BeleidsonderzoekOnline

DOI: 10.5553/BO/221335502018000008001

Maryland 2.0: Evidence-based onderzoek en de Maryland-schaal

Jos Mevissen

Aanbevolen citeerwijze bij dit artikel
Jos Mevissen, 'Maryland 2.0: Evidence-based onderzoek en de Maryland-schaal', *Beleidsonderzoek Online* november 2018, DOI: 10.5553/BO/221335502018000008001

Evidence-based evaluation, overal kom je de laatste jaren deze term tegen. Iedereen die onderzoek uitzet, ongeacht of het een kleine ex post evaluatie van een pilot betreft of een omvangrijke on going evaluatie van een nieuw instrument, vindt dat die evaluatie evidence-based moet zijn, 'want dan is het goed'. Wat de term feitelijk betekent, en wat de toepassing van het principe betekent voor een onderzoek, is veel minder bekend.

Ervaren evaluatoren weten dat 'evidence' verschillende gradaties kent. Vaak wordt daarbij verwezen naar de zogeheten Maryland Scientific Methods Scale (SMS), in 2002 geïntroduceerd door Farrington et al. in een studie naar evidence-based misdaadpreventie. Met de SMS kan de kwaliteit van een onderzoek worden bepaald langs een schaal van vijf gradaties: van niveau 1 (slechtste kwaliteit) tot 5 (optimaal).

Een optimaal onderzoek is volgens deze schaal een kwantitatief onderzoek naar het effect van een beleidsmaatregel door gebruik te maken van een random samengestelde onderzoeks- en controlegroep, een zogenoemde 'randomized controlled trial' (rct-studie). Het laagste niveau, het slechtste onderzoek, is volgens deze schaal een onderzoek naar de samenhang tussen een beleidsmaatregel en een beoogde uitkomst, zonder dat deze kan worden gemeten en zonder gebruik te maken van een controlegroep of van gerandomiseerde toepassing van het instrument. Wat door gebruikers van deze Maryland-schaal meestal uit het oog wordt verloren, is dat het beleidsvraagstuk waarop deze is gebaseerd tamelijk eendimensionaal was: loont het geven van meer straf om tot minder overtredingen te komen?

Nu is algemeen bekend dat een rct geen een-tweetje is. Deze vorm van onderzoek is relatief kostbaar, duurt relatief lang en vergt veel organisatievermogen van de onderzoekers (oplossen van praktijkproblemen, overredingskracht om medewerking te krijgen, controle op uitvoering, etc.) en de uitvoerders van het betreffende beleid (afwijkende procedures, extra workload, minder/geen discretionaire ruimte, etc.). Maar dan heb je ook wat: optimaal onderzoek en onomstotelijke conclusies, direct bruikbaar voor beleid.

Maar is dat wel zo? Ik geef drie redenen om daaraan sterk te twijfelen:

1. De bruikbaarheid in de praktijk van sociaal beleid is betrekkelijk. In het verleden zijn al kritische kanttekeningen geplaatst bij de gedachte dat de uitkomsten van rct's zo bruikbaar zijn voor beleid, omdat ze gemakkelijk leiden tot statistische fouten van het type II. Dat wil zeggen dat ze niet in staat zijn om foutieve 0-hypothesen (lees: uitgangspunten, logic models, beleidstheorieën) waarop de rct gericht is, als foutief te herkennen (Hope, 2005). De reden hiervoor ligt volgens mij in het gegeven dat met een rct het (gewenste) effect van één specifiek beleidsinstrument moet worden gemeten. Of dat beleidsinstrument een goed of het beste instrument is om dat doel te bereiken wordt niet onderzocht, evenmin als de achtergronden van het al dan niet optreden van gewenste en ongewenste effecten. Meer in het algemeen is causaliteit in een beleidstheorie een ingewikkelde kwestie, en als verschillende effecten op elkaar inwerken binnen een rct, wordt het experimentdesign zeer ingewikkeld.

2. Ook het instrument rct als zodanig is bekritiseerd. Bijvoorbeeld vanwege de kosten ervan of het feit dat men bepaalde personen iets onthoudt wat andere wel krijgen, of omdat de uitvoering van een experiment toch vaak weer anders verloopt dan verwacht, waardoor het niet meer een zuiver experiment is.[1] In dit verband geven onderzoekers van sociaaleconomisch beleid er zich te weinig rekenschap van dat rct's ontwikkeld zijn in het medische domein ten behoeve van het testen van een medicijn in een zeer gecontroleerde omgeving.[2] Omdat sociaaleconomisch beleid niet uitgevoerd wordt in een klinische omgeving, is de bruikbaarheid van de uitkomsten in het geding. Na introductie van een 'beproefd' beleidsinstrument kunnen de implementatie en toepassing ervan niet meer zo worden gecontroleerd als tijdens het experiment, of is de werkelijkheid inmiddels alweer een geheel andere dan tijdens het experiment.
   Op deze en andere kritische punten bij een rct wordt door dedicated rct-onderzoekers vaak gereageerd met aanpassingen in de gehanteerde econometrische analysemethodieken (introductie van proxy's, correctie voor *unobserved heterogenity*

of correcties voor multicollineariteit). Aanpassingen die voor de leek niet te begrijpen en dus niet te controleren zijn op hun betekenis en daarom maar als gegeven c.q. oplossing worden geaccepteerd.
Tot hier heb ik niets nieuws verteld, althans voor de goed ingevoerde lezers. Dat geldt niet voor het derde argument om te twijfelen aan de bruikbaarheid van rct's voor beleidsonderzoek.

3. Rct's doen onder voor meta-evaluaties. Recent liep ik tegen een interessante meta-evaluatie over actief arbeidsmarktbeleid aan. Ik las het artikel als arbeidsmarktonderzoeker, maar in dit artikel werd mijn aandacht vooral getrokken door een methodologisch tekstdeel, waarin de volgende conclusie is te lezen: 'We find that the estimated impacts derived from randomized controlled trials, which account for one-fifth of our sample, are not much different on average from the nonexperimental estimates' (Card et al., 2018). Ik teken hierbij aan dat het niet gaat om een vergelijking van 20 of 30 studies, maar om 207 studies uit verschillende landen (dus met een grote variatie in tijd, plaats en variabelen). Deze conclusie is meer dan de hiervoor geschetste kritiek de bijl aan de wortel van de rct. Waarom immers zoveel geld, tijd en energie besteden aan een rct als goedkopere, korter durende en eenvoudiger uit te voeren onderzoeken tot vergelijkbare conclusies leiden?

Toen ik deze overdenkingen deelde met een bevriend methodoloog, vertelde hij mij over onderzoek naar (de beïnvloedbaarheid van) rokersgedrag. Men overwoog een rct uit te gaan voeren, maar kwam tot de conclusie dat de tientallen onderzoeken die over de hele wereld al daarnaar waren uitgevoerd, voldoende 'bewijs' opleverden over welke factoren op dat gedrag van invloed waren. Wat uit deze anekdote voor mij spreekt, is dat men in beleidsevaluaties, ook als die economisch is ingestoken, wat meer los moet komen van het idee dat meten het begin is van betrouwbare kennis. Het begin van betrouwbare kennis over beleid is het doorgronden van de manier waarop dat beleid doorwerkt op gedrag van doelgroepen van beleid en van de factoren die daarop van invloed zijn. Met andere woorden: een echte, goede beleidsevaluatie moet zijn gebaseerd op de beginselen van de *realistic evaluation* (zie Pawson & Tilley, 1997) en begint met een gedegen en onderbouwde interventielogica. Daarbij gaat het niet alleen om de uitkomsten van beleid in een gecontroleerde context. Om een realistische evaluatie te kunnen uitvoeren moet men zich er rekenschap van geven dat er niet één in een experiment geïsoleerde context is en dat vooral de mechanismen moeten worden gekend die van invloed zijn op hoe beleid uitpakt. Het blootleggen van de relaties tussen contexten, mechanismen en uitkomsten rond beleid vergt een breder palet aan methoden en technieken dan rct's om tot valide en bruikbare data te komen. Gebeurt dat niet en is rct een doel op zich,

dan is kwantificering alleen een vorm van camouflage van het niet écht willen kennen van de werkelijkheid. Wordt het misschien tijd voor een Maryland Scientific Methods Scale 2.0?

Literatuur

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association, 16*(3), 894-931. DOI:10.1093/jeea/jvx028

De Koning, J., Mallee, L., De Hek, P., & Groenewoud, M. (2015). *Zijn experimenten een panacee voor de effectiviteitsmeting van re-integratiemaatregelen?* Paper voor de Nederlandse Arbeidsmarkt Dag 2015 (zie www.arbeidsconferentie.nl).

Farrington, D., Gottfredson, D., Sherman, L.W., & Welsh, B. (2002). Maryland Scientific Methods Scale. In L.W. Sherman et al. (Eds.), *Evidence-based crime prevention* (p. 13-21). London: Routledge.

Hope, T. (2005). Pretend it doesn't work: The 'anti-social' bias in the Maryland Scientific Methods Scale. *European Journal on Criminal Policy and Research, 11*(3-4), 275-296.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Noten

1 Zie voor kanttekeningen bij experimenten onder andere: De Koning et al. (2015).

2 Overigens is er ook binnen het medische domein kritiek op rct's.