



HET ONDERZOEKSDESIGN

Op de TIER-BEE site staan studies die laten zien wat wel en niet effectief is in het onderwijs. Het onderzoeksdesign van de studies is een belangrijk selectie criterium. We tonen vooral studies die overtuigend laten zien of iets wel of niet werkt. Het bewijs (de 'evidence') staat centraal, wat betekent dat we alleen studies met een hoge interne validiteit tonen. Andere 'best evidence' of 'what works' initiatieven kennen vergelijkbare design eisen (zie bijvoorbeeld Slavin, 2008).

Wat betekent dit nu? Wanneer bevat een studie overtuigend bewijs? Dit is over het algemeen het geval in studies waar causale effecten worden getoetst. Soms gebeurt dit met een experiment, soms met een quasi-experiment en soms met een design waarbij overtuigende controlegroepen op een andere manier zijn gereconstrueerd. In deze notitie leggen we uit welke onderzoeksdesigns gericht zijn op het toetsen van causale effecten en verantwoorden we waarom we voor deze designs kiezen.

Het principe: een vergelijkbare treatment en controlegroep

Studies die zich richten op het toetsen van causale effecten kijken of iets werkt bij een groep leerlingen of scholen. Vaak doen ze dat door te toetsen of de uitkomsten voor leerlingen of scholen waar een effect verwacht wordt (de experimentele of 'treatment' groep) verschillen van de uitkomsten bij een groep leerlingen of scholen waar geen effect verwacht wordt (de controle groep). De mate waarin de uitkomsten van de treatmentgroep verschillen van die van de controlegroep wordt hierbij gezien als het effect. Cruciaal is dat effecten gemeten worden bij vergelijkbare treatment- en controlegroepen. Er mag dus geen vertekening optreden doordat beide groepen van elkaar verschillen.

Experimentele designs

Het experimentele design is de beste garantie voor een vergelijkbare treatment en controlegroep. Toewijzing van personen of organisaties aan de experimentele of controlegroep gebeurt in experimenten over het algemeen willekeurig ('random'). Experimenten worden hierom 'random control trials' genoemd. Door de random toewijzing zijn experimentele en controlegroep in principe gelijk aan elkaar, zowel waar het gaat om observeerbare als om niet-observeerbare kenmerken. Een betere garantie voor vergelijkbaarheid tussen de twee groepen bestaat er niet. Vandaar dat het experimentele design door sommige wetenschappers de 'gouden standaard' wordt genoemd.

Experimenteel onderzoek is beter in staat causale effecten te toetsen dan andere onderzoeksdesigns, maar weinig onderwijsonderzoek is experimenteel. Experimenten zijn

soms eenvoudigweg niet mogelijk, soms niet praktisch en soms onethisch. Cook (2002) en Schlotter, Schwert & Woessmann (2009) geven een aardig overzicht van mogelijke knelpunten in experimenteel onderzoek. Deze zijn terug te voeren op drie hoofdpunten, namelijk:

- *de kwaliteit van het experiment en de uitvoering*: zorgpunten zijn meetfouten, een imperfecte uitvoering en het feit dat alleen al deelnemen aan een experiment soms een ander gedrag uitlokt (het zogenaamde Hawthorne-effect).
- *de externe validiteit*: experimenteel onderzoek kent soms gebrekkige generalisatiemogelijkheden (zeker bij labexperimenten) en een gebrek aan opschaalbaarheid (bijvoorbeeld omdat de treatments te kostbaar zijn).
- *niet altijd kunnen toetsen van wat we echt willen weten*: knelpunten zijn het gebrek aan een programma theorie, het niet altijd kunnen toetsen van de onderliggende mechanismen en het slecht in staat zijn vragen van scholen te beantwoorden.

Designs met vergelijkbare groepen

De afgelopen twee decennia zijn er verschillende designs ontwikkeld die op een andere manier garanderen dat de vergelijkingsgroepen identiek zijn. Deze designs zijn ook in staat gebleken effecten te isoleren van toevallige verschillen tussen de treatment en controlegroep. We doelen hier op studies met een hoge interne validiteit (Shadish et. al, 2002) en studies die op diverse methodologie schalen hoog gewaardeerd worden (op de Maryland Scientific Methods Scale tenminste op niveau 3 zitten (MSMS3+)).

Quasi-experimenteel onderzoek is onderzoek dat de experimentele setting zo goed mogelijk nabootst. In quasi-experimentele designs wordt op een andere manier dan random toewijzing gegarandeerd dat de treatment en controlegroep identiek zijn. Shadish, Cook and Campbell (2002) onderscheiden verschillende methoden van quasi-experimenteel onderzoek, waaronder:

- *de natuurlijke experimenten*: door een 'natuurlijke' gebeurtenis is sprake van min of meer randomisatie over een treatment en controle groep (zie bijvoorbeeld Webbink, 2008)
- het '*regressie discontinuïteits' design*: hierbij worden discontinuïteiten in (beleids)regels gebruikt om een vergelijkbare treatment en controlegroep te construeren (zie bijvoorbeeld Leuven et al, 2007)
- *instrumentele variabele design*: hier wordt gebruik gemaakt van een (instrumentele) variabele die de treatment en controlegroep onderscheidt, maar los staat van de treatment en het te meten effect

Naast deze quasi-experimentele studies, wordt *matching* vaak gebruikt om vergelijkbare treatment en controlegroepen te creëren. Bij matching creëert de onderzoeker een controlegroep zodat deze op verschillende kenmerken identiek is aan de treatmentgroep. De resultaten van dit design zijn vaak vergelijkbaar met die van een experimenteel design (zie Lalonde, 1986). Bij matching is het wel relevant dat de controle en treatmentgroep overtuigend aan elkaar zijn.

Een *pretest-posttest design* wordt ook vaak gezien als acceptabel alternatief als het gaat om het meten van causale effecten. Panelstudies zijn regelmatig in staat een dergelijk design te hanteren. Cruciaal daarbij blijft natuurlijk wel dat er geen ongemeten

systematische verschillen zijn tussen de treatment- en controle-groepen. De onderzoekers moeten hierbij dus aantonen dat er geen indicaties zijn voor bias en selectie-effecten.

Meta-analyses en reviewstudies

De TIER-BEE site bevat ook enkele samenvattende studies. Hier zijn twee soorten designs gebruikelijk, namelijk de meta-analyse en de systematische reviewstudie. Beide soorten beperken zich idealiter tot studies met een onderzoeksdesign met een hoge interne validiteit. Omdat deze studies in aantal vaak gering zijn, zeker als het gaat om een specifiek thema, wordt hier in de praktijk nogal eens vanaf geweken en kiest met ervoor zoveel mogelijk studies mee te nemen. Een voorbeeld van dit laatste is Hattie (2010) die een fraaie overzichtsstudie heeft gemaakt op basis van ruim 800 meta-analyses.

References

Cook (2002) Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. In: *Educational Evaluation and Policy Analysis*, Vol. 24(3), 175-199.

Hattie, J. (2010) *Visible learning. A Synthesis of over 800 meta-analyses relating to achievement*. London/New York: Routledge.

LaLonde, R.J. (1986) Evaluating the Econometric Evaluations of Training Programs with Experimental Data. In: *The American Economic Review*, 76(4), 604-620.

Leigh, A. & Ryan, C. (2008) Estimating returns to education using different natural experiment techniques. In: *Economics of Education Review*, 27, 149-160.

Leuven, E., Lindahl, M., Oosterbeek, H. & Webbink, D. (2007) The effect of extra funding for disadvantaged students on achievement. In *Review of Economics and Statistics* 89(4). 721-736.

Schlotter, M., Schwert, G & Woessmann, L. (2009) *Methods for causal evaluation of education policies and practices; an econometric toolbox*. Analytical report for the European Commission. European Network on Economics of Education.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont: Wadsworth, Cengage Learning.

Slavin, R.E. (2008) Perspectives on Evidence Based Research in Education. What Works? Issues in Synthesizing Educational Program Evaluations. In: *Educational Researcher*, 37(1), 5-14.

Webbink, H.D. (2008) The effect of local calamities on educational achievement. In: *Disasters*, 32(4), 499-515.