

Fairbrother, Gerry, Michele J. Siegel, Stephen Friedman, Pierre D. Korv en Gary C. Butts, 2001, Impact of Financial Incentives on Documented Immunization Rates in the Inner City: Results of a Rrandomized Controlled Trial*, Ambulatory Pediatrics*, vol. 1(4): 206-12.

Falk, Armin en Michael Kosfeld, 2006, The Hidden Cost of Control, *American Economic Review*, vol. 96(5): 1611-30.

Figlio, D. en S. Loeb, 2011, School Accountability. In E. Hanusek, S. Machin, L. Woessmann, editor: Handbook of the Economics of Education, vol. 3: 383-421.

Figlio, D.N. en L.W. Kenny, 2007, Individual Teacher Incentives and Student Performance, *Journal of Public Economics*, vol. 91: 901-14.

Figlio, D.N. en J. Winicki, 2005, Food for Thought: The Effects of School Accountability Plans on School Nutrition, *Journal of Public Economics*, vol. 89: 381-94.

Francois, Patrick en Michael Vlassopoulos, 2008, Pro-social Motivation and the Delivery of Social Services, *CESifo Economic Studies*, vol. 54(1): 22-54.

Fryer, Roland G., 2010, Financial Incentives and Student Achievement: Evidence from Randomized Trials, NBER Working Paper 15898.

Fryer, R. G., 2011, Teacher Incentives and Student Achievement: Evidence from New York City Public Schools , NBER Working Paper 16850.

Gerhards, Leonie en Neelie Siemer, 2014, Private versus Public Feedback: The Incentive Effects of Symbolic Awards, Economics Working Papers 2014-01.

Glewwe, Paul, Ilias Nauman en Michael Kremer, 2003, Teacher Incentives, NBER Working Paper 9671.

Gneezy, Uri en Aldo Rustichini, 2000a, Pay Enough or Don't Pay at All, *Quarterly Journal of Economics*, vol. 115(3): 791-810.

Gneezy, Uri en Aldo Rustichini, 2000b, A Fine Is a Price, *Journal of Legal Studies*, vol. 29(1): 1-18.

Gneezy, Uri, Stephan Meier en Pablo Rey-Biel, 2011, When and Why Incentives (Don't) Work to Modify Behavior, *Journal of Economic Perspectives,* vol. 25(4): 191-210.

Goodman, S. en L. Turner, 2010, Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program (Working Paper). PEPG Conference "Merit Pay: Will It Work? Is It Politically Viable?". Harvard Kennedy School, June 3-4, 2010.

Grady, Kathleen E., Jeanne Parr Lemkau, Norma R. Lee en Cheryl Caddell, 1997, Enhancing Mammography Referral in Primary Care, *Preventive Medicine*, vol. 26(6): 791-800.

Hasnain, Zahid, Nick Manning en Henryk Pierskalla, 2014, The Promise of Performance Pay? Reasons for Caution in Policy Prescriptions in the Core Civil Service, *World Bank Research Observer.*

Hanushek, Eric A., Susanne Link en Ludger Woessmann, 2013, Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA, *Journal of Development Economics,* vol. 104: 212-32.

Heyman, James en Dan Ariely, 2004, Effort for Payment a Tale of Two Markets, *Psychological Science*, vol. 15(11): 787-93.

Hillman, A. M. Pauly, K. Kerman en C.R. Martinek, 1991, HMO Manager's Views on Financial Incentives and Quality, *Health Affairs*, vol. 10(4): 207-19.

Hillman, Alan L., Kimberly Ripley, Neil Goldfarb, Isaac Nuamah, Janet Weiner en Edward Lusk, 1998, Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care, *American Journal of Public Health*, vol. 88(11): 1699-1701.

Hillman, Alan L., Kimberly Ripley, Neil Goldfarb, Isaac Nuamah, Janet Weiner en Edward Lusk, 1999, The Use of Physician Financial Incentives and Feedback to Improve Pediatric Preventive Care in Medicaid Managed Care, *Pediatrics*, vol. 104(4): 931-35.

Hölmstrom, Bengt, 1979, Moral Hazard and Observability, *Bell Journal of Economics*, vol. 10(1): 74-91.

Hölmstrom, Bengt, 1982, Moral Hazard in Teams, *Bell Journal of Economics*, vol. 13(2): 324-40.

Hölmstrom, Bengt en Paul Milgrom, 1991, Multitask Principal-Agent Analysis: Incentive Contracts, Asset Ownership, and Job Design, *Journal of Law, Economics and Organization*, vol. 7: 24-52.

Jacob, B.A., 2005, Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools, *Journal of Public Economics*, vol. 89:, 761-96.

Jacob, B.A. en S.D. Levitt, 2003, Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *Quarterly Journal of Economics*, vol. 118 (3): 843-77.

Jenkins, G.D., A. Mitra, N. Gupta en J.D. Shaw, 1998, Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research, *Journal of Applied Psychology*, vol. 83 (5): 777-87.

Kandel, Eugene en Edward P. Lazear, 1992, Peer Pressure and Partnerships, *Journal of Political Economy*, vol. 100(4): 801-17.

Kouides, Ruth W., Nancy M. Bennett, Bonnie Lewis, Joseph D. Cappuccio, William H. Barker en Marc LaForce, 1998, Performance-based Physician Reimbursement and Influenza Immunization Rates in the Elderly, *American Journal of Preventive Medicine,* vol. 14(2): 89-95.

Kuhnen, Camelia M. en Agnieszka Tymula, 2012, Feedback, Self-Esteem and Performance in Organizations, *Management Science*, vol. 58(1): 94-113.

Ladd, H.F., 1999, The Dallas School Accountability and Incentive Program: Evaluation of Its Impacts on Student Outcomes, *Economics of Education Review*, vol. 18: 1-16.

Lavy, V., 2008, Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments among Teachers, NBER Working Paper 14338.

Lazear, Edward P., 2000, Performance Pay and Productivity, *American Economic Review*, vol. 90(5): 1346-61.

Leuven, Edwin, Hessel Osteerbeck, and Bas van der Klaauw, 2010, The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment, *Journal of the European Economic Association*, vol. 8(6): 1243-65.

Levitt, Steven D., John A. List, Susanne Neckermann en Sally Sado, 2012, The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance, ZEW Discussion Paper 12-038.

Lindenauer, Peter K., Denise Remus, Sheila Roman, Michael B. Rothberg, Evan M. Benjamin, Allen Ma en Dale W. Bratzler, 2007, Public Reporting and Pay for Performance in Hospital Quality Improvement, *New England Journal of Medicine*, vol. 357(25): 2589-2600.

Newhouse, Joseph P., 1973, The Economics of Group Practice, *Journal of Human Resources,* vol. 8(1): 37-56.

Norton, Edward C., 1992, Incentive Regulation of Nursing Homes, *Journal of Health Economics*, vol. 11(2): 105-28.

Perry, James L. en Lois R. Wise, 1990, The Motivational Bases of Public Service, *Public Administration Review*, vol. 50(3): 367-373.

Perry, James L., Annie Hondeghem en Lois R. Wise, 2010, Revisiting the Motivational Bases of Public Service: Twenty Years of Research and an Agenda for the Future, *Public Administration Review*, vol. 70(5): 681-90.

Petersen, Laura A., LeChauncy D. Woodard, Tracy Urech, Christina Daw en Supicha Sookanan, 2006, Does Pay-for-performance Improve the Quality of Health Care?, *Annals of Internal Medicine*, vol. 145: 265-72.

Prendergast, Canice, 1999, The Provision of Incentives in Firms, *Journal of Economic Literature*, vol. 37(1): 7-63.

Proper, Carol, Matt Sutton, Carolyn Whitnall en Frank Windmeijer, 2010, Incentives and Targets in Hospital Care: Evidence from a Natural Experiment, *Journal of Public Economics,* vol. 94(3): 318-35.

Roski, Joachim, Robert Jeddeloh, Larry An, Harry Lando, Peter Hannan, Carmen Hall en Shu-Hong Zhu, 2003, The Impact of Financial Incentives and a Patient Registry on Preventive Care Quality: Increasing Provider Adherence to Evidence-based Smoking Cessation Practice Guidelines, *Preventive Medicine*, vol. 36(3): 291-99.

Shen, Y., 2003, Selection Incentives in a Performance-based Contracting System, *Health Services Research*, vol. 38(2): 535-52.

Silverman, Elaine en Jonathan Skinner, 2004, Medicare Upcoding and Hospital Ownership*, Journal of Health Economics*, vol. 23: 369-89.

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

Steel, N., S. Maisey, A. Clark, R. Fleetcroft en A. Howe, 2007, Quality of Clinical Primary Care and Targeted Incentive Payments: An Observational Study, *British Journal of General Practice*, vol. 57(539): 449-54.

Steijn, Bram, 2008, Person-Environment Fit and Public Service Motivation, *International Public Management Journal*, vol. 11(1): 13-27.

Thiel, Sandra van en Frans L. Leeuw, 2002, The Performance Paradox in the Public Sector, *Public Performance and Management Review*, vol. 25(3): 267-81.

Tran, Anh en Richard Zeckhauser, 2012, Rank as an Inherent Incentive: Evidence from a Field Experiment, *Journal of Public Economics*, vol. 96(9): 645-50.

Vaghela, P., M. Ashworth, P. Schofield and M.C. Gulliford, 2009, Population Intermediate Outcomes of Diabetes Under Pay-for-performance Incentives in England from 2004-2008, *Diabetes Care*, vol. 32(3): 427-9.

Wilson, J.Q., 1989, *Bureaucracy: What Government Agencies Do and Why They Do It*. New York, Basic Books.

Weibel, A., K. Rost en M. Osterloh, 2009, Pay for Performance in the Public Sector - Benefits and (Hidden) Costs, *Journal of Public Administration Research and Theory,* vol. 20(2): 387-412.

Bijlage 1 Overzicht studies in de zorg

Auteur (jaar)	Instrument	Ontvanger	Methode	Omschrijving	Resultaat	Averechts effect	Effect
Norton (1992)	Beloning voor opname ($3-$28), verbetering in gezondheid na 90 dagen ($126-$370) en voor ontslag ($60-$230).	Verzorgingshuis	RCT	Verzorgingshuizen werden beloond voor het opnemen van patiënten, de verbetering in gezondheid tijdens het verblijf en het tijdig ontslaan van patiënten (mits patiënten niet binnen 90 dagen weer heropgenomen moesten worden).	Er werd een positief effect op de kwaliteit en efficiency gevonden. Bovendien namen verzorgingshuizen zwaardere patiënten aan en was de behandeltijd voor deze patiënten korter.	Geen.	Positief
Kouides et al. (1998)	Beloning van $0,80 per vaccinatie bij een vaccinatiegraad (op praktijkniveau) van ≥70% of een beloning van $1,60 per vaccinatie bij een vaccinatiegraad (op praktijkniveau) van ≥85%.	Huisartsenpraktijk	RCT	Huisartsenpraktijken werden beloond voor de griepvaccinatie van ouderen (≥65 jaar) die vanwege ambulante zorg de praktijk bezochten.	De vaccinatiegraad was met 7% gestegen als gevolg van de incentive.	Geen.	Positief
Fairbrother et al. (1999)	Beloning voor de praktijk oplopend tot $8000 als 80% van de vaccinaties up-to-date was, hogere stuksbeloning ($5) en feedback.	Kinderarts/praktijk	RCT	Kinderartsen (en praktijken) werden beloond voor het up-to-date houden van vaccinaties voor kinderen (DTP etc.). Onderzoekers verzamelden informatie uit de vastgelegde patiëntenstatus.	Alleen de beloning op praktijkniveau had het gewenste effect. Het aantal kinderen waarvan de registratie en vaccinatie up-to-date was steeg met 25% in deze groep.	Onderzoekers vinden aanwijzingen dat kinderartsen de vaccinaties beter zijn gaan registreren. Effect wordt mogelijk (deels) verklaard door een verbetering in registratie.	Positief
Christensen et al. (2000)	Beloning van $4 voor <6 minuten mondeling advies of $6 voor ≥6 minuten mondeling advies.	Apothekers	RCT	Apothekers werden beloond voor het geven van advies over medicijngebruik wanneer medicijnen werden afgehaald.	Apothekers gaven vaker adviezen bij de uitgifte van medicijnen wanneer zij daarvoor beloond werden dan apothekers die daar	Geen.	Positief

Auteur (jaar)	Instrument	Ontvanger	Methode	Omschrijving	Resultaat niet voor beloond werden.	Averechts effect	Effect
Fairbrother et al. (2001)	Beloning oplopend tot $7500 als 90% van de vaccinaties up-to-date waren of een hogere stuksbeloning per vaccinatie ($5).	Arts/Praktijk	RCT	Artsen en hun praktijk werden beloond voor het verbeteren en up-to-date houden van vaccinaties.	Het bonussysteem had een positief effect van 6% op de registratie van vaccinaties en de hogere stuksbeloning een positief effect van 7%.	Een deel van het effect is te verklaren doordat de artsen de vaccinaties die kinderen ergens anders kregen beter registreerden.	Positief
Roski et al. (2003)	Beloning van $5000 voor praktijken met ≤7 werknemers en $8000 voor praktijken met ≥8 medewerkers als ≥75% van de rokers geregistreerd werd en ≥65% een advies kreeg om te stoppen.	Praktijk	RCT	Praktijken werden beloond voor het vastleggen van rokers onder patiënten en het adviseren over het stoppen met roken.	Verbetering van de registratie van het aantal rokers maar geen effect op het daadwerkelijke aantal personen dat stopt met roken.	Geen.	Positief
Lindenauer et al. (2007)	Prestatiebeloning van ziekenhuizen op basis van kwaliteitsindicatoren.	Ziekenhuis	DiD	Ziekenhuizen konden zich aanmelden om deel te nemen aan een project met resultaatafhankelijke beloning (zelfselectie naar de treatmentgroep kan dus een probleem zijn). Deze ziekenhuizen werden beloond voor de voortgang op 4 geaggregeerde en 10 individuele indicatoren. Totale steekproef (inclusief controle groep) bestond uit ziekenhuizen die vrijwillig indicatoren openbaar maakten.	Ziekenhuizen die deelnamen aan het project met resultaatafhankelijke beloning en transparantie lieten een sterkere verbetering zien op de kwaliteitsindicatoren dan ziekenhuizen die alleen deelnamen aan het project met transparantie.	Geen.	Positief
Propper et al. (2010)	Opstellen van targets voor maximale wachttijden voor een ingreep en het monitoren van de ontwikkeling van wachttijden.	Ziekenhuis	DiD	Auteurs maken gebruik van een natuurlijk experiment waarbij maximale wachttijden voor niet spoedeisende zorg werden opgelegd aan ziekenhuizen. Managers van ziekenhuizen die de target niet haalden konden worden ontslagen en managers van ziekenhuizen die de targets wel haalden kregen meer	De wachttijd nam gemiddeld genomen af met 13 dagen (met grote uitschieters aan de bovenkant van de verdeling). De strenge targets hadden geen negatief effect op andere taken zoals spoedeisende hulp of de behandelduur.	Mogelijke aanwijzingen gevonden van manipulatie van de wachtlijsten. Sommige patiënten werden tijdelijk of permanent van de wachtlijst gehaald.	Positief

Auteur (jaar)	Instrument	Ontvanger	Methode	Omschrijving autonomie.	Resultaat	Averechts effect	Effect
Grady et al. (1997)	Beloning van $50 als 50% van de relevante patiënten werd doorverwezen voor borstkanker screening.	Huisartsenpraktijk	RCT	Huisartsen werden beloond voor het doorverwijzen van oudere vrouwen voor screening op borstkanker. Bovendien werden er posters opgehangen in wachtruimtes en reminders gegeven.	Het geven van een beloning of het ophangen van posters had geen effect op het aantal doorverwijzingen van huisartsen.	Geen.	Geen
Hillman et al. (1998)	Beloning van $570-$1260 bij het voldoen aan richtlijnen en regelmatige feedback over het voldoen aan richtlijnen.	Huisarts	RCT	Huisartsen werden beloond voor het voldoen aan richtlijnen (o.a. richtlijnen voor doorverwijzen) voor het screenen op borstkanker bij vrouwen van boven de 50 jaar.	Het bieden van een bonus had geen effect op het gedrag van de huisartsen.	Geen.	Geen
Hillman et al. (1999)	Beloning (gemiddeld $2000) bij het voldoen aan richtlijnen voor vaccinatie van kinderen of feedback over het voldoen aan deze richtlijnen.	Arts	RCT	Artsen werden beloond voor het voldoen aan richtlijnen voor het vaccineren van kinderen, daarnaast was er een extra groep waarbij alleen feedback werd gegeven over het voldoen aan richtlijnen.	Het bieden van een bonus en feedback of het bieden van alleen feedback had geen effect op het gedrag van artsen.	Geen.	Geen
Dranove et al. (2003)	Vrijgeven van informatie over prestaties (transparantie).	Ziekenhuis	DiD	Er werd informatie vrijgegeven op ziekenhuisniveau over de historische prestaties met betrekking tot bypassoperaties.	Het vrijgeven van informatie over de prestaties van de afdeling chirurgie leidde tot selectie van minder ernstige patiënten (patiënten die beter te behandelen zijn). Positief effect was dat de variatie in ernstigheid van aandoeningen binnen ziekenhuizen afnam: patiënten waren beter in staat te sorteren naar de juiste ziekenhuizen.	Selectie op minder ernstige gevallen die makkelijker te behandelen waren.	Negatief

Bijlage 2 Overzicht studies in het onderwijs

Auteur (jaar)	Instrument	Ontvanger	Methode	Omschrijving	Resultaat	Averechts effect	Effect
Fryer (2011)	Geldbonus van $3000 per leraar voor het behalen van targets, bonus van $1500 per leraar voor behalen 75% van de target.	School (school vertaalt prikkel door naar leraren, meeste kiezen voor teambeloning)	RCT	Openbare scholen in de Verenigde Staten ontvangen schoolbonus voor behalen van targets gebaseerd op school report cards.	Voor basisscholen zijn de effecten niet significant. Voor middelbare scholen gingen wiskunde resultaten met 0.048σ en voor lezen met 0.032σ per jaar achteruit. De effecten op aanwezigheid, goed gedrag, alternatieve prestatie-indicatoren, cijfers en slagingskans zijn nihil. Geen indicatie heterogene effecten. Geen effect gevonden op gedrag van leraren in termen van mobiliteit, verlof of antwoorden op leeromgeving enquête.	Outputindicator te complex, leraar weet niet hoe te beïnvloeden. Meeliftergedrag.	Geen
Goodman en Turner (2010)	Zie Fryer (2011)	Zie Fryer (2011)	RCT	Zie Fryer (2011)	Voor kleine scholen positief effect op inspanningsniveau van leraren gemeten aan de hand ziekteverlof. Effect is echter te klein om toetsresultaten te verbeteren. Geen aanwijzingen dat schoolbeleid of lesmethode wordt aangepast. Geen effect op mobiliteit van leraren of kwalificaties van nieuwe aangenomen leraren.	Mogelijk meeliftergedrag	Geen
Springer et al. (2010)	Geldbonus van $15,000 als testresultaten tot 95ste percentiel	Leraar	RCT	296 wiskundeleraren op middelbare scholen uit de Metropolitan Nashville School System hebben zich opgegeven voor het	Geen effect gevonden op toetsresultaten van scholeren of gedrag van leraren (geen aanpassingen lesmethode of	Geen	Geen

	behoren binnen de regio, $10,000 als testresultaten in 90ste percentiel vallen en $5,000 als resultaten binnen 80ste percentiel vallen.			experiment. Een willekeurige groep ontving de prikkel gekoppeld aan de testresultaten van wiskunde.	samenwerking).		
Buurman et al. (2014)	Geven van feedback aan leraren.	Leraar	RCT	Analyseren bij een MBO school het effect van het geven van feedback door studenten aan docenten op de prestaties van docenten.	Het geven van feedback aan docenten heeft geen effect op latere prestaties. Vinden ook geen effect op de zelfevaluatie of arbeidstevredenheid van docenten.	Geen.	Geen
Atkinson et al. (2009)	Geldbonus van £2000 en een hogere salaristrede (meer salaris en betere groeimogelijkheden salaris).	Leraar	DID	Evalueren de introductie van een systeem met resultaatafhankelijke beloning. Beloning afhankelijk van de ontwikkeling van de leraar en de testscores van de leerlingen.	Positief effect op de toetsresultaten van leerlingen. Effect gelijk aan een halve punt hogere score (met uitzondering van het vak wiskunde).	Geen (wel toets op aanwijzingen voor gaming, niet gevonden)	Effectief
Clotfelter et al. (2008)	Geldbonus van $1800 voor leraar die bij een zwakke school blijft werken.	Leraar	DID	Introductie van financiële prikkel om leraren (wiskunde, natuurkunde of bijzonder onderwijs) te behouden bij scholen in achterstandswijken of scholen met zwakke toetsresultaten.	Leraren op zwakke scholen hadden een 17% lagere kans om te vertrekken bij een school ten opzichte van de situatie zonder geldbonus. Effect sterker voor ervaren leerkrachten.	Geen (wel aanwijzing dat criteria om mee te mogen doen onduidelijk of onvoldoende is gecommuniceerd).	Effectief
Cooper en Cohn (1997)	Pakket 1: geldbonus gekoppeld aan aanwezigheid, evaluatie, zelfstudie, leerprestaties leerlingen. Bij goede beoordeling bonus van $2000-	Leraar/school	Matching	Analyse van verschillende beloningsstructuren in de VS.	Beide beloningspakketten hebben een significant positief effect, maar het eerste pakket lijkt effectiever.	Niet van toepassing	Effectief

	$3000. Pakket 2: in aanvulling op pakket 1 ook een schoolbonus.						
Figlio en Kenny (2007)	Financiële- en niet financiële prikkels	Leraar	Cross sectie	Analyse van data afkomstig uit de National Education Longitudinal Survey (VS) gekoppeld aan survey-data onder 2000 scholen over beloningsstructuur.	Toetsresultaten beter bij scholen met resultaatafhankelijke beloning. Omvang effect van een relatief hoge prikkelstructuur is vergelijkbaar met het effect van 3-jaar extra onderwijs genoten door moeder. Een jaarlijkse evaluatie van de prestaties van de leraar blijkt effectief, maar meer frequent evalueren draagt verder niet bij aan de verbetering van de prestaties. Effecten sterkst in achtergestelde regio's.	Niet van toepassing	
Jacob (2005)	Targets en transparantie	School	DID	Evalueert programma geïmplementeerd bij publieke scholen in Chicago in 1996-1997 naar aanleiding van de No Child Left Behind Act dat alle staten in de VS verplicht om scholieren van groep 3, 6 en 8 jaarlijks te toetsen en de resultaten transparant te maken. Slecht presterende scholen liepen de kans om gesloten te worden. Leerlingen die de toets niet haalden moesten verplicht in de zomer lessen volgen. Als tweede toets weer niet succesvol, dan blijven zitten.	Toetsresultaten in wiskunde en lezen verbeteren aanzienlijk, maar een groot deel de resultaten is toe te schrijven aan strategisch gedrag aan de kant van de leraren en hogere inspanning door leerlingen.	Aanwijzingen dat strategisch gedrag door leraren het beeld bepalen: 1) prestaties op high stakes toetsen verbeteren, maar low stakes toetsen niet, 2) zwakkere leerlingen worden doorverwezen naar bijzonder onderwijs, 3) zwakkere leerlingen jaar laten overdoen en 4) ontmoedigen van het nemen van de toets op low stakes vakken zoals natuurkunde en maatschappijleer	Effectief
Ladd (1999)	Geldbonus van $1000 voor	School	DID	Evalueert de Dallas Independent School District programma van	Positief effect op slagingskans scholieren bij de vakken wiskunde	Geen	Effectief

	schooldirecteuren en leraren, $500 voor ondersteunend personeel en een bijdrage van $2000 ten behoeve van schoolactiviteiten.			1991. Scholen worden geëvalueerd op verschillende prestatie-indicatoren (om teaching to the test tegen te gaan), beoordeling corrigeert voor socio-economische status en compositie van studenten die de toetsen nemen (om strategisch gedrag leraren te minimaliseren). De best presterende scholen (top 20%l) ontvangen ook een financiële bonus.	en lezen. Positief effect niet gevonden voor afro-Amerikaans scholieren, wel bij Latijns-Amerikaanse en blanke scholieren. Mogelijk ook vermindering van schooluitval.		
Lavy (2008)	Geldbonus voor prestatie relatief ten opzicht van de prestatie van collega's.	Leraar	RCT	Evaluatie van een experiment onder 9 scholen in Israël waar leraren (wiskunde en talen) een geldbonus konden verdienen als hun prestatie beter was dan vergelijkbare leraren (zelfde vak/school). Studie legt nadruk op verschillen in effecten tussen mannelijke en vrouwelijke leraren.	Prikkels zijn effectief in verhogen slagingskans en toetsresultaten. Prikkels blijken even effectief voor mannen als voor vrouwen, maar vrouwen blijken meer pessimistisch over de effectiviteit van prikkels en meer realistisch over de eigen kans op het krijgen van de bonus.	Geen	Effectief
Figlio en Winicki (2005)	Prestaties scholen werden gemonitord, slecht presterende scholen liepen het risico gesloten te worden.	School	DID	Analyseert een willekeurige steekproef van scholen in Virginia met gegevens over toetsresultaten en voedingswaarde van de school lunch.	Scholen die in de gevarenzone zitten voor wat betreft mogelijke sancties, verhogen het calorie gehalte van de lunch van scholieren (17% meer calorieën dan normaal). Hogere inname calorieën is een bekende stimulans voor verbeteren cognitieve vaardigheden op de korte termijn. Scholen in de gevarenzone die de lunch aanpasten hadden een hoger slagingspercentage (11% voor wiskunde, en 6% voor Engels en geschiedenis/ maatschappijleer).	Gaming in de vorm van manipuleren van samenstelling van de lunch.	Effectief
Jacob en Levitt (2003)	Slecht presterende scholen worden	Leraar/school	NVT	Prestatie scholen werd geanalyseerd om fraude te	Prikkels leiden tot vals spelen. Bijvoorbeeld, in een klas waar	Fraude door leraren (bewerken van toetsen	Geen

	onder toezicht gesteld en lopen het risico op sluiting. Scholieren moeten een bepaald niveau in wiskunde of lezen behalen voordat ze mogen doorstromen naar een hogere klas.			detecteren.	toetsresultaten het vorige jaar een standaard deviatie onder het gemiddelde lagen, is de kans op tentamenfraude 23% hoger.	van leerlingen).	
Eberts et al. (2002)	Basisbedrag per lesuur plus een bonus van 12,5% als 80% van de gestarte leerlingen de cursus afmaakt. Verhoging van 5% van het basisbedrag en 10% van de bonus bij constant hoge studentevaluaties (4 kwartalen lang).	Leraar	DID	Vergelijkt een school in Michigan met individuele resultaatafhankelijke beloning met een andere school zonder resultaatafhankelijke beloning. Doel van maatregel was om uitval van vakken te beperken.	Het beoogde doel is behaald: prikkels leiden tot minder uitval van leerlingen. Geen positief effect gevonden op toetsresultaten, aanwezigheid of slagingskans.	Aanpassing lesmethode zodat het aantrekkelijker is voor leerlingen (e.g. meer studiereizen). Mogelijk ten koste van kwaliteit van de lessen. Het behouden van leerlingen die anders met vak gestopt waren, drukt op toetsresultaten.	Effectief
Van Elk en Kok (2014)	Evaluatie leraren, feedback, coaching en invoeren nieuwe methoden.	Leraar/school	DID	Kwaliteitsaanpak Basisonderwijs Amsterdam is een integrale aanpak waarbij verschillende maatregelen worden gecombineerd, zoals evaluaties van de kwaliteit van leraren op basis van lesobservaties, scholing en coaching van onderwijspersoneel, en de invoering van nieuwe lesmethoden.	Het beleid heeft, in vergelijking met de controlegroep, geleid tot een daling van de CITO-scores met 1,7 punt in de eerste vier jaar na invoering.	Het intensieve en veeleisende karakter van het programma heeft geleid tot weerstand onder en vertrek van leraren. Mogelijk gaat het om aanpassingskosten.	Negatief

Dit is een uitgave van:

Centraal Planbureau
Van Stolkweg 14
Postbus 80510 | 2508 GM Den Haag
T (070) 3383 380

info@cpb.nl | www.cpb.nl

November 2014