Veldexperimenten in de praktijk: opzet, uitvoering en analyse

Nadine Ketel en Sandra Vriend

Veldexperimenten vormen, mits correct uitgevoerd, een overtuigende methode om het causale effect van een interventie te schatten. In dit praktisch ingestoken artikel worden aandachtspunten besproken die van belang zijn bij de opzet en uitvoering van experimenten. Aan de hand van een drietal veldexperimenten met uiteenlopende beleidsvraagstukken worden deze aandachtspunten toegelicht en wordt besproken hoe hiermee, zowel in de opzet als in de data-analyse, rekening gehouden kan worden.

1 Inleiding

Steeds vaker wordt, bijvoorbeeld door overheden, gevraagd om 'evidence-based' beleid, met een wetenschappelijke onderbouwing van de (te verwachten) causale effecten. Om het causale effect van een bepaald beleid te meten, dient de uitkomst met de interventie vergeleken te worden met de uitkomst zonder de interventie. Per definitie wordt altijd slechts één van deze twee uitkomsten geobserveerd. Er zijn in de afgelopen decennia tal van quasi-experimentele econometrische methoden ontwikkeld om empirisch het effect van een beleidsinterventie te kunnen schatten, maar die berusten vaak op (strenge) aannames. Daarom wordt in de economische wetenschap steeds vaker gebruik gemaakt van gerandomiseerde veldexperimenten. Daarbij worden participanten willekeurig ingedeeld in een groep die wordt blootgesteld aan de interventie en een controlegroep die niet wordt blootgesteld aan die interventie. De randomisatie maakt het mogelijk om verschillen in uitkomsten tussen deze twee groepen toe te wijzen aan de interventie.

Een veldexperiment is, mits correct uitgevoerd, een overtuigende manier om het effect van beleid te meten. Er zijn bij de uitvoering echter verschillende valkuilen die de resultaten kunnen vervuilen. In dit praktisch ingestoken artikel bespreken wij enkele hordes bij de opzet van veldexperimenten en ervaringen daarmee in een drietal recentelijk uitgevoerde veldexperimenten. We bespreken hoe rekening kan worden gehouden met deze hordes bij de opzet van het experiment en wat de gevolgen zijn voor de data-analyse. De aandachtspunten die uitgebreid aan bod komen, zijn:

1. de informatievoorziening rondom een experiment en het *Hawthorne*-effect;

2. naleving van de opdrachten van een experiment door deelnemers en uitvoerders van het experiment (*compliance*);
3. het niveau van randomisatie en;
4. de externe validiteit van experimenten.

De besproken experimenten bestrijken een breed beleidsspectrum met toepassingen in de zorg, het onderwijs en de sociale zekerheid. Bovendien laten we zien dat, wanneer de genoemde hordes bij de opzet en uitvoering van veldexperimenten omzeild kunnen worden door de juiste technieken te benutten, veldexperimenten belangrijke beleidslessen kunnen opleveren en invloed kunnen hebben op de beleidspraktijk.

In de volgende sectie worden eerst diverse algemene aandachtspunten bij de opzet en uitvoering van veldexperimenten besproken. Een drietal van die aandachtspunten wordt vervolgens in de toepassingen uitgebreid toegelicht en gekoppeld aan analyses. Sectie 3 kijkt naar een experiment in de markt voor langdurige zorg. In sectie 4 wordt een veldexperiment in de sociale zekerheid besproken en sectie 5 beschrijft een experiment over collegegelden. Ten slotte geeft sectie 6 de conclusies van het artikel weer.

2 Aandachtspunten bij de opzet van een experiment

Bij het opzetten van een experiment dienen enkele belangrijke keuzes te worden gemaakt, die deels afhankelijk zijn van het effect waarin men geïnteresseerd is, maar die ook gedreven kunnen worden door de context waarin het experiment zal worden uitgevoerd.[1] In sommige gevallen is het mogelijk om perfect te randomiseren, maar dat is niet altijd aan te houden in een complexe omgeving. Indien hiervan wordt afgeweken, moet worden bekeken wat de gevolgen daarvan zijn voor de analyse van de data en of de experimentele opzet in dat geval nog steeds tot valide resultaten kan leiden. In deze sectie bespreken we in het algemeen een aantal belangrijke aspecten, te weten:

- de methode van randomisatie;
- de eenheid waarop wordt gerandomiseerd;
- de benodigde steekproefomvang en looptijd van het experiment;
- de informatievoorziening rondom een experiment;
- de benodigde gegevens en daaraan gerelateerd mogelijke uitval van deelnemers aan het experiment;
- de externe validiteit van een veldexperiment.

[1] Een uitgebreide behandeling van de afwegingen bij het maken van deze keuzes is te vinden in Duflo et al. (2008) en List (2011). Kooreman en Potters (2011) noemen in hun uiteenzetting van het wetenschappelijke en maatschappelijke belang van veldexperimenten nog enkele andere aandachtspunten.

Een eerste keuze betreft de methode van randomisatie. Een veel gebruikte methode is om volledig willekeurig te bepalen wie de interventie krijgt en wie niet, of om gestratificeerd te randomiseren. Dit is echter niet in elke situatie haalbaar. Bijvoorbeeld in het geval van aanvragers van een bijstandsuitkering, een groep individuen die centraal staat in het in sectie 4 besproken experiment, kunnen er ethische bezwaren zijn tegen het onthouden van bepaald beleid aan een deel van de klanten. In 2008 werd dit bezwaar door Freek van Ossel, toenmalig wethouder van 'Werk en Inkomen' in Amsterdam, nog gebruikt als reden om een veldexperiment voor deze groep af te wijzen: "*Het idee om bijstandsgerechtigden (…) te selecteren om ze moedwillig de mogelijkheid te onthouden het traject te volgen, is om logische redenen afgewezen. De inspanningen van gemeenten en het re-integratiebudget zijn er om mensen met een uitkering aan een baan te helpen, niet om wetenschappers van de straat te houden."* (*NRC Handelsblad,* 27 november 2008). In 2011 was het klimaat in zoverre veranderd dat de Dienst Werk en Inkomen in Amsterdam wel open stond voor een veldexperiment. In dit experiment werd echter geen gebruik gemaakt van volledige randomisatie maar van een zogenaamd *encouragement design* (Behaghel et al. 2013). Klantmanagers kregen in dit experiment een standaardkeuze, waar ze alleen van mochten afwijken als toepassing van de standaardkeuze tot schrijnende situaties zou leiden. Deze uitwijkmogelijkheid was essentieel voor de acceptatie van het experiment binnen de organisatie. In sectie 4 bespreken we uitgebreid welke gevolgen deze opzet heeft voor de analyse van de data.

Een tweede keuze is de eenheid waarop randomisatie plaatsvindt. Zo kan bijvoorbeeld op individueel niveau, op instellingsniveau of op gemeenteniveau worden gerandomiseerd. In sommige gevallen wordt de randomisatie-eenheid direct bepaald door de interventie en het effect waar men in geïnteresseerd is. In het onderzoek naar de effecten van het toetsingsmoment op het gedrag van zorgaanbieders in de markt voor langdurige zorg bepaalt de interesse in gedragseffecten bij zorgaanbieders dat zorgaanbieders moeten worden gerandomiseerd over *treatment*- en controlegroepen. In andere situaties speelt in de keuze voor de eenheid van randomisatie mee in welke mate *treatment*- en controlegroep elkaar kunnen beïnvloeden en daarmee *spillover*-effecten veroorzaken. De randomisatie-eenheid dient zo te worden gekozen dat de mogelijkheden voor *spillover*-effecten worden beperkt. Gegeven de gekozen randomisatie-eenheid is het essentieel om in de analyses de uitkomsten op dat niveau te vergelijken. Alleen op dat niveau is immers sprake van (volledige) randomisatie.

Ten derde is het van belang om vooraf te bepalen welke looptijd van het experiment en hoeveel deelnemers nodig zijn om met voldoende zekerheid te kunnen vaststellen of de interventie een effect van een bepaalde omvang heeft. Daarvoor kan gebruik worden gemaakt van zogenaamde *power*-analyses, waarbij sprake is van een afruil tussen de omvang van het te vinden effect en de schaal van het experiment. De *power* van een bepaald experimenteel ontwerp hangt af van een aantal factoren, waaronder de methode van randomisatie en de randomisatie-

eenheid in combinatie met de eenheid waarop de gegevens beschikbaar zijn. Wanneer er geen volledige randomisatie plaatsvindt, zoals in het eerder besproken *encouragement design*, moet in de *power*-analyse ook rekening worden gehouden met *compliance*. Als er vaak wordt afgeweken van de initiële randomisatie is een grotere steekproef nodig.

Een vierde aandachtspunt heeft te maken met de informatievoorziening rondom het experiment, zowel aan deelnemers van het experiment als aan de uitvoerders. Wanneer deelnemers wordt verteld dat zij onderdeel zijn van een experiment, kan dit hun gedrag direct beïnvloeden (het *Hawthorne*-effect). Generalisatie van gemeten effecten naar andere situaties kan daardoor worden bemoeilijkt. Bij voorkeur worden de participanten daarom niet geïnformeerd, maar dit is in de praktijk niet altijd mogelijk. Zo was het bijvoorbeeld in het experiment in de langdurige zorg, dat in sectie 3 wordt besproken, noodzakelijk om zorgaanbieders in te lichten over het experiment. In sectie 3 bespreken we hoe kan worden bekeken of het informeren op zichzelf gevolgen heeft gehad voor het gedrag van zorgaanbieders in het experiment.

Wijzigingen in de werkwijze van uitvoerders als gevolg van het experiment vereisen eveneens zorgvuldige communicatie. Soms wordt bijvoorbeeld van een uitvoerder gevraagd om willekeurig te bepalen welke behandeling een deelnemer krijgt, terwijl de uitvoerder gewend is hier een afgewogen keuze in te maken. Daarnaast kan bij uitvoerders het idee ontstaan dat een experiment niet opgezet is om het beleid maar om de uitvoerders zelf te evalueren. Steun van de uitvoerders is essentieel, maar in de praktijk blijkt dat deze er alleen zal komen als de top van de organisatie zich committeert aan het experiment. Naast deze *commitment* is controle van de uitvoering essentieel om er voor te zorgen dat er uiteindelijk daadwerkelijk verschil is tussen de behandeling die verschillende *treatment*-groepen krijgen. In sectie 4 bespreken we wat voor controlemechanismen waren ingebouwd in het experiment met een *encouragement design* om voldoende verschillen tussen *treatment*-groepen te krijgen.

De voorgaande aandachtspunten hadden elk te maken met interne validiteit, maar ook externe validiteit, dus de mate waarin resultaten kunnen worden gegeneraliseerd naar andere situaties, is van essentieel belang om tot relevante conclusies en beleidsaanbevelingen te kunnen komen. De aanwezigheid van een *Hawthorne*-effect, zoals hiervoor al benoemd, kan een belangrijke bedreiging vormen voor de externe validiteit van een experiment. In sectie 5 bespreken we een experiment waarbij het initiatief voor het experiment niet direct vanuit de beleidsmakers kwam, maar was gemotiveerd vanuit een wetenschappelijke vraag. De uitdaging hier was om een opzet te vinden waarin deze vraag goed kon worden beantwoord, en waarin mensen bereid waren om mee te werken. Ook hier is de externe validiteit belangrijk: de setting moet niet te kunstmatig zijn zodat de resultaten generaliseerbaar zijn.

In de volgende secties behandelen we drie van de hiervoor genoemde aandachtspunten in detail. We bespreken telkens hoe dit aandachtspunt in het betreffende experiment naar voren is gekomen en wat voor gevolgen dit heeft

gehad voor de analyses en de effectmeting. In een eerste toepassing is onderzocht wat het effect is van het toetsingsmoment op de kwaliteit en kwantiteit van aanvragen voor langdurige zorg ingediend door zorgaanbieders. Hierbij was de informatievoorziening een belangrijk aandachtspunt. In een tweede toepassing kijken we naar het effect van re-integratieinstrumenten, specifiek het opleggen van een zoekperiode, op de uitstroom van bijstandsgerechtigden. Naleving van de uitvoering van het experiment bleek in dit onderzoek een belangrijk aspect te zijn. Een derde toepassing kijkt ten slotte naar het effect van het geven van een korting op de kosten van bijlessen op de aanwezigheid en prestatie van studenten. Voor dit experiment gaan we in op de externe validiteit en het niveau van randomisatie.

3 Een experiment in de markt voor langdurige zorg

Achtergrond. Sinds 2005 functioneert het Centrum Indicatiestelling Zorg (CIZ) als poortwachter voor de toegang tot (een deel van) de AWBZ-gefinancierde langdurige zorg. Om gebruik te kunnen maken van langdurige zorg, moet een aanvraag bij het CIZ worden ingediend. Dit wordt in de regel gedaan door de aanbieders van langdurige zorg. Voordat zorg kan worden ingezet, dient de aanvraag door het CIZ te worden omgezet in een indicatie. Voor een willekeurige steekproef van de aanvragen gaat dit gepaard met een toetsingsprocedure waarin wordt bekeken of de aangevraagde zorg (bijvoorbeeld type, hoeveelheid en leveringsvorm) overeenkomt met regels en richtlijnen en de benodigde zorg voor de cliënt. Iedere getoetste aanvraag krijgt een label ‘conform’ of ‘niet conform’. Niet conform betekent dat er een verschil is tussen de aangevraagde zorg enerzijds en de regels en richtlijnen en benodigde zorg anderzijds. Een niet-conforme toetsing kan ertoe leiden dat de inzetbare zorg afwijkt van de aangevraagde zorg.

Aangezien de beschikbare middelen voor het uitvoeren van toetsingen beperkt zijn, kwam vanuit het CIZ de vraag op hoe de kleine geldstroom (de operationele kosten die gemoeid zijn met de verwerking en toetsing van aanvragen) de grote geldstroom (de AWBZ-uitgaven) kan beïnvloeden. Enerzijds kan een bijstelling van de aangevraagde zorg bij een niet-conform toetsingsresultaat voor een directe verandering van de AWBZ-uitgaven zorgen. Anderzijds kan de inrichting van de toetsingspraktijk invloed hebben op het aanvraaggedrag van zorgaanbieders, bijvoorbeeld de kwaliteit van aanvragen, en daarmee indirect, zelfs als geen toetsing plaatsvindt, de AWBZ-uitgaven beïnvloeden.

Er zijn verschillende instrumenten in het toetsingsbeleid waarmee kan worden gevarieerd. Een daarvan is het moment van toetsing. Van september 2012 tot april 2013 hebben Lindeboom et al. (2013, 2015) in een veldexperiment onderzocht wat het effect is van het toetsingsmoment op het aanvraaggedrag van zorgaanbieders (het aantal aanvragen en de kwaliteit van de ingediende aanvragen, waarbij het laatste wordt gemeten door de behaalde fractie conforme toetsingen). Het moment van toetsing bepaalt of bijstelling van de aangevraagde zorg mogelijk is bij een

niet-conform resultaat. In dit experiment werd onderscheid gemaakt tussen drie groepen:

1. een groep met *ex-ante* (vooraf) toetsing waarbij de mogelijkheid tot correctie van de aangevraagde zorg bestaat;
2. een groep met *ex-post* (achteraf) toetsing, in welk geval het niet mogelijk is om de inhoud van de zorgvraag aan te passen naar aanleiding van het toetsingsresultaat, en
3. een groep waarin het moment van toetsing werd aangepast op basis van behaalde toetsingsresultaten.

In deze laatste groep vertaalde een hoog conformpercentage zich in *ex-post* toetsing, terwijl zorgaanbieders met lage conformpercentages vooraf werden getoetst (Lindeboom et al. 2015).

Informatievoorziening en het *Hawthorne*-effect. In het experiment werden zorgaanbieders willekeurig toegewezen aan een van de drie bovengenoemde groepen. De zorgaanbieder als randomisatie-eenheid is een logische keuze gezien de interesse in gedragseffecten van zorgaanbieders. Alle deelnemende zorgaanbieders konden volledig willekeurig over de groepen worden verdeeld. Deelname van deze zorgaanbieders was bovendien verplicht. De steun voor het experiment (van de top) binnen het CIZ in combinatie met de verplichte deelname van het merendeel van de zorgaanbieders, maakt dat in dit onderzoek geen sprake is geweest van problemen omtrent de naleving van de experimentele variatie (*noncompliance*) en dat voldoende aanvragen werden ingediend om effecten met voldoende statistische betrouwbaarheid te kunnen vinden.

De informatievoorziening omtrent het experiment is bij dit experiment in het bijzonder een aandachtspunt geweest. Omdat de onderzoekers geïnteresseerd waren in het gedragseffect van een verandering in het toetsingsbeleid en omdat het experiment voor sommige zorgaanbieders directe gevolgen had voor de toetsingsprocedures waarmee zij werden geconfronteerd, was het noodzakelijk om zorgaanbieders vooraf op de hoogte te brengen van het experiment. Zorgaanbieders werden geïnformeerd over de verschillende groepen die binnen het experiment werden onderscheiden. Bovendien werd te kennen gegeven welke verandering in het moment van toetsing de zorgaanbieders tijdens het experiment zouden ondervinden. De uitvoerders van het experiment, dus relatiebeheerders en toetsers bij CIZ, zijn vooraf op de hoogte gebracht van de inhoud van het experiment en de implicaties voor hun werk. Tijdens het experiment is voor hen een presentatie gegeven over het verloop van het onderzoek en enkele tussentijdse resultaten, om zo problemen vroegtijdig te kunnen onderkennen en uitvoerders te motiveren om het experiment zo goed mogelijk te laten verlopen.

Dat de onderzoekers genoodzaakt waren om zorgaanbieders te informeren over het experiment en de veranderingen in het toetsingsmoment die dat voor hen opleverde, kan directe gedragsveranderingen van zorgaanbieders tot gevolg hebben

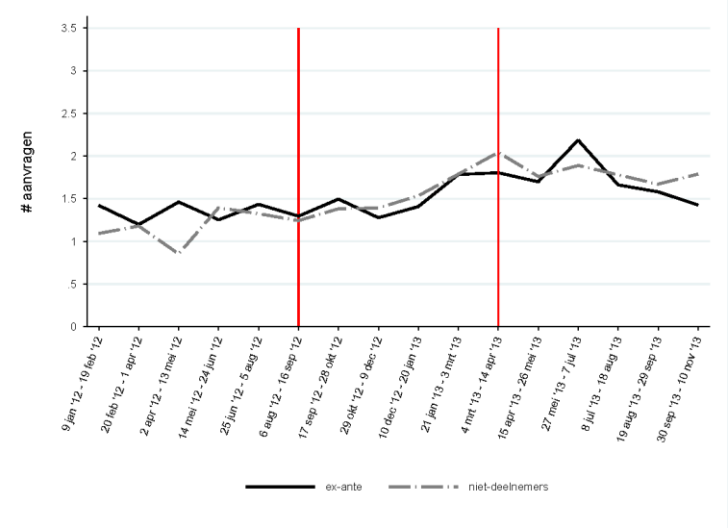
(het eerder benoemde *Hawthorne*-effect). Daarom dient in de analyses te worden bekeken of sprake is van een *Hawthorne*-effect. Dat kan worden gedaan door gebruik te maken van gegevens over niet-deelnemende zorgaanbieders en de uitkomst voor deze groep te vergelijken met de uitkomst voor de groep waarvoor het toetsingsmoment niet wijzigt ten opzichte van de situatie voorafgaand aan het experiment (in dit geval de groep met vooraf toetsing). De niet-deelnemende zorgaanbieders zijn niet direct geïnformeerd over het experiment en voor deze groep valt dus ook geen gedragsreactie op basis daarvan te verwachten. Wanneer de trend in de uitkomst voor de niet-deelnemende zorgaanbieders gelijk is aan de trend voor de zorgaanbieders in de groep met vooraf toetsing, is het niet waarschijnlijk dat het gedrag van zorgaanbieders in de groep met vooraf toetsing gedurende het experiment toegeschreven kan worden aan een *Hawthorne*-effect.

Figuur 1 laat de trend in het aantal aanvragen zien voor de groep zorgaanbieders die tijdens het experiment, net als voorafgaand aan het experiment, vooraf getoetst werd en de niet-deelnemende zorgaanbieders.[2] Voorafgaand aan het experiment was het aantal aanvragen dat werd ingediend constant. Tijdens het experiment is een stijgende trend te zien in het aantal aanvragen. Dit zou (gedeeltelijk) het gevolg kunnen zijn van gedragsaanpassingen vanwege deelname in een experiment. Echter is een soortgelijke trend ook waarneembaar voor niet-deelnemende zorgaanbieders.[3] Dat maakt het onwaarschijnlijk dat de geschatte effecten van de variatie in het toetsingsmoment vervuild zijn door de aanwezigheid van een *Hawthorne*-effect.

[2] De niet-deelnemende zorgaanbieders waren gemiddeld kleiner in termen van het aantal ingediende aanvragen dan de zorgaanbieders in de groep met vooraf toetsing. De grafiek gaat uit van niet-deelnemende zorgaanbieders met een aantal aanvragen in dat gebied waarin ook zorgaanbieders in de andere groep vallen, en *vice versa* voor de zorgaanbieders in de groep met vooraf toetsing. Ten slotte zijn de niet-deelnemende zorgaanbieders gewogen om de gewogen verdeling van het aantal aanvragen van niet-deelnemende en deelnemende zorgaanbieders gelijk te maken. De grafiek geeft het gewogen gemiddelde aantal aanvragen weer voor de niet-deelnemende zorgaanbieders en het aantal aanvragen voor de selectie van zorgaanbieders in het regime met vooraf toetsing (groep 1).

[3] Formeel kan worden getest of sprake is van een verandering in het aantal aanvragen voor de niet-deelnemende zorgaanbieders ten opzichte van de zorgaanbieders in de groep met vooraf toetsing tijdens de experimentperiode door een gewogen panel *fixed effects* model te schatten. Hieruit volgt dat er geen sprake is van een significant verschil in het aantal aanvragen tijdens het experiment voor deze twee groepen (zie Lindeboom et al. (2015) voor meer details).

Figuur 1 Trend in het (gewogen) gemiddelde aantal aanvragen per zorgaanbieder per week.



Noot: De verticale lijnen geven de start- en einddatum van het experiment weer.

Naast informatievoorziening voorafgaand aan het experiment, was in deze studie ook informatievoorziening tijdens het experiment van belang. Daarbij is het essentieel om vooraf na te denken welke informatie moet worden verschaft, in welke opzet en aan welke groepen in het experiment, zonder dat dit een op zichzelf staand effect kan hebben op de uitkomst waarin de onderzoekers geïnteresseerd zijn. In dit experiment vereiste de periodieke verandering in het toetsingsmoment op basis van recente conformpercentages in de derde *treatment*-groep dat zorgaanbieders in deze groep tijdens het experiment van feedback werden voorzien over de voor hen uitgevoerde toetsingen en de resulterende aanpassing in het toetsingsmoment. Wanneer echter geen soortgelijke informatie zou worden verschaft aan zorgaanbieders in de andere groepen, zou in een effectmeting niet alleen het effect van variatie in toetsingsmoment worden opgepakt, maar eveneens het effect van variatie in de verschafte feedback.

Resultaat. De resultaten van het veldexperiment laten zien dat een regime met vooraf toetsing, en dus de mogelijkheid tot correctie van de aangevraagde zorg, 10% minder aanvragen tot gevolg heeft ten opzichte van een regime met achteraf toetsing. Een conditioneel toetsingsregime gaat gepaard met een significante 20%-afname van het aantal aanvragen ten opzichte van de groep met achteraf toetsing. Echter worden deze afnames voor een belangrijk deel verklaard door substitutie naar soorten aanvragen die niet onderhevig zijn aan de experimentele variatie in het moment van toetsing, zoals beschreven in Lindeboom et al. (2015). Daarnaast vinden we negatieve effecten op de kwaliteit van aanvragen: het gemiddelde conformpercentage daalt voor zowel de groep die vooraf wordt getoetst als de

groep waarin het toetsingsmoment prestatie-afhankelijk is gemaakt. Het gaat hierbij om afnames met 4 procentpunt en 3 procentpunt respectievelijk. Dit lijkt echter gedeeltelijk verklaard te worden uit een verschil in ex-ante en ex-post toetsing bij de toetsers.

Samengevat is ondanks de noodzakelijke informatievoorziening aan deelnemende zorgaanbieders, de effectmeting in dit experiment niet gehinderd door de aanwezigheid van een *Hawthorne*-effect. Mede naar aanleiding van de resultaten van het veldexperiment is begin 2014 het toetsingsbeleid voor sommige typen aanvragen aangepast. Het resultaat dat variatie in het toetsingsmoment effect heeft op het gedrag van zorgaanbieders en dat er minder aanvragen worden ingediend lijkt te suggereren dat vooraf toetsing als kritisch wordt ervaren door zorgaanbieders. In het nieuwe toetsingsbeleid wordt het toetsingspercentage conditioneel, op basis van geobserveerde prestatie, periodiek aangepast.

4 Zoekperiode voor aanvragers van een bijstandsuitkering

Achtergrond. De bezuinigingen van de afgelopen jaren op het re-integratiebudget dwingen uitkeringsinstanties terughoudender te zijn met het aanbieden van re-integratie-instrumenten. Daarnaast dragen gemeentes sinds de invoering van de Wet Werk en Bijstand de financiële verantwoordelijkheid voor de uitvoering van de bijstandswet. Uitkeringsinstanties zijn daarom steeds meer op zoek naar trajecten die daadwerkelijk effectief zijn. Zo kwam in het voorjaar van 2011 de vraag vanuit de Dienst Werk en Inkomen van de gemeente Amsterdam om de effectiviteit te meten van de re-integratie-instrumenten die zij inzetten voor de groep met de relatief kortste afstand tot de arbeidsmarkt. Dit zijn klanten waarvoor ingeschat wordt dat ze binnen zes maanden werk kunnen vinden.

Een van de gebruikte re-integratie-instrumenten is de zogenaamde zoekperiode. Een zoekperiode wordt opgelegd tijdens de aanvraag van een uitkering en stelt behandeling van de uitkeringsaanvraag met maximaal vier weken uit. Gedurende deze vier weken is het de bedoeling dat de klant actief naar werk zoekt. De uitkeringsaanvraag wordt alleen geactiveerd als de klant na de zoekperiode terugkeert. Indien de aanvraag wordt toegekend, krijgt de klant met terugwerkende kracht een bijstandsuitkering vanaf de dag van melding. De facto vertraagt een zoekperiode dus alleen het moment van de eerste uitbetaling. Daarom mogen klantmanagers geen zoekperiode opleggen aan klanten met ernstige financiële problemen. Ook als iemand al aantoonbaar vaak heeft gesolliciteerd voor de uitkeringsaanvraag wordt vaak geen zoekperiode opgelegd.

De effectiviteit van het opleggen van een zoekperiode is in een veldexperiment, dat in maart 2012 van start is gegaan, door Bolhaar et al. (2014) onderzocht. Voorafgaand aan het experiment werd bij ongeveer 40% van de klanten een zoekperiode ingezet. De beslissing om een zoekperiode op te leggen wordt gemaakt door de klantmanager die het intakegesprek voert. Zonder een experimentele opzet is het moeilijk de effectiviteit van een zoekperiode te bepalen.

De zoekperiode werd niet voor alle klanten ingezet, en het is heel aannemelijk dat de groep klanten die wel een zoekperiode krijgt verschilt van de groep die geen zoekperiode krijgt. Verschillen in uitkomsten tussen beide groepen kunnen dan net zo goed het gevolg zijn van het krijgen van een zoekperiode als van al bestaande (voor een deel niet-geobserveerde) verschillen tussen beide groepen. Een veldexperiment maakt het mogelijk om deze laatste verschillen uit te sluiten en daarmee de effectiviteit van de zoekperiode te meten.

Methode van randomisatie: *encouragement design.* De voornaamste zorg die bij de Dienst Werk en Inkomen Amsterdam leefde, was dat een veldexperiment met volledig gerandomiseerd toewijzen van zoekperiodes tot schrijnende gevallen zou kunnen leiden. Daarom is gebruik gemaakt van een zogenaamd *encouragement design*. Randomisatie in deze opzet was op het niveau van de klantmanager. Elke klantmanager kreeg gedurende een periode van drie maanden een standaardkeuze, waarvan alleen mocht worden afgeweken als daar een goede reden voor was. Op deze manier behield de klantmanager de mogelijkheid om ongewenste situaties te voorkomen. Het bieden van deze uitwijkmogelijkheid bleek cruciaal in het overtuigen van de klantmanagers om mee te werken aan het experiment. Omdat klanten willekeurig werden toegewezen aan klantmanagers, leidde dit er automatisch toe dat ook klanten willekeurig werden blootgesteld aan verschillend beleid. Een voordeel van de randomisatie op het niveau van de klantmanager was dat de klanten niet geïnformeerd werden dat ze onderdeel van een experiment waren. Hierdoor was er in dit experiment, in tegenstelling tot het hiervoor besproken experiment, geen zorg over een potentieel *Hawthorne*-effect.

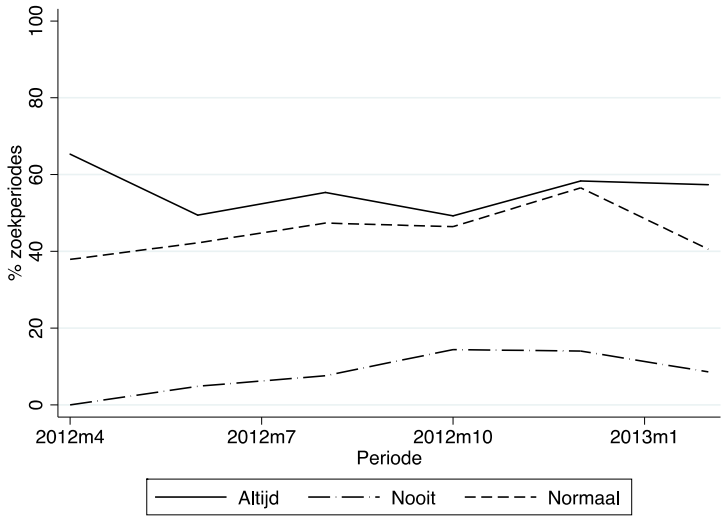
Cruciaal voor het slagen van een experiment met een *encouragement design* is dat het gedrag van de uitvoerders onder de standaardkeuzes voldoende verschilt. Klantmanagers kregen drie verschillende standaardkeuzes: (1) leg zo vaak mogelijk een zoekperiode op, (2) leg nooit een zoekperiode op en (3) bepaal zelf of een zoekperiode opgelegd wordt (normaal beleid). Deze standaardkeuzes worden in de rest van dit stuk aangegeven met respectievelijk ‘altijd’, ‘nooit’ en ‘normaal beleid’.

Naleving van de standaardkeuzes. In de uitvoering werden een aantal elementen ingebouwd die konden bijdragen aan een hogere naleving van de standaardkeuzes. Zo vulden klantmanagers voor iedere klant een formulier in waarop hun standaardkeuze voor die periode voorgedrukt was. Op deze manier werden zij continu herinnerd aan de standaardkeuze. De formulieren boden daarnaast de mogelijkheid om tussentijds bij de klantmanagers langs te komen om de formulieren op te halen. Daardoor bleven de klantmanagers zich steeds bewust van het lopende onderzoek en kon tijdens het experiment worden bijgehouden of ze zich aan de standaardkeuzes hielden.

In Figuur 2 zien we het percentage klanten aan wie een zoekperiode is opgelegd onder de verschillende opdrachten over tijd. Onder de opdracht ‘nooit’ werden bijna geen zoekperiodes opgelegd. Gedurende het jaar neemt het langzaam

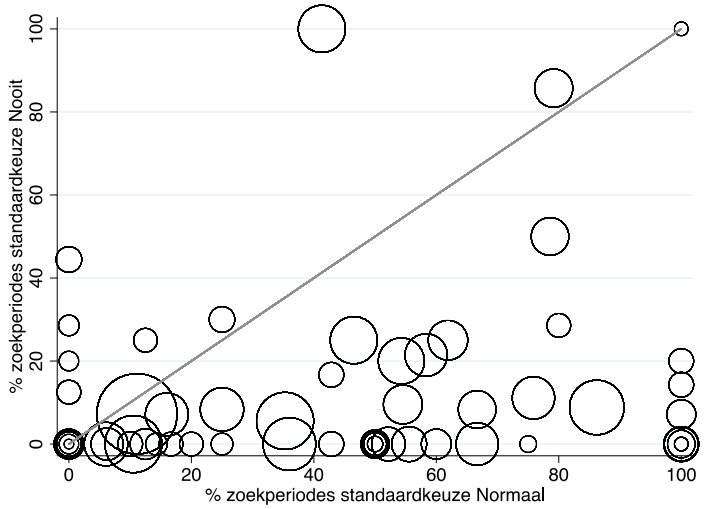
toe, maar het komt nooit boven de 15%. Onder de opdracht ‘normaal beleid’ was het aantal opgelegde zoekperiodes aan het begin van het onderzoek ongeveer 40%, om toe te nemen tot bijna 60% in december 2012. Het verschil tussen de opdracht ‘altijd’ en ‘normaal beleid’ was hierdoor het grootst bij de start van het onderzoek. Het aantal zoekperiodes onder de opdracht ‘altijd’ is redelijk constant over tijd (ongeveer 60%) en ligt over de gehele periode genomen negen procentpunt boven het aantal zoekperiodes onder normaal beleid. Er zijn (zoals verwacht) geen verschillen in de kenmerken van de klanten in de drie groepen, wat betekent dat de groepen alleen verschillen in de kans dat ze een zoekperiode krijgen.

Figuur 2 Aantal opgelegde zoekperiodes over tijd, uitgesplitst per standaardkeuze.



Niet alle klantmanagers hielden zich echter even goed aan het onderzoek. Figuur 3 geeft per klantmanager aan hoe goed de naleving van de opdrachten was. Op de horizontale as is aangegeven welk percentage van de klanten een zoekperiode krijgt van een klantmanager onder ‘normaal beleid’. De verticale as geeft dat onder de standaardkeuze ‘nooit’. De grootte van een cirkel geeft aan hoeveel klanten een klantmanager heeft gehad. Indien de standaardkeuzes niet nageleefd waren en klantmanagers hun gedrag niet veranderen naar aanleiding van de standaardkeuzes, zouden alle bollen zich rond de 45°-lijn bevinden. Als de klantmanagers de opdracht ‘nooit’ volledig naleven zouden alle bollen zich op de horizontale as bevinden. In Figuur 3 is te zien dat een groot deel van de klantmanagers zich (redelijk) aan de opdracht houdt, en een stuk minder zoekperiodes oplegt onder de opdracht ‘nooit’ dan onder de opdracht ‘normaal beleid’. Daarnaast zijn er ook een paar klantmanagers die zich geheel niet aan de opdracht houden en voor wie de bollen op of boven de 45°-lijn liggen.

Figuur 3 Aantal opgelegde zoekperiodes onder de standaardkeuzes 'normaal beleid' en 'nooit', uitgesplitst per klantmanager.



Bij het analyseren van de data zijn er twee manieren om met de niet volledige naleving van de standaardkeuzes om te gaan. Een vergelijking van de uitkomsten van de drie groepen, zonder mee te nemen of er daadwerkelijk een zoekperiode is opgelegd, geeft het zogenaamde *intention-to-treat*-effect. In dit geval is dat te interpreteren als het effect van het hebben van een klantmanager met een bepaalde opdracht. Omdat klantmanagers soms afwijken van de standaardkeuze is dit niet gelijk aan het effect van het opgelegd krijgen van een zoekperiode. Om dit laatste effect te schatten, wordt gebruik gemaakt van een instrumentele-variabelenaanpak, waarin de standaardkeuze van de klantmanager het opleggen van de zoekperiode instrumenteert. Dit geeft het effect van de zoekperiode voor mensen die daadwerkelijk een zoekperiode hebben gekregen. Het is essentieel om hier een instrumentele-variabelenaanpak te gebruiken en niet simpelweg de klanten waarvoor klantmanagers hun *opt-out*-mogelijkheid hebben ingezet weg te laten. Dat is namelijk een keuze van de klantmanager en betreft daarom een (zeer) selectieve groep klanten, waardoor niet het gewenste *treatment*-effect wordt geschat. Of het *intention-to-treat*-effect of het instrumentele-variabeleneffect de voorkeur heeft, hangt af van de beleidsvraag. Indien wordt overwogen om de zoekperiode voor iedereen in te voeren, is het tweede effect het meest interessant. Wanneer ook bij bredere invoering van de zoekperiode de klantmanager de mogelijkheid zal behouden om geen zoekperiode op te leggen, is het *intention-to-treat*-effect een betere weergave van het te verwachten effect.

Resultaat. Omdat de opdrachten tot flinke verschillen leidden in het aantal opgelegde zoekperiodes was het mogelijk om tot een goede schatting van het effect

van een zoekperiode te komen. Het geschatte instrumentele-variabeleneffect van de zoekperiode geeft aan dat een zoekperiode leidt tot 20 procentpunt minder toekenningen van bijstandsuitkeringen. Het gaat hierbij niet alleen om een korte termijn effect, ook na zes maanden is door de zoekperiode het percentage individuen met een bijstandsuitkering nog altijd 12 procentpunt lager. Door het opleggen van de zoekperiode zijn er dus mensen niet in de bijstand terecht gekomen die deze uitkering anders minstens zes maanden hadden ontvangen. De totale bespaarde uitkeringslast is meer dan 800 euro per opgelegde zoekperiode. Klanten compenseren wat ze minder aan bijstandsuitkering ontvangen volledig met meer inkomen uit werk: in de eerste zes maanden wordt gemiddeld 914 euro extra uit werk verdiend. Klanten wijken niet uit naar andere uitkeringen en er is ook geen effect op het gemiddeld uurloon, wat er op wijst dat een zoekperiode er niet toe leidt dat mensen een baan met een lager uurloon accepteren.

Samengevat was het gebruik van het *encouragement design* in dit experiment belangrijk voor de acceptatie van het experiment op de werkvloer en was er, door het inzetten van veel monitoren en communicatie, genoeg naleving van de standaardkeuzes om tot een goede effectmeting te komen. Met de resultaten van het experiment werd voor de Dienst Werk en Inkomen van Amsterdam duidelijk dat een zoekperiode leidt tot een besparing op uitkeringslasten, zonder dat mensen er in inkomen op achteruit gaan. Het gebruik van de zoekperiode is daarom geïntensiveerd. Daarnaast wordt momenteel gekeken om de zoekperiode op een bredere groep klanten toe te passen.

5 Collegegeldexperiment

Achtergrond. In veel landen is onderwijs sterk gesubsidieerd. De argumenten hiervoor zijn onder andere de aanwezigheid van positieve externaliteiten, de imperfectie van kapitaalmarkten (het is niet voor iedereen mogelijk om geld te lenen voor onderwijs) en gelijkheidsoverwegingen. De subsidies voor onderwijs kunnen echter een negatief effect hebben op de tijd en moeite die studenten in hun studie steken. Dit kan op twee manieren: ten eerste trekt een lager collegegeld studenten aan die minder waarde hechten aan het onderwijs; ten tweede kan een lager collegegeld er via het *sunk-cost*-effect voor zorgen dat studenten zich minder verplicht voelen om tijd en moeite in de studie te steken. In het tweede mechanisme kan het bedrag dat wordt betaald dienen als een vrijwillige verbintenis (*commitment*) om meer tijd en moeite in de studie te steken. Het eerste effect (het selectie-effect) is al veel onderzocht, maar studies naar het *sunk-cost*-effect zijn nog schaars.[4]

Om te testen voor het *sunk-cost*-effect is een opzet nodig waarin mensen gemiddeld dezelfde bereidheid hebben om te betalen voor onderwijs, maar in werkelijkheid verschillende prijzen betalen. Omdat deze situatie ongebruikelijk is,

[4] Voor een volledige beschrijving van het onderzoek en de resultaten, zie Ketel et al. (2015).

is gekozen voor een veldexperiment. Idealiter zou de hoogte van het collegegeld hierin worden gevarieerd. Omdat dit een dure aangelegenheid is, is gekeken naar variatie in de kosten van bijlessen. In dit experiment werden kortingen van verschillende omvang gegeven aan studenten die zich hadden ingeschreven voor bijlessen en zich dus al hadden gecommitteerd aan het betalen van het totale bedrag voor de bijles (65 tot 75 euro voor vier of vijf lessen). Er waren vier verschillende *treatments*. Studenten konden een korting krijgen voor het volledige bedrag, moesten 10 euro betalen, kregen een korting van 10 euro, of moesten het volledige bedrag betalen. Op deze manier kon onderscheid worden gemaakt tussen het effect van het geven van een korting, het effect van de grootte van de korting en het effect van niets hoeven betalen (alles ten opzichte van betaling van het volledige bedrag). De omvang van de korting werd bepaald door studenten een envelop te laten trekken uit een stapel. Daardoor waren studenten op de hoogte van het feit dat ze deelnamen in een experiment. Om te voorkomen dat ze zich anders zouden gaan gedragen door het experiment, werd het doel van het experiment zo vaag mogelijk gehouden.

Externe validiteit. In tegenstelling tot de eerder besproken veldexperimenten was er bij de opzet van dit experiment geen directe betrokkenheid van beleidsmakers. De uitdaging was daarom om een opzet te vinden waarin deze vraag goed beantwoord kon worden en waarin mensen bereid waren om mee te werken. De voornaamste zorg hierbij is de externe validiteit van de bevindingen. Eerdere studies naar het *sunk-cost*-effect werden vaak in het lab uitgevoerd, met als voordeel dat de omgeving goed controleerbaar is. Zo is bijvoorbeeld te voorkomen dat studenten met elkaar praten, en worden mogelijke *spillover*-effecten voorkomen. De vraag is echter of resultaten uit zulke experimenten te generaliseren zijn naar een onderwijssituatie waar studenten tijd en moeite in hun studie steken. Een veldexperiment biedt hierbij uitkomst. Het experiment is uitgevoerd in samenwerking met een bedrijf dat bijlessen aanbiedt aan bachelorstudenten (de Bijlespartner). Door deze samenwerking is er een opzet met de juiste doelgroep (universitaire studenten), die geld uitgeven aan het juiste product (onderwijs), en met duidelijk observeerbare uitkomsten (de aanwezigheid bij de lessen). Het bedrijf had zelf geen geldelijk voordeel van de medewerking aan het experiment, maar kon met de experimentele gegevens wel meer inzicht krijgen in de ervaringen van studenten met de bijlessen. Als de onderzoeksvraag niet bij de organisatie zelf vandaan komt kan het zijn dat er minder betrokkenheid is voor een juiste uitvoering van het experiment. Daarom is voor aanvang van het experiment een contract getekend met de belangrijkste afspraken over de uitvoering en over het naar buiten brengen van de resultaten.

Niveau van randomisatie. In dit experiment is er voor gekozen om te randomiseren binnen de bijlesgroepen, in plaats van op groepsniveau. Beide benaderingen hebben voor- en nadelen. Een nadeel van randomisatie binnen een groep is dat er *spillovers* kunnen zijn: als van twee vrienden in een groep de een

wel korting heeft en de ander niet en de persoon zonder korting haalt degene met korting over om toch naar de les te komen, dan zal de onderzoeker geen effect vinden van de korting. Het nadeel van randomisatie op groepsniveau is dat een groep die weinig heeft betaald zou kunnen denken dat de kwaliteit van de docent of de lessen lager is. Dit zou een onafhankelijk effect kunnen hebben op de aanwezigheid in de lessen. Door studenten met verschillende kortingen samen in een klas te hebben, kan dit nooit het geval zijn. Een tweede reden om niet op groepsniveau te randomiseren is *power*; er zouden bij randomisatie op groepsniveau veel meer groepen en dus observaties nodig zijn om een effect te kunnen bepalen. In dit geval wogen de nadelen van randomisatie op groepsniveau zwaarder en is gekozen voor randomisatie binnen groepen.[5]

Resultaat. De voornaamste uitkomstmaat in dit onderzoek was de aanwezigheid in de lessen.[6] Het *sunk-cost*-effect voorspelt dat studenten die een hogere korting krijgen minder vaak naar de bijlessen zullen komen. Dit is het geval voor de studenten die niks betalen: zij waren minder vaak bij alle lessen aanwezig dan studenten in de overige drie groepen. Het patroon in aanwezigheid over de groepen is echter niet geheel zoals verwacht, aangezien de studenten in de groep met een grote korting het meest aanwezig zijn. In de volledige steekproef lijkt daarom geen sprake te zijn van een *sunk-cost*-effect. Een onderdeel van de vragenlijst was een hypothetische *sunk-cost*-vraag.[7] Deze vraag dient om te kijken of bepaalde studenten gevoeliger zijn voor het *sunk-cost*-effect dan anderen. In de gehele steekproef is 45% van de studenten gevoelig voor het *sunk-cost*-effect. Voor die groep studenten lijkt er in het experiment wel een *sunk-cost*-effect te zijn: des te meer een student betaalt voor de cursus, des te vaker de student bij alle lessen aanwezig is. Al met al is er geen onomstotelijk bewijs dat het subsidiëren van onderwijs ertoe leidt dat studenten minder tijd en aandacht aan hun studie besteden, maar suggereren de resultaten voor de groep van *sunk-cost*-gevoelige studenten dat dit effect voor een subgroep wel aanwezig is. Dat de vraag niet direct van beleidsmakers afkomstig was, betekende overigens niet dat er geen interesse voor was bij beleidsmakers. De resultaten van dit veldexperiment zijn regelmatig gepresenteerd voor een beleidspubliek. Bij zulke presentaties heeft een veldexperiment als bijkomend voordeel dat de opzet transparant is en de resultaten voor een breed publiek begrijpelijk zijn.

Samengevat was een belangrijke keuze bij dit experiment, naast het vinden van een partner voor de samenwerking, de keuze voor het randomisatie-niveau. In dit

[5] Studenten werden wel gevraagd op te schrijven met welke mensen in de groep ze bevriend waren, om te kunnen controleren voor eventuele *spillovers*.

[6] In Ketel et al. (2015) worden ook de effecten op andere uitkomstmaten besproken, zoals het behaalde cijfer, of een student geslaagd is voor het vak en het aantal uur dat aan het vak is besteed.

[7] De vraag was: "Stel je hebt een fles sap gekocht voor €2,-. Zodra je begint te drinken merk je dat je de smaak van het sap niet lekker vindt. Drink je de fles leeg?". Daarna werd dezelfde vraag nog twee maal gesteld maar nu met de bedragen €5,-en €1,-. Een participant is gevoelig voor het *sunk-cost*-effect als hij de fles altijd leeg drinkt, of alleen leeg drinkt voor de hoge prijs en niet voor een lagere prijs.

geval was er geen ideale keuze, zowel randomisatie op groepsniveau als randomisatie binnen de groep hadden voor- en nadelen. Hierbij is het belangrijk van te voren te bepalen welke potentiële vervuilende effecten de grootste invloed zullen hebben op de bevindingen, en op basis hiervan de afweging te maken.

6 Conclusie

Veldexperimenten vormen, mits correct uitgevoerd, een overtuigende en transparante methode om beleid te toetsen. Bij de opzet en uitvoering van zulke experimenten moet echter rekening worden gehouden met een aantal potentiële valkuilen die wij in dit artikel aan de hand van veldexperimenten in de zorg, het onderwijs en de sociale zekerheid bespreken. De belangrijkste lessen die we in dit artikel hebben besproken zijn, allereerst, dat het van belang is om vooraf te bepalen welke communicatie noodzakelijk is en wat de mogelijke gevolgen hiervan zijn voor het gedrag van deelnemers aan het experiment. Ten tweede is besproken welke overwegingen relevant zijn wanneer methodes anders dan volledige randomisatie worden ingezet en wat de gevolgen zijn voor de analyses. Ten derde dient de randomisatie-eenheid zo gekozen te worden dat *spillovers* worden beperkt. Ten slotte dient de externe validiteit van het experiment bij de opzet in de gaten te worden gehouden.

Een voordeel van veldexperimenten is dat de resultaten voor beleidsmakers eenvoudig te interpreteren zijn, wat de vertaling in daadwerkelijk beleid bevordert. Zo heeft het experiment in de langdurige zorg er mede toe geleid dat na beëindiging van het onderzoek grootschalige prestatie-afhankelijke variatie in toetsingspercentages bij *ex-ante* toetsing werd ingevoerd. Bij de Dienst Werk en Inkomen in Amsterdam is het gebruik van de zoekperiode naar aanleiding van de onderzoeksresultaten geïntensiveerd. Daarnaast wordt er momenteel naar gekeken om de zoekperiode op een bredere groep klanten toe te passen.

Auteurs
Nadine Ketel (e-mail: n.ketel@vu.nl) is promovenda bij de Universiteit van Amsterdam en de Vrije Universiteit Amsterdam. Sandra Vriend (e-mail: s.vriend@vu.nl) is promovenda bij de afdeling Algemene Economie bij de Vrije Universiteit Amsterdam. Beiden zijn ook verbonden aan het Tinbergen Instituut.

Literatuur

Behaghel, L., B. Crepon, en M. Gurgand, 2013, Robustness of the encouragement design in a two-treatment randomized control trial, IZA Discussion Papers 7447.

Bolhaar, J., N. Ketel en B. van der Klaauw, 2014, Onderzoek naar effectiviteit inzet re-integratieinstrumenten DWI, VU/UvA, Amsterdam.

Duflo, E., R. Glennerster, en M. Kremer, 2007, Using Randomization in Development Economics Research: A Toolkit in Schultz, T. P. and Strauss, J. A., eds., *Handbook of Development Economics*, vol. 4: 3895-3962, North Holland, Amsterdam.

Ketel, N., J. Linde, H. Oosterbeek en B. van der Klaauw, 2015, Tuition Fees and Sunk-Cost Effects, *The Economic Journal*, forthcoming.

Koning, P., 2011, Experimenten in de sociale zekerheid, *Economische Statistische Berichten*, vol. 96(4605): 150-153.

Kooreman, P. en J. Potters, 2011, De gouden standaard: Veldexperimenten in de voorbereiding en evaluatie van beleid, *TPEdigitaal*, vol. 5(3): 76-90.

Lindeboom, M., B. van der Klaauw en S. Vriend, 2013, Proeftuin onderzoek 'Meten is Weten': Eindrapportage, in opdracht van Centrum Indicatiestelling Zorg.

Lindeboom, M., B. van der Klaauw en S. Vriend, 2015, The effect of audit regimes on applications for long-term care, CEPR Discussion Paper No. 10572.

List, J. A., 2011, Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off, *Journal of Economic Perspectives*, vol. 25(3): 3-16.

Ossel, F., 2008, Effect van reïntegratie meten we juist wel, *NRC Handelsblad*, 27-11-2008:7