Van evidentie naar impact

Dinand Webbink

Ik wil dit betoog beginnen met een merkwaardig feit. Over de Goede Doelen sector in Nederland en in Europa weten we veel. We weten hoeveel geld er omgaat, in Nederland ongeveer 4,5 miljard euro, in Europa ongeveer 50 miljard euro. We weten waar het geld vandaan komt, wie de gevers zijn, en ook waar het naar toe gaat. We weten ook veel over de Goede Doelen organisaties, vaak kennen we het salaris van de directeur of de directie, en soms weten we ook nog waar de directie gedineerd heeft en welke wijn ze heeft gedronken. Het merkwaardige feit is dat we nauwelijks weten of al die middelen van de Goede Doelen sector goed worden besteed of dat er veel geld over de balk wordt gegooid. Wat zijn eigenlijk de resultaten van de vele projecten/activiteiten die worden gefinancierd door de Goede Doelen sector? Het antwoord daarop is: we weten nog weinig. Daarom is in september 2009 het Erasmus Centre for Strategic Philanthropy (ECSP) opgericht. Het centrum doet (onder meer) onderzoek naar de effecten van de projecten/activiteiten van Goede Doelen organisaties met als doel het vergroten van de impact van de Goede Doelen sector.

1 Inleiding

En daarmee heb ik het woord geïntroduceerd dat in de Goede Doelen sector momenteel heel veel aandacht trekt: impact. Sinds ik me ben gaan verdiepen in de Goede Doelen sector is me opgevallen dat aan het woord 'impact' een heleboel verschillende betekenissen wordt toegekend. Niet alleen door mensen uit de Goede Doelen sector maar ook door consultants en wetenschappers vanuit verschillende disciplines. Over impact worden veel vragen gesteld en er worden heel moeilijke en wellicht onmogelijke opdrachten verbonden aan impactmeting. Welke uitkomsten moet je meten? Wanneer moet je meten? Impact voor wie, de geldgever of de ontvanger? Zijn er ongewenste effecten of effecten op andere doelgroepen? Gaat het om de tevredenheid van de doelgroep? Je zou het ook anders kunnen zeggen: er bestaat nogal wat verwarring. U begrijpt dat ik me geen beter moment had kunnen wensen om me met impactmeting in de Goede Doelen sector bezig te gaan houden. In de komende veertig minuten zal ik uit de doeken doen hoe ik dat wil gaan aanpakken. Daarbij zal ik veel voorbeelden gebruiken van onderzoek dat ik de afgelopen jaren heb gedaan, vooral op het terrein van onderwijs. Impactmeting wordt daar meestal beleidsevaluatie genoemd. De komende jaren mag ik me met een veel

TPEdigitaal 2011 jaargang 5(3) 60-75

breder terrein dan onderwijs gaan bezighouden. De methoden en ervaringen die ik ga bespreken zijn echter even zeer bruikbaar voor de vele andere gebieden die bestreken worden door Goede Doelen organisaties en publieke overheden.

2 Waarover gaat impactmeting?

Het eerste antwoord van vanmiddag geef ik meteen. Impactmeting is voor mij het vaststellen van het oorzakelijk effect van een bepaald project of van een bepaalde beleidsinterventie.[1] Het gaat om de vraag in hoeverre we bepaalde uitkomsten kunnen toeschrijven aan het gevoerde beleid (of aan bepaalde projecten). Als we weten wat het effect van bepaald beleid is kunnen we deze kennis gebruiken om het beleid te versterken en daarmee meer impact te halen uit de ingezette middelen. Impactmeting / beleidsevaluatie zie ik daarom als een middel om evidentie te verzamelen, te leren en de prestaties te verbeteren. Het belangrijkste probleem bij impactmeting wil ik introduceren met een voorbeeld.

Enkele jaren geleden stelde de Europese Commissie een vraag aan het ministerie van Sociale Zaken en Werkgelegenheid over het werken van jeugdigen in Nederland. Op grond van artikel 7, lid 3 van het Europees Sociaal Handvest mogen leerplichtige jeugdigen geen zodanige arbeid verrichten dat zij niet ten volle het onderwijs kunnen volgen. In Nederland mogen 15-jarigen 's morgens vanaf zes uur ochtendkranten bezorgen. Om te kunnen beoordelen of Nederland daarmee handelt in overeenstemming met het verdrag is Nederland gevraagd om te onderzoeken of het bezorgen van ochtendkranten van invloed is op de schoolprestaties van 15-jarigen. Onderzoekers van de Universiteit Nijmegen hebben vervolgens een studie uitgevoerd naar de effecten van het bezorgen van ochtendkranten door 15-jarigen (Vrieze et al. 2001) dat de titel kreeg 'Vroege Vogels'. De onderzoekers concludeerden dat er geen aanwijzingen zijn 'dat het bezorgen van ochtendkranten door 15-jarigen van negatieve invloed is op de schoolprestaties'. Bovendien vonden zij 'geen aanwijzingen dat de lichamelijke en psychische conditie van ochtendkrantenbezorgers negatief wordt beïnvloed door het kranten bezorgen. Op alle aspecten komen de ochtendkrantbezorgers er positiever uit dan de niet-bezorgers. Krantenbezorgers voelen zich beter uitgerust en ze kunnen zich beter concentreren. Krantenbezorgers gaan met meer plezier naar school dan de controlegroep. Krantenbezorgers willen nog meer dan de controlegroep met hard werken veel bereiken in hun leven. Anders gezegd, het bezorgen van ochtendkranten leidt tot veel goeds. Dat is goed nieuws voor de ministers van Onderwijs, Cultuur en Wetenschappen, en van Sociale Zaken en Werkgelegenheid. Laten we alle middelbare scholieren een baantje geven vroeg in de ochtend, dan slaan we twee vliegen in één klap. Met die baantjes kunnen we een hoop mooie dingen produceren en waarschijnlijk nog belangrijker, we geven een belangrijke bijdrage aan de vorming van 15-jarigen.

[1] Deze definitie is niet door mij bedacht en wordt bijvoorbeeld ook gehanteerd door de Wereldbank (zie http://www.worldbank.org/oed/ie/) of door Howard White, de Excutive Director van *het International Initiative for Impact Evaluation* (3ie).

Dat is pas effectief beleid. Ik heb de indruk dat u toch enige twijfels hebt over dit beleid. Of wellicht over het onderliggende onderzoek. Gelooft u de resultaten niet? Dat lijkt me helemaal terecht. De controlegroep die wordt gebruikt in het 'Vroege Vogels' onderzoek' lijkt namelijk niet heel geloofwaardig. Ik kom daar straks op terug.

Binnen de economische wetenschap is de afgelopen twintig jaar geweldig veel aandacht besteed aan het vaststellen van oorzakelijke effecten. Er zijn nieuwe methoden ontwikkeld en tal van toepassingen gevonden. Dit terrein heeft de naam Program Evaluation gekregen. Het raamwerk waarbinnen dit onderzoek plaatsvindt is verrassend eenvoudig en wordt aangeduid als het Potentiële Uitkomsten Model (Rubin 1974, 1977; Holland 1986). De leidende vraag binnen dit model is: wat zou er gebeurd zijn als het beleid niet was ingevoerd? Of op individueel niveau: wat zou de uitkomst voor het individu zijn als dit individu niet te maken had gehad met het beleid. Voor een individu zijn er derhalve twee potentiële uitkomsten: een uitkomst in de situatie zonder het beleid en een uitkomst in de situatie met het beleid. Het oorzakelijk effect van het beleid voor dit individu is dan eenvoudig te bepalen, namelijk het verschil tussen de twee potentiële uitkomsten. Voor de hele populatie kan het oorzakelijk effect dan bepaald worden door het gemiddelde te nemen over alle individuen. Het probleem is echter dat we maar één uitkomst waarnemen. Die andere uitkomst kunnen we nooit waarnemen en hiervoor zijn we aangewezen op een vergelijking met een andere groep die het beleid niet heeft ondergaan. Voor het bepalen van het oorzakelijk effect van het beleid maken we in dat geval de aanname dat de uitkomst voor die andere groep gelijk is aan de potentiële uitkomst die we niet kunnen waarnemen. We nemen aan dat de uitkomst voor die andere groep de uitkomst is die we hadden gekregen als het beleid niet was uitgevoerd. Deze aanname speelt een cruciale rol in de evaluatie. In normaal Nederlands hebben we het dan over de geloofwaardigheid van de controlegroep.

Bij welke aannamen krijgen we een geloofwaardige controlegroep en bij welke aannamen is er reden tot twijfel? Het onderzoek op het terrein van Program Evaluation heeft hierover veel helderheid verschaft. De afgelopen jaren zijn we steeds beter gaan begrijpen welke aannamen we maken bij verschillende technieken gericht op het bepalen van de effecten van beleid of interventies. Een van de belangrijkste inzichten is dat de geloofwaardigheid van de controlegroep afhangt van de vraag of we begrijpen waarom een bepaalde groep wel de beleidsinterventie heeft gehad en de andere groep niet. In de literatuur over Program Evaluation wordt dit aangeduid als toewijzing aan de interventie (*assignment to treatment*). Aan de hand van het voorbeeld van de studie naar de Vroege Vogels wil ik dit verduidelijken.

In het 'Vroege Vogels' voorbeeld zijn we geïnteresseerd in het effect van een krantenwijk op de resultaten van scholieren. De onderzoekers hebben daarvoor de schoolresultaten van een groep leerlingen met een krantenwijk vergeleken met de schoolresultaten van een groep leerlingen zonder krantenwijk. En daarbij is rekening gehouden met een aantal verschillen tussen de groepen zoals geslacht, leeftijd, etnische herkomst en schooltype. De aanname die gemaakt wordt is dat de potentiele schoolresultaten van de 'Vroege Vogels' als ze geen krantenwijk zouden heb-

ben, gelijk zijn aan de schoolresultaten van de leerlingen zonder krantenwijk, rekeninghoudend met de genoemde verschillen. Is deze aanname geloofwaardig? Dat hangt af van de vraag welke leerlingen een krantenwijk hebben en welke leerlingen niet. Als de krantenwijk door loting wordt toegewezen aan bepaalde leerlingen zouden we geen verschillen verwachten tussen leerlingen met en leerlingen zonder krantenwijk. Als leerlingen echter bewust kiezen voor een krantenwijk kunnen er vele verschillen zijn tussen de twee groepen die niet zo gemakkelijk te observeren zijn. 'Vroege Vogels' zijn wellicht energieke ambitieuze types die ook de eerste uren van de dag goed willen besteden. De controlegroep bestaat wellicht uit 'Late Vogels' die vooral de late uurtjes goed willen besteden. Als deze verschillen tussen de groepen ook belangrijk zijn voor de schoolresultaten zal een vergelijking van de uitkomsten van deze groepen niet het oorzakelijk effect van de krantenwijk opleveren. Een onderzoeker die deze verschillen waarschijnlijk niet kan waarnemen zal dan de verkeerde conclusies trekken. U begrijpt inmiddels dat ik de aanname uit het 'Vroege Vogels' onderzoek niet erg geloofwaardig vind. Maar is het dan ook belangrijk dat die aanname niet geloofwaardig is? Mijn antwoord daarop is ja. Als die aanname niet geloofwaardig is kunnen bepaalde effecten geheel ten onrechte worden toegeschreven aan een beleidsinterventie.

Het probleem met 'de Vroege Vogels' studie is dat leerlingen zelf kiezen voor de krantenwijk. In de sociale werkelijkheid is zelfselectie eerder regel dan uitzondering. Bij vrijwel alle beleidsinterventies of projecten is sprake van selectie van deelnemers. En dat is ook logisch, mensen verschillen immers in voorkeuren en mogelijkheden, en dat leidt tot verschillen in keuzes. Deze zelfselectie is het belangrijkste probleem uit de beleidsevaluatie. Hoe weten we bij een vergelijking tussen twee groepen of het verschil in uitkomsten wordt veroorzaakt door de beleidsinterventie en niet het gevolg is van andere niet geobserveerde verschillen tussen deze groepen? Als een beleidsmaker wil weten wat het effect van een specifiek programma voor jongeren is, bijvoorbeeld het programma 'Meedoen, Leren en Winnen' van de Johan Cruijff Foundation, zal selectie een rol spelen bij de evaluatie. We mogen immers verwachten dat jongeren die willen deelnemen een andere groep zullen zijn dan jongeren die niet willen deelnemen aan het project. Het economisch onderzoek van de afgelopen jaren heeft laten zien dat een onderzoeker die geen rekening houdt met het selectieprobleem tot de conclusie kan komen dat een effect positief is terwijl het ware effect negatief is. Als we echt het effect van beleid willen weten zullen we een oplossing moeten vinden voor het selectieprobleem. Anders, zullen we nooit met zekerheid een bepaalde uitkomst kunnen toeschrijven aan een bepaalde beleidsinterventie.

3 Hoe kunnen we de impact van beleid op een geloofwaardige manier vaststellen?

Het onderzoek van de afgelopen jaren heeft veel inzicht opgeleverd voor het oplossen van het selectieprobleem. De meest overtuigende oplossingen zijn gebaseerd

op een experimentele of quasi-experimentele opzet. Cruciaal daarbij is dat we inzicht hebben in de toewijzing van de beleidsinterventie. We begrijpen waarom sommige individuen wel te maken hebben met de beleidsinterventie en anderen niet. Dit wordt ook wel *design based* onderzoek genoemd. [2] De impact van beleid kan worden vastgesteld door gebruik te maken van transparante onderzoeksdesigns.

Het gecontroleerde sociale experiment. De eerste oplossing voor het selectieprobleem is het gecontroleerde sociale experiment. Door loting wordt bepaald wie de beleidsinterventie wel of niet krijgt. De loting zorgt ervoor dat elk individu evenveel kans heeft op het krijgen van de beleidsinterventie. We mogen daarom verwachten dat de controlegroep zowel op geobserveerde als niet geobserveerde kenmerken vergelijkbaar is met de experimentele groep. Het effect van de beleidsinterventie kan dan worden bepaald door de uitkomsten in de experimentele groep te vergelijken met die in de controlegroep. Tot zover is het allemaal heel eenvoudig. Waarom zien we dan nog weinig sociale experimenten in Nederland? Ik denk dat daarvoor twee redenen zijn aan te wijzen. De eerste reden gaat over tijd en middelen. Sociale experimenten kosten geld en vergen tijd, en beleidsmakers kunnen niet zo lang op antwoorden blijven wachten. De tweede reden, en dat geldt zeker voor de wereld van het onderwijsbeleid en die van Ontwikkelingssamenwerking, is dat beleidsmakers moeite hebben met loting. Ik heb de indruk dat voor veel beleidsmakers het 'L-woord' nog steeds emotioneel beladen is. Ongelijke behandeling stuit velen tegen de borst. Het bezwaar blijkt dan meestal te zijn dat een bepaalde groep een kansrijke interventie wordt onthouden. Het blijft echter altijd de vraag of deze interventie wel werkt. Hoewel de interventie vooraf kansrijk wordt geacht is nooit uit te sluiten dat het uiteindelijke effect nul of zelfs ongunstig is. Als we een boer in een ontwikkelingsland microfinanciering aanbieden kan dit de start van een mooi bedrijf zijn, maar we kunnen niet uitsluiten dat deze boer hierdoor juist een grotere schuldenlast krijgt. De keuze om niet te experimenteren betekent bovendien vaak dat nieuw beleid moet worden ingevoerd zonder goede onderbouwing. Liever experimenteren met de hele populatie dan een deel van de populatie anders behandelen. Ik ben geen ethicus, maar mijn indruk is dat de ethische bezwaren van dit alternatief minstens even groot zijn en waarschijnlijk groter dan die van het 'L-woord'.

Ondanks deze bezwaren vinden wel degelijk sociale experimenten in Nederland plaats, en het aantal experimenten neemt ook toe. Zo voer ik samen met mijn CPB-collega's Marc van der Steeg en Roel van Elk een experiment uit waarbij een coach wordt toegewezen aan leerlingen in het Middelbaar Beroepsonderwijs. De toewijzing van leerlingen aan de experimentele en controlegroepen en ook de toe-

[2] De laatste jaren is ook kritiek gekomen op deze methoden. Deze zouden vooral gericht zijn op het vergaren van zo hard mogelijk bewijs maar daarmee de grote vragen uit het oog verliezen: 'Good answers instead of good questions'. Zie onder andere Heckman (2010), Deaton (2010), Imbens (2010), Angrist & Pischke (2010).

wijzing van de docent is door loting tot stand gekomen. In dit experiment waren onvoldoende middelen beschikbaar voor alle leerlingen. Door te loten kreeg elke leerling evenveel kans om tot de experimentele groep te behoren. In dit geval kan ik helemaal geen ethische bezwaren bedenken tegen loten. Dit soort situaties doet zich overigens veel vaker voor. Zo komt het vaak voor dat bij subsidiemaatregelen de aanvragen het beschikbare budget overschrijden. Bij aanvragen van gelijke kwaliteit kan in dat geval worden geloot, met als nevenopbrengst dat het effect van de subsidie na enige tijd kan worden vastgesteld. Ook bij Goede Doelen organisaties zullen de middelen vaak niet toereikend zijn voor alle aanvragers, en kan loting soms ook worden toegepast. Dit geeft dan niet alleen een eerlijke verdeling van middelen maar biedt bovendien een prachtige kans om impact te meten. De afgelopen jaren is ook een groot aantal experimenten gestart binnen het kader van 'Onderwijsbewijs', een fonds voor experimenten in het onderwijs. In de eerste ronde zijn achttien experimenten gefinancierd, in de tweede ronde zullen nog eens negentien experimenten worden gefinancierd. En natuurlijk worden binnen het *Top Institute for Evidence Based Education Research* (TIER) al enkele jaren sociale experimenten uitgevoerd.

Internationaal is er de afgelopen jaren sprake van een toename van experimenten. Opvallend daarbij is dat veel gecontroleerde sociale experimenten plaatsvinden in ontwikkelingslanden. Zo zijn er onder de vlag van het zogenoemde *Action Poverty Lab* 245 gerandomiseerde evaluaties uitgevoerde in 43 landen en op veel verschillende thema's, zoals gezondheid, microfinanciering, landbouw, arbeidsmarkt, onderwijs, milieu en bestuur.[3]

Natuurlijke experimenten. Gecontroleerde experimenten ontstaan door de hand van de onderzoeker. Er ontstaan echter ook regelmatig experimenten door toevallige situaties in de werkelijkheid, dit worden natuurlijke experimenten genoemd. Natuurlijke experimenten bieden ook een oplossing voor het selectieprobleem.

Bestaande lotingen benutten

In verschillende situaties in de werkelijkheid wordt loting toegepast, bijvoorbeeld bij de toelating tot bepaalde populaire middelbare scholen of bij de toelating tot studies zoals geneeskunde. Deze lotingen bieden vaak de mogelijkheid om oorzakelijke effecten vast te stellen, ook als de loting niet werd geïnitieerd met als doel het uitvoeren van een gecontroleerd experiment.[4] Samen met Rob Luginbuhl en Inge de Wolf heb ik gebruik gemaakt van een loting die door de Inspectie van het Onderwijs wordt toegepast (Luginbuhl et al. 2009). De Inspectie van het Onderwijs trekt, in het kader van het Onderwijsverslag, elk jaar een steekproef van scholen. Deze scholen worden bezocht door een Inspecteur. De steekproef van scholen

[3] Zie www.povertyactionlab.org

[4] In Nederland is de gewogen loting voor geneeskunde benut om het effect van deze opleiding op verschillende uitkomsten vast te stellen (Leuven, et al. 2009). Voor de VS is het effect van schoolkeuze op schoolprestaties vastgesteld door gebruik te maken van de loting bij toelating tot bepaalde scholen (Cullen et al. 2006).

wordt aselect getrokken, eigenlijk net als bij een echt gecontroleerd experiment. Door de resultaten van deze steekproef van scholen te vergelijken met de resultaten van andere scholen konden wij het effect van een schoolbezoek door een Inspecteur vaststellen. Wij vonden een kleine verbetering van de leerprestaties na het bezoek van de Inspecteur.

Regressie discontinuïteiten

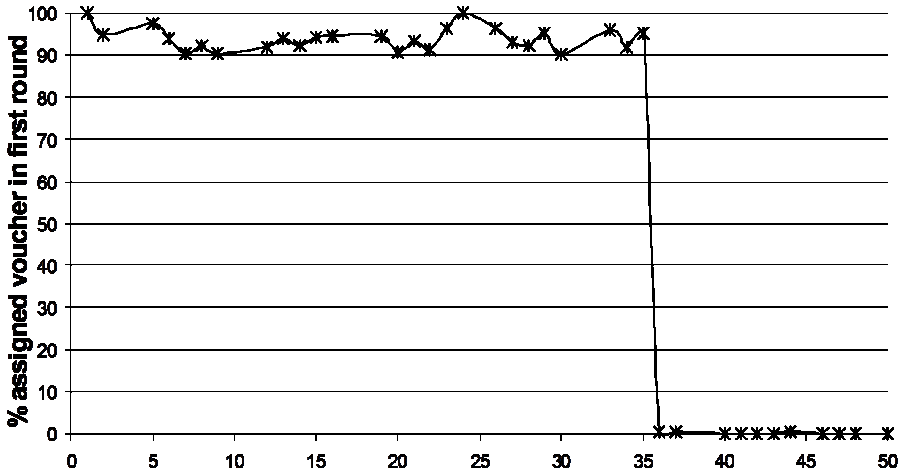
Behalve loting zijn er in de werkelijkheid nog veel meer situaties te ontdekken die sterk lijken op een gecontroleerd experiment. Een van de meest prominente situaties in de evaluatieliteratuur is de zogenoemde regressiediscontinuïteit. Deze techniek is al in de jaren zestig ontwikkeld binnen de psychologie, maar is de afgelopen tien jaar opnieuw uitgevonden binnen de economische wetenschap en voorzien van een theoretisch fundament (Cook 2008). Regressie discontinuïteiten zijn situaties waarbij de toewijzing van de beleidsinterventie afhangt van een score op een bepaalde variabele. Individuen net boven een bepaalde grenswaarde krijgen de beleidsinterventie, individuen net beneden deze grenswaarde krijgen de beleidsinterventie niet. De belangrijkste aanname is dat individuen aan beide zijden van de grenswaarde goed vergelijkbaar zijn.[5] Aan de hand van een voorbeeld wil ik deze aanpak toelichten.[6]

Vanaf 2008 zijn door het ministerie van OCW beurzen beschikbaar gesteld voor leraren. Deze beurzen zijn bedoeld voor het verhogen en verbreden van kwalificaties van leraren. Het is bijvoorbeeld mogelijk om een complete Bachelor of Master's studie te volgen met deze beurs. Een belangrijke vraag is echter of het ontvangen van een beurs ook daadwerkelijk tot extra deelname aan onderwijs leidt of dat de beurs gebruikt wordt voor het financieren van een opleiding die men ook zonder beurs wel was gaan volgen. De beurs vervangt dan de eigen middelen of de middelen vanuit de school. Begin vorig jaar heb ik hier samen met Marc van der Steeg en Roel van Elk naar gekeken (Van der Steeg et al. 2010). Voor het vaststellen van het effect van de beurs op de deelname aan hoger onderwijs hebben we gebruik gemaakt van een toevalligheid bij de toedeling van de beurs. In de eerste ronde waren er ongeveer 7500 leraren die een beurs wilden en er waren slechts 5000 beurzen beschikbaar. De toewijzing van de beurs ging op basis van het moment van aanmelding *(First come, First serve*). In Figuur 1 is de kans op het krijgen van een beurs afgezet tegen de dag van aanmelding.

[5] De veronderstelling is dat de relatie tussen de onderliggende toewijzingsvariabele en de potentiële uitkomsten continu is rond de grenswaarde van de toewijzingsvariabele.

[6] Andere toepassingen zijn Leuven et al. (2007) of Oosterbeek en Webbink (2010).

Figuur 1 Kans op een beurs in de eerste ronde naar dag van aanmelding eerste ronde



Vanaf de eerste dag van de aanmelding is de kans op het toegewezen krijgen van een beurs erg hoog, ongeveer 95%. Een klein deel van de aanmeldingen is afgewezen omdat de aanvraag niet voldeed aan bepaalde criteria.[7] De kans op een beurs blijft hoog tot en met dag 35. Dan zien we een plotseling daling van de kans op een beurs tot nul procent. Deze daling is het gevolg van het feit dat het geld op was. In ons onderzoek gebruiken we deze plotselinge daling van de kans op een beurs. Wij vonden dat de lerarenbeurs de kans op deelname aan het hoger onderwijs verhoogt met 10%-punt. Dat betekent dat de overheid, bij deze opzet van de lerarenbeurs, 10 beurzen moet verstrekken om één leraar extra te verleiden tot deelname aan het hoger onderwijs. Eén beurs wordt niet gebruikt en acht beurzen worden gebruikt voor opleidingen die ook zonder beurs zouden zijn gevolgd. Economen noemen dit een hoge *dead weight loss*.

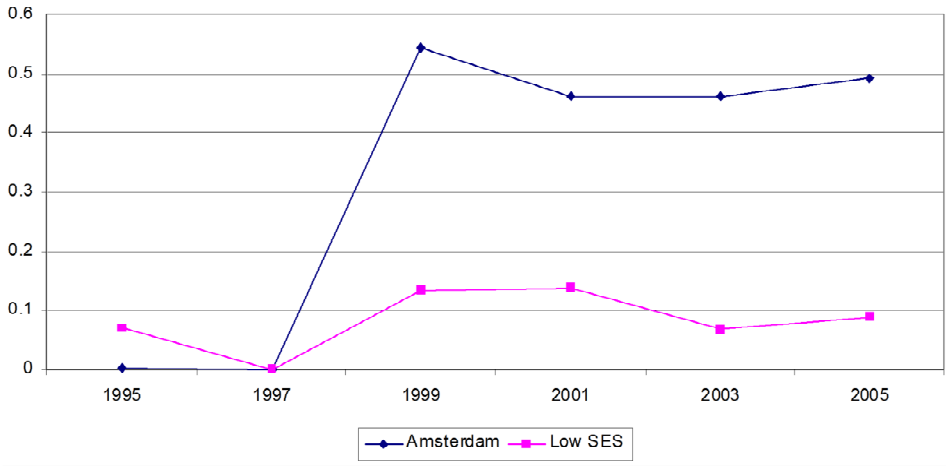
Difference-in-differences modellen

Een derde prominente techniek voor het oplossen van het selectieprobleem is het zogenoemde difference-in-differences model. In dit model wordt gebruikt gemaakt van een experimentele en een controlegroep en is sprake van een voor- en nameting. De belangrijkste aanname is dat de trend in de controlegroep, dat is het verschil tussen de voor- en nameting, gelijk is aan de trend in de experimentele groep als de beleidsinterventie niet zou hebben plaatsgevonden. In de Verenigde Staten zijn DD-modellen vaak toegepast door het beleid in een bepaald gebied (staat, stad of deelgemeente) te analyseren waarbij andere gebieden als controlegroep gebruikt worden.[8] Samen met Victoria Chorny (oud-CPB) heb ik deze techniek toegepast voor een analyse van het zogenoemde 'accountability-beleid' in het Amsterdamse

[7] Zoals het aantal contacturen, de accreditatie van de aanbieder, en de bevoegdheid van de docent.
[8] Zie bijvoorbeeld Card (1990), Jin and Leslie (2003), Meyer et al. (1995), Eissa and Liebman (1996).

basisonderwijs vanaf het midden van de jaren negentig (Chorny en Webbink 2010). De gemeente Amsterdam bemoeide zich intensief met het basisonderwijs en maakte afspraken over deelname aan de te behalen resultaten op de Cito-toets. Scholen moesten plannen opstellen voor het behalen van schoolspecifieke doelen. Aan de uitvoering van de plannen en het behalen van de resultaten werden middelen gekoppeld. Aan de hand van een groot gegevensbestand van leerlingen in het basisonderwijs, het zogenoemde PRIMA-onderzoek, hebben we de ontwikkeling van de scores op de Cito-toets in Amsterdam geanalyseerd. We hebben de trend in Amsterdam vergeleken met die in heel Nederland en ook met de trend in een specifieke steekproef met veel achterstandsleerlingen (en met de trend in de andere grote steden). Figuur 2 laat zien hoe de trend in Amsterdam (de blauwe lijn) en de trend in de steekproef met veel achterstandsleerlingen (de paarse lijn) is gaan afwijken van de landelijke trend (de x-as).

Figuur 2 Trend in Cito-score in Amsterdam en in de lage SES steekproef ten opzicht van de landelijke trend



We zien dat de trend in Amsterdam ten opzichte van de landelijke trend geheel vlak is tot 1997, en daarna doet zich een spectaculaire stijging voor van de Amsterdamse resultaten. De ontwikkeling in de steekproef met veel achterstandsleerlingen is vlak. In onze schattingen vinden we een verbetering van de Amsterdamse toetsresultaten met ongeveer 0,5 standaarddeviatie. Dat zijn ongeveer 5 punten op de Cito-toets en dat is een hele sterke stijging. Het zal u waarschijnlijk niet zijn ontgaan dat er in de media ook veel aandacht is besteed aan deze opvallende progressie. Er zijn veel manieren om de toetsresultaten te verbeteren. Zo gaan de scores fors omhoog als de zwakke leerlingen uit de toets worden gelaten, en veel oefenen op de toets helpt ook. In ons onderzoek hebben we naar verschillende kanalen gekeken die zouden kunnen leiden tot een onbedoelde stijging van de toetsresultaten. Zo hebben we gekeken naar het uitsluiten van leerlingen, verwijzing naar het spe-

ciaal onderwijs of zittenblijven. Deze analyses hebben geen aanwijzingen opgeleverd voor strategisch gedrag van Amsterdamse scholen.[9]

Deze methoden, gebaseerd op experimentele of quasi-experimentele onderzoeksdesign, maken het mogelijk om oorzakelijke effecten van beleidsinterventies of projecten vast te stellen. De toepassing van deze methoden kan derhalve evidentie opleveren die beleidsmakers kan helpen om de effectiviteit van hun beleid te vergroten. Anders gezegd, dit type onderzoek kan het beleid *evidence based* maken. Echter, het gebruik maken van onderzoeksresultaten in beleid is in de praktijk niet vanzelfsprekend.

4 Gebruik maken van evidentie: Evidence Based Beleid

Het doel van impactmeting is tweeledig: het verantwoorden van de inzet van middelen en het leren over de effecten. Allereerst bestaat er altijd de behoefte om vast te stellen of de middelen goed worden ingezet. Dit geldt voor de overheid en nog sterker voor Goede Doelen organisaties. Zij zijn immers opgericht om 'goed te doen', dus om impact te hebben. En voor de gevers aan Goede Doelen organisaties is het ook weer belangrijk om te zien dat de organisaties daadwerkelijk resultaten boeken. In de tweede plaats, en in mijn ogen het meest belangrijk, is het doel van beleidsevaluatie om te leren over de effecten van beleid en deze kennis te gebruiken voor het versterken van de effectiviteit van het beleid. Instrumenten die niet werken kunnen worden gestopt, beleid dat wel werkt kan worden uitgebreid. De baten van beleid dat werkt kunnen geweldig hoog zijn. Neem het onderwijs. Inmiddels weten we dat onderwijs grote opbrengsten levert voor zowel individuen als landen. Hogere testscores zijn belangrijke voorspellers voor economische groei. Beleid dat de Nederlandse onderwijsresultaten blijvend kan verhogen kan op termijn geweldige productiviteitseffecten opleveren. Dit betekent tegelijkertijd ook dat beleid dat niet werkt geweldig hoge kosten heeft. Het goed gebruik maken van de beschikbare evidentie is daarom heel belangrijk, maar, in de praktijk is de relatie tussen onderzoek en beleid verre van eenvoudig.

Spanning tussen onderzoek en beleid. Beleid maken is niet eenvoudig. Er zijn veel belangen, er is weinig tijd, de politieke arena heeft een geweldige dynamiek, politici moeten scoren en journalisten moeten ook scoren. Alle beleidsproblemen hebben een hoge *sense of urgency*, beleidsmakers moeten 'meters maken' en 'steden staan in brand'. Slecht nieuws lijkt in de media beter te scoren dan goed nieuws. Een aantal jaren geleden heb ik met een aantal CPB-collega's een vergelijking uitgevoerd van de prestaties van het Nederlands onderwijs met die van het onderwijs in een aantal andere rijke landen. Onze conclusie was, en dat zal u mogelijk verbazen, dat het Nederlands onderwijs er niet slecht voor staat. De kop van

[9] Een ander toepassing van DD-modellen is het onderzoek naar de effectiviteit van het beleid gericht op het verminderen van voortijdig schoolverlaten (Van der Steeg et al. 2008).

ons persbericht was dan ook: 'Nederlands onderwijs niet onder de maat'. De volgende dag stond echter in een vooraanstaand landelijk dagblad: 'CPB: Nederlands onderwijs onder de maat'. De teneur van de berichtgeving over de Goede Doelen sector lijkt niet heel anders. In deze beleidsomgeving leidt nieuwe evidentie niet automatisch tot aanpassingen van beleid. Goed nieuws wordt omarmd, slecht nieuws over de resultaten van het beleid wordt bestreden. Als je een tijd in Den Haag hebt gewerkt herken je de beleidsreflexen bij 'slecht nieuws':
1. Het onderzoek deugt niet;
2. Het onderzoek richt zich slechts op een deel van het beleid;
3. De doelen van het beleid waren heel anders;
4. Het beleid is al bijgesteld, het onderzoek is achterhaald.

Al deze beleidsreflexen heb ik de afgelopen jaren gezien, en ze zijn ook heel goed te begrijpen. Ik heb niet de illusie dat de relatie tussen onderzoek en beleid ooit zonder problemen zal zijn. Wel denk ik dat er mogelijkheden zijn om het beleid verder te versterken met evidentie. In dat verband is de laatste jaren de term *evidence based beleid* in zwang geraakt. Ik ben daarvan een sterk voorstander. Maar laat ik eerst aangeven wat *evidence based* beleid niet is.

Ongewenste 'beleidsonderbouwing'. Nadat de recessie in Nederland hard had toegeslagen nam het aantal studenten in het hoger onderwijs fors toe. In onderzoek naar de stijging van de deelname aan hoger onderwijs werd echter de conclusie getrokken dat de stijging niets te maken had met de daling van de conjunctuur (Berger en Broek 2010). Met als directe implicatie, de stijging van de deelname aan hoger onderwijs zou wel eens structureel kunnen zijn en derhalve structureel meer middelen vereisen. Hoe hebben de onderzoekers het effect van de conjunctuur vastgesteld? Dit hebben ze gedaan door studenten te vragen naar de motieven voor hun deelname. Het belangrijkste motief was dat studenten zichzelf wilden ontwikkelen en dat heeft niet te maken met de conjunctuur. Als we dit onderzoek bekijken vanuit het potentiële uitkomsten model dan moeten we echter constateren dat dit onderzoek ons geen enkel inzicht verschaft in de vraag wat deze studenten zouden hebben gedaan als de conjunctuur veel beter was geweest. Dit onderzoek lijkt uitsluitend bedoeld om middelen te claimen van de Rijksbegroting. Dit type onderbouwing van beleid heeft niets te maken met *evidence based* beleid. Dit soort onderzoek kunnen we beter achterwege laten.

Hoe dan wel? Maar hoe zou *evidence based* beleid er dan wel moeten uitzien? Hoe slagen we erin om beleid beter gebruik te laten maken van onderzoek. Uiteindelijk hebben beleidsmakers en onderzoekers toch hetzelfde doel: het verbeteren van de impact van de ingezette middelen. *Evidence based* beleid zou ik willen definiëren als beleidskeuzen baseren op geloofwaardige wetenschappelijke evidentie. Serieus proberen vast te stellen wat wel of niet werkt, en deze informatie gebruiken bij beleidsbeslissingen. *Evidence based* beleid betekent niet alleen goed onderzoek doen maar vooral ook onderzoek beschikbaar hebben voor de beleidsbeslissingen. Om

dit te bereiken is een goede afstemming tussen beleid en onderzoek noodzakelijk. Deze afstemming zou er al moeten zijn vanaf de start van het beleid. Ik wil ingaan op twee mogelijkheden

Serieus werk maken van de eerste beleidsfase. De eerste stadia van het beleidsproces bieden in mijn ogen de meeste mogelijkheden om gebruik te maken van wetenschappelijke kennis. Bij de keuze van beleidsinstrumenten kan gekeken worden naar de internationale ervaringen met deze instrumenten. Vervolgens komt de fase aan de orde dat de nieuwe instrumenten in de Nederlandse context getest worden. In het onderwijs zien we dan vaak allerlei proefprojecten plaatsvinden. Dit is een periode die zich uitstekend leent voor het vergaren van kennis. Bijvoorbeeld, als 'pilots' worden uitgebreid met controlegroepen kunnen eerste effecten worden vastgesteld. Na de testfase van het beleid komt de implementatiefase. Ook hier zijn kansen voor beleidsevaluatie. Het komt immers vaak voor dat beleid niet direct over de hele linie wordt ingevoerd. Zo werd bij de introductie van de Tweede Fase in het voortgezet onderwijs gestart met ongeveer 25% van de scholen. Een gefaseerde invoering leidt ertoe dat sommige scholen al wel en sommige scholen nog niet te maken hebben met het nieuwe beleid, hetgeen kansen schept om goede controlegroepen te vormen. Ik denk dat we in de eerste fase van het beleidsproces veel kansen laten liggen om al te leren over de effecten van het beleid.

Alleen projecten financieren die zicht geven op de resultaten. De tweede mogelijkheid om beleid meer *evidence based* te maken is om gebruik te maken van de financiering. Dit kan door alleen projecten te financieren die zicht geven op de resultaten. Een mooi voorbeeld hiervan is het *Social Innovation Fund* van president Obama. Dit fonds financiert projecten die bijdragen aan sociale innovatie op het terrein van gezondheid, werkgelegenheid of jeugd. Projecten komen alleen voor financiering in aanmerking als ze gericht zijn op meetbare uitkomsten en de effecten op een geloofwaardige manier zichtbaar maken. Daarmee wordt bereikt dat de middelen impact kunnen hebben voor veel mensen en tegelijkertijd wordt een catalogus verkregen van 'benaderingen die werken'. Ook interessant aan dit initiatief is dat het fonds wordt gevuld met zowel publieke middelen als met private middelen afkomstig uit de Goede Doelen sector. Deze aanpak lijkt nu ook navolging te krijgen binnen de Nederlandse Ontwikkelingssamenwerking, en dat vind ik een heel goede zaak. Het beschikbaar stellen van middelen voor Ontwikkelingssamenwerking lijkt momenteel veel minder vanzelfsprekend dan in het verleden. Ontwikkelingsorganisaties in Nederland wordt steeds vaker gevraagd hun resultaten te laten zien. Dit heeft in de afgelopen jaren geleid tot vele 'evaluatierapporten' maar de focus lag daarin nog niet op het vaststellen van de oorzakelijke effecten van projecten of programma's. De komende jaren is het echter te bedoeling om bij belangrijke evaluaties de focus te verleggen en gebruik te maken van een *counterfactual* en een nulmeting. Daarmee wordt de financiering van ontwikkelingsorganisaties verbonden aan de zichtbaarheid van de resultaten, in lijn met het fonds van Obama. Deze benadering maakt het mogelijk om antwoord te krijgen op de cruciale vraag

welke vormen van ontwikkelingshulp daadwerkelijk impact hebben. En dat is toch uiteindelijk wat we willen weten.

Ik ben ook erg blij met een aantal nieuwe inspanningen die zicht geven op de effecten van beleid. Allereerst natuurlijk Onderwijsbewijs. Een tweede inspanning die ik bijzonder vind is een project dat ik zelf de afgelopen maanden heb mogen doen samen met Marc van der Steeg, Roel van Elk en Frans-Bauke van der Meer. Wij hebben meegedacht over de opzet van een aantal nieuwe beleidsmaatregelen op het terrein van onderwijs. Het doel daarvan is het beleid zodanig vorm te geven dat we evidentie kunnen genereren over de effecten. Ik ben ook erg benieuwd of deze ontwerpen stand zullen houden in het geweld van de dagelijkse beleidsdynamiek. Ook de aanpak bij de zogenoemde Wijkscholen in Rotterdam verdient lof. De beslissing over voortzetting van de financiering is bij dit project afhankelijk gemaakt van de resultaten die worden vastgesteld met een serieuze evaluatie.

Serieus werk maken van beleidsevaluaties leidt onvermijdelijk tot 'slecht nieuws' voor verschillende dossiers. Evaluaties zullen laten zien dat sommige beleidsinstrumenten slecht werken of helemaal niet werken. Hoe goed het beleid ook is voorbereid, er is nooit een garantie dat dit beleid daadwerkelijk effectief is in een nieuwe context. Echter, als tijdig kan worden vastgesteld dat de resultaten tegenvallen, kan grote schade worden vermeden. Een goede opzet geeft dus ook een *early warning* systeem. Ministers of bestuurders van Goede Doelen organisaties die zorgen voor zicht op de effectiviteit van beleid verdienen daarvoor in mijn ogen applaus.

5 Onderzoeksagena

De onderzoeksagenda die ik de komende jaren wil uitvoeren bestaat uit het toepassen van moderne econometrische evaluatietechnieken voor beleid en projecten op het terrein van filantropie, gezondheid en onderwijs. Samen met Karen Maas, Kellie Liket, Frank Hubers, Job Harms en Lara Hemmes en ondersteund door de Vereniging Fondsenwervende instellingen (VFI) hebben we het afgelopen jaar een start gemaakt met diverse evaluaties waaronder projecten rond de Johan Cruijff Courts, micro-financiering, trainingsprogramma's voor jongeren en maatschappelijke dienstplicht. Ik zie dit als een prachtig begin dat smaakt naar veel meer. Ik wil evaluaties uitvoeren die daadwerkelijk inzicht bieden in de effecten van beleidsinterventies en die bruikbaar zijn voor beleidsmakers. Om dit te kunnen bereiken zoek ik naar afstemming tussen beleid en onderzoek. Als bij nieuwe projecten of beleidsinstrumenten vanaf de start de insteek wordt gekozen om te leren over de effecten, kunnen evaluaties van hoge kwaliteit tot stand komen. Met deze evidentie kan de impact van de ingezette middelen worden vergroot. In het beleid lijkt het tonen van acties vaak belangrijker dan het vaststellen van de daadwerkelijke resultaten van beleid. Ik denk dat het beleid gebaat is bij een verschuiving van actieplannen naar serieuze evaluaties. Het gaat uiteindelijk niet om de acties maar om de echte resultaten van het beleid. Aan het begin van dit betoog stelde ik vast dat

we nog weinig weten over de resultaten van de inspanningen van de Goede Doelen sector. Dat gaan we de komende jaren veranderen.

Auteur

Dinand Webbink (e-mail: webbink@ese.eur.nl) is hoogleraar Policy Evaluation aan de Erasmus School of Economics van de Universiteit van Rotterdam. Zijn leerstoel is verbonden aan het *Erasmus Centre for Strategic Philanthropy.*

Literatuur

Angrist, J.D, en J Pischke, 2010, The credibility revolution in empirical economics: How better research design is taking the con of out econometrics, *Journal of Economic Perspectives*, vol. 24 (2): 3-30.

Antenbrink, P., K. Burger, M. Cornet, M. Rensman en D. Webbink, 2005, Nederlands onderwijs en onderzoek in internationaal perspectief, CPB Document 88.

Beerends, H., en S. van der Ploeg, 2001, Onderzoek vergoeding schoolspecifieke knelpunten, Regioplan, Report OA-230.

Berger, J.H.J. en S. Broek, 2010, Aanmeldingsgolf door crisis?, Analyse van studentenaantallen en studiemotieven 2009-2010, Research voor Beleid.

Black, D.A., T.G. McKinnish en S.G. ., 2005, Tight labor markets and the demand for education: Evidence from the Coal Boom and Bust, *Industrial and Labor Relations Review*, vol. 59 (1): 3-16.

Card, D., 1990, The Impact of the Mariel boatlift on the Miami Labor Market, Industrial and Labor Relations Review, vol. 43(2): 245-57.

Chorny, V., en D. Webbink, 2010, The effect of accountability policies in primary education in Amsterdam, CPB Discussion Paper 144.

Cook, T. D., 2008, Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics, *Journal of Econometrics*, vol. 142(2): 636-54.

Cullen, J.B., B.A. Jacob en S. Levitt, 2006, The effect of school choice on participants: evidence from randomized lotteries, *Econometrica*, vol. 74(5): 1191-1230.

Dearden, L., C. Emmerson, C. Frayne en C. Meghir, 2009, Conditional Cash Transfers and School Dropout Rates, *Journal of Human Resources*, vol. 44(4): 827-857.

Deaton, Angus, 2010, Instruments, Randomization, and Learning about Development, *Journal of Economic Literature*, vol. 48(2): 424-55.

Eissa, N., en J. Liebman, 1996, Labor Supply Response to the Earned Income Tax Credit, *Quarterly Journal of Economics*, vol. 111(2): 605-37.

Hanushek, E.A., 1986,'The economics of schooling: production and efficiency in public schools, *Journal of Economic Literature*, vol. 24 (3): 1141-77.

Hanushek, E.A., 2003, The failure of input-based schooling policies, *The Economic Journal*, 113(485): F64-F98.

Hanushek, E.A. en L. Woessmann, 2010, The Economics of International Differences in Educational Achievement, NBER Working Paper 15949.

Heckman, James J., 2010, Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy, *Journal of Economic Literature*, vol. 48(2): 356-98.

Holland, P, 1986, Statistics and causal inference (with discussion and rejoinder). Journal of the American Statistical Association, vol. 81(396): 945-70.

Imbens, Guido W., 2010, Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),. Journal of Economic Literature, vol. 48(2): 399-423.

Jin, G., en P. Leslie, 2003, The effect of information on product quality: Evidence from restaurant hygiene grade cards, *Quarterly Journal of Economics*, vol. 118(2): 409-51.

Lalonde, 1986, Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review*, vol. 76(4): 604-20.

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 I. de Wolf, 2009, The effects of health education on health outcomes: Evidence from a natural randomized experiment, in mimeo.

Luginbuhl, R., D. Webbink en I. De Wolf, 2009, Do inspections improve primary school performance?, *Educational Evaluation and Policy Analysis*, vol. 31 (3): 221-37.

Meyer, B.D., W.K. Viscusi en D.L. Durbin, 1995, Workers' Compensation and Injury Duration : Evidence from a Natural Experiment, *American Economic Review*, vol 85(3): 322-40.

Oosterbeek, H., en D. Webbink, 2010, Does studying abroad induce a brain drain? *Economic*a, vol. 78: 347-66.

Rivkin, S.G., 1995, Black/White differences in Schooling and Employment, *Journal of Human Resources*, vol. 30 (4): 826-52.

Rubin, D.B., 1974, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, vol. 66(5): 688-701.

Rubin, D.B., 1977, Assignment to treatment group on the basis of a covariate. *Journal of . Educational Statistics,* vol. 2: 1-26.

Steeg, M.W.van der, R. van Elk en D. Webbink, 2010, Het effect van de lerarenbeurs op scholingsdeelname docenten, CPB Document 205.

Steeg, M.W.van der, R. van Elk en D. Webbink, 2008, Did the 2006 covenant program reduce school dropout in the Netherlands? CPB Document 177.

Vrieze, G., R. Kloosterman en N. van Kessel, 2001, Vroege Vogels, Onderzoek naar de gevolgen van het 's ochtends kranten bezorgen voor de schoolprestaties en schoolbeleving van 15-jarige ochtendkrantbezorgers, ITS Nijmegen.