De gouden standaard: Veldexperimenten in de voorbereiding en evaluatie van beleid*

Peter Kooreman en Jan Potters

Effecten van interventies kunnen het meest geloofwaardig worden gemeten met behulp van een gerandomiseerd veldexperiment. Denk bijvoorbeeld aan het toedienen van medicijnen, het uitdelen van malarianetten in een ontwikkelingsland, of het veranderen van de default-inleg in een spaarplan. Bij zulke experimenten wordt een willekeurig gekozen deel van de te onderzoeken personen aan de interventie blootgesteld. De overige personen dienen als controlegroep. We bespreken het wetenschappelijke en maatschappelijke belang van veldexperimenten, mede aan de hand van voorbeelden zoals onderzoek naar het effect van de default-inleg in spaarplannen op besparingen en het ontwerpen van online tools die klanten helpen inzicht te krijgen in de afruil tussen risico en rendement. We gaan in op veel gehoorde bezwaren tegen veldexperimenten en bespreken organisatorische, juridische en ethische aspecten.

1 Inleiding

Wat gebeurt er met de studieprestaties van eerstejaars studenten als hen een beloning in het vooruitzicht wordt gesteld? Wat is het effect van micro-kredietfaciliteiten op het inkomen en de levensstandaard van Mexicaanse gezinnen? Wat gebeurt er met het spaargedrag en het consumptiepatroon van Nederlandse werknemers als zij het vakantiegeld voortaan maandelijks ontvangen in plaats van eenmaal per jaar?

Deze vragen lijken eenvoudig, maar ze zijn dat allerminst. Want in feite moeten we telkens twee situaties met elkaar vergelijken, waarvan per definitie zich er slechts één tegelijk kan voordoen. In het voorbeeld van de werknemer: hij krijgt in een bepaald jaar het vakantiegeld maandelijks uitbetaald óf eenmaal per jaar.

De vraag hoe we in complexe situaties een oorzakelijk verband kunnen vaststellen, heeft wetenschappers eeuwenlang beziggehouden. Het boek *The*

* Dit artikel is gebaseerd op een voor Netspar geschreven paper over veldexperimenten (NEA nr. 38). De auteurs bedanken Lans Bovenberg, Gijsbert van Lomwel en vier anonieme referenten voor commentaar en Anton Vedder voor discussie over de juridische aspecten van veldexperimenten.

Design of Experiments van de Engelse geneticus en statisticus Ronald A. Fisher (1890-1962), dat vijfenzeventig jaar geleden verscheen, kan worden gezien als de voltooiing van die wetenschappelijk zoektocht. De kern van Fishers methode – de 'gouden standaard' voor het vaststellen van causale verbanden – is in beginsel eenvoudig: stel een willekeurig gekozen deel van de te onderzoeken eenheden bloot aan een interventie (*treatment*), en gebruik de overige eenheden als controlegroep. Fisher liet zien dat we op die manier kunnen bepalen of waargenomen verschillen tussen de *treatment* en de controlegroep na de interventie louter het gevolg van toeval óf systematisch van aard zijn.

Economen en beleidsmakers hebben lange tijd het werk van Fisher als niet-relevant beschouwd, of hooguit als een onbereikbaar ideaal. In de economie kun je immers geen experimenten doen, was lange tijd de overtuiging. In een lange traditie van empirisch onderzoek is daarom vaak geprobeerd om op basis van niet-experimentele gegevens oorzakelijke verbanden vast te stellen. Het belangrijkste gereedschap daarbij was het regressiemodel of een vergelijkbare econometrische methode. Maar zoals de theorie achter het regressie-model aangeeft, kunnen oorzakelijke verbanden daarmee alleen worden vastgesteld wanneer aan strenge voorwaarden is voldaan. De belangrijkste voorwaarde komt erop neer dat de variabele waarvan men het effect wil meten (x) onafhankelijk (ongecorreleerd) dient te zijn van alle ongeobserveerde variabelen (ε) die ook van invloed zijn op de te verklaren variabele (y). Is daar niet aan voldaan, dan wordt het effect van ε ten onrechte aan x toe geschreven.

Stel bijvoorbeeld dat y weergeeft of een kostwinner een lijfrenteverzekering heeft afgesloten (zo ja: y=1; zo nee: y=0). De variabele x geeft aan of de kostwinner zijn vakantiegeld in mei uitbetaald krijgt (x=1) of gekozen heeft voor maandelijkse uitbetaling (x=0). ε geeft de (ongeobserveerde) mate van ongeduld weer van de kostwinner. Hoe hoger ε , hoe kleiner de kans dat y=1. Ook is het waarschijnlijk dat een lage waarde van ε samengaat met x=1; x en ε zijn dus negatief gecorreleerd. Het regressiemodel overschat dan het effect van x op y, omdat x niet alleen het effect van de jaarlijkse in plaats van maandelijkse uitbetaling weergeeft, maar ook de lage (niet-geobserveerde) waarde van ε. Om het pure effect van de betalingsfrequentie (x) op het afsluiten van de verzekering (y) te kunnen meten, moet de variatie in x dus min of meer willekeurig zijn en niet afhangen van ε. Een veldexperiment maakt dat mogelijk.

Lalonde (1986) liet zien dat niet-experimentele methoden een vertekend beeld opleverden van het effect van activerend arbeidsmarktbeleid, door ze te vergelijken met de resultaten van een veldexperiment.

2 De methodologische kracht van veldexperimenten

Stel: een werkgever die het vakantiegeld nu eenmaal per jaar uitbetaalt, biedt zijn werknemers de optie over te stappen op maandelijkse uitbetaling. Er zijn dan dus twee groepen werknemers: zij die willen overstappen op maandelijkse betaling (groep M) en de andere werknemers, die het vakantiegeld jaarlijks blijven ontvangen (groep J). Na enige tijd kan dan gekeken worden of er een verschil is in bijvoorbeeld het gemiddelde spaarsaldo in de twee groepen (M_1-J_1). Maar omdat werknemers zelf konden kiezen, geeft dit verschil niet het causale effect weer van de frequentie-verandering. Het is bijvoorbeeld denkbaar dat ongeduldige werknemers eerder kiezen voor maandelijkse uitbetaling. Het verschil (M_1-J_1) geeft dan zowel het effect van de frequentie-verandering weer als het effect van een verschil in de mate van ongeduld. Het is lastig daarvoor te corrigeren, onder meer omdat geduld moeilijk meetbaar is.[1]

De sleutel voor de oplossing van dit probleem is een veldexperiment met *randomisatie*. In ons voorbeeld betekent randomisatie dat een willekeurig gekozen deel van de werknemers het vakantiegeld voortaan maandelijks uitbetaald krijgt. Met willekeurig wordt bedoeld: door het lot bepaald. Randomisatie garandeert dat er vooraf geen enkel systematisch verschil bestaat tussen de groep met maandelijkse en jaarlijkse uitkering van het vakantiegeld. Bijgevolg weten we zeker dat een eventueel verschil in spaarsaldi tussen groep M en groep J alleen door het verschil in betalingsfrequentie van het vakantiegeld veroorzaakt kan zijn.[2]

De methode van de randomisatie is wetenschappelijk onbetwist door zijn eenvoud en transparantie. Wel kunnen aan randomisatie ethische, juridische of praktische bezwaren kleven. In Sectie 4 gaan we daar op in. Naast randomisatie is een tweede kenmerk van veldexperimenten dat deelnemers beslissingen nemen in hun natuurlijke omgeving. Anders dan bij onderzoek op basis van niet-experimentele gegevens, theoretische analyses, of laboratoriumexperimenten, hebben de resultaten van veldexperimenten dan ook vaak een grote overtuigingskracht.

Een onderscheid kan nog gemaakt worden tussen veldexperimenten waarbij de deelnemers niet weten dat ze aan een experiment meedoen en die waarbij ze dat wel weten. In het eerste geval spreken we wel van een *natuurlijk* veldexperiment,

[1] Een stap in de goede richting zou zijn eerst het verschil in gemiddelde spaarsaldi voor de frequentie-verandering in beeld te brengen, en vervolgens te kijken of het verschil tussen de verschillen (double difference) veranderd is: (M_1-J_1) - (M_0-J_0). Toch meet ook dit dubbele verschil zonder aanvullende voorwaarden niet het pure oorzakelijke effect van de frequentie-verandering.

[2] Bij toewijzing die volledig random is kan het voorkomen dat er toevalligerwijs bijvoorbeeld veel meer vrouwen in treatment M zitten dan in treatment J. Om dat te voorkomen kan de onderzoeker ervoor zorgen dat van de totale sample precies de helft van de vrouwen (willekeurig gekozen) treatment M krijgt en de andere helft treatment J. Zo kunnen gebalanceerde (gestratificeerde) subsamples worden opgezet om de vergelijkbaarheid te vergroten en de statistische ruis te verminderen. Een formelere bespreking van de methodologie van veldexperimenten is te vinden in bijvoorbeeld Duflo et al. (2007).

omdat verwacht kan worden dat deelnemers zich volstrekt natuurlijk zullen gedragen.[3] In het tweede geval spreken we van een *gekaderd (framed)* veldexperiment; deelnemers weten dat ze 'in beeld zijn'. Als deelnemers weten dat hun gedrag geobserveerd wordt, kan dit hun gedrag beïnvloeden. Zo vindt List (2006) bijvoorbeeld dat handelaren zich in een experimentele markt socialer gedragen dan in hun natuurlijke omgeving. Omdat dit effect (dat wel het 'Hawthorne-effect' wordt genoemd) zich voordoet voor zowel de *treatment*groep als de controlegroep, hoeft dat een zuivere vergelijking tussen de twee niet in de weg staan.[4] Net zo min als een placebo-effect het toetsen van medicijnen zinloos maakt. Een probleem kan wel zijn dat de waargenomen effecten niet zonder meer generaliseerbaar zijn. Dit mogelijke probleem kan versterkt worden als de deelnemers zich op vrijwillige basis opgeven om aan een experiment mee te doen. Mensen die kiezen om mee te doen aan een experiment zijn niet noodzakelijkerwijs representatief voor de gehele populatie (Gautier en Van der Klaauw 2011). [5] Om deze redenen is het *natuurlijke* veldexperiment het ideaaltype. Toch zijn er ook veel zinvolle en valide veldexperimenten waarbij deelnemers weten dat ze deelnemen en dit op vrijwillige basis doen. Het vergt dan vaak wel meer inspanning en vernuft om te zorgen dat de *treatment*- en controlegroep onderling vergelijkbaar zijn en ook zoveel mogelijk representatief voor de doelpopulatie.

3 Het maatschappelijke belang van veldexperimenten

Fisher hield zich vooral bezig met landbouwkundige en biologische toepassingen, maar gerandomiseerde veldexperimenten worden nu toegepast in tal van wetenschapsgebieden.

Klinische experimenten in de geneeskunde, bijvoorbeeld om de werking van nieuwe geneesmiddelen te onderzoeken, zijn een belangrijk toepassingsgebied van Fischers methode. Daar geldt dan nog een extra eis: zowel de patiënt als degene die direct het medicijn toedient, weet of het nieuwe medicijn dan wel een placebo wordt toegediend ('dubbelblind'). Dit om te voorkomen dat die wetenschap de uitkomst beïnvloedt. Dit onderzoekparadigma heeft geleid tot de opkomst van de zogeheten *Evidence Based Medicine* (EBM). Binnen EBM worden therapieën pas als werkzaam beschouwd als dat (herhaaldelijk) is gebleken uit dubbelblinde, gerandomiseerde experimenten.[6]

[3] Het verschil tussen een natuurlijk veldexperiment en een natuurlijk experiment is dat in het laatste geval de randomisatie niet door de onderzoeker maar door toevallige omstandigheden tot stand komt.

[4] Er zijn echter veldexperimenten waarvoor het cruciaal is dat deelnemers niet weten dat ze aan een experiment meedoen of op z'n minst niet weten wat de onderliggende vraagstelling is, bijvoorbeeld veldexperimenten over discriminatie op basis van geslacht of etniciteit.

[5] Er zijn ook studies die suggereren dat zelfselectie niet altijd problematisch is (Benz en Meier 2008; Cleave et al. 2011; Haan en Kooreman 2002).

[6] Overigens wordt een strikte toepassing van EBM in de medische praktijk van alledag als te rigide beschouwd, omdat resultaten uit de klinische experimenten die voor populaties gelden niet op

Veldexperimenten zijn inmiddels ook toegepast op een groot aantal deelgebieden binnen de economische wetenschap. Voorbeelden zijn ontwikkelingseconomie (zie bijvoorbeeld Duflo 2008), onderwijseconomie (Leuven et al. 2007; Leuven et al. 2010), sociale interacties (Soetevent 2004; Soetevent 2005; Duflo en Saez 2006) en spaar- en leengedrag (Ashraf et al. 2006; Ausubel 1999; Bertrand et al. 2010). De snel toenemende betekenis van veldexperimenten in empirisch onderzoek blijkt ook uit het groeiende aantal overzichtsartikelen, zoals Harrison en List (2004), Duflo (2008), Levitt en List (2009), Angrist en Pischke (2010) en Charness en Kuhn (2010). Hier beperken we ons tot een aantal voorbeelden.

Een zeer populair onderzoeksterrein voor veldexperimenten is liefdadigheid. Welke factoren bepalen of, en hoeveel mensen doneren aan een goed doel? Een veldexperiment bestaat er dan vaak uit dat aan verschillende mensen verschillende 'bedelbrieven' worden gestuurd. De inhoud van de brief en de condities voor een donatie worden dan gerandomiseerd. Zo wordt bijvoorbeeld aan één groep een klein presentje meegestuurd (zoals een setje ansichtkaarten) en aan de andere groep niet. De vraag is dan of mensen uit wederkerigheid geneigd zullen zijn om het presentje met een donatie te belonen en of dat de kosten van het presentje meer dan compenseert. Het antwoord hierop is een overtuigend 'ja' (Falk 2007). Een andere vraag: hoe kan een bijdrage van een grote donor (denk aan Bill Gates) het best worden aangewend? Is het bijvoorbeeld beter om te 'matchen' (voor elke euro van u past de grote donor een euro bij tot een maximum van X) of het bedrag in één keer op tafel te leggen (een grote donor heeft al een bedrag X gestort)? Theoretisch zou het eerste veel effectiever moeten zijn, maar veldexperimenten laten zien dat matching veel minder effectief is dan, door economen, vaak wordt gedacht (Karlan en List 2007).

Veldexperimenten worden ook veelvuldig gebruikt om te onderzoeken hoe mensen reageren op productaanbiedingen die ze via de post ontvangen. Hoe gevoelig zijn mensen voor de prijs? Kijken ze dan naar het gehele kostenplaatje of vooral naar de kosten op de korte termijn? In hoeverre spelen ook andere factoren een rol, zoals het aantal aanbiedingen en de aanwezigheid van een presentje? Zo blijkt uit een veldexeperiment uitgevoerd in Zuid-Afrika dat het – naast de aangeboden rente – belangrijk is wie er op de foto staat bij de aanbieding. Als dit een vrouw is, zijn mannen gemiddeld bereid om een veel hogere rente te accepteren. De huidskleur van de persoon op de foto doet er echter niet toe.

Veldexperimenten hebben soms tot nieuwe inzichten geleid die haaks staan op de *conventional wisdom.* Een pijler in de standaard economische theorie is dat individuen harder werken naarmate de financiële beloning hoger is. In hun studie onder eerstejaars economiestudenten vonden Leuven et al. (2010) echter dat het

(heterogene) individuele patiënten van toepassing hoeven te zijn; zie bijvoorbeeld Smulders et al. 2010. Dit ontkracht niet het belang van experimenten, maar benadrukt dat een interventie voor verschillende deelpopulaties (bijvoorbeeld laagopgeleide mannen, hoogopgeleide vrouwen, etc.) een verschillend effect kan hebben. Om dat te kunnen vaststellen moet het aantal observaties per deelpopulatie in het onderzoek uiteraard voldoende groot zijn.

beloven van een forse financiële bonus gemiddeld niet leidt tot betere studieprestaties. Het effect van de financiële prikkel blijkt genuanceerder te liggen. Studenten met veel talent (afgemeten aan middelbareschoolcijfers voor wiskunde) bleken wel beter te presteren wanneer hen een bonus in het vooruitzicht werd gesteld. Maar studenten met minder talent (relatief lage middelbareschoolcijfers voor wiskunde) bleken met beloofde bonus juist slechter te presteren dan de controlegroep met dezelfde middelbareschoolcijfers zonder beloofde bonus.

Zonder veldexperiment zou het vrijwel onmogelijk zijn geweest om deze verrassende effecten te meten. In de eerste plaats komt het in een niet-experimentele context weinig voor dat aan studenten een bonus wordt beloofd. In de tweede plaats zal in een niet-experimentele context het beloven van een bonus samenhangen met ongeobserveerde karakteristieken van de student (bijvoorbeeld een gebrek aan motivatie) die ook direct van invloed zijn op studieprestaties. Daardoor kan binnen een niet-experimentele context het causale effect van een bonus op studieprestaties niet zuiver worden gemeten.

Een andere studie met een verrassende uitkomst: het veldexperiment dat Gneezy en Rustichini (2000) uitvoerden bij een aantal kinderdagverblijven. Ouders kwamen vaak pas na de officiële sluitingstijd hun kind ophalen, dit tot ergernis van de staf. Als remedie voerde een aantal kinderdagverblijven een (geringe) financiële boete in voor te laat komen. Het gevolg was echter dat nog meer ouders te laat kwamen. Bovendien keerde het gemiddelde 'ophaaltijdstip' niet meer terug op het oude niveau toen de boete weer was afgeschaft. Een mogelijke verklaring voor deze uitkomsten is dat de boete de sociale norm verdrong. Het slechte geweten kon immers worden afgekocht tegen een relatief lage prijs.

Andere veldexperimenten, of sterk verwante methodes, met verrassende resultaten zijn: Leuven et al. (2007), computers in het basisonderwijs hebben een negatief effect op rekenprestaties van leerlingen; Kastoryano en Van der Klaauw (2011), re-integratietrajecten voor werkloze onderwijzers werken averechts'.[7] Mogelijk zijn deze onverwachte bevindingen een verklaring voor de weerstand die soms tegen veldexperimenten bestaat (zie Sectie 4).

Aan complexe situaties is in de economie geen gebrek. Bovendien is de wens om te interveniëren groot. Toch kent Nederland geen traditie om beleid op basis van gecontroleerde experimenten voor te bereiden en te evalueren. Overigens gaat het hierbij om experimenten in de economische werkelijkheid van alledag; niet om die in de kunstmatige omgeving van het laboratorium. Veldexperimenten kunnen worden gezien als *pilots* met een wetenschappelijk zorgvuldig onderbouwd

[7] Een methode die verwant is aan een veldexperiment maakt gebruik van een zogenaamde *beleidsdiscontinuïteit* (zie ook van der Klaauw 2010). Werkloze leraren in het primair onderwijs komen in aanmerking voor een gesubsidieerd re-integratietraject, maar alleen als ze 50 jaar of ouder zijn. Het idee is dat 49-jarige werkloze leraren, die (nog) niet aanmerking komen voor het re-integratietraject, niet veel verschillen van hun 50-jarige collega's in termen van kansen op de arbeidsmarkt. De vergelijking tussen 49- en 50-jarige werkzoekenden laat vervolgens zien dat de re-integratietrajecten, die ongeveer 4000 euro per deelnemer kostten, een *averechts* effect hadden op werkhervatting. Een verklaring is dat werklozen minder gebruik maakten van hun eigen netwerk om een nieuwe baan te vinden.

onderzoeksdesign. Het kiezen van de juiste controlegroep staat daarbij centraal. Met een zorgvuldig onderzoeksdesign kan uit een beleidswijziging optimaal lering worden getrokken.

Veldexperimenten kunnen niet alle onderzoeksvragen beantwoorden. Bij de Betuwelijn en de introductie van de euro zijn ze moeilijk uitvoerbaar. Maar in veel andere gevallen is de gouden standaard van het gecontroleerde experiment de aangewezen onderzoeksmethode. Zoals ook de besproken voorbeelden illustreren, kan daarbij worden gedacht aan een groot aantal verschillende onderwerpen: het meten van de effecten van financiële prikkels (arbeidsmarkt, sociale zekerheid, zorg, mobiliteit, onderwijs, R&D, milieu), de effecten van onderwijsinterventies (klassenverkleining, computers), of de rol van informatie in de marketing van pensioenproducten.

Veldexperimenten leveren geloofwaardig bewijs over de werkzaamheid van beleid. Daarin ligt het maatschappelijk belang. Wanneer beleid niet grondig wordt geëvalueerd – en grondig impliceert meestal een veldexperiment – is er een gerede kans dat beleid ineffectief is of, erger, averechts werkt.

4 Juridische, ethische en praktische aspecten van veldexperimenten

Op juridische en ethische gronden kan het wenselijk zijn individuen toestemming te vragen voor deelname aan een veldexperiment. Of hen ten minste te informeren over het feit dat ze deelnemen aan een wetenschappelijk experiment. Maar vanuit methodologisch perspectief is het meestal wenselijk dat individuen daarover niet worden geïnformeerd. Die kennis zou sociaal wenselijk gedrag kunnen uitlokken en zo het werkelijke effect van de interventie vertroebelen.

In een opiniebijdrage in *NRC Handelsblad* van 24 november 2008 stelde de Amsterdamse hoogleraar Mirjam van Praag dat ambtenaren en politici heel wat beleid ontwikkelen en uitvoeren, zonder zich af te vragen of het wel werkt. Nadat de Nederlandse overheid twee miljard euro had uitgegeven aan re-integratietrajecten voor bijstandsgerechtigden, werd een voorstel van Van Praag om de effecten te meten met behulp van een veldexperiment afgewezen. In een reactie (*NRC Handelsblad* 27 november 2008) licht de Amsterdamse wethouder Frank Ossel van 'Werk en Inkomen' deze weigering toe:

> *"Het was de gemeente Amsterdam die Van Praag in maart 2008 benaderde om dit project te evalueren, juist omdat Amsterdam beseft hoe belangrijk de meting is van de effectiviteit van re-integratietrajecten. Van Praag stelde toen voor om twintig uitkeringsgerechtigden te selecteren voor dit project, maar in werkelijkheid niets te doen aan hun re-integratie, bij wijze van controlegroep. [...] Klanten in de controlegroep mocht volgens deze opzet niet worden verteld dat zij slechts als proefkonijn dienden. Het idee om bijstandsgerechtigden met deze vaardigheden voor spek en bonen 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."

Wanneer experimenten klinisch van aard zijn (dat wil zeggen: in directe relatie staan tot de gezondheid van de deelnemers) is wettelijk toestemming vereist van de Centrale Commissie Mensgebonden Onderzoek (CCMO). Op de meeste veldexperimenten is dit niet van toepassing. Wanneer op voorhand niet bekend is wat het effect van een bepaalde behandeling is, zijn er geen strikte juridische belemmeringen om gelijksoortige personen verschillend te behandelen. Zoals het citaat van de Amsterdamse wethouder illustreert, kunnen er wel (vermeende) praktische of ethische bezwaren bestaan.

Maar, ethische kanttekeningen kunnen ook worden gemaakt wanneer beleid wordt veranderd zonder goed te weten hoe dit uitwerkt. In feite dient dan de gehele bevolking als proefkonijn.

Een optie is om een gedragscode te ontwikkelen (bijvoorbeeld via de VSNU[8]) met richtlijnen voor het uitvoeren van veldexperimenten. Daarin moet een balans gevonden worden tussen juridische en ethische voorwaarden enerzijds, en wetenschappelijk belang en uitvoerbaarheid anderzijds. De internationale code voor markt- en sociaalwetenschappelijk onderzoek van ICC/ESOMAR[9] zou daarbij als vertrekpunt kunnen dienen. Elementen daarin zijn onder meer zorgvuldigheid en transparantie in opzet, uitvoering en rapportage en het niet benadelen of schaden van privépersonen, met als doel het bevorderen van het publieke vertrouwen in dit type onderzoek.

Een veldexperiment betekent inderdaad dat beleid willekeurig wordt toegewezen aan gelijksoortige individuen. In geval van bijvoorbeeld een spaarplan betekent het niet per se dat sommigen helemaal van deelname worden uitgesloten. Variatie tussen individuen in het moment waarop de interventie plaatsvindt, is soms voldoende om de effecten van een interventie te kunnen meten.

Naast ethische bezwaren zijn er ook praktische hindernissen. De mogelijkheid dat deelnemers uitvallen gedurende het experiment kan een probleem zijn. Zeker als het gaat om een experiment dat een langere periode duurt. Uitval (*attrition*) leidt ertoe dat de verdeling van deelnemers over *treatment*- en controlegroep niet langer willekeurig (*random*) hoeft te zijn. Het feit dat iemand niet meer meedoet of wil doen, kan namelijk samenhangen met zijn of haar indeling in een van de twee groepen. Dit kan leiden tot zowel een onder- als overschatting van effecten. Dus is het belangrijk om een experiment zo op te zetten dat deelnemers van begin tot eind kunnen worden gevolgd. Dat onderstreept ook het belang van het gebruik van administratieve data, zoals inkomensgegevens via de Belastingdienst, claimgedrag via (zorg)verzekeringsmaatschappijen of leerprestaties via de Onderwijsinspectie. Individuen kunnen dan lange tijd worden gevolgd. Daardoor is het beter mogelijk de effecten van uitval te analyseren en het verschil tussen korte- en

[8] Vereniging van Samenwerkende Nederlandse Universiteiten.

[9] ICC: International Chamber of Commerce; ESOMAR: European Society for Opinion and Market Research.

langetermijneffecten van interventies in kaart te brengen. In diverse onderzoeken worden grote effecten op de korte termijn gevonden, die echter verdwijnen op de langere termijn(bijvoorbeeld Leuven et al.2011).

Een ander mogelijk probleem is dat de *treatment*groep en de controlegroep elkaar beïnvloeden (*spill-over* effecten). Stel bijvoorbeeld dat de groep studenten die een beloning krijgt voor het behalen van veel studiepunten de groep studenten zonder beloning 'meetrekt'; het effect van belonen zal worden onderschat, omdat het verschil tussen de twee groepen kleiner is door die beïnvloeding. Dit probleem kan soms worden vermeden door beide groepen letterlijk ver genoeg uit elkaar te zetten.

5 Voorbeelden van mogelijke veldexperimenten[10]

Voorbeeld 1. De frequentie van betalingen en inningen. Een factor die mogelijk grote invloed heeft op spaargedrag, consumptiepatroon en problematische schuldsituaties, is de frequentie waarmee salaris wordt uitbetaald. In Nederland ontvangen vrijwel alle werknemers hun salaris maandelijks. Tot in de jaren zestig was het gebruikelijk het salaris wekelijks (en contant in plaats van giraal) uit te betalen. Wat doet de maandelijkse betalingsfrequentie met het gedrag van de consument? Wordt het makkelijker of moeilijker om de eigen financiën te beheren, treden er vaker of minder vaak problematische schuldsituaties op? Wordt er meer of minder gespaard? Worden grote uitgaven langer uitgesteld? Vergelijkbare vragen kunnen gesteld worden over de frequentie waarmee AOW of bijstand worden uitgekeerd, en de frequentie waarmee (spaar)premies geïnd worden.

Zonder veldexperiment is het vrijwel onmogelijk deze vragen te beantwoorden, omdat variatie in de frequentie van salarisbetaling ontbreekt. Voor zover werknemers hun salaris niet maandelijks uitbetaald krijgen, betreft dit atypische groepen, zoals vakantiekrachten of Amerikaanse militairen (die hun salaris tweewekelijks uitbetaald krijgen). Alleen wanneer *binnen* een relatief homogene groep werknemers de betalingsfrequentie niet-systematisch varieert, kan een oorzakelijk effect worden vastgesteld.

We schetsen in grote lijnen hoe een veldexperiment om het effect van betalingsfrequentie te meten, er uit zou zien:

1. Het uitwerken van de onderzoeksvraag en het identificeren van de variabelen waarover gegevens nodig zijn. In dit voorbeeld omvat dat minimaal a) de frequentie van salarisbetaling; b) gegevens over inleg en onttrekken van (spaar)rekeningen van de werknemer;
2. Een *nulmeting* onder de werknemers met informatie over inkomen, spaargedrag en achtergrondkarakteristieken (leeftijd, geslacht, burgerlijke staat, opleiding en functie);

[10] Koning (2011) beschijft een aantal veldexperimenten in de sociale zekerheid.

3. Een werkgever kan het ongewenst vinden om zonder toestemming van de werknemers de frequentie van de salarisbetaling te wijzigen, daarom wordt de volgende procedure gevolgd: een procedure die kan worden gezien als een compromis tussen praktische uitvoerbaarheid en wetenschappelijk eisen: De werkgever biedt alle werknemers de *optie* de frequentie van de salarisbetaling te wijzigen in wekelijks. Daarbij wordt vermeld dat het praktisch niet mogelijk is om de frequentie direct voor iedereen te wijzigen die dat wenst. Noem de mensen die een wekelijkse frequentie wensen groep A. Bij een willekeurige helft van hen wordt de frequentie direct gewijzigd (groep A+). Bij de andere helft van groep A (groep A-) verandert de frequentie (vooralsnog) niet. Het eventuele verschil in bijvoorbeeld spaargedrag tussen de groepen A- en A+ na de frequentieverandering geeft dan het causale effect weer van de frequentieverandering *voor mensen die die frequentieverandering wensen*. Doordat zowel A- als A+ werknemers bevat die graag een wekelijkse frequentie willen, wordt gewaarborgd dat de twee groepen (afgezien van feitelijke betalingsfrequentie) volkomen vergelijkbaar zijn. Dat zou niet het geval zijn wanneer alle werknemers direct volgens hun voorkeur bediend zouden worden (uiteraard kunnen er nog meer frequenties aangeboden worden, zoals een tweemaandelijkse salarisbetaling.) Werknemers wordt meegedeeld dat een eenmaal gekozen frequentie pas na een bepaalde periode weer veranderd kan worden (bijvoorbeeld na 3 maanden, 6 maanden of een jaar; uit onderzoeksoogpunt zo lang mogelijk). De wetenschappelijk prijs die wordt betaald voor het bieden van de optie is dat voor werknemers die hun salaris maandelijks willen blijven ontvangen, *niet* kan worden vastgesteld wat het effect is van een frequentieverandering;
4. Een longitudinaal databestand wordt opgebouwd met per werknemer informatie over spaargedrag;
5. Analyse van de het databestand en publicatie over de resultaten.

Voorbeeld 2. Analyse van online tools. Een punt van zorg is dat veel mensen slechts zeer beperkte financiële kennis hebben (Lusardi en Van Rooij 2010; Van Rooij, Kool en Prast 2007). Een van de suggesties om deze situatie te verbeteren is gebruikmaken van interactieve hulpmiddelen die kunnen worden aangeboden via het internet (Dellaert 2010). Met behulp van zulke *online tools* kunnen mensen zich een beter beeld vormen van hun financiële voorkeuren. Ook kan het ze een beter inzicht geven in de randvoorwaarden en de afruilmogelijkheden op dat gebied (Goldstein et al. 2008). Een andere toepassing is het communiceren van risico. Hoe kun je over risico's communiceren zonder dat mensen deze onnodig zwaar wegen en zonder dat ze hun capaciteit om risico's te absorberen, onderschatten.

Er is echter weinig tot niets bekend over de effectiviteit van *online tools* op het gebied van pensioenen (Dellaert 2010).[11] Bovendien is de vraag wat een goede

[11] Recentelijk is bij PGGM een eerste succesvolle proef op dit gebied uitgevoerd (zie Verbaal, 2011).

vormgeving van dergelijke *tools* is. Hoeveel tijd mag het gebruik ervan kosten? Hoe complex mag de *tool* zijn?

Om dergelijke vragen te beantwoorden, zijn veldexperimenten uitermate geschikt. Ter illustratie stellen we de vraag hoe een *online tool* mensen een beter inzicht kan geven in de uitruil tussen risico en rendement en de prijs van zekerheid. Risico is een moeilijk begrip voor veel mensen en wordt vaak uitsluitend begrepen in negatieve zin (verlies), zonder oog voor de positieve kant (opwaarts potentieel) en zonder begrip van de risicopremie. In een *online tool* kan men deze premie op (minstens) twee manieren in beeld brengen: via het heden (hoeveel bespaar ik op mijn inleg als ik minder zekerheid eis) of via de toekomst (hoeveel levert het me in de toekomst naar verwachting op als ik minder zekerheid eis). In beginsel zijn deze twee manieren equivalent. Er moet alleen op een correcte manier worden verdisconteerd. Er zijn aanwijzingen dat veel mensen hier moeite mee hebben (Binswanger en Carman 2010). De vraag is dan of de presentatie verschil maakt en zo ja, welke het meest effectief is? Om deze vraag te beantwoorden, kan men als volgt te werk gaan:

1. Deelnemers aan een pensioenfonds of bezoekers van een website als pensioenkijker.nl worden uitgenodigd de *online tool* te doorlopen;
2. Eerst wordt een nulmeting gedaan over financiële kennis ("Een pensioen met meer zekerheid geeft gemiddeld een hogere of lagere uitkering?") en risicovoorkeuren ("Vindt u dat uw pensioenfonds meer of minder risicovol zou moeten beleggen?"). Ook worden verschillende achtergrondvariabelen gemeten (geslacht, inkomen, opleiding);
3. Deelnemers worden willekeurig toegewezen aan een van de varianten van de online tool: risicopremie vertaald naar het heden of naar de toekomst;
4. Na afloop wordt een nieuwe (identieke) meting gedaan van financiële kennis en risicovoorkeuren;
5. Een vergelijking van de metingen voor en na het gebruik van de online tool geeft aan in hoeverre kennis en voorkeuren zijn veranderd. Ook kan worden gekeken welke variant van de tool het meest effectief is. Daarnaast kan worden geanalyseerd in hoeverre dit afhangt van achtergrondkenmerken van de deelnemers.

Voorbeeld 3. Vrije keuze met defaults. Twee begrippen die vaak opduiken in de pensioendiscussie zijn keuzevrijheid en *defaults*. Er is een toenemende roep om mensen meer keuzevrijheid te geven bij het invullen van hun pensioencontract: zie bijvoorbeeld Nijman en Oerlemans (2008), Bodie en Prast (2010), Kooreman en Prast (2010). Individualisering vraagt om maatwerk, omdat mensen steeds meer van elkaar verschillen in hun omstandigheden en voorkeuren. Bovendien is het risico van pensioenen in de tweede pijler steeds meer verschoven van werkgever naar werknemer. Deze ontwikkelingen vragen om differentiatie en keuzevrijheid. De vraag is echter of mensen zulke keuzes wel willen maken (keuzestress) en kunnen maken (begrensde rationaliteit). Zogeheten *defaults* zouden aan beide problemen enigszins tegemoet kunnen komen. *Defaults* zijn keuzes die voor

iemand gelden, tenzij hij of zij ze zelf verandert. Naarmate deze *defaults* een sterkere aantrekkingskracht hebben, wordt het belangrijker om ook deze *defaults* toe te spitsen op de persoonlijke situatie.

Het introduceren van meer keuzevrijheid en het opzetten van geschikte *defaults* is een operatie met ingrijpende maatschappelijke en persoonlijke gevolgen. Het lijkt dan ook onverstandig dit te doen zonder beter inzicht in bijvoorbeeld de werkingskracht van *defaults* en de geneigdheid van mensen om ervan af te wijken. Afwijken kan goed zijn (als mensen een betere optie vinden), maar houdt ook het risico in dat slechte keuzes worden gemaakt.

Een veldexperiment is een uitstekende manier om te onderzoeken hoe sterk de werking van *defaults* is en hoe groot het risico is dat mensen uiteindelijk toch 'verkeerde' keuzes maken.[12]

Men zou er voor kunnen kiezen om de experimenten in eerste instantie te beperken tot keuze-opties die deelnemers nu soms ook al hebben, zoals de pensioneringsbeslissing, de uitruil van het partnerpensioen, en de keuze voor een hoog-laagconstructie in de uitkeringsfase. Vervolgens zou men de werking van *defaults* kunnen onderzoeken voor eventueel nieuw te introduceren opties, zoals de hoogte van de premie, het kiezen voor garanties (reëel dan wel nominaal), en de verdere invulling van de uitkeringsfase (annuïteit).

Zo'n experiment zou in grote lijnen kunnen bestaan uit de volgende stappen:

1. Bepaal voor welke parameters van een bestaand of nieuw pensioencontract – zoals pensionering, premiehoogte, risicoprofiel of invulling van de decumulatie – een keuze geïntroduceerd moet worden;
2. Een groep deelnemers van een pensioenfonds krijgt te horen dat ze vanaf heden de vrijheid hebben om, binnen bepaalde grenzen, het betreffende onderdeel van hun pensioencontract zelf in te vullen;
3. Eén groep, willekeurig gekozen, krijgt de huidige situatie als *default*; de andere groep krijgt een persoonlijke *default* waarbij de keuzeopties zo goed mogelijk zijn aangepast aan de persoonlijke omstandigheden, zoals leeftijd, pensioenopbouw carrièrepad, levensverwachting en risicovoorkeur (Bovenberg et al. 2007; Kortleve en Slager 2010; Nijman en Oerlemans 2008). Mensen krijgen tot een bepaalde datum de vrijheid om van de default af te wijken en een eigen keuze te maken;
4. Overwogen kan worden om (een deel van de) deelnemers een soort PGB[13] voor persoonlijk financieel advies te geven. Op deze manier kan onderzocht worden of, en in welke richting, zulke adviezen de uiteindelijke keuzes beïnvloeden;
5. Als alle deelnemers hun keuzes hebben gemaakt kan onder meer wordt geanalyseerd (a) hoeveel mensen kiezen voor de *default*, (b) in hoeverre dit afhangt van de geïmplementeerde *default* (huidige situatie versus 'persoonlijke' *default*), (c) in welke richting mensen afwijken van de *default*

[12] Uiteraard is niet altijd duidelijk hoe'goed' of 'slecht' een keuze is. Dit hangt mede af van voorkeuren en verwachtingen, welke moeilijk waarneembaar zijn.

[13] Persoonsgebonden Budget.

als ze dat doen, en (d) in hoeverre dit afhankelijk is van persoonlijke kenmerken van de deelnemer (zoals leeftijd, inkomen en geslacht).

De resultaten van een dergelijk experiment geven aan hoe belangrijk *defaults* in de betreffende context zijn en hoeveel zorg besteed moet worden aan het opstellen ervan. Ook geven de resultaten een indruk van het risico dat deelnemers – ondanks de aanwezigheid van *defaults* – slechte keuzes zullen maken (dat wil zeggen keuzes waarvan moeilijk kan worden aangenomen dat ze in het belang van de deelnemer zijn) en hoe dit risico varieert met achtergrondkenmerken zoals leeftijd en opleidingsniveau (Agarwal et al. 2009).

6 Besluit

Een kleine tien jaar geleden adviseerde een internationale visitatiecommissie die het CPB doorlichtte "*to make far greater use of [...] experimentation to pursue evaluation of policies*".

Deze aanbeveling sloot aan bij een ontwikkeling in de academische economische tijdschriften, waarin het belang van experimenten om geloofwaardig conclusies over causaliteit te kunnen trekken steeds meer werd benadrukt.

Sindsdien dringt het belang van veldexperimenten langzaam door in de Nederlandse beleidswereld. Inmiddels zijn er in Nederland en daarbuiten veelbelovende veldexperimenten uitgevoerd, bijvoorbeeld rond de relatie tussen uitkeringsduur en het gedrag van werkzoekenden, over de invloed van financiële prikkels op studieresultaten, over de rol van anonimiteit bij collectes, en over de rol van informatie bij het aanbieden van kredieten. Het ministerie van Sociale Zaken en Werkgelegenheid heeft onlangs een groot experiment naar de effecten van re-intregatiebeleid gestart en het ministerie van Onderwijs, Cultuur en Wetenschap heeft via *OnderwijsBewijs* een groot aantal kleinschalige experimenten naar de effecten van interventies in het onderwijs in gang gezet.

Het belang van deze studies wordt nog versterkt door het feit dat breedgedragen beleid in een aantal gevallen een averechts effect blijkt te hebben.

Een prominentere rol voor veldexperimenten in de beleidsvoorbereiding en –evaluatie is een vanzelfsprekende stap voor tal van maatschappelijke organisaties. Daarbij draagt nauwe samenwerking tussen die organisaties en universitaire onderzoekers bij aan praktische uitvoerbaarheid, onderzoekskwaliteit en relevantie van vraagstelling.

Auteurs
Peter Kooreman (e-mail: p.kooreman@uvt.nl en Jan Potters (e-mail:j.j.m.potters@uvt.nl) zijn beiden hoogleraar Economie aan de Universiteit van Tilburg en verbonden aan Netspar.

Literatuur

Agarwal, S., J. Driscoll, X. Gabaix en D. Laibson, 2009, Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation, *Brookings Papers on Economic Activity*, Fall 2009, 51-117.

Angrist, Joshua, en Jörn-Steffen Pischke, 2010, The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics, NBER Working Paper 15794.

Ashraf, N., D. Karlan en W. Yin, 2006, Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines, *Quarterly Journal of Economics*, vol. 121(2):635-72.

Benz, M., en S. Meier, 2008, Do people behave in experiments as in the field? Evidence from donations, *Experimental Economics*, vol. 11(3): 268-81.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir en J. Zinman, 2010, What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment, *Quarterly Journal of Economics*, vol. 125(1): 263-305.

Binswanger, J., en K.G. Carman, 2010, The miracle of compound interest: Does our intuition fail?, CentER Discussion Paper 2010-137.

Bodie, Z. en H. Prast, 2010, Rational Pensions for Irrational People, paper gepresenteerd op de Netspar-conferentie Macroeconomics of Pension Reform.

Bovenberg, A., R. Koijen, T. Nijman, en C. Teulings, 2007, Saving and investing over the life cycle and the role of collective pension funds, Netspar Panel Paper 1.

Charness, G., en P. Kuhn, 2010, Lab Labor: What Can Labor Economists Learn from the Lab?, in: Orley Ashenfelter en David Card (eds.), Handbook of Labor Economics, volume 4.

Cleave, B., N. Nikiforakis en R. Slonim, 2011, Is there selection bias in laboratory experiments? The case of social and risk preferences, IZA Discussion Paper 5488.

Dellaert, B., 2010, Interactive online decision aids for complex consumer decisions Benedict Dellaert, Netspar Panel Paper 19.

Duflo, E., R. Glennerster en M. Kremer, 2007, Using Randomization in Development Economics; a Toolkit, CEPR Discussion Paper 6059.

Duflo, E., M. Kremer en J. Robinson, 2008, How High Are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya, *American Economic Review: Papers & Proceedings*, vol. 98(2): 482-88.

Duflo, E., en E. Saez, 2006, The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment, *Quarterly Journal of Economics,* vol. 118(3): 815-42.

Falk, A., 2007, Gift exchange in the field, *Econometrica*, vol. 75(5): 1501-11.

Fisher, R.A., 1935, The Design of Experiments, Oliver and Boyd, Edinburgh.

Gautier, P., en B. van der Klaauw, 2011, Selection in a field experiment with voluntary participation, *Journal of Applied Econometrics*, forthcoming.

Gneezy, U. en A. Rustichini, 2000, A Fine is a Price, *Journal of Legal Studies*, vol. 29(1): 1-17.

Goldstein, D., E. Johnson en W. Sharpe, 2008, Choosing Outcomes versus Choosing Products: Consumer-Focused Retirement Investment Advice, *Journal of Consumer Research,* vol. 35(4): 440-56.

Haan, M.A., en P. Kooreman, 2002, Free Riding and the Provision of Candy Bars, *Journal of Public Economics*, vol. 83(2): 279-93.

Harrison, G.W., en J. List, 2004, Field Experiments, *Journal of Economic Literature*, vol. 42(4): 1009-55.

Karlan, D., en J. List, 2007, Does Price Matter in Charitable Giving? Evidence from a large-scale natural field experiment, *American Economic Review*, vol. 97(5): 1774-93.

Kastoryano, S., en B. van der Klaauw, 2011, Dynamic evaluation of job search assistance, IZA Discussion Paper 5424.

Klaauw, B. van der, 2010, Aan het werk, *TPEdigitaal*, jaargang 4(2): 130-47.

Klaauw, B. van der, E. Leuven en H. Oosterbeek, 2004, Financiële prikkels voor studenten, *Economisch Statistische Berichten*, vol. 89(4430): 156-57.

Koning, P., 2011, Experimenten in de sociale zekerheid, *Economisch Statistische Berichten*, vol. 96(4605): 150-53.

Kooreman, P., en H.M. Prast, 2010, What Does Behavioral Economics Mean for Policy? Challenges to Savings and Health Policies in the Netherlands, *De Economist*, vol. 158(2): 101-22.

Kortleve, N., en A. Slager, 2010, Consumenten aan het roer. Strategische toekomstvisie voor de Nederlandse pensioensector, NEA Paper 27.

LaLonde, R.J., 1986, Evaluating the Econometric Evaluations of Training Programs with Experimental Data, *American Economic Review*, vol. 76: 604-20.

List, J.A., 2006, The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions, *Journal of Political Economy,* vol. 114(1): 1-37.

Leuven, E., M. Lindahl, H. Oosterbeek en D. Webbink, 2007, The effect of extra funding for disadvantaged students on achievement, *Review of Economics and Statistics*, vol. 89(4): 721-36.

Leuven, E., H. Oosterbeek en B. 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: 1243-65.

Leuven, E., H. Oosterbeek, J. Sonnemans en B. van der Klaauw 2011, Incentives versus sorting in tournaments: evidence from a field experiment, *Journal of Labor Economics* vol. 29(3): 637-58.

Levitt, S.D., en J.A. List, 2009, Field Experiments in Economics: The past, the present, and the future, *European Economic Review*, vol. 53(1): 1-18.

Lusardi A., en M. van Rooij, 2010, Financial Literacy: Evidence and Implications for Consumer Education, Netspar Panel Paper 16.

Nijman, T., en A. Oerlemans, 2008, Maatwerk in Nederlandse Pensioenproducten, NEA Paper 8.

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

Praag, M. van, 2008, Politici willen niet weten of beleid werkt, *NRC Handelsblad*, 24-11-2008:7.

Rooij, M. van, C. Kool en H. Prast, 2007, Risk-Return Preferences in the Pension Domain: Are People Able to Choose?, *Journal of Public Economics*, vol. 91: 701-22.

Smulders, Y. M., M. Levi, C.D.A. Stehouwer, M. H.H. Kramer en A. Thijs, 2010, De rol van epidemiologisch bewijs in de zorg voor individuele patiënten, *Nederlands Tijdschrift voor Geneeskunde*, vol. 154(19): 892-96..

Soetevent, A., 2004, Een duit in het mandje - de rol van anonimiteit bij kerkcollectes, *Economisch Statistische Berichten*, vol. 89(4435): 275-77.

Soetevent, A., 2005, Anonymity in Giving in a Natural Context – A Field Experiment in 30 Churches, *Journal of Public Economics*, 89(11-12): 2301-23.

Verbaal, G., 2011, The Preference Indicator; An Online Tool for Closing the Pension Expectation Gap, Master thesis Economics and Finance of Aging, Tilburg University.