Wat bezielt economen?

Mijnheer de Rector, Dames en Heren,

Op de vraag van mijn titel hebt u vast al uw eigen antwoord bedacht. Dat niet al die antwoorden complimenteus voor mijn vakgebied zijn, besef ik terdege. Maar, zoals u natuurlijk al vermoedde, is mijn vraag niet retorisch bedoeld: in dit college wil ik met u bespreken wat economen bezielt.

Het antwoord is al te vinden in de titel van het beroemde boek dat Adam Smith, één van de grote figuren van de Schotse Verlichting, in 1776 publiceerde: *An Inquiry into the Nature and the Causes of the Wealth of Nations*. Adam Smith was een ontwikkelingseconoom *avant la lettre*, want hij vroeg zich af waarom in zijn tijd landen als Engeland en de Republiek rijk waren terwijl andere landen nog vastzaten in diepe armoede. Die vraag is nog steeds actueel, al was in de achttiende eeuw rijkdom vooral een verbazingwekkende uitzondering. Adam Smith zag armoede dan ook als een ellendige, maar normale toestand en was gefascineerd door de vraag hoe een land daaruit kon ontsnappen. Economen hebben in de volgende twee eeuwen naar antwoorden op dergelijke grote vragen gezocht, gedreven door intellectuele nieuwsgierigheid, maar ook bezield door de overtuiging dat hun analyse de wereld beter kon maken.

Smith dacht dat hij causale verbanden had gevonden, dus dat hij de oorzaken (de *causes* in zijn titel) van ontwikkeling kon identificeren. Anderen zijn hem daarin gevolgd. Voor Max Weber was het duidelijk waarom landen als Nederland en Engeland rijk waren geworden: het protestantisme leidde tot werklust en spaarzin. Later gaven de Nederlander J.H. Boeke en de Zweed Gunnar Myrdal culturele verklaringen voor de economische stagnatie in Azië. Daarover werd niets meer vernomen toen juist de onstuimige groei daar moest worden verklaard. Toen de koloniën in Afrika en Azië hun onafhankelijkheid kregen, zag men het feit dat die nieuwe landen niet konden lenen op de internationale kapitaalmarkt en over weinig technische expertise beschikten, als de kern van het probleem. Daarmee was de rol van ontwikkelingshulp meteen duidelijk: die moest technische kennis en kapitaal overhevelen van rijk naar arm. Nog

weer later werd armoede in ontwikkelingslanden toegeschreven aan hun koloniale verleden, aan de dalende prijzen van hun uitvoerproducten of aan hun geografische positie.

Dergelijke zeer algemene en tegenstrijdige ontwikkelingstheorieën konden lang gedijen omdat zij niet getoetst werden: de gegevens die daarvoor nodig waren, bestonden nog niet. Sinds die data er wel zijn, trachten economen de vraag van Adam Smith empirisch te beantwoorden: zij gebruiken regressiemethoden om na te gaan in hoeverre verschillen tussen landen in economische groei of in welvaartsniveau kunnen worden verklaard uit allerlei determinanten: het rechtssysteem van de kolonisator, de geografische ligging van een land, de gezondheid en het opleidingsniveau van de bevolking, de investeringsgraad, de ontvangen ontwikkelingshulp of het gevoerde handelsbeleid. Centraal in dit onderzoek staat de vraag of de gevonden verbanden wel causaal zijn.

Die vraag wordt mooi geïllustreerd door de anekdote waarmee Hans Clevers vorig jaar zijn jaarrede als president van de KNAW begon.[1] Hij vertelde dat hij eens met zijn zoontjes, die toen zeven en vijf jaar waren, op de Randmeren langs een rij windmolens zeilde en dat de jongste vroeg: “Als die wieken niet meer draaien, stopt onze boot dan?” Dat is een uitstekende vraag. Natuurlijk kan met een simpel experiment, het stilzetten van de windmolens, worden vastgesteld dat er geen causaal verband is tussen het draaien van de wieken en de voortstuwing van de zeilboot, maar de econoom die de oorzaken van welvaartsverschillen tussen landen onderzoekt, heeft die optie niet.

Geografie en instituties

Hij zou bijvoorbeeld graag het effect van een open handelsbeleid, een kwestie die nog steeds omstreden is, onderzoeken door landen willekeurig in te delen in twee groepen, de ene met, de andere zonder een dergelijk beleid. Dat is natuurlijk nooit gedaan, want geen enkele regering zal willen meedoen aan dat experiment. Het tijdschrift *The Economist* dacht dat de vraag al was beantwoord door een “natuurlijk experiment”. Immers, een kopgroep van Aziatische landen die hun grenzen in de jaren zestig van de vorige eeuw openden, liep in economisch opzicht heel snel uit op het peloton. Die redenering deugt natuurlijk niet: het verschil in economisch succes mag niet worden toegeschreven aan het handelsbeleid, want de kopgroep en de achterblijvers

verschilden ook in tal van andere opzichten. In regressies moet dan ook recht worden gedaan aan het feit dat de verdeling van de twee soorten beleid over landen zeker niet willekeurig is. Voor dat probleem bestaan er ingenieuze statistische technieken, maar geheel overtuigend zijn die meestal niet; of een gevonden verband werkelijk causaal is, blijft daardoor omstreden.

Het empirische onderzoek richt zich op twee soorten verklaringen voor de economische verschillen tussen landen: geografie en instituties. Die tweedeling doet in veel opzichten denken aan het bekende *nature* versus *nurture* debat.[2]

Geografie is hier een verzamelterm voor temperatuur, regenval en grondkwaliteit, de beschikbaarheid van delfstoffen, de mogelijkheden voor binnenlands vervoer en buitenlandse handel, kortom voor factoren die grotendeels vastliggen.[3] Die factoren kunnen van doorslaggevend belang zijn. Zo woont in Afrika ongeveer een derde van de bevolking in landen die niet aan zee liggen en ook geen hulpbronnen hebben.[4] In veel van die landen biedt alleen migratie hoop: een zandvlakte leent zich nu eenmaal niet voor economische ontwikkeling.

Als welvaart vooral het directe effect van geografische omstandigheden weerspiegelt, zou de economische situatie aan weerskanten van een grens (als die bijvoorbeeld een breedtegraad volgt) niet moeten verschillen. Het bekendste voorbeeld van een dergelijk natuurlijk experiment is dat het verschil tussen Noord- en Zuid-Korea zelfs vanuit de ruimte duidelijk te zien is: het is donker ten noorden en licht ten zuiden van die grens. Kennelijk zijn instituties die per land verschillen, belangrijker dan geografie.[5]

Jared Diamond geeft in zijn beroemde boek *Guns, Germs and Steel* een mooi voorbeeld van het belang van geografische factoren: dat de Europees-Aziatische landmassa zich langs een Oost-Westas uitstrekt, maakte de snelle verspreiding van de in het Midden Oosten ontwikkelde landbouw mogelijk: men bleef in geografische zin op ongeveer dezelfde breedte. Verspreiding van landbouwgewassen langs een Noord-Zuidas, zoals in Afrika, is uiteraard veel lastiger: een gewas komt dan snel in een gebied waar het niet kan gedijen.[6]

Met instituties wordt het overheidsbeleid aangeduid, maar vooral het kader waarbinnen dat beleid wordt gekozen: de grenzen aan de macht. Is een land een kleptocratie, wordt het dus

geregeerd door dieven, of een *developmental state* waarin machthebbers onderworpen zijn aan (al of niet democratisch) toezicht en daarom het algemeen belang nastreven? Voor Adam Smith was dit vertrouwd terrein: hij wist dat overheden ontwikkeling vaak hinderden, maar die ook konden stimuleren, door het garanderen van de randvoorwaarden van een markteconomie: rechtszekerheid, onderwijs, mededingingsbeleid, een eerlijk belastingstelsel. We zouden nu zeggen: door het bieden van een *enabling environment* voor ondernemerschap. De spectaculaire veranderingen in China vanaf het eind van de jaren zeventig en in veel Afrikaanse landen sinds de jaren negentig laten zien hoe belangrijk die rol is.

Net zoals *nature* en *nurture* zijn ook geografie en instituties niet strikt te scheiden. Zo hebben geografische omstandigheden vaak de aard van een koloniaal regime bepaald en daarmee van de instituties die ontstonden. Die erfenis kan lang doorwerken: daar waar een Europeaan een hogere kans liep te sterven, zoals in het zeer ongezonde Batavia van de VOC, zijn nu nog steeds zwakkere instituties te vinden.[7]

Het belang van instituties, van de aard van het regime in een ontwikkelingsland, wordt steeds duidelijker. Dat het aantal heel arme mensen in de wereld in één generatie dramatisch is gedaald, met ongeveer 600 miljoen volgens de schattingen van de Wereldbank, is naar mijn mening de belangrijkste gebeurtenis in de recente geschiedenis.[8] Dat die verheugende ontwikkeling veroorzaakt is door beleidswijzigingen, vooral in China, staat buiten kijf. Dat brengt ons bij een kernvraag voor de ontwikkelingseconomie. Als er in een ontwikkelingsland een slecht beleid wordt gevoerd, komt dat dan doordat het regime niet beter weet of omdat het geen prikkel heeft om het algemeen belang te dienen?

In het eerste geval is de oplossing eenvoudig: gebrek aan kennis kan worden verholpen. De adviserende rol van donoren, van giganten zoals de Wereldbank tot de kleinste boetiek hulporganisatie, is daarop gericht. Wetenschappelijk onderzoek heeft daarbij een zeer belangrijke bijdrage geleverd, die echter vrijwel nooit de media heeft gehaald. Ik geef enkele voorbeelden:

- de herziening van het wisselkoersbeleid in Ethiopië na de burgeroorlog;
- het opgeven door landen die landbouwproducten uitvoeren, van de stabilisatie van producentenprijzen;
- het in plaats daarvan aanbieden van asymmetrische financiële contracten (in feite *put* opties) aan boeren in Afrika die zich daarmee tegen prijsdalingen kunnen beschermen;
- de ontmanteling van het Tanzaniaanse systeem van prijsbeheersing.

Dit waren stuk voor stuk beleidsveranderingen met enorme gevolgen die waren gebaseerd op grondige, landspecifieke analyses. Economen gebruikten hierbij geen experimenten, maar deductieve methoden, onder andere door de theorie van het "*second best*" toe te passen.[9]

Maar in veel gevallen kan ook het beste advies niets uithalen: de regering kiest een beleid dat vooral het eigen belang dient, niet omdat hij niet beter *weet*, maar omdat hij niet anders *wil*. Bij de analyse van de reactie van overheden op prikkels werken politicologen en economen tegenwoordig zeer vruchtbaar samen. De oude naam voor de economische wetenschap, politieke economie, krijgt zo een nieuwe betekenis. De overheid is in deze visie niet de dienaar van het algemeen belang, zoals in de welvaartstheorie, maar een economisch subject met eigen doelstellingen, net zoals een bedrijf of een consument. Die overheid kan in bijzondere omstandigheden een groter belang dienen, maar alleen omdat hij anders zijn positie verliest door een verkiezingsnederlaag of een staatsgreep of omdat de kosten van het beleid voor het regime zelf te hoog worden.[10] Een regime van uitzuigers kan een stabiele oplossing zijn: er bestaat wel een alternatief dat verkiesbaar is in de zin van Pareto, dat wil zeggen een situatie waarin niemand slechter af is en sommigen beter af zijn, maar vanwege coördinatieproblemen blijft het regime toch aan de macht. Dan rijst de vraag of er een kleine externe verandering te bedenken is, een *nudge* of stootje in de juiste richting, die de stabiliteit van dat slechte evenwicht doorbreekt en daarmee ontwikkeling mogelijk maakt.[11]

Impact evaluatie: de nieuwe ontwikkelingseconomie?

Het politiek-economische onderzoek naar het ontstaan van normen en instituties en de prikkels voor machthebbers om ontwikkeling te bevorderen bloeit, maar heeft last van een methodologisch probleem: er is weinig ruimte voor een experimentele opzet. Dat werd vroeger

niet als een bezwaar gezien, maar causaliteit is inmiddels voor economen een obsessie geworden.[12] Dat siert hen natuurlijk als wetenschappers, maar heeft er toe geleid dat sommigen zich afkeren van de grote vragen, want methoden die causale relaties kunnen blootleggen zijn daarvoor zelden geschikt. Dat leidt tot verdeeldheid: de methodologisch rekkelijke onderzoekers houden zich bezig met grote vragen, de preciezen zijn noodgedwongen minder ambitieus in hun vraagstelling.

In de ontwikkelingseconomie werd dat duidelijk bij de opkomst van de *randomized controlled trials*, de RCTs. U kent die uit het geneesmiddelenonderzoek. Een groep patiënten wordt willekeurig (vandaar: *randomized*) ingedeeld in een *treatment* groep die de nieuwe medicijn krijgt en een *control* groep die een placebo (of een ander medicijn) krijgt. Noch de patiënten, noch de behandelaars weten wie in welke groep zit.[13] Aangezien de verdeling over de groepen is gerandomiseerd, kan een verschil in uitkomst worden toegeschreven aan de medicijn: andere verschillen tussen de twee groepen kunnen als de steekproef voldoende groot is, niet systematisch zijn.

Economen hebben deze methode overgenomen, allereerst in analyses van onderwijs en van de arbeidsmarkt.[14] Tijdens de hoge werkloosheid in de jaren tachtig omarmden politici allerlei werkgelegenheidsprogramma's die effectief *leken*, bijvoorbeeld omdat de deelnemers aan een omscholingsprogramma snel een baan vonden. Arbeidseconomen maakten korte metten met dergelijke beweringen: RCT-onderzoek liet meestal zien dat wat een succes leek, vooral een selectie-effect was. Het "succes" zei dus meer over de deelnemers dan over het programma zelf.

Sinds tien jaar worden RCTs op grote schaal door ontwikkelingseconomen gebruikt. Zij hebben geprobeerd met experimenten vragen te beantwoorden zoals: Hoe kunnen arme Mexicaanse ouders ervan worden overtuigd dat zij hun kinderen naar school moeten sturen? Kan verkiezingsgeweld worden bestreden door informatie daarover te verspreiden? Kan het informeren van burgers over de slechte kwaliteit van overheidsdiensten hen ertoe brengen om politieke actie te ondernemen? Welke maatregelen zijn het meest effectief voor het verbeteren van leerprestaties op Keniaanse basisscholen? Leidt de beschikbaarheid van schoon pompwater

in Benin tot minder diarree bij kinderen? Deze voorbeelden illustreren dat economen zich ver buiten hun vakgebied wagen en dat zij een verbijsterende diversiteit aan vragen aanpakken. Over de laatste twee voorbeelden wil ik wat meer zeggen.

In Kenia werd onderzocht welke maatregelen het meest effectief zijn om leerprestaties te verbeteren, onder andere betere schoolgebouwen, meer lesmateriaal, het inzetten van beter opgeleide onderwijzers. De theorie geeft hier nauwelijks houvast en het antwoord kwam als een grote verrassing: een ontwormingsprogramma waardoor kinderen vaker en in betere conditie naar school gingen en zich beter konden concentreren, was veel effectiever dan de meer conventionele maatregelen.[15]

In de Beninstudie, waarin vanuit onze groep Youdi Schipper een grote rol speelde, werden dorpen met en zonder waterpompen vergeleken. De waterkwaliteit werd geanalyseerd, zowel op het punt waar vrouwen water haalden (voor de *treatment* dorpen was dat de pomp, waaruit schoon water kwam) als binnen het huishouden, voordat het gedronken werd of gebruikt bij het koken. Er was uiteraard een groot verschil tussen de twee groepen bij de eerste meting, maar dat bleek bij de tweede meting verdwenen: er was geen significant verschil tussen de twee groepen dorpen in de kwaliteit van het drinkwater vlak voor gebruik. Nader onderzoek wees uit dat het water, meestal door een vrouw in een schaal op haar hoofd gedragen, door haar handen waarmee zij de schaal vasthield, werd vervuild. Zo werd een voortreffelijke interventie in het allerlaatste deel van de keten volledig teniet gedaan. De onderzoekers losten dit probleem op door het aanbieden van een gesloten watercontainer. Daarmee werd de vervuiling tussen pomp en huis op eenvoudige en doeltreffende wijze geëlimineerd.

Deze twee voorbeelden illustreren dat de vaakgehoorde kritiek dat men van impactevaluaties niet kan leren, onzinnig is. Ze laten ook zien dat ontwikkelingshulp vaak verkeerde doelstellingen nastreeft: bij onderwijs ligt de nadruk op het naar school gaan in plaats van op wat er daar geleerd wordt, bij drinkwater op de vraag of arme huishoudens toegang hebben tot een bron van schoon water in plaats van op de kwaliteit van het water dat zij drinken. De veelgeroemde Millenniumdoelstellingen geven in beide gevallen verkeerde prikkels.

Het gebruik van RCTs om vast te stellen wat wel of niet werkt bij ontwikkelingsprojecten, werd al snel heel populair onder wetenschappers en beleidsmakers. De Nederlandse overheid liep daarbij niet voorop: Jacques van der Gaag en ik probeerden in een vroeg stadium het ministerie van Buitenlandse Zaken te overtuigen van het belang van impactevaluatie bij ontwikkelingsprojecten, maar dat ging niet van een leien dakje. Inmiddels heeft de evaluatieafdeling van het ministerie (IOB) een groot aantal impactevaluaties uitgevoerd en speelt IOB in internationaal verband een voortrekkersrol. De VU en de UvA hebben een gemeenschappelijke leerstoel op dit gebied, die van Menno Pradhan. De Rekenkamer heeft alle ministeries aanbevolen om hun beleid met experimentele methoden te evalueren en verwees onlangs instemmend naar de Beninstudie.[16] Het gebruik van impactevaluaties om te bepalen welk beleid effectief is in het bestrijden van armoede, heeft ook in de Nederlandse pers veel aandacht gekregen.[17]

Sommige voortrekkers op het gebied van impactevaluatie hebben de naam randomista's gekregen vanwege hun fundamentalistische positie dat RCTs de gouden standaard in de wetenschap zijn, dat vragen die zich niet voor RCTs lenen, niet wetenschappelijk zijn en dat op de algemene geldigheid van een RCT-resultaat mag worden vertrouwd.[18] De aan het MIT verbonden economen Banerjee en Duflo vinden dat ontwikkelingseconomie impactevaluatie moet zijn en niets anders. Het vak moet zich volgens hen zeker niet met "grote vragen" bezighouden:[19]

> "*Instead of discussing how best to fight diarrhea or dengue, many of the most vocal experts tend to be fixated on the "big questions": what is the ultimate cause of poverty? How much faith should we place in free markets? Is democracy good for the poor? Does foreign aid have a role to play? and so on.*"

Hun critici hebben natuurlijk geantwoord dat je dergelijke belangrijke vragen niet uit de weg kunt gaan alleen maar omdat ze zich niet lenen voor je favoriete methode.[20] Hier komt methodologische zuiverheid op gespannen voet te staan met maatschappelijke relevantie.[21]

Dat is niet de enige reden om het overdreven enthousiasme voor "de nieuwe ontwikkelingseconomie" wat te temperen: op de methodologische zuiverheid van RCTs valt wel

degelijk wat aan te merken. Angus Deaton, hoogleraar aan Princeton, wil zelfs niets weten van de RCT als "gouden standaard": "*experiments have no special ability to produce more credible knowledge than other methods*".[22] Die merkwaardige uitspraak geeft wel aan dat de emoties ook onder econometristen hoog kunnen oplopen. Deaton maakt zich vooral zorgen over niet-representatieve steekproeven in situaties waarin het effect van een maatregel sterk kan verschillen tussen groepen.[23] Die kritiek geldt zeker voor ontwikkelingseconomen, die vaak een RCT uitvoeren op een plaats waar dat goed uitkomt, dus zonder speciale aandacht voor representativiteit. Er is echter een veel fundamentelere kritiek op RCTs mogelijk.

De glans van de gouden standaard: externe validiteit en diversiteit

Elke onderzoeker hoopt dat een experimenteel resultaat ook geldt buiten de oorspronkelijke context, dus dat het extern valide is. De hoop op die ruimere geldigheid, de mogelijkheid om te generaliseren van het bijzondere naar het algemene, heeft een betoverend effect op onderzoekers. Het is de Sirenenzang die hen aantrekt en motiveert, maar soms ook bedwelmt en misleidt.

Context kan slaan op wie er behandeld worden, dus op de populatie, maar ook op de manier waarop de behandeling wordt uitgevoerd. In het eerste geval is de vraag of wat in Indonesië is gevonden, ook geldt in Oeganda; of wat waar blijkt te zijn voor schoolmeisjes op het platteland van Kenia, geëxtrapoleerd mag worden naar schoolkinderen in Nairobi. In het tweede geval gaat het om de vraag of een resultaat verkregen in de door de onderzoekers gekozen experimentele opzet blijft staan als de behandeling "in het echt" wordt uitgevoerd.

We beginnen met externe validiteit wat betreft de populatie. De *treatment* en *control* groepen in een RCT vormen een steekproef, getrokken uit een bepaalde populatie, bijvoorbeeld alle kinderen van schoolgaande leeftijd in een bepaald gebied. Er is meestal geen reden om te veronderstellen dat het effect dat voor *die* populatie wordt gevonden, ook geldt voor een andere populatie, ook al suggereren onvoorzichtige onderzoekers dat wel eens.[24] Die onvoorzichtigheid leidt nogal eens tot een gemakzuchtige beperking van het empirische onderzoek tot een bepaald gebied. Zo werden in Kenia allerlei onderwijshervormingen met RCTs in één provincie, Western Kenya, onderzocht. De minister van onderwijs verzuchtte dan

ook dat hij nu heel veel wist over wat wel en niet werkt in dat deel van Kenia, maar dat hij verantwoordelijk was voor het onderwijsbeleid in *alle* provincies. Hij zette dus een vraagteken bij de externe validiteit van de RCTs in geografische zin.

Die minister had natuurlijk groot gelijk: externe validiteit kan geen geloofsartikel zijn, maar zal door nieuw onderzoek moeten worden aangetoond. Dat gebeurt steeds vaker. Zo is nu in heel veel landen de effectiviteit onderzocht van *conditional cash transfers*, het geven van geld aan arme ouders op voorwaarde dat zij hun kinderen naar school laten gaan. Die financiële prikkel blijkt in allerlei landen, in heel verschillende situaties verrassend goed te werken. Er is geen ontkomen aan: om dergelijke conclusies te kunnen bereiken moet onderzoek onder allerlei verschillende omstandigheden worden herhaald. Kennis accumuleert dan, "*one experiment at the time*", zoals Martin Ravallion dat treffend formuleerde.

De zaak ligt minder eenvoudig als externe validiteit niet slaat op de populatie, maar op de interventie of behandelwijze. Dat AIDS-remmers werken als zij op de juiste manier worden ingenomen, staat buiten kijf. Maar in Afrika is het beoogde effect vaak niet bereikt, meestal omdat wie de medicijnen kreeg, die deelde met anderen. Medici maken dan ook al heel lang onderscheid tussen de effectiviteit van een geneesmiddel of behandelingsmethode in het laboratorium of in een RCT (dat noemen zij *efficacy*) en in de praktijk (*effectiveness*). In de wereld van impactevaluaties verliezen economen dat onderscheid wel eens uit het oog. Het belang ervan wordt goed geïllustreerd door de eerder genoemde studie van drinkwatervoorziening in Benin: in dat geval was er niets mis met de *efficacy* van het programma, maar in termen van *effectiveness* faalde het volledig.

Iets dergelijks doet zich vaak voor als er wordt geprobeerd een veelbelovende aanpak op grote schaal toe te passen. Zo werd onlangs in Kenia onderzocht of de leereffecten op basisscholen verbeterden door het inzetten van extra onderwijzers op tijdelijke contracten. Een RCT waarbij een zeer competente NGO die inzet regelde, liet een overtuigend positief effect zien.[25] De Keniaanse regering besloot daarom deze maatregel in het hele land in te voeren. Onderzoekers vergeleken bij die opschaling twee varianten: de inzet van de onderwijzers werd in het ene geval door een internationale NGO, zoals in de oorspronkelijke RCT, in het andere door de

overheid georganiseerd. Dat maakte veel verschil: als de overheid die rol speelde, bleef er van het eerder gevonden effect niets over. Het resultaat dat geldt als een NGO als uitvoerder optreedt, heeft dus geen externe validiteit voor de Keniaanse overheid. *Scaling up* impliceert hier kennelijk een beslissende verandering van context.[26] Dit is een ernstige waarschuwing voor degenen die een positief RCT-resultaat voldoende basis vinden voor een grootschalige beleidsverandering.

Twijfel over externe validiteit is de meest bekende vorm van kritiek op RCTs. Maar er is meer. Zo faalt de RCT als de behandeling niet alleen de *treatment* groep, maar ook de *control* groep beïnvloedt. Dat mag natuurlijk niet in een laboratorium, maar is daarbuiten soms niet te vermijden. Zo kunnen bij een programma van de centrale overheid lagere overheden besluiten om dorpen in de *control* groep te compenseren voor het feit dat zij niet profiteren van het programma.[27] De RCT zal het effect van de behandeling dan onderschatten, zodat interne validiteit verloren gaat: het RCT-resultaat is dan zelfs voor de oorspronkelijke context niet juist.

De Achilleshiel van RCTs: wie profiteert er van de interventie?

We gaan nog één stap verder. Bij een medisch experiment ligt uiteraard vast wie wordt behandeld (of althans de behandeling aangeboden krijgt), namelijk alle leden van de *treatment group*. Maar in de ontwikkelingspraktijk zien we vaak een heel andere situatie: er is niet precies vastgelegd voor wie de interventie geldt, dus wie de “behandeling” zal ontvangen en in welke mate: wie daarover beslist, heeft binnen de grenzen van formele criteria vaak veel vrijheid. De projectverantwoordelijke zal zijn beslissing deels baseren op informatie die ook beschikbaar is voor de onderzoeker (bijvoorbeeld als de doelgroep is gedefinieerd door een objectief armoedecriterium), maar hij kan ook gebruikmaken van privé-informatie, bijvoorbeeld zijn eigen inschatting van de effectiviteit van een project voor een bepaalde locatie of een bepaalde persoon.[28]

Dit geval (in het jargon aangeduid als *selection on unobservables*) doet zich in de praktijk vaak voor.[29] Hulporganisaties zoals Hivos, Plan of Oxfam-Novib gebruiken algemene criteria om te bepalen waar en hoe hulp moet worden ingezet. Een partnerorganisatie in een ontwikkelingsland zal daaraan een nadere invulling geven. Maar ook dan blijven de richtlijnen

meestal tamelijk vaag: er wordt bijvoorbeeld gekozen voor dorpen in arme districten die geen toegang tot schoon drinkwater hebben. Binnen dat ruime kader wordt de concrete invulling bepaald door iemand die laag in de organisatie zit en de plaatselijke situatie goed kan beoordelen. Die persoon mag beslissen om het project in bepaalde dorpen uit te voeren, bijvoorbeeld omdat hij denkt dat daar de bereidheid om zich in te zetten voor betere leefomstandigheden groter is dan in andere dopen. De projectverantwoordelijke baseert die beslissing dan op een persoonlijke inschatting van verschillen tussen dorpen in de mate waarin het project effectief zal zijn.[30] Zoiets is kenmerkend voor maatschappelijke ontwikkelingsorganisaties, NGOs, die vaak geen strakke hiërarchie kennen en veel ruimte geven voor initiatief aan de basis.[31]

Een soortgelijke situatie doet zich voor als er tussen individuen of bedrijven moet worden gekozen, zoals bij kredietverlening. Ook dan zullen formele criteria die de aflossing en rentebetaling moeten veiligstellen, worden aangevuld door persoonlijke inschattingen van degene die over kredietverlening beslist.

In beide gevallen is er sprake van “essentiële heterogeniteit”: de effectiviteit van de behandeling verschilt tussen individuen of locaties en is bovendien gecorreleerd met de behandeling zelf.[32] Die correlatie impliceert een probleem dat econometristen endogeniteit noemen. De gevolgen daarvan zijn ernstig: de schattingen kunnen ook in grote steekproeven fout zijn. De onderzoeker zou daardoor kunnen concluderen dat de interventie werkt terwijl dat niet zo is, of omgekeerd: hij zou een effectieve methode kunnen verwerpen. Meestal gaat de onderzoeker endogeniteit te lijf door te proberen een “instrumentele variabele” te vinden. Maar dat is hier een doodlopende weg, want de twee eisen waaraan een instrumentele variabele moet voldoen, zijn in dit geval strijdig. Een zoektocht naar een dergelijke variabele heeft dus geen zin: wat gezocht wordt, bestaat niet.

Wie in deze situatie het effect probeert te schatten door een uitkomstvariabele (bijvoorbeeld schoolprestaties van individuele kinderen) te regresseren op de ontvangen “behandeling” en andere verklarende variabelen, stuit dus op een onoplosbaar endogeniteitsprobleem. RCTs bieden hier geen uitkomst. Wie het effect van een onderwijsprogramma wil schatten door de

schoolprestaties te vergelijken van leerlingen binnen en buiten het programma, trekt verkeerde conclusies, ook al zijn de schoolkinderen keurig gerandomiseerd over die groepen verdeeld. Ook voor RCTs is hier een fundamentele grens bereikt. Zij geven een correct antwoord op de vraag hoe effectief dit programma zou zijn *als alleen het toeval bepaalt wie behandeld wordt en in welke mate.* Maar dat is natuurlijk niet de vraag, want in werkelijkheid speelt het toeval slechts een bijrol; de hoofdrol is weggelegd voor de inschatting door de projectverantwoordelijke van de individuele reacties op de interventie.[33] In deze situatie kan de beleidsmaker niet afgaan op een RCT, want die weerspiegelt dat toewijzingsproces niet.

Zo blijft er van de glans van de "gouden standaard" weinig over: met een RCT krijgt de onderzoeker in dit geval een correct antwoord op een irrelevante vraag. Mijn indruk is dat er niet alleen in mijn vak, maar ook in de medische wetenschap onvoldoende aandacht is voor deze beperking van RCTs.[34]

We komen hiermee tot een deprimerende conclusie: in een veelvoorkomende situatie kan de effectiviteit van een interventie noch met traditionele regressiemethoden, noch met RCTs worden vastgesteld.

De oplossing: terug naar regressies met observationele data

Gelukkig is er wel een oplossing. De onderzoeker moet dan geen experimentele gegevens gebruiken, maar observationele data, waarnemingen uit de praktijk. Vroeger gebruikten economen dergelijke data even onbekommerd in regressies als tegenwoordig menig politicoloog, historicus of socioloog, maar zij zijn daarin vanwege hun grote nadruk op causaliteit zeer terughoudend geworden. Regressies met observationele data hebben een slechte naam gekregen, omdat de onderzoeker daarmee snel wegzakt in een moeras van endogeniteit. Het is echter tijd voor eerherstel, althans in deze situatie: hier falen RCTs, maar bieden juist deze data een oplossing. Mijn collega Chris Elbers en ik laten dat zien in een artikel dat binnenkort verschijnt.[35]

Ik schets u de essentie van dat verhaal. Allereerst moeten we de vraag scherp krijgen. Wat willen we weten als we een project evalueren? Op individueel niveau is het effect van de

interventie natuurlijk het product van de variabele die de interventie meet en de impactparameter (die tussen mensen kan verschillen en door de onderzoeker meestal niet kan worden waargenomen).[36] Bijvoorbeeld in een onderwijssituatie: het aantal lesuren dat een schoolkind krijgt, maal het (individuele) leereffect per uur. In een impactevaluatie zijn we op zoek naar de verwachte waarde van dat product in de relevante populatie. De onderzoeker is geneigd eerst de verwachte waarde van de impactparameter te schatten.[37] We hebben al gezien dat die weg doodloopt. Met conventionele regressiemethoden stuiten we op het genoemde, onoverkomelijke endogeniteitsprobleem. Met een RCT kan men die verwachte waarde wel vinden, maar alleen door op te leggen wie behandeld wordt (of althans de behandeling aangeboden krijgt). Dat maakt de uitkomst irrelevant: want wat in het experiment wordt opgelegd, wordt in werkelijkheid bepaald door de projectverantwoordelijke en die doet dat niet op basis van randomisatie. Kortom: het zoeken naar de verwachte waarde van de impactparameter is zinloos.

Het mooie is dat de tussenstap die niet mogelijk blijkt te zijn, kan worden overgeslagen. De verwachte waarde van het product (wij noemen dat het *total program effect*) kan namelijk rechtstreeks worden geschat met observationele data door in de regressievergelijking een mogelijk verband tussen de impactparameter en de behandeling te incorporeren.[38] Het gevolg daarvan is dat we als verklarende variabelen voor de uitkomst niet alleen de behandelingsvariabele en de *controls* opnemen, maar ook hun producten (de interactietermen), het kwadraat van de behandelingsvariabele en eventueel ook nog termen van hogere orde.[39] Dat de oplossing voor een belangrijk probleem zo eenvoudig kan zijn, verbaast ons nog steeds.

De vakgenoten zullen tegenwerpen dat een impactevaluatie die deze methode hanteert, endogeniteit accepteert in plaats van ervoor te corrigeren. Zij hebben volkomen gelijk, maar wat een nadeel van deze methode lijkt, is juist een voordeel: als we willen weten of een programma effectief is, moeten we juist *niet* voor deze vorm van endogeniteit corrigeren. Immers, als het programma zo is georganiseerd dat de middelen vooral daar worden ingezet waar het verwachte rendement volgens ingewijden hoog is, dan is het juist door die ongelijke

inzet effectief.[40] We moeten dat effect niet elimineren, maar opnemen in onze evaluatie. De neiging tot correctie is natuurlijk begrijpelijk: we weten dat door endogeniteit de zuiverheid en consistentie van de schatters verloren gaat. Onze methode heeft daarvan geen last: in de schattingsvergelijking is er geen endogeniteit.

Hoe ernstig is het probleem van diversiteit in de effectiviteit van de behandeling eigenlijk? In het artikel laten wij met een voorbeeld, een impactevaluatie van een ziektekostenverzekering in Vietnam, zien dat het al of niet rekening houden met heterogeniteit een enorm verschil kan maken. Het effect van die verzekering op de gezondheid (gemeten als de verandering in lichaamsgewicht) bleek twee keer zo groot als wat men zou vinden met de traditionele methode. Met andere woorden: wie homogeniteit oplegt, vindt een veel te lage schatting van het effect van de verzekering. Dit voorbeeld is misschien uitzonderlijk, maar in elke evaluatie met observationele data kan statistisch worden getoetst of verscheidenheid een verwaarloosbaar probleem is. Het zou verstandig zijn om die toets steeds uit te voeren.

Het is tijd voor een samenvatting. Dat het in een RCT gemeten effect kan worden toegeschreven aan de interventie, de interne validiteit, wordt zelden betwist.[41] De externe validiteit van RCTs is echter door velen in twijfel getrokken, ondanks de grandioze pretenties van de randomista's. Die kritiek is gemakkelijk op te vangen als het gaat om de vraag of het gevonden resultaat ook geldt voor andere populaties: dan moeten er ook voor die populaties RCTs worden gedaan. Maar als het gaat om de vraag of wat werkt onder toezicht van een stel bevlogen wetenschappers of van een zeer competente NGO, ook werkt op grotere schaal (bijvoorbeeld bij de hele overheid), dan kan de kritiek niet makkelijk worden gepareerd. Een positief RCT-resultaat is dan een uiterst nuttige aanwijzing, een *proof of principle*, maar zonder verdere informatie geen betrouwbare basis voor beleid.

De laatste vorm van kritiek is dat de selectie van de treatment group in een RCT vaak niet overeenkomt met de wijze waarop in de praktijk beslissingen worden genomen. Die kritiek raakt de RCT-methodologie in het hart, want in dat geval is het experimentele resultaat zelfs voor *dezelfde* populatie niet bruikbaar.[42] Het is verrassend dat er een oplossing is voor dit probleem. Die oplossing is niet zonder ironie. Om zeker te zijn van causaliteit omarmden

economen enthousiast de RCT-methodologie. Diezelfde overweging leidt nu tot een oplossing waarbij zij experimentele gegevens weer moeten inruilen voor observationele data.

Nogmaals: de nieuwe ontwikkelingseconomie

De suggestie van Banerjee en Duflo dat met impactevaluatie een nieuwe, betere ontwikkelingseconomie is ontstaan, is gretig overgenomen in de pers, ook in Nederland. Ik beschouw die visie op het vak als een verarming. Economen zien zichzelf graag in het voetspoor van Keynes als "*the trustees not of civilization, but of the possibility of civilization*".[43] Dat hun vak dat mogelijk maakt, is wat economen werkelijk bezielt. Dat geldt zeker voor ontwikkelingseconomen: voor hen gaat het bij beleid niet om een kleine verandering in een inkomensplaatje, maar om de mogelijkheid van een menswaardig bestaan.

Economen vervullen die rol door beleidsmakers, die Keynes met zijn scherpe tong "*madmen in authority who hear voices in the air*" noemde, te bestoken met ongevraagd advies en door materiaal aan te dragen voor wie deelneemt aan het debat, zo u wilt: het discours, over economisch beleid. Welnu, in het publieke debat is het niet voldoende om aan te tonen wat wel of niet werkt. Er moet ook aandacht zijn voor kosten en voor prikkels.

Een halve eeuw geleden ontwikkelden enkele welvaartseconomen de kosten-baten analyse, een vorm van *ex ante* impactevaluatie die vandaag in de wetenschap haast vergeten lijkt.[44] Bij een RCT wordt zelden naar de kostenkant gekeken, alsof een beleidsmaker alleen zou willen weten of iets werkt, maar niet wat hij daarvoor moet opgeven. Als kosten wel worden geschat, dan gebeurt dat meestal alleen op basis van marktprijzen. Daarmee raakt een kernbegrip van de economie, *opportunity costs*, buiten beeld. Wie schaarse middelen inzet, maakt een ander gebruik van die middelen onmogelijk. Dat zijn de echte kosten, maar die verschijnen zelden op het prijskaartje. De ontwerpers van de kosten-baten analyse beseften heel goed dat in een markteconomie, zeker in ontwikkelingslanden, prijzen vaak geen betrouwbare signalen geven omdat allerlei markten gebrekkig functioneren of zelfs ontbreken. Ruerd Ruben, de directeur van de evaluatieafdeling van het ministerie van Buitenlandse Zaken, bepleit terecht dat evaluaties niet alleen baten, maar ook kosten, gemeten met schaduwprijzen, beoordelen.

Economen bestuderen hoe individuen en instituties reageren, al of niet rationeel, op prikkels. Dat moeten zij vooral blijven doen, want daar ligt hun kracht. Het lijkt alsof experimentele methoden, gericht op de vraag *of*, niet *waarom* iets werkt, daarbij weinig te bieden hebben, maar die schijn bedriegt. De laatste jaren worden experimenten vaak zo opgezet dat zij gebruikt kunnen worden om theorieën over reacties op economische prikkels te toetsen. Een mooi voorbeeld is het veldexperiment dat Berber Kramer en Wendy Janssens in Tanzania uitvoerden. Zij onderzochten daarmee of het bestaan van een informeel verzekeringsstelsel individuen prikkels geeft die het opzetten van een formele verzekering onmogelijk kunnen maken.[45] Zij gebruikten hun speltheoretische analyse van prikkels om hypothesen af te leiden die zij in het veldexperiment konden toetsen. Dit werk laat zien dat de afstand tussen theorie en experiment niet zo groot is als soms wordt gesuggereerd.

Besluit en dankwoord

Hoe zullen onze opvolgers later terugkijken op deze periode waarin voor experimenten het alleenrecht voor het vaststellen van causaliteit werd opgeëist? Met verbazing over de *hubris*, de overmoed, van de voortrekkers, vermoed ik, maar ook met respect voor een methode die grote vooruitgang bracht en leidde tot een constructieve bezinning op kernvragen.

Dit afscheidscollege is een prachtige gelegenheid voor een openbaar dankwoord. Ik maak daarvan graag gebruik, want mijn dankbaarheid is groot. Die geldt allereerst het bestuur van de Stichting VU-VUmc, het College van Bestuur en het bestuur van onze faculteit. De VU is onherkenbaar veranderd sinds 1982, toen ik hier aantrad. Zij is een grote, sterke, naar buiten gekeerde universiteit geworden, met een economische faculteit waar voortreffelijk onderzoek wordt gedaan. De VU heeft mij niet alleen veel ruimte gegeven om te doen wat ik belangrijk vond - dat is gelukkig de regel in de wetenschap - maar steunde mij ook actief, ook al zal men mij lang hebben gezien als een vreemde eend in de VU-bijt.

In het faculteitsbestuur werkte ik jarenlang samen met Jan Klaassen, Janny Westra, Peter Sneep, Frans Snijders en Harmen Verbruggen. Ik bewaar daaraan heel goede herinneringen. Dat de faculteit er goed voor staat is in hoge mate te danken aan hun bekwaamheid, inzet en betrokkenheid. De bestuursstijl die zij hebben nagelaten, is doordesemd van die betrokkenheid.

Hoewel dit een heel grote faculteit is, voelt wie hier werkt zich lid van een gemeenschap en wordt er zelden geklaagd over een managerscultuur of bestuurlijke vervreemding. Dat is een erfenis die we moeten koesteren.

Harmen was bijna tien jaar decaan. Hij heeft in die lange periode heel veel gedaan voor de faculteit en voor de universiteit, niet alleen door wat hij tot stand bracht, maar ook door wat hij wist te voorkomen. Ik heb grote bewondering voor zijn standvastigheid in moeilijke omstandigheden en voor het morele fundament van zijn leiderschap. Dat hij mij jarenlang als informeel adviseur gebruikte, heb ik zeer gewaardeerd. Maar bovenal: ik ben erg op hem gesteld.

Dat een eerdere decaan, Nol Merkies, aan wie ik veel te danken heb, eens tegen mij zei dat ik alles mocht doen, als ik maar niet een samenwerking met de UvA zou beginnen, past nu in de categorie “omzien in verwondering”. Jaren later, toen Jacques van der Gaag decaan aan de UvA was, begonnen wij met een intensieve VU-UvA samenwerking. Wat eerst ondenkbaar leek: dat UvA masters studenten in de tram zouden stappen om onderwijs aan de VU te volgen en omgekeerd, was al snel werkelijkheid. Bij het onderzoek leidde onze samenwerking tot de oprichting van het *Amsterdam Institute for International Development*, in de wandeling het *AIID*. Dat zou niet gebeurd zijn zonder Jacques’ vele initiatieven, zijn energie, enthousiasme en vindingrijkheid. Ik ontleen veel werkvreugde aan onze samenwerking en ben dan ook blij dat die gewoon doorgaat.

Niek Urbanus, Taede Sminia en Jaap Zwemmer stonden al aan de wieg van het *AIID*. Zij gaven Jacques en mij veel ruimte, ondersteunden ons met grote hartelijkheid, maar waren ook heel duidelijk in hun vermaningen. Niek, onze voorzitter, is een buitengewoon bekwaam en wijs, maar ook een heel vrolijke bestuurder. Dat hij in zijn AMC-tijd mij eens een hoogleraarschap in de interne geneeskunde beloofde, is tekenend voor onze relatie.[46] Ik ben hem zeer erkentelijk, ook al ging die benoeming niet door.

Michiel Keyzer speelt al sinds onze studententijd een grote rol in mijn leven: als vriend, als de erudiete geleerde die mij de weg naar Spinoza wees en als inspirerende collega. Dat alles, maar vooral die vriendschap, was en blijft belangrijk voor mij.

Met Chris Elbers, de eerste van mijn promovendi, heb ik vooral nadat ik terugkwam uit Oxford heel veel samengewerkt, hier op de VU en op allerlei plaatsen in Afrika. Chris is een onconventioneel onderzoeker, die liever ongehinderd door de vakliteratuur zijn eigen weg zoekt dan gebaande paden te volgen. Dat maakt hem een voortreffelijk wetenschapper. Hij is ook een heel sympathieke collega. Onze vrijwel dagelijkse discussies, vaak staande bij het bord waar we proberen een vraag duidelijk te krijgen, een probleem te formaliseren, vormen mijn beste universitaire herinneringen.

Onze groep ontwikkelingseconomie, waarvan de kern wordt gevormd door Chris, Menno Pradhan, Remco Oostendorp, Wendy Janssens en Trudi Heemskerk, is een hechte club die een grote aantrekkingskracht uitoefent op heel bekwame jonge onderzoekers die graag methoden aan de grens van de wetenschap willen toepassen op vraagstukken van wezenlijk belang. Wat deze onderzoekers inspireert, misschien zelfs bezielt, is dat er uiteindelijk in ons werk veel op het spel staat: een goede economische analyse kan in ontwikkelingslanden vele mensenlevens aangenaam maken of zelfs redden. Het werken met deze collega's is een groot genoegen en meer dan dat: een voorrecht.

Ten slotte: velen van u weten of vermoeden wat Louise, onze kinderen en kleinkinderen voor mij betekenen. Daarover zeg ik - onze traditie getrouw - in een openbaar college niets. Maar wie zich afvraagt wat *deze* econoom bezielt, kan het antwoord raden.

Literatuur

Acemoglu, D., S. Johnson en J.A. Robinson (2001), `The Colonial Origins of Comparative Development: an Empirical Investigation', *American Economic Review*, vol. 91, pp. 1369-1401.

Adam, C.S. en S.A. O'Connell (1999), `Aid, Taxation and Development in Sub-Saharan Africa', *Economics & Politics*, vol. 11, pp. 225-253.

Algemene Rekenkamer (2012), *Effectiviteitsonderzoek bij de Rijksoverheid*, rapport.

Angrist, J.D. en V. Lavy (1999), `Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement', *Quarterly Journal of Economics*, vol. 114, pp. 533-575.

Banerjee, A.V. en E. Duflo (2009), `The Experimental Approach to Development Economics', *Annual Review of Economics*, vol. 1, pp. 151-178.

Banerjee, A.V. en E. Duflo (2011), *Poor Economics: a Radical Rethinking of the Way to Fight Global Poverty*, New York: Public Affairs.

Banerjee, A. en R. He (2008): `Making Aid Work', in: W. Easterly (red.), *Reinventing Foreign Aid*, Cambridge, Mass.: MIT Press.

Bevan, D., P Collier en J.W. Gunning (1990), *Controlled Open Economies: a Neoclassical Approach to Structuralism*, Oxford: Oxford University Press (Clarendon Press).

Bold, T., M. Kimeny, G. Mwabu, A. Ng'ang'a en J. Sandefur (2013), `Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education', Working Paper 321, Center for Global Development, Washington, DC.

Chen, S. en M. Ravallion (2012), `More Relatively-Poor People in a Less Absolutely-Poor World', World Bank Staff Working Paper 6114.

Clevers, H. (2013), `Wetenschap op gevoel', jaarrede van de president van de Koninklijke Nederlandse Akademie van Wetenschappen, uitgesproken op 27 mei 2013;

htpps://www.knaw.nl/shared/resources/actueel/publicaties/pdf/130527_KNAW_jaarrede_Hans_Clevers.pdf

Collier, P. (2007), *The Bottom Billion. Why the Poorest Countries Are Falling Behind and What Can Be Done About It*, Oxford: Oxford University Press.

Collier, P. en J.W. Gunning (1999), 'Why Has Africa Grown Slowly?', *Journal of Economic Perspectives*, vol. 13, 1999, pp. 3-22.

Deaton, A. (2010), `Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, vol. 28, pp. 424-455.

Duflo, E., P. Duplas en M. Kremer (2012), `School Governance, Teacher Incentives and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools', NBER Working Paper 17939.

Duflo, E., R. Glennerster en M. Kremer (2008), `Using Randomization in Development Economics Research: a Toolkit', in T.P. Schultz en J. Strauss (red.), *Handbook of Development Economics,* Amsterdam: North-Holland, pp. 3895-3962.

Elbers, C. en J.W. Gunning (2014), `Evaluation of Development Programs: Randomized Controlled Trials or Regressions?', verschijnt in de *World Bank Economic Review*.

Elbers, C. en J.W. Gunning (2014a), `What Do Development NGOs Achieve?', verschijnt in J.Y. Lin and C. Monga (red.), *The Oxford Handbook of Africa and Economics*, Oxford: Oxford University Press.

Gunning, J.W. (2012), `Evaluating Development NGOs', Policy Brief nr. 56, Clermont-Ferrand: FERDI.

Günther, I. en Y. Schipper (2013), `Pumps, Germs and Storage: the Impact of Improved Water Containers on Water Quality and Health', *Health Economics*, vol. 22, pp. 757-774.

Harrod, R. (1951), *The Life of John Maynard Keynes*, New York: Harcourt.

Heckman, J.J., S. Urzua en E. Vytlacil (2006), `Understanding Instrumental Variables in Models with Essential Heterogeneity', *Review of Economics and Statistics*, vol. 88, pp. 389-432.

Hoop, J.J. de (2011), *Keeping Kids in School: Cash Transfers and Selective Education in Malawi*, Tinbergen Institute PhD Theses Series nr. 504.

Hoover, K.D. (2008), `Causality in Economics and Econometrics', *The New Palgrave Dictionary of Economics*, Macmillan.

Janssens, W. en B. Kramer (2013), `The Social Dilemma of Microinsurance: a Framed Field Experiment with Microcredit Groups in Tanzania', ongepubliceerd, FEWEB, Vrije Universiteit.

MacLeod, W.B. (2013), `On Economics: a Review of *Why Nations Fail* by D. Acemoglu and J. Robinson and *Pillars of Prosperity* by T. Besley and T. Persson', *Journal of Economic Literature*, vol. 51, pp. 116-143.

Miguel, E. en M. Kremer (2004), `Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica*, vol. 72, pp. 159-217.

Pritchett, L. en J. Sandefur (2013), `Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix', Working Paper 336, Center for Global Development, Washington, DC.

Ravallion, M. (2012), `Fighting Poverty One Experiment at a Time: a Review of Abhijit Banerjee and Esther Duflo's *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*', *Journal of Economic Literature*, vol. 50, pp. 103-114.

Ridder, G. (1995), *Oorzaak en gevolg*, oratie Vrije Universiteit.

Rodrik, D. (2008), `The New Development Economics: We Shall Experiment But How Shall We Learn?', John F. Kennedy School of Government, Harvard University, HKS Working Paper RWP 08-055.

Rodrik, D., A. Subramanian en F. Trebbi (2004), `Institutions Rule: the Primacy of Institutions over Geography and Integration in Economic Development', *Journal of Economic Growth*, vol. 9, pp. 131-165.

Spolare, E. en R. Wacziarg (2013), `How Deep Are the Roots of Economic Development?', *Journal of Economic Literature*, vol. 51, pp. 325-369.

Thaler, R.H. en C.R. Sunstein (2008), *Nudge: Improving Decisions about Health, Wealth and Happiness*, Yale University Press.

Voors, M., T. Turley, A. Kontoleon, E. Bulte en J.A. List (2012), `Exploring whether Behavior in Context-free Experiments is Predictive of Behavior in the Field: Evidence from Lab and Field Experiments in Rural Sierra Leone', *Economics Letters,* vol. 114, pp. 308-311.

Wennberg, D.E., F.L. Lucas, J.D. Birkmeyer, C.E. Bredenberg en E.S. Fisher (1998), `Variation in Carotid Endarterectomy Mortality in the Medicare Population: Trial Hospitals, Volume, and Patient Characteristics', *Journal of the American Medical Association*, vol. 279, pp. 1278-81.

[1] Clevers (2013).
[2] Jeffrey Sachs heeft vaak een tamelijk extreem standpunt van geografisch determinisme verdedigd. Rodrik *et al.* (2004) laten zien dat relatief meer gewicht moet worden toegekend aan de institutionele verklaring. Zij onderscheiden handelsbeleid als een derde categorie, naast geografie en instituties.
[3] De beschikbaarheid van delfstoffen wordt deels bepaald door beleid. Veel Afrikaanse regeringen hebben decennialang olie-exploratie onaantrekkelijk gemaakt. Sinds het beleid op dat punt is veranderd, zijn er grote olievoorraden gevonden.
[4] Collier (2007).
[5] Zo bleek ook de stelling, die in Nederland lang populair was, dat armoede vooral te maken heeft met grondstoffenprijzen, niet houdbaar: landen die met dezelfde prijzen werden geconfronteerd in wereldmarkten, vertoonden grote verschillen in uitkomsten: Collier en Gunning, 1999; MacLeod, 2013.
[6] Deze theorie lijkt moeilijk te toetsen, maar inmiddels gebeurt dat wel: Spolare en Wacziarg (2013).
[7] Acemoglu *et al.* (2001).
[8] Chen en Ravallion (2012) schatten het aantal armen (met minder dan $ 1,25 per dag in 2005 *purchasing power parity* prijzen) op ongeveer 1,9 miljard in 1981 en 1,3 miljard in 2008.
[9] Bevan *et al.* (1990) beschrijven hoe hervormingen in zwaar gereguleerde economieën (zoals Tanzania vanaf 1974) averechts kunnen werken. Ervaringen in Afrika in de jaren tachtig met hervormingen ontworpen op basis van de theorie van het "*second best*" werden helaas genegeerd in Oost-Europa na de val van de muur. Daar werd vaak elke liberalisatie als een stap in de juiste richting gezien, met voorspelbare rampzalige gevolgen, vooral bij de privatisering van staatsbedrijven.
[10] In het model van Adam en O'Connell (1999) heft de overheid belastingen om met de opbrengsten de eigen achterban (politieke partij, stam of regionale groep) te kunnen bevoordelen. Als die achterban relatief klein is, worden de economische kosten van die belastingheffing vooral door anderen gedragen. Bij een breed gedragen regime, dus als de achterban relatief groot is, moet die groep een groot deel van de kosten zelf dragen. Als een kritieke grens wordt overschreden, heeft het regime daardoor geen prikkel meer om belasting te heffen voor overdrachten aan de eigen achterban. Zo ontstaat een *developmental state*: de overheid stimuleert economische groei, niet uit overtuiging, maar uit eigenbelang.
[11] Over het begrip *nudge* in deze context: Thaler en Sunstein (2008).
[12] Hoover (2008) wijst erop dat na het beroemde werk van de Cowles Commission kort na de Tweede Wereldoorlog de aandacht van economen voor causaliteit verdween: "*causal language in economics virtually collapsed between 1950 and about 1990*".
[13] In de economie zijn RCTs zelden "*double blind*" en dat is in sommige situaties ook niet wenselijk is; zie Voors *et al.* (2012).
[14] Ridder (1995) geeft een mooi overzicht van deze vroege literatuur.
[15] Miguel en Kremer (2004).
[16] Algemene Rekenkamer (2012), `Rekenkamer: Effecten van ontwikkelingshulp zijn vaak onduidelijk', *Het Financieele Dagblad*, 14 november 2013.
[17] Zie bijvoorbeeld: Hanneke Chin-a-fo, `Fles chloor bij elke dorpspomp', *NRC Handelsblad*, 14 december 2011 en Tonie Mudde, `Experiment armoede', *De Volkskrant*, 23 november 2013, pp. V 2-3.
[18] Het standpunt over externe validiteit van Banerjee en He (2008) is extreem, maar dat van Banerjee en Duflo (2011) heel redelijk.
[19] Banerjee en Duflo (2011), p. 3.
[20] Ravallion (2009, p. 108) formuleerde die kritiek door retorisch te vragen: "*Small is beautiful, but is it big enough?*".

[21] Zie hierover Banerjee en Duflo (2009), Ravallion (2009), Deaton (2010) en vooral Rodrik (2008).
[22] Deaton (2010), p. 424.
[23] Deaton verwijst naar het voorbeeld van Wennberg *et al.* (1998): de ziekenhuizen die deelnamen aan een *trial*, waren verre van representatief, zodat in de praktijk het sterftecijfer veel hoger bleek te liggen dan in de RCT.
[24] Banerjee en He (2008) betoogden dat een beroemde studie naar het effect van klassengrootte op leerprestaties in Israëlische scholen (Angrist en Lavy, 1999) geldig zou zijn in de hele wereld. Pritchett en Darfur (2013) maken die stelling terecht belachelijk.
[25] Duflo *et al.* (2012).
[26] Bold *et al.* (2013).
[27] Ravallion (2009), Deaton (2009).
[28] Als de beslissing uitsluitend afhangt van informatie waarover ook de onderzoeker beschikt, dan kan het proces in principe worden gemodelleerd. In dat geval is de verbazing van Ravallion (2009, p. 104) over de obsessie van economen met endogeniteit dan ook terecht: naarmate meer data beschikbaar komen, wordt het probleem minder ernstig. Als deelname echter afhangt van privé-informatie, het geval dat Elbers en Gunning (2014) behandelen, dan kan het probleem uiteraard niet worden opgelost met meer data.
[29] Onderzoekers die zich met impactevaluatie bezighouden, hebben sinds kort oog voor de veranderingen in menselijk gedrag waartoe interventies kunnen leiden en ook voor verscheidenheid: voor verschillen tussen individuen of gemeenschappen in het effect van interventies. Die twee aandachtspunten vallen hier, waar het nu juist gaat om de reactie van de projectverantwoordelijke op de diversiteit in de populatie, samen.
[30] De projectverantwoordelijke hoeft de locatiespecifieke effectiviteit uiteraard niet precies te kennen. We veronderstellen alleen dat de behandeling (opgevat als een variabele die niet noodzakelijk binair is) gecorreleerd is met de effectiviteit; die correlatie hoeft zelfs niet positief te zijn.
[31] Gunning (2012), Elbers en Gunning (2014a).
[32] Zie Heckman *et al.* (2006) over essentiële heterogeniteit. Een duidelijkere terminologie is *selection on the gain*: verschillen tussen individuen in het effect van de interventie (de *gain*) bepalen wie aangewezen wordt (of zichzelf aanmeldt) als deelnemer aan het programma. In het tweede geval, dus als de correlatie het gevolg is van zelfselectie, dan is het resultaat van de RCT (mits de behandeling niet wordt opgelegd, maar aangeboden aan de leden van de *treatment* groep) wel een consistente schatter van het (*intention to treat*) effect.
[33] Men kan tegenwerpen dat RCTs op deze manier standrechterlijk worden veroordeeld, namelijk door te veronderstellen dat randomisering plaats zou vinden op het niveau van de interventie, in het voorbeeld dus op dorpsniveau. Een RCT-onderzoeker die het probleem ziet, zou natuurlijk op een hoger niveau kunnen randomiseren: over projectverantwoordelijken in plaats van over dorpen. Dat leidt echter niet alleen tot verlies van statistische kracht, maar ook tot een ernstiger probleem. De standplaats van projectverantwoordelijken is immers niet willekeurig gekozen, maar op basis van criteria zoals hun lidmaatschap van een etnische groep, kennis van een taal, ervaring met bepaalde landbouwtechnieken of kennis van een nieuwe onderwijsmethode. Daardoor zullen individuele kenmerken van de projectverantwoordelijken die hun beslissingen beïnvloeden, gecorreleerd zijn met de *controls*. Dit kan leiden tot systematische verschillen tussen de *treatment* en *control* groepen zodat zelfs interne validiteit verloren gaat. Zie hierover Elbers en Gunning, 2014, sectie IV.
[34] Daarbij moet worden aangetekend dat in de medische praktijk richtlijnen de ruimte voor individuele oordelen inperken.
[35] Elbers en Gunning (2014).
[36] Deze formulering impliceert lineariteit, maar daartoe hoeven we ons niet te beperken.

[37] Uiteraard heeft dat alleen zin als de verwachte waarde van het product van de twee termen gelijk is aan het product van hun verwachte waarden, dus als de impactparameter en de interventievariabele onafhankelijk zijn, maar in dit geval zijn ze juist gecorreleerd.

[38] In het artikel houden we rekening met de mogelijkheid dat de projectverantwoordelijke *selection on the gain* toepast en daarbij ook privé-informatie gebruikt: we veronderstellen dat de variabele die de behandeling meet, niet alleen bepaald wordt door de *controls*, de variabelen die kunnen worden waargenomen, maar ook door de impactparameter, die nu ook een variabele is, maar niet kan worden waargenomen. De impactparameter kan dan worden geschreven als een functie van de behandelingsvariabele en de *controls*. Chris Elbers had de briljante inval dat als we in de oorspronkelijke vergelijking voor de uitkomstvariabele de impactparameter vervangen door een polynomische benadering van die functie, de endogeniteit wordt geëlimineerd. De vergelijking kan dan ook met gewone kleinste kwadraten worden geschat. De som van de steekproefgemiddelden van de termen waarin de behandelingsvariabele voorkomt (d.w.z. die variabele zelf, haar kwadraat, de interactietermen met de *controls* en eventueel termen van hogere orde), elk gewogen met de bijbehorende regressiecoëfficiënt, is een consistente schatter van het *total program effect*.
Deze methode kan ook worden toegepast als de behandeling bestaat uit een combinatie van verschillende interventies, zoals vaak het geval is in ontwikkelingsprojecten. In de analyse is de behandelingsvariabele dan een vector.

[39] Als de behandelingsvariabele alleen de waarden 0 en 1 kan aannemen, wordt de kwadratische term natuurlijk niet meegenomen, maar wij beperken ons niet tot dat binaire geval.

[40] Die deskundigen kunnen het mis hebben. Dan is de interventie door hun gebrekkig oordeel minder effectief dan mogelijk zou zijn. Die lagere effectiviteit moet uiteraard blijken uit de evaluatie: een "correctie" voor selectiviteit zou misleidend zijn.

[41] Dat zou soms wel moeten gebeuren, bijvoorbeeld in het eerder genoemde geval van *spillovers* van de *treatment* naar de *control* groep.

[42] Of interne validiteit daarmee is verworpen, is een semantische kwestie. Wie in dit geval het experimentele resultaat intern valide wil noemen, moet aanvaarden dat het resultaat van de RCT dan zelfs voor *dezelfde* populatie in de praktijk irrelevant is. Wie ook dan nog interne validiteit als een groot goed beschouwt, kan met recht een fundamentalist worden genoemd.

[43] Keynes' beroemde toast op economen bij een diner van de *Royal Economic Society*; Harrod (1951), pp. 193-194.

[44] De grondleggers van die analyse: Ian Little, James Mirrlees, Partha Dasgupta en Amartya Sen (van wie Mirrlees en Sen later de Nobelprijs wonnen) speelden een belangrijke rol in de ontwikkelingseconomie en zagen in dat vak de belangrijkste toepassingsmogelijkheden voor de kosten-baten analyse.

[45] Janssens en Kramer (2013).

[46] Die benoeming zou niet eens zo heel vreemd zijn geweest vergeleken bij mijn aanstelling in 1973 aan de Rijksuniversiteit te Groningen voor "het onderwijs en het wetenschappelijk onderzoek in de experimentele natuurkunde" .