Eindevaluatie Onderwijsbewijs

Samenvatting

Deze notitie geeft een evaluatie van Onderwijsbewijs, een regeling die is opgezet om experimenten in het onderwijs te genereren die overtuigende kennis opleveren over de effectiviteit van verschillende onderwijsinterventies. De regeling is gestart in 2008 en afgesloten in 2016.

De belangrijkste algemene lessen van Onderwijsbewijs zijn als volgt:

- Het uitvoeren van experimenteren in het onderwijs met random toewijzing van interventies blijkt goed mogelijk.
- De bereidheid van scholen (en onderzoekers) om deel te nemen aan experimenten is groot.
- Een behoorlijk aantal onderzoekers heeft meer ervaring op kunnen doen met experimenteel onderzoek.
- De onderzoeksresultaten bieden overtuigende en nuttige evidentie over de effectiviteit van interventies.

Meer specifieke lessen zijn de volgende:

Selectie van projecten:
- Een helder format voor indienen van onderzoeks-ideeën en voorstellen met daarop expliciet de belangrijkste vragen helpt een selectiecommissie om een beter beeld te krijgen van de voorstellen en dwingt indieners specifiek te zijn.
- Het hanteren van random toewijzing van interventies zou een go no-go beslissing moeten zijn bij de selectie.
- De expertise van onderzoekers met gecontroleerde experimenten (RCTs) zou moeten worden meegewogen bij besluitvorming over honorering.
- Het zou wenselijk zijn om experimenten meer te richten op de fundamentele en grote beleidsvragen.
- (Te) sterke verwevenheid van indieners/onderzoekers met de te beproeven interventie is een aandachtspunt. Randomisatie door een externe partij verdient aanbeveling, evenals het verplicht ter beschikking stellen van de data na afloop van het experiment.

Opzet van experimenten:
- Er dient expliciet aangegeven te worden wat de omvang van het beoogde minimale verschil in effect tussen de experiment en controle groep is.
- Een *poweranalyse* als onderdeel van het onderzoeksplan is belangrijk om vooraf expliciet te bepalen hoeveel scholen nodig zijn om dat minimale verschil in effect aan te tonen. Hierbij dient rekening gehouden te worden met clustering van leerlingen in klassen of scholen.
- Beginnen met een pilot periode en een concrete interventie die goed gepland is en bij voorkeur reeds is uitgetest zijn belangrijke succesfactoren.
- Er zou zo veel mogelijk aangesloten moeten worden bij reeds bestaande metingen i.p.v. uitsluitend werken met zelf geconstrueerde uitkomstmaten.
- Het verdient aanbeveling om niet slechts korte termijn effecten maar ook de langere termijn effecten in kaart te brengen.

Uitvoering van experimenten:
- Onderzoeken worden te optimistisch gepland waardoor het bereiken van het beoogde aantal deelnemende scholen (of klassen of leerlingen) vaak tegen valt.
- Het monitoren van de mate en kwaliteit van de uitvoering van de vergeleken interventies, in het bijzonder de experimentele interventie, is belangrijk.

- Het daadwerkelijk realiseren van voldoende vergelijkbare experiment- en controlegroepen is een aandachtspunt.
- Het afhaken van deelnemende scholen (of klassen of leerlingen) gedurende het project is een belangrijk risico.
- Een onrealistische planning en te weinig voorbereidingstijd zijn veelgenoemde bedreigingen voor adequate uitvoering van onderzoek.
- Een goed draagvlak creëren binnen de aan onderzoek deelnemende scholen en een aanvaardbare extra belasting voor deze scholen zijn belangrijke succesfactoren.
- In het speciaal onderwijs blijkt uitvoeren van experimenten lastiger dan in het reguliere onderwijs.

Effectanalyse en follow-up

- Het is van belang om zo mogelijk alle originele toegewezen scholen (of klassen of leerlingen) mee te nemen in de effectschattingen, ook die tussentijds reeds zijn afgehaakt of uitgevallen.
- Er dient bij de effectanalyse rekening gehouden te worden met clustering van leerlingen in klassen of scholen. De standaardfouten van de effectschattingen worden hierdoor over het algemeen groter.
- In totaal zijn 37 experimenten uitgevoerd die een veelheid aan resultaten hebben opgeleverd. Deze resultaten zijn te vinden op de website van Onderwijsbewijs.
- De resultaten van de experimenten zijn wisselend. Er zijn ongeveer evenveel experimenten die een duidelijk effect laten zien als experimenten waar geen effect is gevonden. Daarnaast is er een wat kleinere groep waarvan hele kleine of andere dan verwachte effecten zijn aangetoond. Alle thema's kennen zowel effectieve als niet effectieve interventies.

Tot slot

- Het zou nuttig zijn als er een leidraad komt voor het opzetten en evalueren van experimenten op het gebied van onderwijs.
- Borging van de onderzoeksuitkomsten van de experimenten in een voor het veld toegankelijke database/website is van belang.
- Onderwijsbewijs heeft laten dat klassieke experimenten mogelijk zijn in het onderwijs en heeft een veelheid aan informatie opgeleverd over kleinere interventies. Experimenten met grotere meer fundamentele beleidsvragen waren binnen Onderwijsbewijs nog niet mogelijk. Als vervolg op Onderwijsbewijs zou gezocht kunnen worden naar mogelijkheden om dit wel te realiseren.

Inleiding

Deze notitie geeft een evaluatie van Onderwijsbewijs op verzoek van OCW. Onderwijsbewijs is een regeling van OCW die als doel heeft om "via wetenschappelijke experimenten kennis te verzamelen over wat wel werkt en niet werkt in het onderwijs". Daarvoor konden onderzoekers en scholen allianties aangaan om experimenten te ontwikkelen en uit te voeren en de effectiviteit van interventies te toetsen.

Binnen Onderwijsbewijs zijn in twee tranches in totaal 37 projecten gehonoreerd. Hiermee is ruim 20 miljoen euro aan middelen gemoeid.[1] De eerste tranche aanvragen vond plaats eind 2008, de tweede tranche medio 2010. Per tranche is gekozen voor een aparte lijst met thema's waarop voorstellen konden worden ingediend. Bij de eerste ronde ging het om de thema's leerlijn taal-rekenen, lerarentekort, hoogbegaafdheid, jeugdzorg, leerlijn voor vve, en verbetering leesonderwijs speciaal onderwijs. Bij de tweede ronde ging het om de thema's dagindeling en opvang, excellentie, burgerschap, gedragsproblemen en pesten, en vermindering van achterstanden. De regeling is opgezet in de vorm van een prijsvraag. Een jury selecteerde de meest veelbelovende voorstellen. Bij de eerste tranche zijn 112 voorstellen binnengekomen, waarvan 18 gehonoreerd. Bij de tweede tranche ging het om 74 voorstellen en 19 honoreringen. De voorstellen met de meeste punten zijn gehonoreerd. De jury heeft gescoord op de criteria haalbaarheid, verwachte effectiviteit, methodologie/onderzoeksdesign en mogelijkheden tot opschaling. Het gaat meestal om driejarige experimenten.

In deze evaluatie verschaffen we inzicht in de belangrijkste lessen uit Onderwijsbewijs. De evaluatie is gebaseerd op ervaringen gedurende de looptijd van Onderwijsbewijs met diverse projecten, informatie uit voortgangsrapportages, meerdere besprekingen van de voortgang van alle projecten binnen de begeleidingscommissie vanaf de start van Onderwijsbewijs, presentaties van indieners bij de Onderwijsresearchdagen, een enquête onder begeleidingscommissieleden, een enquête onder de indieners, gesprekken met enkele begeleidingscommissieleden (Dinand Webbink en Lex Borghans) en met de verantwoordelijke vanuit Dienst Uitvoering Onderwijs (Imro Simmelink). De enquêtes zijn uitgezet in het najaar van 2012. De enquête onder indieners heeft een response opgeleverd van 33 uit 37 op projectniveau. Het betreft een response van 90 procent. Bij de begeleidingscommissie heeft iedereen de enquête ingevuld.
De lessen zijn gebaseerd op wat het vaakst of het meest nadrukkelijk genoemd is, waarbij ervaringen en constateringen van de begeleidingscommissieleden relatief zwaar hebben meegewogen. In de volgende paragrafen noemen we alleen de belangrijkste lessen.

De opzet van de rest van dit document is als volgt. Paragraaf 2 geeft de lessen ten aanzien van de selectie van de projecten. Paragraaf 3 gaat in op de opzet van de experimenten. Paragraaf 4 evalueert de uitvoering van de experimenten. Paragraaf 5 zet de belangrijkste aandachtspunten bij de effectanalyse op een rij. Paragraaf 6 sluit af met enkele noties.

[1] Het grootste toegekende bedrag aan een voorstel bedroeg 1.5 miljoen euro, het kleinste 150 duizend euro.

Selectie projecten

- Een helder format voor indienen met de belangrijkste vragen helpt een selectiecommissie om een beter beeld te krijgen van de voorstellen en dwingt indieners specifiek te zijn. Bij de eerste aanvraagronde stond er bijvoorbeeld nog geen blokje in waarin gevraagd werd naar de poweranalyse. Het feit dat dit er bij de tweede ronde wel in stond heeft waarschijnlijk ertoe bijgedragen dat toen veel meer indieners een poweranalyse hebben uitgevoerd. Andere belangrijke vragen waren naar aantallen betrokken scholen en leerlingen, naar bewijzen van reeds aanwezig commitment om mee te doen, naar de wijze van randomisatie, en naar de inschatting van de belangrijkste risico's en hoe men deze denkt te gaan ondervangen.

- Het zou ter bepaling van wetenschappelijke onderbouwde beleidsbeslissingen op gebied van onderwijs wenselijk zijn om experimenten meer te richten op de fundamentele en grote beleidsvragen. Nu hadden veel experimenten maar relatief beperkte reikwijdte omdat ze gericht waren op betrekkelijk gedetailleerde interventies en vragen op een soms heel specifiek gebied.

- Random assignment zou een go no-go beslissing moeten zijn bij selectie van experimenten en bij de besluitvorming over de doorgang van experimenten. Sommige projecten zijn uiteindelijk toch uitgemond in een controlegroep op basis van matching, wat een veel minder stevig onderzoeksdesign oplevert.[2]

- Het zou goed zijn als de expertise van onderzoekers met het opzetten, uitvoeren en evalueren van experimenten wordt meegewogen bij besluitvorming over honorering. Deze expertise blijkt vooralsnog relatief schaars in Nederland en onderzoekers blijken regelmatig begeleiding nodig te hebben bij de uitvoering van een experiment. Er was (met name in de eerste ronde) een redelijk aantal projecten met problemen met de randomisatie, de power (voldoende aantallen deelnemende scholen/klassen/leerlingen), de statistische analyses en met verstorende invloeden op het onderzoeksdesign. Deze problemen konden soms gedurende de rit nog hersteld worden, maar soms ook niet meer. 4 op de 10 indieners had voorafgaande aan Onderwijsbewijs nog geen actieve betrokkenheid gehad bij een gecontroleerd experiment (RCT).

- Er is geconstateerd dat onderzoekers soms erg verweven zijn met de te beproeven interventie en een sterk geloof hebben in de positieve werking ervan. Soms spelen ook commerciële belangen een rol, bijvoorbeeld bij een ICT-methode die op de markt gebracht kan worden. Dat is op zichzelf geen reden om projecten niet te selecteren, maar vereist wel extra kritische aandacht. Het verdient bijvoorbeeld aanbeveling om de loting zeker in dat soort gevallen door een externe partij te laten verrichten. Het is overigens regelmatig gebeurd dat het CPB op verzoek van onderzoekers de loting heeft verricht. Ook zou het goed zijn als er een verplichting komt de data na afloop van het project ter beschikking te stellen. Idealiter zijn degenen die de interventie hebben ontwikkeld niet direct betrokken bij de evaluatie ervan.

Opzet experimenten

- Een degelijke poweranalyse is nodig om te beoordelen of men met het beoogde aantal scholen wel de verwachte effecten kan aantonen. Dit vereist een realistische en onderbouwde inschatting van

[2] Agodini en Dynarksi (2001) bijvoorbeeld onderzoeken of matching technieken, waarbij behandelde leerlingen gematcht worden met niet-behandelde leerlingen met vergelijkbare karakteristieken, niet-vertekende effectschattingen oplevert. Ze doen dit door experimentele schattingen te vergelijken met schattingen op basis van matchingtechnieken (propensity score matching). Ze vinden geen consistent bewijs dat schattingen op basis van matchingtechnieken de experimentele schattingen repliceren, zelfs als er veel data beschikbaar zijn om op te matchen. De auteurs concluderen dat niet-geobserveerde factoren een grote impact kunnen hebben op de uitkomsten, waarmee lastig valt om te gaan met niet-experimentele schattingsmethoden zoals matching. (R. Agodini & M. Dynarksi, 2001, Are experiments the only option? A look at dropout prevention programs, Mathematica Policy Research, Princeton (New York).

verwachte effectgroottes, bijvoorbeeld op basis van literatuur waarin vergelijkbare interventies in het buitenland of een andere onderwijssector zijn geëvalueerd. De indruk bestaat dat effectgroottes soms te optimistisch zijn ingeschat. Respondenten geven vijf keer zo vaak aan dat de gevonden effecten kleiner zijn dan vooraf ingeschat dan dat ze aangeven dat de effecten groter zijn dan vooraf ingeschat. Ongeveer driekwart van de respondenten op een evaluatie-enquête onder indieners geeft aan vooraf een poweranalyse te hebben uitgevoerd. De kwaliteit van de uitgevoerde poweranalyses was wisselend. Uit de poweranalyses volgen minimaal benodigde aantallen deelnemers. Het is verstandig om rekening te houden met enige uitval van scholen gedurende het experiment en meer deelnemers te werven, zodat men niet al bij de minste geringste uitval in de problemen komt met de power.

- Beginnen met een pilotperiode komt de kwaliteit van de experimenten ten goede, zeker als het interventies betreft die nog in ontwikkeling zijn of zelfs nog helemaal niet beproefd zijn. Het zorgt ervoor dat allerlei kinderziektes kunnen worden verwijderd, zowel rondom de interventie als rondom de meetinstrumenten. Ongeveer tweederde van de indieners heeft een pilotfase ingebouwd in het project. Een van de meest genoemde kritische succesfactoren door indieners betreft een uitgekristralliseerde goed geplande interventie met een pilotfase om deze nog uit te proberen.

- Het verdient aanbeveling zo veel mogelijk aan te sluiten bij reeds bestaande metingen (reguliere toetsen, doorstroomcijfers, etc) in plaats van eigen/extra uitkomstmaten. Aansluiten bij bestaande metingen zorgt voor een hogere response, levert minder extra lasten op voor de school en de resultaten sluiten dan ook direct aan bij uitkomstmaten waar de scholen zich op richten. Beter een paar uitkomstmaten goed en volledig gemeten dan een hele batterij half gemeten.

- Het is belangrijk om zo mogelijk ook de wat langere termijn effecten in kaart te brengen (in plaats van alleen direct na afloop van de interventie). Hiermee kan meer inzicht worden verkregen in de persistentie van de eventuele gevonden effecten. Het uitdoven van effecten is een veelvoorkomend fenomeen bij onderwijsinterventies. Andersom zijn er ook voorbeelden van interventies die op de korte termijn geen of beperkte effecten laten zien, maar op de wat langere termijn wel (zie bijvoorbeeld Taylor en Tyler, 2011[3]). 60 procent van de indieners geeft aan ook langere termijn effecten te onderzoeken (minstens een half jaar na afloop van de interventieperiode).

Uitvoering experimenten

- Het behalen van het beoogde aantal deelnemende scholen / klassen / leerlingen valt soms tegen. Zowel commissieleden als indieners geven dit aan. Dit wordt door indieners gezien als gemiddeld het belangrijkste probleem dat men is tegengekomen. Een derde van de indieners beschouwt dit als een zeer groot probleem (score 8-10 op een 1-10 schaal).

- Het monitoren van de mate en kwaliteit van de uitvoering van de experimentele interventie is belangrijk. Hiermee kan meer inzicht worden verkregen in de vraag of een experiment geen effect oplevert omdat de interventie op zichzelf niet effectief is, of omdat de interventie gebrekkig is uitgevoerd of zelfs helemaal niet is uitgevoerd. Vier op de tien indieners geven aan ten minste enige problemen te hebben ondervonden met interventies die niet werden uitgevoerd zoals beoogd/afgesproken.[4] Monitoring kan er ook toe bijdragen dat er tussentijds nog kan worden bijgestuurd, zodat de interventie wel wordt uitgevoerd zoals beoogd. Dat is bij diverse projecten het geval geweest.

- Het daadwerkelijk realiseren van voldoende vergelijkbare experiment- en controlegroepen is in sommige gevallen een probleem. De helft van de indieners geeft aan tenminste enig probleem

[3] Taylor, E.., en J. Tyler, 2011, The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers, NBER Working Paper no. 16877

[4] We definiëren tenminste enige problemen als een indiener een score geeft van minimaal 4 op een schaal van 1 tot 10.

hiermee te hebben. Twintig procent van de indieners heeft hier zelfs zeer grote problemen mee ondervonden. Een mogelijkheid om dit te ondervangen is een zogenoemde gestratificeerde loting, waarbij steeds binnen subgroepen (bijvoorbeeld naar startniveau van leerlingen of naar onderwijstype als een experiment met meerdere onderwijstypen) in gelijke mate leerlingen of scholen worden toebedeeld aan de experimentele groep of de controlegroep. Bij experimenten met beperkte hoeveelheden deelnemende scholen of leerlingen wordt de kans in zijn algemeenheid groter dat de groepen minder goed vergelijkbaar zijn, ook al wordt er gerandomiseerd.

- Afhaken van deelnemende scholen is een belangrijk risico. Circa de helft van de indieners geeft aan hier problemen van enige omvang mee te hebben ondervonden. Tien procent geeft aan hier zeer grote problemen mee te hebben ondervonden. Dit afhaken verlaagt niet alleen de power van het onderzoek, maar kan ook nog eens selectief zijn (bv de weinig gemotiveerde of slechtere scholen haken af). Dit kan er voor zorgen dat experimentele en controlegroepen die bij aanvang nog goed vergelijkbaar waren, dit niet meer zijn als de eindmetingen worden verricht onder de scholen die nog zijn overgebleven. Met andere woorden, het experimentele design waarin juist het uitgangspunt was om gelijke groepen te creëren wordt bedreigd. Redenen voor uitval die genoemd worden zijn onder meer te veel inspanningen die gevraagd worden van scholen en te weinig betrokkenheid bij het project. De belangrijkste succesfactoren die door indieners zijn genoemd, kunnen dit risico waarschijnlijk inperken. Het gaat dan om een goed draagvlak creëren bij deelnemende scholen (niet alleen bij schoolleiders, maar ook bij leerkrachten en andere betrokkenen), een interventie die aansluit bij de problemen die op de school leven, een geringe extra belasting van het experiment voor leerkrachten en scholen, duidelijke communicatie over wanneer wat verwacht wordt en zichtbare opbrengsten voor de deelnemers. Dit laatste is vooral van belang voor de controlegroepen, die geen speciale interventie krijgen aangeboden.

- Een onrealistische planning en te weinig voorbereidingstijd wordt als een belangrijke bedreiging gezien. Ruim twintig procent van de indieners geeft aan dit als 1 van de 3 belangrijkste bedreigingen te zien. Problemen die we relatief vaak zijn tegengekomen zijn het niet tijdig kunnen starten met nulmetingen, waardoor nulmetingen nog moeten worden afgenomen als de interventieperiode al gestart is.

- Een goed draagvlak creëren en een aanvaardbare extra belasting voor de scholen is een belangrijke succesfactor. Scholen zijn soms afgehaakt vanwege het gevoel te veel inspanning te moeten leveren. Dit kan voorkomen worden door kritisch te beoordelen of de meerwaarde van meerdere of lange vragenlijsten of metingen opweegt tegen de extra lasten, duidelijk te communiceren naar scholen vooraf wat er wanneer van hen verwacht wordt, en zo veel mogelijk aan te sluiten bij metingen die toch al verricht worden. Ook kan het helpen als leraren bijvoorbeeld een kleine attentie of cadeaubon krijgen voor hun medewerking, met name als er een grotere inspanning van hen wordt verwacht.

- Een nulmeting op de uitkomstvariabelen kan de precisie van de effectschattingen behoorlijk vergroten en zorgt ervoor dat minder deelnemers nodig zijn om effecten van een bepaalde omvang aan te tonen dan wanneer geen nulmeting wordt verricht (of dat met hetzelfde aantal deelnemers effecten van kleinere omvang kunnen worden aangetoond).

- In het speciaal onderwijs blijkt uitvoeren van experimenten lastiger dan in het reguliere onderwijs. Er is relatief veel meer verloop van leerlingen, verloop en verzuim van leraren, er zijn moeilijkheden met het afnemen van testen en het uitvoeren van observaties in verband met afwezigheid en concentratieproblemen. Ook is er meer heterogeniteit in de leerlingpopulaties tussen klassen en scholen, waardoor de statistische power bij een gelijk aantal scholen of klassen lager is dan in het reguliere onderwijs.

Effectanalyse en follow-up

- **Het is van belang om zo mogelijk alle originele toegewezen scholen mee te nemen in de effectschattingen, ook de scholen (of klassen of leerlingen) die tussentijds zijn afgehaakt. Er bestaan dan econometrische technieken om toch het effect van het ontvangen van de interventie goed te schatten. Mochten van de afhakers geen uitkomsten meer te achterhalen zijn, dan dient tenminste melding te worden gemaakt van hoeveel er zijn afgehaakt. Ook kan op basis van de nulmetingen en andere relevante kenmerken worden onderzocht in hoeverre er sprake is van selectieve uitval (of non-response). Men kan dan op basis daarvan uitspraken doen of de schattingen een over- of onderschatting geven van het werkelijke effect van de interventie.**

- **Er dient rekening gehouden te worden met clustering van leerlingen in klassen of scholen. Leerlingen in klassen of scholen kunnen niet als onafhankelijke waarnemingen worden beschouwd. Er kunnen gemeenschappelijke factoren zijn (zoals bijvoorbeeld de kwaliteit van de leraar of het gemiddelde aanvangsniveau van de klas) die invloed hebben op de prestaties van leerlingen die in dezelfde klas of school zitten. Correctie voor clustering leidt tot andere standaardfouten van de effectschattingen. Dit zorgt over het algemeen voor grotere standaardfouten (en dus zijn effecten minder snel statistisch significant).**

- In totaal zijn 37 experimenten uitgevoerd die een veelheid aan resultaten hebben opgeleverd. Deze resultaten zijn te vinden op de website van Onderwijsbewijs. De resultaten zijn ook gepresenteerd op wetenschappelijke conferentie en gepubliceerd in wetenschappelijk tijdschriften.
- **De resultaten van de experimenten zijn wisselend. Er zijn ongeveer evenveel experimenten die een duidelijk effect laten zien als experimenten waar geen effect is gevonden. Daarnaast is er een wat kleinere groep waarvan hele kleine of andere dan verwachte effecten zijn aangetoond. Alle thema's kennen zowel effectieve als niet effectieve interventies.**

Tot slot

Onderwijsbewijs is een regeling die goed past binnen het streven naar evidence based beleid. In een paar jaar tijd hebben tientallen onderzoekers en honderden scholen meer ervaring opgedaan met het opzetten, uitvoeren en evalueren van experimenten op een groot aantal uiteenlopende themas. Deze evaluatie laat zien dat gecontroleerde experimenten goed mogelijk zijn in het onderwijs en dat er een behoorlijke belangstelling voor is onder onderzoekers en scholen. Indieners geven echter ook in grote meerderheid aan dat gerandomiseerde experimenten veel nieuwe inzichten kunnen opleveren over wat werkt en niet werkt in het onderwijs en dat hun expertise op het vlak van gecontroleerde experimenten is vergroot. Het voordeel van experimenten met random toewijzing weegt voor de meeste indieners (80 procent) op tegen de nadelen, zoals de extra inspanningen die moeten worden gedaan om te kunnen loten en om verstoringen van het experimentele design te voorkomen.[5] Eenzelfde percentage is van mening dat het feit dat bij experimenten effecten geloofwaardig kunnen worden vastgesteld een positieve invloed heeft op de acceptatie en verspreiding van de onderzochte interventies. De helft van de indieners is van plan de komende drie jaar opnieuw een of meerdere gecontroleerde experimenten op te gaan zetten (en 15 procent niet), wat erop duidt dat Onderwijsbewijs een duidelijke spinoff heeft richting meer experimenteel onderwijsonderzoek in Nederland.

Tegenover deze positieve constateringen staat dat een succesvol experiment uitvoeren nog niet zo eenvoudig is in de praktijk. Dit geldt zowel voor het ontwerpen, uitvoeren als evalueren van experimenten. Indieners beamen dit ook.

[5] Merk op dat bij een meer kwalitatieve aanpak ook extra inspanningen komen kijken, zoals het organiseren en uitwerken van case studies, interviews, etcetera. De stelling doet geen uitspraken over welk type onderzoek per saldo meer inspanning vergt.

Een handzame en toegankelijke leidraad voor experimenteel onderzoek zou nuttig kunnen zijn, waar onderzoekers en scholen gebruik van kunnen maken als ze een experiment aan het voorbereiden, uitvoeren en evalueren zijn. Enkele aandachtspunten bij het opzetten van experimenten zijn al wel verschenen in buitenlandse papers, zie bijvoorbeeld List et al., 2010[6]. Ook valt te denken aan gerichte sessies waarbij uitvoerders van experimenten met elkaar in contact kunnen treden om te leren van elkaars ervaringen, zowel de valkuilen als de succesfactoren.[7] Het lijkt verstandig om bij het toekennen van onderzoeksgeld voor experimenten meer aandacht te besteden aan de uitvoering in het veld. De vertaling van het idee van de onderzoeker naar de scholen is vaak ingewikkeld voor beide partijen. Zonder goede uitvoering heeft een effectmeting weinig zin en leidt dit tot een inefficiënt gebruik van middelen. Het vrijmaken van middelen binnen de begroting voor het aanstellen van een persoon die de dagelijkse gang van zaken monitort lijkt nuttig. In de medische wetenschappen is dit een gebruikelijke gang van zaken, want de arts kijkt niet elke dag mee of de patiënten wel de juiste acties uitvoeren. Daar heeft de arts assistenten voor, die zowel kennis hebben van het veld als het onderzoek. Bij experimenteel onderzoek in het onderwijs is het ook aan te bevelen een of meerdere personen met een dergelijke rol aan te stellen, afhankelijk van de grootte van het experiment.

Voor de verspreiding en borging van de opgedane kennis is opname van de uitkomsten in een database waardevol. Het Nederlands Jeugdinstituut heeft een overzicht van meerdere interventies gericht op jeugd, niet specifiek alleen onderwijsinterventies (http://www.nji.nl/watwerkt). In het buitenland zijn hier ook al enkele voorbeelden van, zoals What Works Clearinghouse (http://ies.ed.gov/ncee/wwc/). De uitdaging is om een dergelijke website levend te houden en ook toegankelijk te maken voor het onderwijsveld, zodat er een grotere verbinding ontstaat tussen onderwijsonderzoek en het onderwijsveld. Ook verdient het aanbeveling strikte eisen te hanteren aan de onderzoeksdesigns voordat studies mogen worden opgenomen in de database, zodat alleen de studies met een geloofwaardig onderzoeksdesign erin terecht komen.

Onderwijsbewijs heeft laten dat klassieke experimenten mogelijk zijn in het onderwijs en heeft een veelheid aan informatie opgeleverd over kleinere interventies. Experimenten met grotere meer fundamentele beleidsvragen waren binnen Onderwijsbewijs nog niet mogelijk. Als vervolg op Onderwijsbewijs zou gezocht kunnen worden naar mogelijkheden om dit wel te realiseren.

[6] List, J., S. Sadoff, en M. Wagner, 2010, So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design, NBER Working Paper nr. 15701.

[7] In 2010 heeft een grote conferentie “Lessen uit Onderzoek” plaatsgevonden, waaraan 375 mensen hebben deelgenomen en waarin enkele sessies waren waarin expliciet uitleg werd gegeven over het opzetten en evalueren van experimenten.