

Beter evalueren van hulp vergroot de impact ervan

Wereldwijd geven donoren voor meer dan 125 miljard euro aan ontwikkelingshulp, waarvan ongeveer een vijfde via non-gouvernementele organisaties. Wetenschappelijke impact-evaluaties – al dan niet via experimenten – hebben veel kennis opgeleverd over wat werkt en wat niet. Desondanks is er in Nederland nog niet veel veranderd in de praktijk van het evalueren van ontwikkelingsprojecten.

IN HET KORT

- Een betere evaluatie van de Nederlandse hulp, bij voorkeur door onafhankelijke partijen, kan de hulp effectiever maken.
- Overheden en ngo's lijken niet geïnteresseerd in de effecten van hulp op het bereiken van de *Sustainable Development Goals*.
- De overheid moet een sterkere regierol op zich nemen in het afdwingen van grondige, prospectieve evaluaties.

ERWIN BULTE

Hoogleraar aan Wageningen University & Research (WUR)

ROBERT LENSINK

Hoogleraar aan de Rijksuniversiteit Groningen en aan de WUR

Nederland heeft zich gecommitteerd aan de zeventien zogeheten *Sustainable Development Goals* (SDG's) die voor 2030 behaald moeten worden. De SDG's betreffen doelstellingen op het gebied van armoedevermindering, ongelijkheid, klimaatverandering, onderwijs, gezondheid en mensenrechten. Het lijkt vanzelfsprekend dat impact-evaluaties, waarmee de bijdrage van projecten en programma's aan het behalen van de SDG's kan worden bepaald, een centrale plaats innemen in het Nederlandse beleid.

Niets is minder waar. Het Ministerie van Buitenlandse Zaken, maar ook veel non-gouvernementele organisaties (ngo's), lijken niet echt geïnteresseerd te zijn in de effecten van hun hulpprogramma's. Er lijkt meer interesse te bestaan voor het *aantal* mensen dat wordt bereikt met een hulpprogramma's dan voor de bijdrage van de programma's aan de levens van de doelgroep of het behalen van de SDG's. Als gevolg hiervan blijft veel beleid en een groot deel van de hulp gebaseerd op veronderstellingen. Hoe komt het dat er zo weinig gebruik wordt gemaakt van impact-evaluaties en wat kunnen we eraan doen?

Onvoldoende evaluaties

Het Nederlandse hulpbeleid besteedt veel aandacht aan monitoring en evaluatie, waarbij er wordt bekeken of de uitvoerende organisatie gedaan heeft wat beloofd werd, of de financiën op orde zijn en of de doelpopulatie is bereikt.

Vaak gebeurt dit achteraf via een zogenaamde *retrospectieve* evaluatie. Een dergelijke evaluatie is belangrijk maar verschaft weinig inzicht in de toegevoegde waarde van een project aan de uiteindelijke doelstelling ervan.

Toch worden grondige evaluaties maar weinig uitgevoerd. Dat komt ten eerste doordat veel grondige evaluaties tot teleurstellende inzichten leiden over het effect ervan.. Ten tweede kost een goede evaluatie tijd, die er bijna nooit lijkt te zijn. Ngo's willen hun projecten snel opschalen, en het ministerie wil snel weten of een project werkt of niet. Ten derde zijn grondige evaluaties vaak kostbaar. Met name het verzamelen van data over uitkomstvariabelen, zoals inkomen of gezondheid, bij een groot aantal huishoudens is duur. En ten derde kan de angst een rol spelen. Als een project geen aantoonbare impact blijkt te hebben, kunnen ngo's hun subsidie verliezen en het draafvlak voor hulp afkalven. Om de effecten van hulp goed te kunnen bepalen is grondig evalueren noodzakelijk. Daarbij is het ontwerp van de evaluatie van belang.

Vorm evaluatie belangrijk

Om de impact van een project te meten is het noodzakelijk te kunnen vaststellen wat er zonder het project zou zijn gebeurd – dus de *counterfactual*. Van de mensen die deelnemen aan een project kunnen we echter niet meten waar zij waren uitgekomen als ze niet hadden deelgenomen. Daarom moet deze behandelgroep vergeleken worden met niet-deelnemers als controlegroep.

De belangrijkste uitdaging hierbij is het vermijden van selectie-effecten, die kunnen ontstaan als de twee groepen niet vergelijkbaar zijn. Een goede impactmeting vereist dus op zijn minst het verzamelen van gegevens over een controlegroep die vergelijkbaar is met de behandelgroep.

Bovendien is het voor een solide impact-evaluatie zeer nuttig als er ook informatie beschikbaar is over de controlegroep en de behandelgroep *voordat* het project wordt gestart. Dan kan er een zogenaamde *double difference*-analyse worden uitgevoerd. Verschillen tussen de twee groepen die constant zijn over de tijd (zoals opleidingsniveau van de ouders) kunnen zo de resultaten niet beïnvloeden.

Belangrijk voor het meten van de impact is dus dat er een evaluatieplan wordt gemaakt alvorens het project wordt opgestart, zodat men een nulmeting kan uitvoeren en er op tijd kan worden nagedacht over selectie-effecten. In de praktijk van evaluaties is er echter zelden sprake van een dergelijke prospectieve evaluatie. Daardoor wordt er bijna altijd onvoldoende aandacht besteed aan het verzamelen van de juiste informatie om het effect van de inspanning te kunnen evalueren. Hoogstens verzamelen ngo's

informatie over de behandelgroep voor en na de interventie. Soms is er ook informatie over een controlegroep na de interventie, maar informatie over de controlegroep vóór de interventie ontbreekt bijna altijd. Op basis van realisaties achteraf kan er een vergelijking worden gemaakt, maar dat levert zelden een betrouwbare counterfactual op. Evaluaties met behulp van gerandomiseerde interventies kunnen hierbij meer inzicht verschaffen.

Evaluaties met gerandomiseerde interventies

Het meten van de impact van hulp kan worden verbeterd door het uitvoeren van evaluaties met behulp van gerandomiseerde interventies (*randomised controlled trials*: RCT's), waarbij er willekeurig individuen, groepen of gemeenschappen worden gekozen om in aanmerking te komen voor een interventie.

Het gebruik van evaluaties met behulp van gerandomiseerde interventies is niet nieuw. Zo worden er al geruime tijd gerandomiseerde studies uitgevoerd voor het testen van medicijnen – dubbelblinde, gerandomiseerde studies gelden hier als de gouden standaard. Ook in onderzoek naar een negatieve inkomensbelasting worden dit soort experimenten uitgevoerd (Ross, 1970).

Het gebruikmaken van gerandomiseerde studies ter vermindering van de mondiale armoede leverde voor Abhijit Banerjee, Esther Duflo en Michael Kremer in 2019 de Nobelprijs voor de Economie op. Het betekende een nieuw paradigma binnen de ontwikkelingseconomie (zie kader 1).

Een groot voordeel van een gerandomiseerde interventie is dat effecten van een (ontwikkelingshulp-) project simpelweg kunnen worden bepaald door een vergelijking van de behandel- en controlegroepen. Randomisatie zorgt er namelijk voor dat, gemiddeld genomen, deze groepen identiek zijn, waardoor we eventuele verschillen kunnen toewijzen aan de interventie. Dit maak randomisatie effectief in het voorkomen van selectie-effecten, zonder dat er – zoals bij studies gebaseerd op niet-experimentele data – complexe statistiek nodig is om te controleren voor selectie-effecten. Economen zijn er steeds beter in geworden, en zijn ook in staat om steeds meer te randomiseren – inclusief complexe zaken als *accountability* en *governance* op lokaal niveau.

Lessen trekken uit gerandomiseerde studies

Afgelopen tien jaar is veel ervaring opgedaan met gerandomiseerde evaluaties, zoals door het Abdul Latif Jameel Poverty Action Lab (J-PAL) en door Innovations for Poverty Action (IPA). Alleen al door IPA zijn er meer dan 800 evaluaties uitgevoerd, in samenwerking tussen wetenschappers en veldwerkers. Maar ook onafhankelijke organisaties en wetenschappers hebben bijgedragen aan de kennis over wat werkt en wat niet. Om inzicht te geven in de resultaten van gerandomiseerde studies, bespreken we drie sectoren waarover veel is geschreven: het onderwijs, de gezondheidszorg en het microkrediet.

Onderwijs

De gangbare visie met betrekking tot onderwijs was dat economische groei in belangrijke mate wordt bepaald door 'menselijk kapitaal' en dat extra financiële ondersteu-

ning voor onderwijs daarom van groot belang is. De eerste RCT's van Michael Kremer richtten zich op onderwijs, en met name op de effecten van het aanbieden van extra inputs (zoals boeken of flip-overs). Tot zijn verrassing leverde dit bijna niets op. De vele gerandomiseerde studies die volgden, bevestigden dit beeld. Het is zinloos om simpelweg meer fondsen aan onderwijs te verstrekken, bijvoorbeeld door meer lesmateriaal aan te bieden. Van veel groter belang is dat curricula goed aansluiten bij het leerniveau van de leerlingen. Dat pleit dus niet voor een standaardcurriculum, maar voor een gedifferentieerd aanbod in de lokale taal. Tevens benadrukt dit type onderzoek het belang van het verbeteren van het management van de scholen, en van de verantwoordelijkheid van de docenten.

Zorg

Op het gebied van gezondheidszorg is er veel onderzoek gedaan naar de effecten van prijsstelling op gezondheidsproducten, met name op het terrein van preventieve gezondheidszorg. Dat betreft de vraag of je, vanuit het principe van financiële duurzaamheid en marktconforme vergoeding moet vragen bijvoorbeeld voor een malarianet of ontwormingstabletten, of dat je deze producten tegen een lage prijs of zelfs gratis zou moeten verstrekken. Veel onderzoek heeft laten zien dat de vraag naar preventieve gezondheidsproducten verdamppt als er zelfs maar een zeer lage (sterk gesubsidieerde) prijs gevraagd wordt (Kremer en Miguel, 2007). Het gratis verstrekken van producten – bijvoorbeeld zodra het malarianet versleten is – heeft daarentegen als mogelijk

Gerandomiseerde studies binnen de ontwikkelingseconomie

KADER 1

De ontwikkelingseconomie richt zich vooral op de vraag hoe we het beste mondiale armoede kunnen verminderen. In het verleden hielden veel ontwikkelingseconomen zich met name bezig met de grote thema's, zoals "wat is de onderliggende oorzaak van armoede?", "wat is het belang van handel?" en "wat is de effectiviteit van ontwikkelingshulp?" Complexe vragen die moeilijk te beantwoorden zijn met het economische standaardinstrumentarium, in ieder geval op de manier die veel economen als voldoende grondig beschouwen. Duflo, Kremer en Banerjee pleitten ervoor om de grote vragen op te splitsen in kleine deelvragen die beter te beantwoorden zijn, bijvoorbeeld middels experimenten. Ze wezen erop dat armoede in de wereld niet moet worden gezien als het gevolg van één overkoepelend probleem, maar als het gecombineerde resultaat van vele kleine problemen.

Tevens merkten ze op dat de vraag of ontwikkelingshulp nu wel of niet werkt een onbelangrijke, misschien zelfs nutteloze vraag is. Waar het om gaat is uit te zoeken welke projecten werken (en deze te steunen en uit te breiden) en welke projecten niet werken (en deze te stoppen). De nadruk op 'randomisatie' zorgde er tevens voor dat ontwikkelingsconomen

in toenemende mate correcte identificatie centraal gingen stellen, wat waarschijnlijk een belangrijke reden is voor de herwonnen positie van ontwikkelingseconomie binnen het standaardcurriculum van de economische wetenschappen. Onder andere door de inspanningen van de Nobelprijswinnaars is de ontwikkelingseconomie als wetenschapsveld gerevitaliseerd.

De experimentele revolutie heeft niet alleen de banden tussen ontwikkelingsconomen en andere economen verstevigd. Er bestond ook weinig tot geen contact tussen ontwikkelingswetenschappers en ontwikkelingswerkers in de praktijk.

De directe implicatie van de zienswijze van Duflo, Kremer en Banerjee is dat ontwikkelingsconomen uit hun ivoren toren moeten komen, en dat ze samen moeten gaan werken met ontwikkelingsorganisaties en overheden om te onderzoeken welke interventies effectief zijn. Gerandomiseerde experimenten zijn bij uitstek toepasbaar bij het evalueren van ontwikkelingsprojecten. De Nobelprijswinnaars hebben ervoor gezorgd dat het doen van veldwerk weer een belangrijk onderdeel van het takenpakket van de ontwikkelingseconoom werd, waarmee de kloof tussen wetenschappers en mensen in de praktijk deels gedicht kon worden.

gevaar dat mensen geen geld meer willen betalen voor het product. Onderzoek heeft echter uitgewezen dat de leereffecten van gesubsidieerde producten groter zijn: door ze gratis te verstrekken leren mensen dat een product werkt en schaffen ze het daarna zelf aan (Dupas, 2014). Veel internationale organisaties – zoals het Britse Department for International Development (DFID), Save the Children UK, en het Millennium Project en Commission for Africa van de Verenigde Naties – hebben hun prijsbeleid voor dergelijke producten herzien en pleiten nu voor het gratis verstrekken van ‘gezondheidsproducten’.

Gerandomiseerde experimenten kennen veel voordelen, maar kunnen niet alle grote vragen beantwoorden

Microkrediet

Er is veel onderzoek gedaan naar de effecten van (micro) krediet. Microkrediet werd een belangrijke rol toegedicht bij het verminderen van het wereldwijde armoedeprobleem. Dat blijkt uit de uitreiking van de Nobelprijs voor de Vrede aan Muhammad Yunus van de Grameenbank, die voornamelijk microkredieten verstrekt. Recente studies schatten het effect van microkrediet echter lager in. Toegang tot microkrediet leidt zelden tot een omvangrijke armoedevermindering, laat staan tot een transformatie van het leven van ontvangers. Een mogelijke reden hiervoor is de hoge rente die vaak gerekend moet worden, en die belangrijke risico's voor lenende klanten met zich meebrengt.

Daartegenover suggereren vele gerandomiseerde studies dat het simpelweg geven van hulp zonder condities (*unconditional cash transfers, zoals 'giving directly'*) juist erg positieve effecten kan hebben. Over het algemeen gaan ontvangers zorgvuldig om met gratis geld – veel wordt gebruikt om te investeren of consumptie uit te smeren ‘gesmeerd te laten verlopen’, en het wordt niet over de balk gegooid zoals paternalistische lieden uit de wereld van de ontwikkelingssamenwerking soms vrezen.

Beperkingen van experimenten

Ondanks de grote voordelen van experimenteren om te leren, hebben gerandomiseerde experimenten ook nadelen. In sommige gevallen zijn andere methoden beter geschikt (of zelfs noodzakelijk) omdat de praktijk niet voldoet aan de voorwaarden voor een ideaal experiment.

Een belangrijk punt is dat veel zaken die ontwikkeling beïnvloeden niet makkelijk op te splitsen zijn in deelvragen. Wij zijn in ieder geval niet bekend met gerandomiseerde experimenten op het gebied van monetair beleid of buitenlandse handel. En zelfs echte micro-onderwerpen, zoals onvolledige mededinging in landbouw-waardeketens, laten

zich moeilijk randomiseren. Dit hoeft natuurlijk geen probleem te zijn, maar de dominantie van RCT's in het onderzoek lijkt ten koste te gaan van andere benaderingen. Het streven naar een *credible counterfactual* kan dan betekenen dat de *'first-order questions'* blijven liggen ten faveure van simpele randomiseerbare vragen. Statistische onderbouwing en publicatiekansen lijken in het economisch onderzoek soms zwaarder te wegen dan het belang van de onderliggende vraag.

Een tweede beperking van gerandomiseerde experimenten is dat ze ons vaak slechts in beperkte mate informatie opleveren over de onderliggende mechanismen die verklaren waarom iets wel werkt of niet werkt. Experimenten zijn primair gericht op het vaststellen van causale effecten, niet op het begrijpen van procedures. Daardoor blijft er vaak een analyse nodig van de economische en sociale context om iets wezenlijks te kunnen zeggen over de onderliggende mechanismen.

Ten derde geven de experimenten geen duidelijkheid over de vraag of een succesvol beleid in het ene land ook succesvol uit te voeren is in een ander land. Liefhebbers van gerandomiseerde experimenten zullen zeggen dat ieder experiment een nuttige observatie is, en onderdeel is van een groter plaatje dat geleidelijk duidelijker wordt. Andere technieken zijn vaak beter om de externe validiteit van een interventie te onderzoeken.

Tot slot zijn er veel gevallen waarbij een afruil tussen bias en precisie bestaat. Gerandomiseerde experimenten vermijden weliswaar selectieproblemen (resultaten zijn niet *biased*), maar bij interventies met een lage adoptie – zoals gebruikelijk in projecten die verzekeringen of microfinanciering betreffen – kan de precisie erg laag zijn. Dit leidt tot het risico van zogenaamde type II-fouten (het risico dat de onderzoeker ten onrechte concludeert dat er geen significant effect is). Een niet-gerandomiseerde studie, waarbij deelnemers achteraf worden vergeleken met niet-deelnemers op basis van *propensity-score matching*, kan tot een veel preciezer resultaat leiden dan een RCT met een hele kleine steekproef.

Soms is een onzuivere maar precieze schatting van het gemiddelde *treatment effect* nuttiger dan een zuivere maar niet-precieze schatting. Soms geldt het omgekeerde. Maar dat laat onverlet dat er van goed vormgegeven experimenten kan worden geleerd in de praktijk.

Evaluatie in de praktijk

In Nederland werd in 2015 een uniek evaluatieproject afgesloten. Door een groep van wetenschappelijke onderzoekers werden acht landenprogramma's van negentien allianties van Nederlandse ngo's gezamenlijk geëvalueerd (de zogenaamde MFS-II-evaluaties; MFS staat voor 'medefinancieringsstelsel'). Het idee was een prospectieve evaluatie uit te voeren, in een samenwerking van wetenschappers en mensen uit de praktijk. Wij zijn bij deze evaluatie nauw betrokken geweest.

De evaluatie was verre van optimaal, en veel minder prospectief dan gehoopt. Zo waren projecten in veel gevallen al gestart voordat we met de 'nulmeting' konden beginnen, en ook waren behandelgroepen door de ngo meestal zorgvuldig geselecteerd (en waren deze vaak al jaren de

vaste klanten van de implementerende organisatie). Zo konden we bijna nooit een gerandomiseerde studie uitvoeren. Daarnaast waren interventies meestal klein en waren we door tijdsdruk gedwongen om, erg snel na het afronden van het project, de impactmeting uit te voeren.

Desalniettemin hebben de meeste betrokkenen veel van deze evaluatie geleerd, al was het maar hoe we dit soort inspanningen in de toekomst anders zouden moeten organiseren. Het is naar onze mening daarom erg jammer dat deze evaluatiemethode geen vervolg heeft gekregen. Het Nederlandse evaluatiebeleid bestaat nu weer grotendeels uit retrospectieve evaluaties, meer gericht op het afleggen van verantwoording dan op leren.

Evalueren noodzaak voor betere hulp

In de praktijk van het Nederlandse ontwikkelingsbeleid wordt er met een combinatie van interventies getracht een bijdrage aan de SDG's te bewerkstelligen. Tegelijkertijd heeft het beleid ten doel de positie van het Nederlandse bedrijfsleven te bevorderen en de migratie te reduceren. Het huidige kabinet heeft hiertoe een aantal veranderingen in de ontwikkelingssamenwerking geïnitieerd, zoals het verleggen van de ontwikkelingssamenwerking naar instabiele regio's minder ver van Europa, zoals de Sahel, de Hoorn van Afrika, het Midden-Oosten, Noord-Afrika en West-Afrika.

De grotere aandacht voor samenwerking met het bedrijfsleven blijkt uit de oprichting van een Nationaal Fonds voor Klimaat en Ontwikkeling, waarmee Nederlandse bedrijven kunnen investeren in klimaatprojecten in ontwikkelingslanden. Tevens zijn er veel subsidies beschikbaar voor bedrijven uit de zogenaamde topsectoren, zoals het *Good Growth Fund*, die willen investeren in ontwikkelingslanden, en wordt er nadruk gelegd op de handelsmissies van de overheid voor Nederlandse bedrijven.

Toch is de bijdrage van dergelijke initiatieven aan het behalen van SDG's hoogst onzeker. In het licht van de aanzienlijke sommen belastinggeld die met de hulpinspanning zijn gemoeid, is het eigenaardig dat het Ministerie van Buitenlandse zaken niet een grotere regierol op zich neemt en de sector meer tot prospectieve impact-evaluaties dwingt. Vanuit de academische wereld staan mensen te trappelen om de handschoen op te nemen en samen met de ambassades en ngo's aan de slag te gaan.

Literatuur

Dupas, P. (2014) Getting essential health products to their end users: subsidize, but how much? *Science*, 345(6202), 1279–1281.

Kremer, M. en E. Miguel (2007) The illusion of sustainability. *Quarterly Journal of Economics*, 122(3), 1007–1065.

Ross, H. (1970) An experimental study of the negative income tax. *Child Welfare*, 49(10), 562–1569.