University of Groningen

Werkzame principes

Rol, Menno

Published in:
Algemeen Nederlands tijdschrift voor wijsbegeerte

DOI:
[10.5117/ANTW2017.1.MROL](https://doi.org/10.5117/ANTW2017.1.MROL)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Rol, M. (2017). Werkzame principes: Hoe te bouwen aan een collectief geheugen bij evaluatieonderzoek. *Algemeen Nederlands tijdschrift voor wijsbegeerte*, 109(1), 123-143.
<https://doi.org/10.5117/ANTW2017.1.MROL>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Werkzame principes

Hoe te bouwen aan een collectief geheugen bij evaluatieonderzoek*

Menno Rol

ANTW 109 (1): 123–143

DOI: 10.5117/ANTW2017.1.MROL

Abstract

Working principles: Building scientific memories in social research

Internal and external validity of experimental and quasi-experimental research are both essential in the evaluation of interventions; this is true in such diverse practices as social policy and pharmacology. However, they may trade-off if the quest for internal validity brings about the treatment of contextual variables – that helped to make or fail the intervention – as competing with the intervention instead of necessary for its success. Investment in internal validity takes the form of many types of correction for bias, often obscuring the role of contextual factors.

As helping factors make or break the intervention, they should rather be treated as causal factors of interest no less than the intervention. I claim that, only then, the conclusions of the (quasi)experiment can serve to gain knowledge of working principles in a variety of contexts.

Correction for bias is useful, but only if the evaluation of the success of an intervention is theory driven. Ex ante explanatory hypotheses must be the starting point of evaluation research and testing them the object of it. Thus, the empirical cycle is covered entirely. If no theoretical hypotheses guide the questions of evaluation research, all we will learn is that ‘most interventions sometimes work’, to speak with Pawson and Tilley (1997). What we are left with is little more than meaningless intervention-impact couples that cover half the empirical cycle lacking any cognitive value.

* Ik dank twee anonieme referees voor hun nuttige aanwijzingen. Die hebben de tekst beter gemaakt. Ik dank Lieuwe Zijlstra voor het plaatsen van zeer kritische opmerkingen. Al heb ik die niet alle verwerkt, ze hebben mij wel opnieuw grondig aan het denken gezet.

Keywords: internal validity, external validity, impact evaluation, explanation, social mechanism

1 Inleiding: De uitruil van interne en externe validiteit

De beleidsevaluatie is een productieve tak van sociale wetenschap, met name van de bestuurskunde en de sociologie. Overheden willen ingrijpen in de sociale werkelijkheid en er is sociale kennis vereist om dit succesvol te doen. Evaluatie van een beleidsinterventie is daarbij ook het produceren van sociale kennis: dit werkte (onder deze omstandigheden) wel en dat niet. Idealiter zou de beleidsevaluatie dus clous moeten opleveren over de geldigheid of de waarheid van sociaalwetenschappelijke claims. Dat valt in praktijk tegen.

Wie in de sociale wetenschap iets wil leren van succesvolle beleidsinterventies moet zorgen dat het gemeten effect dat met de interventie gepaard gaat ook werkelijk aan de interventie en alleen daaraan toe te schrijven is. Het onderzoek naar zulke interventies stelt dus de vraag naar effectiviteit, niet per se naar causaliteit. Maar als het lukt om het effect van de interventie te isoleren, is er dan wel genoeg geleerd om een dergelijke interventie in een andere context weer te vertrouwen?

Onenigheid over het antwoord op deze vraag leidt tot debat over de vermeende uitruil van interne en externe validiteit van sociaalwetenschappelijk onderzoek.¹ Dat gaat meestal over experimenten in het veld of in een laboratoriumomgeving, maar de standpunten in dit debat zijn even relevant voor de effectmeting van een sociaalwetenschappelijke interventie. Jiménez-Buedo en Miller (2010) leveren een goed voorbeeld van zulke standpunten. Het debat wordt mijns inziens niet helder gevoerd, verdient heroverweging en een aantal filosofisch scherpe onderscheidingen die spraakverwarring kunnen wegnemen. Mijn doel is hier om dergelijke onderscheidingen voor te stellen. Met name maak ik het belangrijke onderscheid tussen externe validiteit van een interventiestudie in de zin van de kwaliteit van de steekproef en die in de zin van bruikbaarheid in nieuwe contexten. Veel literatuur maakt tussen deze twee interpretaties geen verschil. Ook is belangrijk verklarende hypothesen achteraf – die slechts vertellen waarom een interventie niet werkte – te onderscheiden van theoretische hypothesen vooraf die ter toetsing in de evaluatie ingebracht wor-

¹ Zie bijvoorbeeld Schram, A. (2005).

den. Ten slotte wil ik helder maken welke rol de ontologie zou moeten spelen in dergelijk onderzoek.

Het meeste dat hier gezegd wordt geldt evenzeer voor andere wetenschappelijke praktijken waar impactmeting de voornaamste machinerie van onderzoek is als voor de beleidsevaluatie. Impactmeting is behalve bij de sociologie ook aan de orde bij biomedisch onderzoek en in de cognitieve psychologie. Zo is bijvoorbeeld de discussie over de cognitieve waarde van *Randomised Controlled Trials* (RCT's) even relevant voor de farmacologie als voor de bestuurskunde.² Buiten echte experimenten, dat wil zeggen RCT's, wordt ook quasi-experimenteel onderzoek gedaan door impactmeting en, zoals in paragraaf 3 zal blijken, waar zulk onderzoek weinig expliciete theoretische hypothesen hanteert, mist het in veel gevallen een scherp oog voor causale verbanden. In dit stuk beperk ik me tot beleidsrelevante sociale wetenschap; maar de opmerkingen, over de problemen die ontstaan bij het nastreven van de interne validiteit, kunnen in principe worden herhaald als het om impactmeting in andere disciplines gaat.

De kwestie begint ermee dat onderzoek naar de werking van beleidsinterventies intern valide dient te zijn. Een interventie – bijvoorbeeld een trainingsprogramma om werklozen eerder de arbeidsmarkt op te helpen of een alternatief voor gevangenisstraf om de recidive van jonge overtredders te verminderen – kan wel samengaan met een zogenaamde meetbare uitwerking (of impact), maar mogelijk is deze fortuinlijke uitwerking mede of slechts veroorzaakt door andere factoren, *confounders* genoemd. Deze worden in het onderzoek opgenomen door de zogenaamde 'covariaten'. De vermindering van werkloosheid kan bijvoorbeeld vooral veroorzaakt zijn doordat de groep die het trainingsprogramma volgde sterker dan gemiddeld gemotiveerd is ten opzichte van de controlegroep, die het programma niet heeft gekregen.³ Of de afname van recidive kan ermee te maken hebben dat de mensen met de alternatieve straf in omstandigheden verkeerden waar zij beter leerden uit handen van de politie te blijven, ondanks voortzetting van hun criminele gedrag. In zulke gevallen is er sprake van een overschatting van het effect van de interventie. Evenzeer is het omgekeerde mogelijk: dat de *confounders* tot een onderschatting van het effect leiden, namelijk als deze op de controlegroep een invloed hebben die op de behandeling lijkt. Te denken valt aan spill-over effecten als de werklozen

2 Zie daarvoor bijvoorbeeld Nancy Cartwright (2011) die de bevindingen afkomstig van haar onderzoek naar de economie in *The Lancet* wist te publiceren.

3 Dit kan het gevolg zijn van een zg. selectie-effect, dat soms optreedt als de toewijzing aan controle- en behandelgroep niet – zoals dat heet – 'door toeval bepaald' is.

die geen behandeling krijgen op de hoogte zijn van het bestaan van een behandelgroep. Als onzeker blijft of deze confounders veel van het werk gedaan hebben, dan kan onmogelijk een conclusie getrokken worden over de werkzaamheid van het beleidsmiddel op zichzelf en zeker niet in een nieuwe context. Een goed ontworpen (quasi-)experiment – dat al deze onzekerheid niet (al te zeer) laat bestaan – heet dus intern valide te zijn. Interne validiteit van effectmeting garandeert dat het gemeten effect een netto uitwerking van de interventie is en niet een bruto uitwerking van de instrumentvariabele – de interventie – plus de andere oorzaken. Voor *confounders* moet dus gecorrigeerd worden en daarvoor moet de omvang van de covariaten geschat worden. Dit zou een noodzakelijke voorwaarde zijn voor zinvolle kennisproductie indien de sprong van meetbare statistische verbanden naar uitspraken over causaliteit gemaakt moet worden.

Er valt echter ook te denken dat het streven naar interne validiteit ten koste kan gaan van bruikbare kennis over wat werkt in nieuwe contexten. Om door een experiment kennis op te doen over de werking van een interventie moeten allerlei omstandigheden gelijk gehouden worden. Dit simpele feit geldt natuurlijk evenzeer voor natuurwetenschappelijke kennisproductie als voor medische of sociale wetenschappen. Het debat over de bovengenoemde uitruil is echter het meest acuut in de sociaalwetenschappelijke praktijk en niet in natuurwetenschappelijke disciplines. Dat heeft mede te maken met de veelheid aan soorten covariaten die in sociaal onderzoek hun rol spelen. Bovendien, waar het constant houden van omgevingsvariabelen uiteraard een zorg is in alle onderzoek, is het voorkomen van een overloop tussen controlegroep en treatmentgroep vooral een specifiek probleem voor biomedisch en sociaal onderzoek. De confounders komen in beleidsrelevant onderzoek uit alle hoeken en gaten van de interventiecontext sijpelen. Er zijn Hawthorne-effecten, de genoemde overloopverschijnselen tussen behandelgroep en controlegroep, maar ook tussen het aantal werkzoekenden en het loon, er is selectiebias, er zijn allerlei soorten ongeobserveerde heterogeniteit.⁴ En dan zijn dit zijn nog slechts

4 *Hawthorne*-effecten treden op als de subjecten wier gedrag onderzocht wordt zich bewust zijn van het onderzoek en zich daardoor al anders gaan gedragen. Overloopverschijnselen op de arbeidsmarkt kunnen bijvoorbeeld zijn dat werkenden hun baan verliezen als werklozen geholpen worden en deze worden wel *general equilibrium* effecten genoemd. Selectiebias ontstaat als de subjecten invloed hebben op de toewijzing aan behandel- en controlegroep, of als anderen die invloed hebben (bijvoorbeeld rechters bij de bepaling van de strafmaat). Ongeobserveerde heterogeniteit betreft ongemeten eigenschappen van subjecten die langs indirecte weg gezuiverd moeten worden; een voorbeeld is motivatie bij het vinden van werk. Daarover verderop meer.

de meest voorkomende kwesties die de interne validiteit in gevaar brengen.

Grofweg is de redenering dat de interne validiteit van een experiment vooral bereikt wordt door op basis van de covariaten te corrigeren. Maar hoe beter het lukt de correcties door te voeren, des te minder het experiment zou lijken op de werkelijkheid: onbruikbaar voor conclusies over de werking van de interventie in een andere dan de specifieke casus waarin de impactmeting werd gedaan. Men zegt dan dat de externe validiteit tekortschiet in de mate dat de interne validiteit is gediend. Dit is evenzeer aan de orde in een laboratorium als bij een veldexperiment, dat geen laboratorium gebruikt maar wel een deel van de werkelijkheid afgrenst: de in toom gehouden werkelijkheid. Beleidsinterventies kun je zien als een speciaal geval van veldexperimenten. In al deze gevallen zou de prijs van een goed uitgevoerd onderzoek de verloren mogelijkheid tot 'extrapolatie' naar nieuwe gevallen zijn. Voor alle andere genoemde vormen van bias geldt precies hetzelfde.

Dat de uitruil onvermijdelijk zou zijn, wordt betwist. Jiménez-Buedo en Miller (2010) zeggen dat de interne validiteit van onderzoek juist een noodzakelijke voorwaarde is voor externe validiteit. Zij benadrukken dat afgrenzing van versturende variabelen zelfs een eerste stap is van het verkrijgen van betrouwbare kennis voor toekomstige generalisering en dat eenvoudige herhaling van (quasi)experimenteel onderzoek voldoende garanties biedt voor het winnen van externe validiteit.⁵ En inderdaad, er is natuurlijk niets mis met goed dichtgetimmerd onderzoek als dat inzicht geeft in de werking van een instrumentvariabele.

In de volgende paragraaf breng ik de kwestie rond validiteit uitgebreid in beeld. Die kwestie is: hoe meer we er in slagen de uitkomst van een interventie louter toe te schrijven aan slechts die interventie, des te minder we te weten komen hoe precies die interventie heeft gewerkt. Dan volgt in paragraaf 3 een illustratie aan de hand van onderzoek naar de werking van arbeidsmarktmaatregelen. In paragraaf 4 beschrijf ik wat volgens mij de voornaamste tekortkoming is van impactonderzoek. In de conclusie, ten slotte, stel ik twee criteria voor waaraan evaluaties van beleid moeten voldoen, willen ze bijdragen aan een collectief sociaalwetenschappelijk geheugen en aan kennis over wat ik noem 'werkzame principes'.

5 *It would seem too demanding of any scientific practice to expect that comprehensive knowledge about a causal relation of interest would stem from a single exercise: in this case, from one particular experiment.* Jiménez-Buedo en Miller (2010), 319.

2 Ruilen de interne en de externe validiteit uit?

Ik zal de stelling verdedigen dat de uitruil wel degelijk bestaat, zij het in één bepaalde betekenis van de term ‘externe validiteit’. Maar deze betekenis is voor de groei van wetenschappelijke kennis nou net de enige die er echt toe doet. Om mijn opvatting te verdedigen stel ik hier twee problemen aan de orde die in het debat over de uitruil over het hoofd gezien worden. Het eerste is dat interne validiteit vergt dat juist de oorzakelijk met de uitkomst van een experiment of quasi-experiment verbonden grote hoeveelheid aan variabelen uit de theoretische belangstelling van de onderzoekers geraken. Het tweede is dat externe validiteit zelf een vaag begrip is dat onopgemerkt in het debat de neiging heeft van betekenis te veranderen.

Het eerste van deze twee problemen betreft de beperking van de onderzoeksvraag, in veel evaluaties, tot de uitwerking van één geïsoleerde variabele, namelijk de interventie zelf. Daartoe zijn methoden nodig die voor bias controleren of die deze bij voorbaat voorkomen. Het tweede probleem gaat over de noodzaak in de toekomst nieuwe, soortgelijke interventies te doen.

2.1 De cognitieve prijs

De bias die in de covariaten te vinden moet zijn bij de bestudering van een interventie levert niet slechts een praktische kwestie. Voor de oplossing van het praktische vraagstuk dat overige omstandigheden getemd moeten worden en dat de controle- en behandelgroepen homogeen moeten zijn, is inmiddels een enorme hoeveelheid methodologische literatuur beschikbaar. Deze strekt van methoden om via *matching* de groepen te homogeniseren op gemeten eigenschappen tot methoden om econometrisch via logistische selectievergelijkingen de heterogeniteit van de ongemeten eigenschappen boven tafel te krijgen en ervoor te corrigeren. Het is in deze praktische context dat de vaak geuite voorkeur voor randomisering – liefst door een RCT – begrepen moet worden. (En in deze context zijn er consistente pleidooien die betwijfelen of randomisering nou zo belangrijk is.⁶) Maar er is een veel principiëlere kwestie aan de orde: al deze oplossingen bieden een verdwijntruc. Zij zijn erop gericht om de zogenaamde overige invloeden buiten beeld te brengen, in plaats van te onderzoeken.

Dat heeft een cognitieve prijs. Immers, alle oorzakelijke factoren die een

6 Een voorbeeld is John Worrall, 2002 en 2007. In de vierde paragraaf zullen we zien dat hij RCT's in voorkomende gevallen zelfs zeer ongewenst vindt.

gewenst effect teweeg brachten – de interventie, de eigenschappen van de omgeving en de causale invloed van een of ander sociaal mechanisme dat met de interventie geëxploiteerd wordt – werkten samen met dit beoogde effect als resultaat. Dan is voor begrip van dit mechanisme juist ook een diepgaande analyse van de aard en werking van alle ‘helpende factoren’⁷ nodig, niet slechts een kwantitatieve schatting van de mate van verstoring die zij teweeg brachten.

Het is noodzakelijk om voor de opbouw van een collectief geheugen sociale theorieën te ontwikkelen op een gematigd niveau van abstractie. Dit is fundamenteel om ‘werkzame principes’ te vinden, principes die de werking van een interventie verklaren en die in een andere casus ook van toepassing zijn. Neem de criminologie. We willen iets te weten komen over de terugkerende patronen in de jeugddelinquentie, maar kennis van zeer globale tendenties in de voorwaarden voor vermindering van recidive is natuurlijk feilbaar. Zulke patronen, gemeten bij groepen, kunnen bij individuele boefjes immers steeds anders uitpakken. Een in de experimentele context verankerde behandeling van een specifieke groep jonge overtredders leidt alleen tot kennis over deze groep, in deze context, op dit betreffende moment. Ook hier is dus een uitruil aan de orde. Enerzijds, hoe abstracter de beschrijving van de werking van een sociale interventiesituatie, des te lastiger de concrete toepassing van de opgedane kennis in een nieuw geval. Hoe meer anderzijds de evaluatie louter de bestudeerde interventiecasus theoriearm en zonder abstractie in kaart brengt, des te minder leerzaam de bestudeerde casus zal zijn voor sociaal beleid in de toekomst en op andere plekken. Sociale kennis kan pas nuttig zijn als deze voldoende theoretisch van aard is, maar zulke kennis mag ook weer niet te abstract zijn. Dat wil zeggen dat de kunst van beleidsrelevante wetenschap is om inzicht in beperkte wetmatigheden te verwerven en tegelijk enigermate terughoudend te zijn in de toepassing ervan. Laten we het een sterk soort fallibilisme noemen. De casus van de volgende paragraaf zal tonen dat het streven naar interne validiteit leidt tot theoriearmoede als dit streven centraal komt te staan. De stap naar de uitruil tussen interne en externe validiteit is dan snel gemaakt.

7 De ongebruikelijke term ‘helpende factoren’ is geïntroduceerd door Nancy Cartwright. (Zie haar 2007.) Hij verwijst nadrukkelijk naar *alle* andere variabelen die met een vermeende oorzaksvariabele samenwerken, zowel vermoede als onvermoede en zowel contextuele als die van het mechanisme dat van de context onderscheiden wordt. Ik kom hier in paragraaf 3 op terug.

2.2 Betekeniswisseling van ‘externe validiteit’

Het tweede probleem is dat externe validiteit nogal onopgemerkt iets anders gaat betekenen dan oorspronkelijk bedoeld. Dit blijkt meer te zijn dan een semantische kwestie. Een steekproef wordt meestal als extern valide beoordeeld als de eigenschappen van deze steekproef dezelfde zijn als die van de gehele populatie waaruit de steekproef is genomen. We willen iets weten over het gedrag van een grote populatie en de steekproef dient om tegen relatief geringe kosten kennis te verwerven over deze populatie. Een kiezersonderzoek geschiedt door het nemen van zo’n steekproef. Als het onderzoek naar stemgedrag niet extern valide is, dan komen we er niets uit te weten over de uitslag van de komende verkiezingen. Als je alleen mensen uit de Randstad bevraagt heb je geen ‘representatieve steekproef’. Externe validiteit gaat dan over representatie.⁸

De beleidsmakers zijn echter geïnteresseerd in de werking van een instrument in nieuwe gevallen. Het oude geval – de eenmalige interventie waarvan een evaluatie is gedaan – is niet een steekproef van alle denkbare gevallen. Zelfs niet van denkbare soortgelijke gevallen. Stel dat in de conclusies van een beleidsevaluatie adviezen staan op basis van de gevonden correlatie tussen interventie en effect. Ondanks dat hier helemaal geen steekproef gedaan is, spreekt men desalniettemin van de externe validiteit van de evaluatie als deze bevindingen elders toepassing vinden. Dat is dus heel wat anders dan de representativiteit van een steekproef.

Hier is meer aan de hand dan louter de betekenisverschuiving van een term. Onderzoekers ontberen naar mijn indruk vaak een helder inzicht in het cruciale verschil tussen een steekproef die iets zegt over een populatie en een onderzoek naar werkzame principes die in andere contexten hun nut bewijzen. De betekenisverschuiving is daar een symptoom van. Heel pregnant is dit verschijnsel terug te zien in een publicatie van een van Neerlands bekendste arbeidsmarktsociologen, Pierre Koning (2005). Op één en dezelfde bladzijde (85) merkt hij eerst op dat ‘de daadwerkelijke participanten en controlegroepen van een specifiek scholingsprogramma een goede dwarsdoorsnede [dienen] te vormen van de beoogde doelgroep’ en daarna dat het maar de vraag is ‘in hoeverre deze onderzoeksresultaten te vertalen zijn naar meer algemene situaties en daarmee de basis kunnen vormen van algemeen scholingsbeleid.’

Onder de mat vinden we de weggeveegde scherven. Wie ze lijmt, ont-

8 Shadish et al. (2002) noemen dit ‘*construct validity*’ en onderscheiden het daarmee wel goed van ‘*external validity*’ in de hierna volgende betekenis. Maar zij begeven zich niet in het debat over de uitruil van interne en externe validiteit waarin de verwarring volgens mij groot is.

dekt dat de sprong van kennis van het effect van de eenmalige – contextueel verankerde – beleidsinterventie zelf, naar denken in termen van sociaalpsychologische mechanismen die iets doen in een reeks van contexten, reusachtig is. De werkzaamheid van zulke mechanismen zal zich moeten bewijzen. Externe validiteit betekent plots niet meer ‘het was een goede dwarsdoorsnede’ maar ‘het toont een werkzaam principe waarvan we in de toekomst gebruik kunnen maken’. De beoogde doelgroep wordt gevormd door alle mensen die je zo’n scholing zou willen aanbieden, in verleden, heden en toekomst. Het ‘specifieke scholingsprogramma’ is één keer ingezet bij een zekere groep mensen. Deze groep wordt nu gezien als een steekproef van de grotere populatie op wie we ooit zo’n programma zouden willen loslaten. Dit is onzin. Nieuwe gevallen in nieuwe contexten zijn geen onderdeel van een grotere populatie waaruit een steekproef is genomen. De wisseling in betekenis van ‘externe validiteit van het onderzoek’ blijft, voor wie slordig leest, onopgemerkt.

De betekenisverandering doet er nogal toe. De tijd-en-plaats-bepaaldheid van elk sociaal programma verhindert het verwerven van verder strekkende kennis als onderzoekers niet méér willen doen dan het leggen van een rechtstreeks en onvervuild verband tussen een zekere interventie en haar impact. Het streven naar interne validiteit leidt dan tot almaar verfijndere methoden om dat verband rechtstreekser te maken en steeds minder vervuild. Het resultaat is soms een positief statistisch significant verband tussen ingreep en uitkomst, soms een negatief verband en vrij vaak een verband dat niet aan de statistische eisen van significantie voldoet.⁹ Alle impactonderzoek in de sociale wetenschap van de afgelopen halve eeuw samengenomen leidt tot de kennis dat, zoals Pawson en Tilley (1997) het nogal sterk stelden, ‘de meeste interventies soms werken’. Er zijn eenvoudigweg te veel soorten situaties, rommelige interventies, onbekende variabelen, selectieproblemen en belangen in een voortdurend veranderende wereld. Zo wordt niets geleerd. Tegelijk wekt deze verfijning in de methoden de indruk dat het met de kwaliteit van de steekproef – van de bestudeerde casus – wel heel goed gesteld moet zijn. De nieuwere betekenis die ‘externe validiteit’ krijgt is dat de conclusies als het ware extrapolbaar zijn naar andere contexten, zoals de eigenschappen van een steekproef dat zijn naar die van de populatie. Doordat deze betekenisverande-

9 Gewoonlijk wordt een significantieniveau van 95% vereist. Valen Johnson (2013) heeft met een Bayesiaanse benadering van de *uniformly most powerful Bayesian* test laten zien dat deze norm twee tot tien keer strenger mag. De voorkeur voor Neyman-Pierson statistiek in de beleids-evaluatie is een heel ander (maar verwant) probleem dat buiten het bestek van dit artikel blijft.

ring onopgemerkt blijft, schuift er een mistbank over het feit dat bij een impactevaluatie helemaal geen steekproef in de klassieke zin is genomen en dat zuivere extrapolatie niet is waarom het gaat.

2.3 Werkzame principes

Waar gaat het dan wel om? Om de bruikbaarheid van theoretische kennis in nieuwe gevallen; dat wil zeggen, om werkzame principes. Dat heeft niets met extrapolatie te maken. We willen weten of bepaalde vormen van sociaal beleid in een verzameling contexten een goede kans maken vruchten af te werpen tegen redelijke kosten. Weer anders gezegd, we willen dat belastinggeld goed besteed wordt. Daarvoor is kennis nodig van de werking van de sociale wereld in die contexten. Zo'n verzameling contexten laat zich kenschetsen door de reikwijdte van de interventies, zoals door het kiezen van het domein 'scholing voor jonge werklozen', 'cameratoezicht in uitgaansgebieden', 'verkeersdrempels in woonwijken', 'begeleiding van voorwaardelijk vrijgelatenen' en dergelijke. Sociaalwetenschappelijke kennis moet dus verder strekken dan slechts die betreffende de werkzaamheid in een bestudeerde casus; de kennis moet voldoende theoretisch en in zekere mate abstract zijn. Maar sociale wetenschappers moeten niet trachten hun kennis zover te laten strekken dat de pretentie ontstaat deze toe te passen op alle samenlevingen, onder willekeurig welke institutionele arrangementen, betreffende uiteenlopende beleidsgebieden. Bescheidenheid is nodig, in overeenstemming met Poppers adagium de wereld hooguit te veranderen door *piecemeal engineering*.¹⁰

Het gaat om niets minder dan de toepassing van de hele empirische cyclus in de zoektocht naar evidence voor beleid. Impactanalyse doorloopt de cyclus van het verwerven van data tot generalisering van deze data en houdt dan op. Een nieuwe interventie herhaalt dit slechts. In wetenschap – echte wetenschap – stelt men verklaringshypothesen op, laat men zien hoe deze samenhangen met bestaande theorieën, ontleent men aan de theorie toetshypothesen, test deze en voegt men zowel bij falsificatie als bij confirmatie nieuwe data toe die de theorie helpen verfijnen.

Elke interventie is immers impliciet of expliciet gebaseerd op vermoedens over de werking van sociale en psychologische mechanismen in een bepaalde context. Deze vermoedens zijn hoe dan ook theoretische hypothesen. Zij zouden zo behandeld moeten worden: als op waarheid te toetsen uitspraken die voorafgaand aan de te bestuderen interventie expliciet gemaakt worden. Het doel van elke beleidsevaluatie zou dus niet eenmalig

¹⁰ Popper (1961: 64).

toepasbare kennis over de werking van een specifiek programma moeten zijn, maar een boekhouding van falsificaties en confirmaties van beleidstheorieën. Op basis van deze kennis moeten onze theorieën met elke studie verfijnd en uitgebreid worden. Met het opgebouwde collectieve geheugen kunnen nieuwe, beter geïnformeerde hypothesen ter toetsing te berde gebracht worden.

Vergelijk dit met de evaluatiepraktijk in veel arbeidsmarktonderzoek, het voorbeeld dat uitgewerkt wordt in de volgende paragraaf. Dat beperkt zich als gezegd tot impactmeting en doorloopt daarmee slechts een klein deel van de empirische cyclus. De hypothesen over de werking komen post hoc, dat wil zeggen, na de verslaglegging van het gekwantificeerde resultaat van de interventie. Zij werkte niet. Waardoor niet? Wellicht was W daarvoor verantwoordelijk. Zij werkte wel, hoe komt dit? Misschien door X of Y, misschien door Z. (Meestal wordt de waardoor-vraag alleen gesteld als de interventie niet werkte, curieus genoeg zelden als zij wel werkte.) De weg van een vermoeden over hoe iets werkt naar fatsoenlijke toetshypothesen en de weg van falsificatie naar herziening en integratie van alternatieve hypothesen in het bestaande theoretische corpus vind je in dergelijk onderzoek nauwelijks terug; zelfs vaak niet in onderzoek dat wel de pretentie heeft 'theoretisch gestuurd' te zijn.

3 Ongeobserveerde heterogeniteit in arbeidsmarktonderzoek

Het meeste onderzoek naar de werking van arbeidsmarktbemiddeling is quasi-experimenteel. Dat wil zeggen, er vindt een interventie plaats en men meet het bruto-effect ervan op de hoogte van de werkloosheid, precies zoals u een pijnstiller slikt tegen hoofdpijn om daarna te constateren of de pijn vermindert. Daarna corrigeert men voor covariaten met oog op de vaststelling van het netto-effect, het effect dat werkelijk aan de interventie en aan niets dan de geïsoleerde interventie kan worden toegeschreven. Het echte experimentele werk, met *Randomised Controlled Trials*, liefst dubbelblind uitgevoerd, wordt het hoogst aangeslagen maar is lastig te doen bij dit soort onderzoek. Toewijzing door toeval is bij tal van interventies niet haalbaar. Er treedt bijvoorbeeld zelfselectie op. Maar ook is er bij veel beleidsgebieden vaak selectie door een ander. Bij de arbeidsmarkt zijn de arbeidsbemiddelaars degenen die bepalen wie meedoet aan een experiment. Bij criminologische interventies, bijvoorbeeld, kan dat de

rechter zijn, de officier van justitie of de politie, actoren met veel jurisdictie als het gaat om arrestatie, vervolging en veroordeling.

Het quasi-experiment vormt het alternatief voor het echte experiment maar kent zelf weer een hele reeks technische moeilijkheden. Dit komt doordat de correctie door toewijzing aan de hand van toeval slechts deels uitgevoerd kan worden. Sommige vormen van heterogeniteit zijn ongeobserveerd – zoals de mentaliteit of de initiatiefrijkeheid van de participanten – waarvoor correctie alleen via uiterst gecompliceerde econometrische methodes te doen is. Bovendien is elk arbeidsmarktprogramma noodzakelijk slechts een marginale interventie, omdat arbeidsbemiddeling een recht is. Het niet geven van de specifieke bemiddeling, dat wil zeggen de beoogde interventie, impliceert nooit het uitblijven van een of andere bemiddeling. Al met al treedt zonder allerlei ingrijpende correcties zowel onderschatting als overschatting van het netto-effect op. De mate waarin deze schattingsfouten elkaar compenseren of versterken is moeilijk vast te stellen. Om kort te gaan, evaluatie van arbeidsmarktbeleid is lastig.

De covariaten representeren dus mede-veroorzakers van een positief of negatief effect. Nancy Cartwright noemde deze mede-veroorzakers met reden *'helping factors'*. Zij moeten in de effectvergelijking meegenomen worden. Elke interventie heeft als afhankelijke variabele de uitkomst van de interventie en als onafhankelijke variabelen de interventie zelf plus de helping factors. Voor zover deze bekend zijn, of op goede gronden vermoed worden, krijgen ze een parameter in de logistische vergelijking die interventie en impact in verband brengt. De ongeobserveerde factoren en de onvermoede variabelen moeten daarentegen tot uiting komen doordat de error-term niet gelijk is aan nul. Econometristen weten daar raad mee.

Voorbeelden van ongeobserveerde factoren in arbeidsmarktbeleid zijn motivatie, flexibiliteit bij het accepteren van werk, soms communicatieve vaardigheden of sociale vaardigheden. Merk op dat we het hier dus over causale factoren hebben die onontbeerlijk lijken voor het doen slagen van om het even welke interventie op de arbeidsmarkt! Stel nu dat we het effect van een interventie, zoals bemiddeling of training, in isolatie willen meten, d.i. dat we het netto-effect willen kennen. Dan moeten we voor verschillen in motivatie – om dat voorbeeld te gebruiken – zien te corrigeren. Juist motivatieverschillen verklaren soms waarom sommige werklozen in de controlegroep komen, de groep die geen interventie ondergaat. Er bestaat dan heterogeniteit in controle- en treatmentgroep als gevolg van een selectie-effect. Het kan zijn dat de werklozen zelf indirect invloed konden uitoefenen op de toewijzing aan deze groepen, maar ook dat er

selectie plaatsvindt door de bemiddelaars, die immers een belang hebben bij succes in hun werk. In het laatste geval spreekt men van 'afroming'.

Hoe corrigeer je voor een verschil in motivatie in de controle- en behandelgroep? Met behulp van de volgende redenering. Neem aan dat dit verschil bestaat. Dan hebben de werknemers die ongunstig bedeed zijn wat betreft de observeerbare variabelen, bijvoorbeeld in de vorm van een laag scholingsniveau, en die desondanks toch een baan vonden, een bijzonder ander attribuut. Die eigenschap zou zo maar de sterke motivatie kunnen zijn. Deze subgroep der ongunstig bedeeden, die niettemin werk vinden, krijgt nu langs econometrische weg een correctiefactor toebedeeld in de logistische vergelijking die instrumentvariabele en afhankelijke variabele in verband brengt. De correctie op de ongeobserveerde variabele wordt dus ontleend aan de observeerbare variabele.¹¹

Als we aannemen dat dit type methoden puur methodologisch gesproken succesvol is, dan zitten we met een merkwaardig epistemologisch en ontologisch probleem. Hier behandel ik eerst het epistemologische probleem, in de volgende paragraaf komt de ontologie aan bod.

In onze pogingen bruikbare kennis over de sociale wereld te krijgen behandelen we onontbeerlijke causale factoren die een interventie succesvol zouden moeten maken als verstoringen van diezelfde interventie. Arbeidsmarktsocioloog Arie Glebbeek noemde dit in een venijnig artikel 'absurd'.¹² Ik vrees dat hij gelijk heeft. Het evaluatieonderzoek zit gevangen in een paradigma waarin impactonderzoek allesbepalend is en het zoeken naar oorzaken en mechanismen van sociale processen weinig prominent. Dit leidt tot vreemde paradoxen. Als een arbeidsmarkt bemiddelaar de opdracht krijgt de slechtst plaatsbare werklozen de arbeidsmarkt op te helpen dan leidt succes van dit beleid tot het naar elkaar toegroeien van de kansen van gemakkelijk plaatsbare mensen en moeilijk plaatsbare mensen. De uitkomstmaat kan namelijk niet slechts gericht zijn op de verandering in de werkloosheidscijfers van kansarmen, want er is altijd een vergelijking nodig met een andere groep, zowel bij de nulmeting als bij de impactmeting. Die andere groep was al kansrijker, dus succes impliceert dat de behandelde groep gaat lijken op de kansrijke groep. Het slagen van de inter-

11 Een econometrisch sterk doorontwikkelde methode hiervoor is de Heckman twee-stapsmethode. Die stelt eerst een selectievergelijking op om het selectie-effect te ontleden, en daarna de substantiële vergelijking die instrument in verband brengt met het selectie-effect en met de netto-*impact*. Zie Heckman 1979. Er zijn ook iets simpeler maar minder precieze methoden in gebruik.

12 Glebbeek (2005: 40). De auteur citeert de voorvechters van wat heet 'realistische evaluatie', Pawson en Tilley (1997), die precies hetzelfde stellen.

ventie gaat samen met een nulcorrelatie! De groep waarop het beleid zich richt, onderscheidt zich immers niet meer zo sterk van de groep die niet bemiddeld wordt. De meting zelf wordt onmogelijk.¹³

De roep om evidence based beleid heeft zo wel tot veel onderzoek geleid, maar niet tot veel kennis. Waar echte experimenten gedaan konden worden, met gerandomiseerde toewijzing aan experiment- en controle-groep, kon van elke afzonderlijke interventie wel redelijk betrouwbaar bepaald worden of deze een gunstige uitkomst had. Maar het wisselende beeld dat overbleef na het overzien van vele soortgelijke interventies leidde niet tot inzicht in wat voor type interventies nou wel en niet werkte. In arbeidsmarktonderzoek spelen echter als gezegd vooral quasi-experimenten een rol, zonder randomisering maar met de toepassing van een heel arsenaal aan methoden om de heterogeniteit tussen de verschillende groepen te verminderen of om voor deze te corrigeren. Terwijl juist diezelfde heterogeniteit kan helpen nagaan of en hoe verschillen in de context het succes van een interventie bepalen, wordt veel intellectueel kapitaal geïnvesteerd in het wegwerken van deze verschillen. Zowel experimenteel als quasi-experimenteel onderzoek is er op gericht om de interventie te isoleren en de rest van de wereld als verstoring te neutraliseren. Wie echter wil weten welk soort interventie op de arbeidsmarkt werkt voor wie en in welke omstandigheden zou beter inzoomen op een causaal verband tussen het gedrag dat specifieke deelnemers aan de interventie vertonen en het resultaat van het zoeken naar werk. Dat brengt ons op het spoor van de eerder genoemde werkzame principes.

Interessant in dit verband is de oplossing die onderzoekers soms voorstellen als ze wel interesse beweren te hebben in de omstandigheden die het succes van ingrijpen op de arbeidsmarkt maken of breken. Karl Ashworth deed met collega's een meta-evaluatie van experimenten in de Verenigde Staten met als doel te achterhalen hoe je nou mensen het beste uit een uitkering krijgt. Op zoek naar werkzame principes, dus, op basis van een grote hoeveelheid verschillende soorten interventies. Zijn conclusie luidt dat de omstandigheden van de implementatie van de programma's minstens zoveel uitmaken als het ontwerp van de programma's zelf. Dat ligt vreselijk voor de hand, maar blijkbaar is er toch een uitvoerige meta-evaluatie nodig om deze conclusie te trekken. Over de onderzoeken die de basis vormden van zijn meta-evaluatie schrijven Ashworth et al.:

13 Het is verleidelijk te denken dat de meting gewoon omgedraaid kan worden: hoe kleiner de correlatie, des te succesvoller het beleid. Maar zo werkt dat niet. Onderzoekers zijn gehouden de significantie van de correlaties aan te tonen. Nulcorrelatie geeft insignificante resultaten.

Social and environmental conditions affecting programme sites, for instance, were seldom taken into account. In fact, they did not need to be because the evaluation designs used by many studies, based on random assignment of welfare recipients into experimental and control groups, guaranteed that programme participants and nonparticipants shared environmental conditions.¹⁴

Er was in de onderliggende studies weinig interesse waar te nemen voor de omgevingsfactoren, maar dat was ten slotte ook niet nodig, want ze werden toch gelijk getrokken over de verschillende deelnemers, zo lezen we. De condities die een succes mogelijk moeten maken worden ook hier behandeld als rivaliserend met de interventie en daarmee als verstoring van het onderzoek. De oplossing die de onderzoekers kiezen is die van de meta-evaluatie: van losstaande onderzoeken die qua impact in alle richtingen wijzen naar een overzicht van deze onderzoeken dat nog steeds in alle richtingen wijst.

Een tegenwerping tegen mijn kritiek ligt hier voor de hand. Als bepaalde omstandigheden zo interessant zijn voor de werking van een of ander sociaal mechanisme, dan moet het toch zinvol zijn om voor alle andere contextuele factoren te corrigeren? Kunnen we niet een interventie loslaten op gemotiveerden door deze observeerbaar te maken? Je introduceert de zelfselectie door het introduceren van drempels; en wel zodanig dat alleen mensen met een sterke motivatie aan het experiment meedoen, om het even in de controle- of behandelgroep. Dan komen we te weten of interventie-plus-motivatie een significante impact heeft. Dat kan inderdaad. Maar dan herhaalt zich het probleem met andere causaal relevante factoren, die nu als rivaliserend moeten worden behandeld. Er is de conjunctuurbeweging – zeer belangrijk voor de arbeidsmarkt – de sociale vaardigheid van de werkzoekenden, er zijn de eigenaardigheden van het experimentontwerp en ga zo maar door.

De kwestie is en blijft: er is een mechanisme aan het werk dat interacteert met de omgeving, die minstens evenveel causale variabelen aanlevert als het mechanisme zelf, en waarschijnlijk meer. Als dat mechanisme ergens anders ook moet werken, dus als we willen weten of het een werkzaam principe is, dan zoeken we ‘externe validiteit’ in de tweede betekenis die hierboven is genoemd, de mogelijkheid om een interventie opnieuw toe te passen in een andere context. We willen dan weten hoe die interactie verloopt om te voorspellen aan welke voorwaarden de omgeving

14 Ashworth, K. et al (2004: 194).

moet voldoen om een soortgelijke interventie succesvol te maken; of om te voorspellen hoe de omgeving kan worden aangepast voor dat doel.

Het beleid dient om de sociale werkelijkheid in gunstige zin te veranderen; werklozen moeten uit de uitkering. Voor andere beleidsgebieden geldt hetzelfde, recidive moet verminderd, gokverslaving voorkomen. Wat daartoe aan interventies bedacht wordt kan werken of niet, waar het om gaat is hoe die interventie de werkelijkheid verandert. Er is, met andere woorden, een beeld van de causale structuur van die werkelijkheid nodig.

4 Ontologische armoede

Een probleem is dat de evidence-based praktijk helemaal geen wereldbeeld postuleert, een wereldbeeld waarvan oorzakelijkheid een integraal onderdeel is. Behalve de gemeten uitstroom uit de uitkering zou bijvoorbeeld de causale samenhang tussen motivatie en arbeidsmarktkansen empirisch te onderzoeken zijn. Mede op basis hiervan kunnen, wederom bijvoorbeeld, hypothesen worden getoetst over de rol die laagopgeleide werklozen zichzelf toekennen en over het gedrag dat zij op basis daarvan vertonen. Er zou, anders gezegd, een sociale theorie kunnen worden opgesteld. Voor wie er zo'n wereldbeeld op na houdt zijn behalve wat meetbaar is ook andere aspecten van de werkelijkheid te zoeken en te vinden, sommige even vatbaar voor onbemiddeld empirisch onderzoek als de gewoonlijk gemeten factoren, andere alleen indirect afleidbaar uit zulk onderzoek.

De groei van (quasi)experimenten en impactanalyse is voortgekomen uit de behoefte de praktijk in biomedisch en farmacologisch onderzoek te imiteren. Juist over zulk onderzoek schreef John Worrall een fascinerend verslag dat aantoonde tot welke dwaasheid het negeren van ons inzicht in de causale structuur van de werkelijkheid leidt. Worrall's doelwit is weliswaar niet speciaal de sociale wetenschap maar alles wat uit deze passage te leren valt, is direct toepasbaar op beleidsexperimenten. Hij schetste de achtergrond van twee experimenten om een uiterst succesvolle therapie tegen *persisterende pulmonale hypertensie* (pphn) bij baby's te testen.¹⁵

De ernstige ziekte pphn is het gevolg van onvolgroeide longen bij de

¹⁵ Worrall (2007: 455-458). Worrall voert een pleidooi tegen RCT's in het algemeen, dat hier verder niet relevant is en waarvan de argumenten ook niet alle even sterk zijn. Uitsluitend van belang zijn de implicaties van de beschreven experimenten voor de ongeschikte rol die de onderzoekers aan hun kennis van causale relaties blijkbaar toekenden.

geboorte. In de zeventiger jaren kwam er een therapie tegen: zuurstoftoediening via een membraam buiten het lichaam (ecmo) die de longetjes de gelegenheid gaven verder te groeien tot zij het lichaam van de baby zelf van zuurstof konden voorzien. De mortaliteit als gevolg van de ziekte bedroeg voor de introductie van deze therapie 80%, erna nog maar 20%. Deze cijfers zijn dus afkomstig uit wat ‘slechts’ een observational study oftewel een *historically controlled trial* was, geen echt experiment. Artsen wisten dit. Groot was dan ook hun aarzeling toen ze de druk voelden om de therapie aan een gerandomiseerd experiment te onderwerpen. Wie het succes van de behandeling heeft gezien, die levert een ernstig zieke baby niet graag uit aan de controlegroep, die de behandeling onthouden wordt. De ethische keus om soms geen experiment uit te voeren is in dit geval innig verbonden met de epistemologische vraag of de kennis over het succes wetenschappelijk betrouwbaar genoeg is. Wie meent dat met historische data zonder controlegroep en nulmeting eenvoudigweg geen informatie van betekenis opgeleverd kan hebben, die ziet geen been in een experiment en heeft reden agnostisch te zijn over de ethische vraag. De baby's die de behandeling niet krijgen lopen dan schijnbaar dezelfde kans op overleven voordat het experiment voltooid is als zij die in de behandelgroep terecht komen. Maar wie onder de indruk is van de ervaringen met de therapie in de niet-experimentele setting (20% versus 80% mortaliteit), die vreest baby's aan een extreem hoog risico bloot te stellen. Ethische keuzes veronderstellen vertrouwen of wantrouwen in de waarde van de reeds aanwezige kennis.

De artsen, die het experiment begin tachtiger jaren deden, hadden op basis van hun professionele ervaring in feite al veel vertrouwen in ecmo. Dat leidde tot een halfslachtige uitvoering van het experiment. De eerste baby kreeg de behandeling en overleefde, de tweede kreeg een alternatieve behandeling en stierf. Acht volgende baby's werden toegewezen aan de behandelgroep en zij overleefden allen. (Nog twee kregen de behandeling en zij overleefden ook, maar zij werden officieel buiten het experiment geplaatst.) Het experiment stopte hier, want de artsen durfden geen kinderen meer in levensgevaar te brengen.

Het experiment voldeed uiteraard niet aan de zware eisen van een echt gerandomiseerd experiment. Daarmee was, ten koste van één leven, een onderzoek uitgevoerd dat als waardeloos moest worden bestempeld. Het moest opnieuw. En zo geschiedde. Maar ook in het tweede experiment, eind tachtiger jaren uitgevoerd door anderen, werd rekening gehouden met het probleem dat uitsluiting van behandeling zou leiden tot de onnodige dood van baby's. De onderzoekers implementeerden de stopregel dat

na het vierde sterfgeval het experiment zou worden afgebroken. Ook hier kun je zeggen dat het experiment werd gedaan ondanks de kennelijk halfbewuste overtuiging dat het niet nodig was: men wist al dat het werkte en men doorzag onmiddellijk de ethische consequentie. Met twintig baby's zou het significantieniveau van $p < 0,05$ net kunnen worden gehaald.¹⁶ Negen baby's kwamen in de behandelgroep, van wie allen overleefden, en tien in de controlegroep, van wie er zes overleefden. De stopregel was juist voor het bereiken van het gewenste significantieniveau in werking getreden.

Het is moeilijk te veel te benadrukken dat het punt van dit relaas beslist niet is om *Randomized Controlled Trials* als zodanig in een kwaad daglicht te stellen, al mogen bij de aanname dat zij zo een enorm geprivilegieerde status hebben vraagtekens gezet worden. Wat deze treurige geschiedenis duidelijk maakt is dat twee generaties onderzoekers er in feite al een wereldbeeld op nahielden waarin oorzaak-gevolgrelaties een cruciale rol spelen. Te kleine longen kunnen de druk, die gepaard gaat met benodigde zuurstofopname, niet aan, met gescheurde longblaasjes als gevolg. De onderliggende oorzaak is onvolgroeidheid. De therapie moet derhalve gericht zijn op longgroei terwijl het kind op een andere manier ademt om de longen te ontzien. Er is dus een mechanistisch beeld van de werkelijkheid aan de orde die een causale keten van eerste oorzaak, intermediaire oorzaak en gevolg postuleert met een helder punt waar de interventie plaatsvindt. Het mechanisme is bovendien van een verassende eenvoud; iedereen snapt het. Als artsen inzicht hebben in de werking van een mechanisme dan is een impactanalyse, uitgevoerd zonder theoretische consequenties voor de aanvaardbaarheid van de hypothesen over dat mechanisme, niet meer relevant.¹⁷ Onderzoek moet zich dan ook richten op toetsing van hypothesen over de werking van dat mechanisme, niet op de platte vraag of de therapie werkt. Kennis vermeerdert zich aan de hand van de vraag hoe het werkt, niet of het werkte bij één bepaald experiment (of twee, of drie). De uitsluitende zoektocht naar het statistische verband

16 Niemand weet precies waarom die 95% betrouwbaarheid zo heilig is. Historisch bepaald als deze regel is, had het elk ander percentage kunnen zijn. Zie ook noot 9.

17 Worrall merkt nog een ander aspect op van de dwaasheid van de artsen. De randomisering wordt nodig geacht om selectie-effecten uit te sluiten. Maar voordat de twee experimenten gedaan werden en na de introductie van de behandelmethode kregen alle baby's, die zonder de beschikbaarheid van de nieuwe techniek de alternatieve oude behandeling zouden hebben gekregen, de nieuwe therapie. Er werd niet geselecteerd, dus kon er ook geen selectie-effect zijn (Worrall 2007: 485).

tussen toediening van een interventie en de uitkomstmaat is een uiting van ontologische armoede.

Dit geldt nu evenzeer voor sociaalwetenschappelijk onderzoek. Dat moet ook ontologisch gestuurd zijn, een wereldbeeld postuleren. Zonder theorie geen inzicht. Pas als een statistisch gecontroleerd verband informatie mag leveren over de plausibiliteit van vooraf opgestelde hypothesen over de werking van de wereld kan er sprake zijn van enige kennisvermeerdering.

5 Conclusie: Twee voorwaarden voor de succesvolle toepassing van sociaalwetenschappelijke kennis

De interne validiteit van een onderzoek naar de werking van een interventie betreft de correctheid van de methode en de betrouwbaarheid van het gemeten effect. De externe validiteit gaat over de mogelijkheid de conclusies naar een grotere populatie te extrapoleren. Een goed onderzoek is intern en extern valide. Als nu de interventie werkt zou de beleidsmaker deze nog eens willen doen, in een andere context, een andere plaats, met een andere populatie. Maar extrapolatie geeft geen enkele garantie dat de successen uit het verleden herhaald kunnen worden. Immers, de situaties in de toekomst vormen niet een populatie waarvan het eenmalige succes uit het verleden een steekproef is. Voor herhaling is het nodig dat onze kennis van het werkzame principe 'op reis' kan gaan naar de nieuwe situatie. Daar homogeniteit van populatie en steekproef enerzijds en het genereren van nuttige kennis voor de toekomst anderzijds met elkaar worden verward, is het begrip 'externe validiteit' ambigue.

Voor kennis van de principes die beleid mogelijk succesvol kunnen maken is een collectief sociaalwetenschappelijk geheugen nodig. Dan hebben we het over de tweede betekenis die vaak aan het concept 'externe validiteit' wordt gegeven, want dat geheugen helpt ons plausibele hypothesen over de werking van nieuwe interventies op te stellen. Een beleidsvoorstel impliceert altijd hypothesen over de werking van de interventie, in de vorm van een impliciete 'beleidstheorie', maar vaak zijn juist die hypothesen behalve slecht gearticuleerd vooral zeer implausibel. Een experiment met een interventie zou juist die hypothesen tot object van toetsing moeten maken, zodat er ook bij falen iets wordt geleerd over de structuur van de sociale werkelijkheid. Dat kan, als aan ten minste twee noodzakelijke voorwaarden wordt voldaan.

In de *eerste plaats*, verklarende hypothesen moeten niet (uitsluitend)

achteraf worden opgesteld indien het beoogde effect uitbleef, maar vooraf, met het doel om deze door de bestudeerde interventie te toetsen. Post hoc hypothesen zouden alleen geformuleerd moeten worden als het om hypothesegenererend onderzoek gaat, dat wil zeggen met het oogmerk nieuwe vermoedens te genereren en om deze in een ander experiment expliciet te toetsen; en niet alleen bij falen maar ook als de interventie in het bestudeerde geval wel succesvol was. De vraag 'hoe het komt' is immers niet alleen relevant als er aan de beleidsmaker uitgelegd moet worden waardoor de interventie niet werkte. In echte wetenschap moet ook succes uitgelegd worden. Het object van onderzoek zou – als aan deze voorwaarde wordt voldaan – verschuiven van de grootte van de impact naar de waarheid van de hypothese over het relevante werkzame principe. De interventie kan plezierige of ongewenste effecten hebben voor de doelgroep, wat blijft is nieuwe kennis over hoe het veronderstelde werkzame principe geworteld is in de context van een concrete casus. Deze kennis kan zijn functie uitoefenen in de empirische cyclus die de sociale wetenschap doorloopt.

De *tweede voorwaarde* is dat a priori aangenomen wordt dat actoren gedrag vertonen dat afgestemd is op de omgeving van deze actoren en dat de interactie tussen die omgeving en dat gedrag structuur heeft. Actoren hebben bijvoorbeeld intenties, verwachtingen en een zekere rekenkracht, maar ook handelen zij regelmatig onbewust. De configuratie die daaruit voortkomt heeft een causale structuur. Wij kunnen zinvolle en toetsbare hypothesen opstellen over deze causale structuur. Wij moeten de moed hebben onze ontologische veronderstellingen te expliciteren en denken in termen van de sociale mechanismen die de impact – of het gebrek eraan – verklaren.

De argumenten voor de stelling dat er géén uitruil is tussen interne en externe validiteit betreffen steeds kwesties rond het probleem van de isolatie van laboratorium- en veldexperimenten in de wereld. Ik meen dat die kwesties inderdaad geen strijd behoeven te geven tussen de kwaliteit van een experiment en de bruikbaarheid van de kennis over de wereld die daaruit voortkomt. Meer veld- en laboratoriumexperimenten in steeds licht gewijzigde omstandigheden kunnen heel leerzaam zijn. De vermaleidige uitruil ontstaat als impactanalyse geacht wordt pas goed uitgevoerd te zijn indien er heel veel correctie is voor precies die causale variabelen die de interventie ondersteunen. Daardoor gaat essentiële informatie voor inzicht in causale samenhangen verloren. Het resultaat is ongeschikt om de reis te maken naar nieuwe contexten doordat de gewonnen gegevens vastzitten aan de specifieke omstandigheden van het experiment.

Bibliografie

- Ashworth, K. et al. (2004) Meta-evaluation: Discovering What Works Best in Welfare Provision. *Evaluation*, 10-2, pp. 193-216.
- Cartwright, N. (2007) *Hunting Causes and Using Them*. Cambridge: CUP.
- Cartwright, N. (2011) The art of Medicine. A Philosophers View of the Long Road from RCT's to Effectiveness. *The Lancet* 377, pp. 1400-1401.
- Cartwright, N. (2012) Will this Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps. *Philosophy of Science* 79(5), pp. 973-989.
- Cartwright, N. en Hardie, J. (2012) *Evidence Based Policy. A Practical Guide to Doing it Better*. Oxford en New York: Oxford University Press.
- Chelmsky, E. (1995) Where We Stand Today in the Practice of Evaluation. *Knowledge and Policy* 8:3, pp. 8-19.
- Glebbeeck, A. (2005) De Onrealistische Evaluatie van Arbeidsmarktbeleid. *Tijdschrift voor Arbeidsvraagstukken*, 21-1, pp. 38-48.
- Heckman, J. J. (1979) Sample Selection Bias as a Specification Error. *Econometrica*, 47, pp. 153-161.
- Jiménez Buedo, M. en Miller, L. M. (2010) Why a Trade-off? The Relationship between the External and Internal Validity of Experiments. *Theoria* 69, pp. 301-321.
- Johnson, V. E. (2013) Revised Standards for Statistical Evidence. *PNAS* 110(48), 19313-19317 [doi/10.1073/pnas.1313476110]
- Koning, P. (2005) De Effectiviteit van Scholing voor Werklozen. *Tijdschrift voor Politieke Economie*, 21(3), pp. 79-99.
- Pawson, R. en Tilley, N. (1997). *Realistic Evaluation*. London: Sage.
- Popper, K. (1961) *The Poverty of Historicism*. London: Routledge & Kegan Paul.
- Rol, M. en Cartwright, N. (2012). Warranting the Use of Causal Claims. *Theoria* 74, pp. 189-202.
- Schram, A. (2005). Artificiality: The Tension Between Internal and External Validity in Economic Experiments. *Journal of Economic Methodology*, 12, pp. 225-237.
- Shadish, W.R., Cook, T.D. en Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Worrall, J. (2002) What Evidence in Evidence-based Medicine? *Philosophy of Science* 69, pp. S316-S330.
- Worrall, J. (2007) Why There's No Cause to Randomize. *British Journal for the Philosophy of Science* 58, pp. 451-488.

Over de auteur

Menno Rol is docent wetenschapsfilosofie van de sociale en culturele wetenschappen bij de faculteit gedrags- en maatschappijwetenschappen en de faculteit godsdienstwetenschappen van de Universiteit Groningen. Hij was vakdidacticus economie bij de Universiteit Twente. Zijn promotieonderzoek (2007) betrof de vorming van theoretische concepten in de theoretische economie en hun rol in de verklarende progressie van deze discipline.