DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 136-148

Jutta Wolff

Das evaluieren wir (mal eben)

Was Auftraggebende über Wirksamkeitsnachweise wissen sollten

Zusammenfassung

Auftraggebende verbinden mit dem Inauftraggeben einer Evaluation oft hohe Erwartungen, wobei die Notwendigkeit eigener vorbereitender Arbeiten teilweise unterschätzt und das Potenzial von Evaluationen, die gewünschten Erkenntnisse zu generieren, vielfach überschätzt werden. So soll mit Evaluationen häufig die Wirksamkeit eines Programms nachgewiesen werden. Verbreitet ist deshalb der Ruf nach (quasi-)experimentellen Designs, die beim Nachweis kausaler Beziehungen oftmals als „Goldstandard" gelten. Im Beitrag wird dargestellt, welchen Begrenzungen diese Designvarianten gerade im pädagogischen Bereich unterliegen.
Schlüsselwörter: Bildungsevaluation, Evaluationsplanung, Evaluationsprozess, Evidenzbasierung, (quasi-)experimentaldesigngesteuerte Evaluation, Wirksamkeit

We Evaluate That (just Quickly)

What Customers Should Know about the Proof of Effectiveness

Summary

Customers often have high expectations when commissioning an evaluation, whereby the own planning needs are partly underestimated and the potential of evaluations to generate the desired knowledge is overestimated. Often an evaluation is intended to demonstrate the effectiveness of a program. Then the call for (quasi-)experimental designs, which are viewed as a "gold standard" in the detection of causal relationships, is widespread. The article illustrates the limits of these design variations in the field of education
Keywords: educational evaluation, evaluation planning, evaluation process, evidence-based, (quasi-)experimental design driven evaluation, effectiveness

1. Einleitung

Evaluationen können verschiedenste Gegenstände in den Blick nehmen wie z.B. Organisationen, Curricula, Projekte, Leistungen. Der Beitrag fokussiert auf *Programmevaluationen*, worunter Beywl und Niestroj (2009, S. 83) folgend die „Evaluation eines Bündels von Maßnahmen (Programm), das basierend auf einem Set von Ressourcen aus einer Folge von Interventionen besteht", verstanden wird, das „auf bestimmte, in der Regel bei bezeichneten Zielgruppen zu erreichende Resultate gerichtet ist."

Nicht immer ist allen Beteiligten bewusst, dass es sich bei Evaluationen um sehr voraussetzungsvolle Prozesse handelt, in denen zahlreiche Entscheidungen zu fällen sind, die den Erkenntnisgewinn der Evaluation und die Nutzung der Ergebnisse beeinflussen: Auftraggebende[1] von Evaluationen nehmen bisweilen an, ihre Aufgabe sei mit der Beauftragung von professionellen Evaluierenden weitestgehend abgeschlossen, und sie könnten nach einigen Monaten oder Jahren den Ergebnisbericht interessiert entgegennehmen. Doch nur, wenn Auftraggebende und ggf. weitere Beteiligte intensiv an der Planung der Evaluation mitwirken, können nützliche und potenziell genutzte Evaluationsergebnisse generiert werden.

Entsprechend wurden in den letzten Jahren Leitfäden für Auftraggebende entwickelt, um diese im Vorfeld einer Evaluation über die auf sie zukommenden Aufgaben und zu klärenden Aspekte zu informieren (vgl. Deutsches Jugendinstitut 2010; DeGEval 2012). Unter anderem wird darauf aufmerksam gemacht, dass Auftraggebende (ggf. mit Unterstützung der Evaluierenden)

- *ihren Eigenaufwand und den weiterer Beteiligter realistisch einschätzen sollten* (z.B. zeitliche und personelle Ressourcen für Informationsweitergabe, Absprache von Evaluationseckpunkten und ggf. Erhebungsinstrumenten, Datenerhebung),
- *eine Evaluation rechtzeitig in Auftrag geben sollten* (z.B. um Dokumentationsverfahren eines Programms so zu konzipieren, dass sie gleichzeitig als Datenquelle für die Evaluation genutzt werden können, oder um Erhebungen der Ausgangslage zu ermöglichen),
- *prüfen sollten, ob und in welchem Umfang wichtige Akteure in den Evaluationsprozess eingebunden werden sollten, und diese ggf. einbeziehen* (z.B. um die Qualität der Evaluation durch fachliche Beratung und Akzeptanzsteigerung zu erhöhen),
- *den Evaluationszweck und die Fragestellungen präzisieren sollten* (u.a. um die Evaluation auf diese fokussieren zu können),
- das Verhältnis von Aufwand und Nutzen abwägen sollten.

1 Mit „Auftraggebenden" sind im Folgenden neben Geldgebern v.a. Projektverantwortliche oder andere Personen gemeint, die als Ansprechpartner für die Evaluierenden fungieren.

„Das evaluieren wir (mal eben)" betrifft jedoch nicht nur die Verkennung der notwendigen Aktivitäten der Auftraggebenden im Planungsprozess einer Evaluation, sondern auch – und das soll hier im Fokus des Beitrags stehen – die oftmals überhöhten Erwartungen an die Aussagekraft von Evaluationsergebnissen bezüglich der Wirksamkeit eines Programms.

Werden neue pädagogische Maßnahmen implementiert, interessiert es sowohl Auftraggebende als auch Verantwortliche von Konzeption und Durchführung, ob die Intervention so, wie gedacht, funktioniert („what works"), d.h., ob eine Maßnahme die erhofften Wirkungen entfaltet. Dabei sollen in der Regel die Annahmen über einen Ursache-Wirkungs-Zusammenhang geprüft werden, um die gemessenen Wirkungen *kausal* auf eine Intervention zurückzuführen (= Wirksamkeit).

Doch wie lässt sich solch ein Wirksamkeitsnachweis führen? In der evidenzbasierten Medizin hat sich ein Stufenmodell des Wirksamkeitsnachweises etabliert. In dieser Evidenzhierarchie gelten Experimente (= randomisiertes Kontrollgruppendesign) als „Goldstandard". Ursprünglich in Studien zur Wirksamkeit von Medikamenten eingesetzt, wurde das randomisierte Kontrollgruppendesign als Forschungsmethode auf weitere Gebiete der Medizin und schließlich auch auf den sozialen Bereich ausgeweitet. Verstärkt wurde dies unter anderem durch die Neue Steuerung im Bildungswesen, in deren Zuge der Generierung von empirischer Evidenz eine immer größere Bedeutung zugeschrieben wurde (vgl. Hammersley 2013; Pant 2014).

Die Annahme einer generellen Überlegenheit des experimentellen Designs gegenüber anderen Designvarianten führt in der Evaluationspraxis häufig dazu, dass bei der Frage nach einem Wirksamkeitsnachweis reflexartig ein (quasi-)experimentelles Design angestrebt wird, z.T. ohne Kenntnis der Implikationen eines solchen Designs.[2] Im Folgenden sollen deshalb Begrenzungen einer (quasi-)experimentaldesigngesteuerten Evaluation[3] angesprochen werden, um Auftraggebenden (und Evaluierenden) eine Abschätzung von Chancen und Risiken einer solchen Designvariante zu ermöglichen. Die „Checkliste" im Anhang ist ein erster Versuch, überblicksartig Aspekte zu benennen, die bei der Erwägung einer (quasi-)experimentaldesigngesteuerten Evaluation bedeutsam sind.

2. Begrenzungen (quasi-)experimentaldesigngesteuerter Evaluation

Ziel des experimentellen Designs ist es, einen Kausalzusammenhang zwischen einer Intervention (unabhängige Variable) und einer Wirkung (abhängige Variable) zwei-

2 Ausgangspunkt der intensiven Beschäftigung mit dieser Thematik waren eigene Erfahrungen bei der Umsetzung eines quasi-experimentellen Designs bei der Evaluation eines Qualifizierungsprogramms für Schulen und ihre Lehrkräfte (vgl. Wolff 2015).

3 Diesen Begriff habe ich Beywl (2006) entnommen.

felsfrei zu belegen (= interne Validität). Hierbei wird das Prinzip des Vergleichs genutzt: Es werden Zielgrößen für die am Programm teilnehmende Gruppe (Experimentalgruppe) gemessen und mit den Werten einer Kontrollgruppe, die nicht an der Intervention teilgenommen hat, verglichen.[4] Die Zuweisung zu Experimental- bzw. Kontrollgruppe erfolgt per Zufall, d.h. durch *Randomisierung*. Dahinter steht die Annahme, dass bei genügend großen Versuchsgruppen Personenvariablen, die ggf. Einfluss auf das Ergebnis haben („Störfaktoren"), ausgemittelt werden und Unterschiede der abhängigen Variablen ausschließlich auf die Intervention zurückgeführt werden können.

Zentrale Komponenten des Experiments sind demnach: (1) Randomisierung: Zuordnung zu Interventions- und Kontrollgruppe per Zufall, (2) eine Intervention, die bestimmte Wirkfaktoren (unabhängige Variablen) umfasst, und (3) die Messung von abhängigen Variablen, d.h. der angestrebten Wirkungen.

Im Folgenden werden Schwierigkeiten bei der Umsetzung dieser zentralen Komponenten im pädagogischen Bereich und der zu erwartende Erkenntnisgewinn angesprochen.[5]

2.1 Probleme der Randomisierung

Der relativ unbestrittene Nutzen des Experiments liegt darin, dass durch die Randomisierung die Gefahr des Selektionsbias reduziert wird: Durch die Zufallszuweisung wird sichergestellt, dass sich die Objekte in den Versuchsgruppen (z.B. Schulen, Lehrkräfte, Schüler und Schülerinnen) nicht systematisch unterscheiden und gemessene Unterschiede tatsächlich auf die Intervention zurückzuführen sind. Aber gerade dieser Vorteil des Experiments kann im pädagogischen Kontext oft nicht zum Tragen kommen, da die Zufallszuordnung oft nicht möglich ist. So stellt es ein ethisches Problem dar, bestimmten Personen eine Intervention zu versagen oder andere mit einer Intervention „zwangszubeglücken".

Stattdessen ist es in der Praxis die Regel, dass die Teilnehmenden der Interventionsgruppe bereits festgelegt sind, u.a. durch Auswahlkriterien des Programms (z.B. Kinder mit erheblichen Sprachdefiziten nehmen am Sprachförderprogramm teil) oder durch das Anmeldeverhalten der Zielgruppe (z.B. Anmeldung zu einer Lehrerfortbildung ⇨ Selbstselektion). Die Schwierigkeit für Evaluierende besteht dann darin, eine möglichst äquivalente Vergleichsgruppe zu finden, die sich nicht durch wesentliche Merkmale von der Interventionsgruppe unterscheidet. Handelt es sich um ein Programm mit partieller Erfassung, nimmt also noch nicht die gesamte

4 Für weitere Spezifikationen dieses Grundgedankens vgl. Frey/Frenz (1982).
5 Für eine intensivere Beschäftigung mit der Thematik siehe Bellmann/Müller (2011).

Zielgruppe am Programm teil, so können „Wartegruppen" als Kontrollgruppe herangezogen werden. Bei Programmen, die die gesamte Zielgruppe umfassen, ist die Suche nach Vergleichsgruppen nochmals schwieriger.

Sobald die Zuweisung zu den Versuchsgruppen nicht per Zufall erfolgen kann (d.h., „nur" ein *quasi*-experimentelles Design umgesetzt werden kann), werden Annahmen darüber benötigt, welche Merkmale der Versuchspersonen ggf. die Wirkungen (abhängige Variable) beeinflussen, um diese Merkmale bei den statistischen Analysen entsprechend berücksichtigen zu können. Da diese Annahmen aufgrund fehlender Theorien eher Vermutungen sind, werden häufig hilfsweise gängige soziodemographische Variablen wie z.B. Alter, Schulform etc. in die Berechnungen einbezogen, ohne sicher zu sein, alle entscheidenden Variablen erfasst zu haben. Entsprechendes gilt für das Matching, bei dem für Objekte der Interventionsgruppe statistische Zwillinge gesucht werden (und das zudem das Vorliegen von entsprechenden Datensätzen voraussetzt).

2.2 *Was* wirkt denn eigentlich?

Eine Prämisse für ein experimentelles Design ist, dass die untersuchte Maßnahme klar identifizierbar sein und relativ standardisiert durchgeführt werden muss. Bei Experimenten zur Wirksamkeit von Medikamenten ist die Intervention in hohem Maße standardisiert: Die *Bestandteile* des Medikaments und die *Dosis* werden kontrolliert. Im pädagogischen Bereich ist die Standardisierung ungleich schwieriger (vgl. Hammersley 2015).

Vielzahl potenzieller Wirkfaktoren
Bereits in der Grundlagenforschung ist es eine kaum zu lösende Aufgabe, die theoretischen Vorstellungen in konkrete Variablen eines Feldexperiments umzusetzen (vgl. Frey/Frenz 1982). Im Kontext von Evaluationen mit starker Anwendungsorientierung gilt dies umso mehr: Der Ausgangspunkt von Evaluationen sind in der Regel Programme, die aus der Praxis heraus als Antwort auf spezifische Anforderungen/Probleme entwickelt wurden. Die Überprüfung der Wirksamkeit stand dabei nicht im Fokus. In der Konsequenz bedeutet dies, dass ein Programm häufig aus einer Vielzahl von Interventionen besteht, d.h., eine Vielzahl von möglichen „Wirkfaktoren" beinhaltet, für deren Zusammenwirken selten explizierte Annahmen bestehen.

Ein Beispiel: Es wird das Problem erkannt, dass Schulen mit niedrigem Sozialindex hinsichtlich der Leistungsergebnisse ihrer Schülerinnen und Schüler hinter anderen Schulen zurückbleiben. Es wird ein Programm zur Unterstützung dieser Schulen aufgelegt, um die fachlichen Kompetenzen der Schülerinnen und Schüler zu verbessern. Dieses Programm wird in der Regel nicht aus einer einzelnen, eng umschriebenen Maßnahme bestehen (z.B. zusätzliche einstündige wöchentliche Förderung in

Mathematik durch Mathematikfachkräfte mit genauer Durchführungsanweisung). Eher wird – ausgehend von Überlegungen zu wichtigen Einflussfaktoren auf Leistungsergebnisse und pragmatischen Erfordernissen – ein Maßnahmenbündel „geschnürt", und es werden verschiedene Maßnahmen ergriffen, die z.B. auf Ebene von Schulleitung (Schulleitungshandeln), Lehrpersonen (Unterrichtsentwicklung), Eltern (elterliche Unterstützung) und Schülerinnen und Schülern (z.B. Lernberatung) ansetzen könnten. Andere, bereits für alle Schulen implementierte Maßnahmen, wie z.B. die Lernförderung bei nicht ausreichenden Leistungen, werden beibehalten.

Wird die Fachleistung der Schülerinnen und Schüler der Programmschulen und möglichst ähnlicher Schulen[6] in Zeitreihen erfasst, so könnte sich herausstellen, dass sich sowohl in der Interventionsgruppe als auch in der Vergleichsgruppe gleichermaßen Schulen mit und ohne deutliche Leistungssteigerungen finden lassen. So berichtet Berliner (vgl. 2002, S. 19) von Programmevaluationen, die ergaben, dass die Schülerleistungen innerhalb eines Programms stärker variierten als zwischen den Programmen. Wie lässt sich das erklären?

Macht des Kontexts/Vielzahl von Interaktionen
Selbst wenn die Kernelemente einer Intervention als „Ausführungshinweise" schriftlich fixiert sind (das ist selten der Fall), ist der konkrete Verlauf im pädagogischen Kontext in verschiedenen Umgebungen nicht identisch, *kann* er nicht identisch sein: U.a. Berliner (2002) macht darauf aufmerksam, dass einerseits zahlreiche (lokale) Kontextbedingungen die konkrete Ausgestaltung pädagogischer Aktivitäten mitbestimmen (*Power of Contexts*) und dass andererseits die Interaktionen zwischen den Beteiligten zu völlig unterschiedlichen Verläufen führen (*Ubiquity of Interactions*).

Als Beispiel führt er das Klassenzimmer an, in dem verschiedene Charakteristika der Lehrpersonen (z.B. Ausbildung, Motivation) mit Eigenschaften der Schülerinnen und Schüler (z.B. IQ, sozioökonomischer Status, Lernmotivation) ebenso interagieren wie mit Merkmalen der Umwelt (z.B. Lehrpläne, Jugendarbeitslosigkeit) und die Wirkungsrichtung oft unklar sei: „Moreover, we are not even sure in which directions the influences work, and many surely are reciprocal. Because of the myriad interactions, doing educational science seems very difficult, while science in other fields seems easier." (Berliner 2002, S. 19)

Kelle (2006) macht auf ein Grundproblem der Kausalanalyse sozialen Handelns aufmerksam: Diejenigen, die das Programm planen oder anordnen, können nur die *Handlungsbedingungen* der Akteure vor Ort *direkt* beeinflussen, nicht aber deren Handlungen selber. Da diese individuellen Akteure ggf. konkurrierende Handlungsziele verfolgen, müsse gerade in Interventionsstudien

6 Wobei sich auch hier die Frage stellt, welches „ähnliche" Schulen sind: ähnlich hinsichtlich von Merkmalen der Schülerinnen und Schüler (z.B. Sozialindex) oder der Lehrpersonen (z.B. Alter, unterrichtsrelevante Einstellungen) etc.

> „damit gerechnet werden, dass die Betroffenen in nicht vorhersagbarer Weise auf Interventionsmaßnahmen reagieren und dabei in kreativer Weise jene Handlungsbedingungen verändern, die durch die Intervention sozialtechnologisch geschaffen und beeinflusst werden sollen." (Ebd., S. 133)

Die hier aus Forschersicht als „Störfaktor" kritisierte Variabilität von pädagogischen Maßnahmen ist aus der Sicht von Pädagogen und Pädagoginnen oftmals erwünscht und für den Erfolg einer Intervention unverzichtbar. So muss z.B. eine Lehrperson ihr eigenes Verhalten an die Klasse adaptieren, um Lernfortschritte zu ermöglichen. In welchem Ausmaß eine Intervention im Feld variiert wird – darauf haben die Evaluierenden in der Regel keinen Einfluss. Es bleibt ihnen lediglich, die *Implementationstreue* zu erheben, um Umfang, Genauigkeit und Qualität der Implementation abschätzen zu können. Evaluationsergebnisse können dann eher eingeordnet werden. Die vergleichsweise geringen Lernfortschritte im Bereich Sprache in einer Kindertagesstätte können dann z.B. evtl. mit der langfristigen Erkrankung der Förderfachkraft oder mit deutlichen Abweichungen vom Ursprungsprogramm in Verbindung gebracht werden – und gehen weniger zulasten des Programms.

Dilemma interne und externe Validität (ökologische Validität)
Forschungen auf der Basis von (Quasi-)Experimenten befinden sich in einem Dilemma zwischen interner und externer Validität: Die interne Validität, d.h. die eindeutige kausale Interpretierbarkeit eines Zusammenhangs (z.B. kann die Veränderung der Fachleistung Mathematik *eindeutig* auf die infrage stehende Maßnahme zurückgeführt werden), ist am ehesten zu erreichen, wenn eine Intervention unter hochkontrollierten und deshalb recht künstlichen (Labor-)Bedingungen durchgeführt wird – die unabhängige Variable somit eindeutig identifizierbar ist. Fraglich ist dann jedoch die externe (ökologische) Validität: Gelten die unter künstlichen Bedingungen erlangten Ergebnisse auch in realen Situationen, können sie auf Kontexte verallgemeinert werden, die nicht untersucht wurden?

Wird ein stärker realitätsnahes Forschungsdesign genutzt, z.B. indem eine Intervention von verschiedenen Personen in verschiedenen Klassen durchgeführt wird, ist die Gefahr unkontrollierbarer oder unbeachteter Störeinflüsse groß und die interne Validität, d.h. die eindeutige Rückführbarkeit einer Wirkung auf die Maßnahme, gefährdet.

2.3 Wie lassen sich Wirkungen *messen*?

Als „Messen" wird Kriz und Lisch (1988, S. 175) folgend „die Zuordnung von Zahlen (numerisches Relativ) zu Objekten und deren Eigenschaften (empirisches Relativ) mit dem Ziel einer isomorphen oder homomorphen Abbildung" verstanden. Diese „Zuordnung von Zahlen" ist je nach Zielvariable unterschiedlich schwierig: Soll die

Wirksamkeit eines blutdrucksenkenden Medikaments untersucht werden, so lässt sich ein Blutdruckmessgerät verwenden. Eine wesentlich höhere Anforderung stellt jedoch die Messung z.B. von Leistung, Lernmotivation, Unterrichtsverhalten, Kooperation der Lehrpersonen etc. dar. Diese Herausforderung gilt zwar nicht nur für (quasi-)experimentelle Designs, beschädigt jedoch deren Anspruch, besonders überzeugende Belege für die Wirksamkeit einer Maßnahme zu liefern.

Frey und Franz (1982, S. 231) sehen es zurecht als eine der zentralen Fragen an, wie man ein Messinstrument „finden (bzw. konstruieren) könne, das in der Lage ist, das in der Hypothese ideell fixierte Verhaltensphänomen O (Observation) tatsächlich empirisch zu erfassen." Hierzu ist zunächst ein umfangreiches Wissen über die zu messenden Variablen notwendig – das auch die Frage umfasst, ob diese überhaupt eine quantitative Struktur aufweisen, also eine metrische Messung angemessen ist (vgl. Hammersley 2013, S. 69f.). Welche Annahmen oder gar Theorien gibt es dazu, was „Unterrichtsqualität", „Lehrerkooperation" etc. ausmacht, bzw. was soll mit dem Programm genau erreicht werden?

Im nächsten Schritt müssen Messinstrumente gefunden bzw. konstruiert werden, die in der Lage sind, diese Konzepte abzubilden. In der Evaluationspraxis gibt es hierzu zwei Vorgehensweisen:

(1) Es wird (ressourcensparend) auf bereits in anderen Studien verwendete Messinstrumente zurückgegriffen. Für den Bereich Schule liegen z.B. zahlreiche Fragebögen zu zentralen Konzepten wie Klassenklima, Unterrichtsverhalten etc. vor.[7] Die Gefahr besteht hier, dass mit Messinstrumenten, die nicht auf das Programm zugeschnitten wurden, der Erfolg der Maßnahme nicht hinreichend erfasst werden kann:

> „Ein dann routinemäßig geübter Rückgriff auf vorhandene und vermeintlich bewährte standardisierte Messinstrumente kann erhebliche Probleme mit sich bringen, wenn die hiermit gemessenen Variablen zu unspezifisch sind, um den Erfolg zu erfassen." (Kelle 2006, S. 127)

(2) Es werden neue Messinstrumente entwickelt. Dieses ist voraussetzungsvoll und i.d.R. mit hohem Aufwand verbunden.

Problematisch erscheint zudem, dass vornehmlich solche Messinstrumente gewählt werden, die angesichts von Zeitknappheit und Kostendruck mit möglichst geringem Aufwand angewendet werden können, wie z.B. Selbsteinschätzungsskalen. Sie haben den Vorteil, dass sie die Untersuchenden „quasi in Sekundenschnelle mit quantitativer Information über Parameter versorgen, auch wenn deren empirische Korrelate

7 Viele Skalen samt Vergleichswerten werden leicht zugänglich vom Deutschen Institut für Pädagogische Forschung in der „Datenbank zur Qualität von Schule" (DaQS; URL: http://daqs.fachportal-paedagogik.de/; Zugriffsdatum: 11.04.2016) vorgehalten.

umstritten oder gänzlich unklar sind." (Frey/Frenz 1982, S. 255) So werden beispielsweise zur Erhebung von Unterrichtsverhalten Einschätzungsskalen für Lehrpersonen oder Schülerinnen und Schüler genutzt – das komplexe Unterrichtsgeschehen somit auf leicht erfassbare und methodisch unkompliziert erhebbare Items reduziert.

Evaluierende befinden sich in einem Dilemma zwischen Wissenschaftlichkeit und Praktikabilität: Die Erhebung eines Konstrukts sollte gegenstandsangemessen erfolgen, was ggf. umfangreichere und methodisch anspruchsvollere Erhebungen erforderlich macht. So ist die Messung einer Variablen in der Regel umso aufwendiger, je höher sie in der Hierarchie der Resultatsarten steht: Outputs (wie z.B. Zahl der Teilnahmen an einer Fortbildung, Zufriedenheit) sind einfacher zu messen als Wissen oder Kenntnisse (Outcome I), diese wiederum einfacher als Verhalten (Outcome II) etc.[8] Andererseits müssen neben dem Zeit- und Kostenrahmen auch datenschutzrechtliche Bestimmungen beachtet werden, die ggf. das Gewünschte einschränken (z.B. die Beobachtung/Befragung von Schülerinnen und Schülern). Auch muss berücksichtigt werden, dass Evaluationen nur dann zu nützlichen Ergebnissen führen, wenn die Datengebenden die Datenerhebungsmethoden akzeptieren. Das heißt: Obwohl die Messgenauigkeit i.d.R. mit der Länge des Messinstruments steigt, darf dieses nicht zu umfangreich sein. Auch ist nicht jede Messmethode überall einsetzbar; so ist beispielsweise die Akzeptanz von Wissenstests bei Lehrpersonen eher gering.

2.4 Erkenntnisgewinn von (Quasi-)Experimenten

Bei Durchführung einer (quasi-)experimentaldesigngesteuerten Evaluation sollten sich Auftraggebende und Evaluierende nicht nur über zahlreiche praktische Probleme und die sehr hohen Kosten im Klaren sein, sondern auch über den potenziellen Erkenntnisgewinn, der mit dieser Designvariante verbunden ist.

Im Idealfall ergeben sich in der Interventionsgruppe gegenüber der Kontrollgruppe deutlich höhere Ausprägungen der Zielvariablen. Unklar bleibt jedoch, welche Wirkmechanismen hierfür ausschlaggebend waren und ob sich dementsprechend die Ergebnisse in anderen Kontexten replizieren lassen. Auch lassen die Ergebnisse selbst bei sehr sorgfältig und aufwendig gestalteten Quasi-Experimenten vielfältigen Interpretationsspielraum. Sind z.B. für die geringfügig höheren Leistungsergebnisse der Schülerinnen und Schüler, die mit der „Gruppenrallye im Mathematikunterricht" unterrichtet wurden, Novitätseffekte der Methode verantwortlich (vgl. Wandeler et al. 2015)?

8 Bei der Benennung der Resultatsarten orientiere ich mich an der Einteilung von W. Beywl. URL: http://www.eval-wiki.org/w_glossar/images/f/fb/Variantentafel-Resultatsarten-fuer_Glossar.pdf; Zugriffsdatum: 29.11.2015.

Nicht selten ergeben sich nur sehr geringe Unterschiede zwischen den Gruppen – was angesichts der multifaktoriellen Bedingtheit vieler Zielvariablen nicht verwundert. Es könnte sogar der Fall eintreten, dass die Ergebnisse nicht in die gewünschte Richtung weisen. Erst mithilfe qualitativer Methoden können Erklärungen für solche unerwünschten Effekte gefunden werden (vgl. Kelle 2006). Ein ausschließlich auf quantitativen Daten fußendes Design ist geeignet, unter bestimmten Bedingungen ein überblicksartiges Wissen über den Erfolg oder Misserfolg eines Programms oder einer Intervention zu generieren. Die Wirkmechanismen bleiben jedoch im Dunkeln; es können keine (erwünschten oder unerwünschten) Nebenwirkungen erfasst und keine Hinweise auf die Verbesserung eines Programms generiert werden.

3. Schluss

Der Beitrag sollte deutlich machen, dass es kein generell überlegenes Erhebungsdesign oder keine überlegenen Erhebungsmethoden zum Nachweis der Wirksamkeit gibt – jedes Design, jede Methode hat Stärken und Schwächen. Insofern schlägt Hammersley (2015) vor, eher von einer Matrix als von einer Hierarchie der Evidenz auszugehen.

Auch Berliner warnt davor, Wissenschaft und Methode zu verwechseln. Ein experimentelles Design sei eine Methode oder Technik, aber nicht mit Wissenschaft gleichzusetzen:

> "But to think that this form of research is the only 'scientific' approach to gaining knowledge – the only one that yields trustworthy evidence – reveals a myopic view of science in general and a misunderstanding of educational research in particular." (2002, S. 18)

Kelle (2006, S. 134) zufolge wird die Analyse kausaler Beziehungen „oft zu Unrecht für ein genuines und exklusives Feld quantitativer Datenanalyse" gehalten. Qualitative Forschung halte Lösungen bereit zur Identifikation unterschiedlicher kausaler Pfade und von (erwünschten und unerwünschten) Nebenwirkungen.

Es kann zusammenfassend festgehalten werden: Wirkungen kausal auf ein Programm zurückzuführen, ist in der Regel aufwendig und teuer. Es ist für jedes einzelne Evaluationsprojekt abzuwägen, welches Design (*one-shot*, mehrere Messzeitpunkte, …) und welche Methoden (quantitative, qualitative) oder welche Kombinationen von Methoden in Anbetracht der Kontextbedingungen (z.B. Ressourcen) am besten geeignet sind, die Wirkungen eines Programms nachzuweisen. Während quantitative Methoden geeignet sind, ein überblicksartiges Wissen zu den Wirkungen eines Programms zu generieren, liegt die Stärke qualitativer Methoden in der Erkundung der zugrunde liegenden Wirkmechanismen. Insofern erscheint eine Kombination

quantitativer und qualitativer Methoden beim Führen eines Wirkungsnachweises geboten – aber angesichts oftmals begrenzter Ressourcen schwer realisierbar.

Und manchmal gilt vielleicht auch die Mahnung Hammersleys (2015, S. 9), wonach Evaluationen nicht auf alle drängenden Fragen Antworten geben und die Validität der Ergebnisse garantieren können. Um nicht falsche Erwartungen zu wecken, muss auch dies manchmal ausgesprochen werden:

> "Unfortunately, we must face the fact that research cannot always provide answers to questions that are seen as pressing by policymakers and practitioners. Nor can we guarantee the validity of our findings. And we need to make this clear to lay audiences. So, there is a danger of over-promising or over-claiming."

Anhang: „Checkliste" für eine (quasi-)experimentaldesign-gesteuerte Evaluation

Ohne Anspruch auf Vollständigkeit werden im Folgenden checklistenartig einige – z.T. bereits angesprochene – Aspekte aufgeführt, die geprüft werden sollten, sobald ein (quasi-)experimentelles Design in Betracht gezogen wird.

Wie sollte das zu evaluierende Programm beschaffen sein?

1) Das zu evaluierende Programm lässt angesichts seines zeitlichen Umfangs nennenswerte Wirkungen erwarten.
2) Das Programm ist teuer, so dass eine Wirkungsanalyse bedeutsam ist.
3) Das Programm hat potenziell eine große Reichweite, betrifft also viele Personen und wird voraussichtlich langfristig eingesetzt.
4) Das Programm beinhaltet wenige eng umschriebene Kernelemente, deren angenommene Wirkungen auf das Outcome expliziert sind (Wirkmodell).
5) Die Ziele des Programms sind schriftlich fixiert und evaluierbar formuliert (*smart*).
6) Die Programmstabilität ist im Rahmen des Möglichen gewährleistet: (a) Das Programm ist schriftlich fixiert, um eine ähnliche Durchführung durch verschiedene Personen wahrscheinlich zu machen. (b) Das Programm ist „ausgereift", d.h., seine Konzeptschwächen wurden bereits behoben, so dass anzunehmen ist, dass künftig keine wesentlichen Konzeptänderungen vorgenommen werden.
7) Das Programm sollte möglichst „immun" gegen (wechselnde) äußere Rahmenbedingungen sein, z.B. gegenüber Änderungen politischer Vorgaben.

Was sollte bei der Evaluation berücksichtigt werden?

1) Der Projektzeitraum der Evaluation ist angemessen: (a) Die Evaluation kann rechtzeitig beginnen, so dass Erhebungen zur Ausgangslage erfolgen können (b) Die

Laufzeit der Evaluation wird so gewählt, dass die erwarteten Wirkungen innerhalb dieser Zeit eintreten können.

2) Es stehen ausreichend Ressourcen für die Evaluation zur Verfügung. Gerade (quasi-)experimentelle Designs sind extrem teuer, da Daten zu mehreren Messzeitpunkten erhoben werden müssen.
3) Es liegt ein möglichst zuverlässiges Instrument zur Messung der Wirkungen vor (damit u.a. auch kleine Wirkungen erfasst werden können) – oder es stehen Zeit, Geld und Expertise zur Verfügung, ein solches zu entwickeln.
4) Es ist möglich, eine zur Versuchsgruppe äquivalente Vergleichsgruppe zu bilden (z.B. bei einem Programm mit partieller Erfassung eine Wartegruppe) oder auf Daten zuzugreifen, auf deren Grundlage statistische Zwillinge gebildet werden können (*matching*). Dabei liegen möglichst Annahmen darüber vor, welche Merkmale der Untersuchten die Wirkungen moderieren.
5) Die geplante Evaluation wird durch datenschutzrechtliche Bestimmungen hinsichtlich des zu erwartenden Erkenntnisgewinns nicht übermäßig eingeschränkt.
6) Der (langfristige) Zugriff auf relevante Daten ist gesichert: Es ist zu erwarten, dass die Datengebenden über einen langen Zeitraum für mehrere Datenerhebungen zur Verfügung stehen, zur Teilnahme bereit/verpflichtet und im Projektverlauf „auffindbar" sind (Gefahr des *Dropout*).
7) Der voraussichtlich durch die Evaluation generierte Nutzen steht in einem angemessenen Verhältnis zu dem (i.d.R. sehr großen) Aufwand.
8) Ergebnisse der Evaluation werden nicht sofort benötigt.

Literatur und Internetquellen

Bellmann, J./Müller, T. (Hrsg.) (2011): Wissen, was wirkt. Kritik evidenzbasierter Pädagogik. Wiesbaden: VS.

Berliner, D.C. (2002): Educational Research: The Hardest Science of All. In: Educational Researcher 31, H. 8, S. 18-20.

Beywl, W. (2006): Evaluationsmodelle und qualitative Methoden. In: Flick, U. (Hrsg.): Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek: Rowohlt-Taschenbuch-Verlag, S. 92-116.

Beywl, W./Niestroj, M. (^{2}2009): Das A-B-C der wirkungsorientierten Evaluation. Glossar – Deutsch/Englisch – der wirkungsorientierten Evaluation. Köln: Univation – Institut für Evaluation.

DeGEval – Gesellschaft für Evaluation (18.06.2012): Empfehlungen für Auftraggebende von Evaluationen. Eine Einstiegsbroschüre für den Bereich der Öffentlichen Verwaltung. Mainz: DeGEval – Gesellschaft für Evaluation. URL: http://www.degeval.de/fileadmin/Publi kationen/Publikationen_Homepage/DeGEval_-_Empfehlungen_Auftraggebende.pdf; Zugriffsdatum 19.12.2015.

Deutsches Jugendinstitut (2010): Vergabe und Begleitung externer Evaluationen in der Kinder- und Jugendhilfe. Ein Leitfaden für Auftraggebende. München: DJI (Projekt exe).

Frey, S./Frenz, H.-G. (1982): Experiment und Quasi-Experiment im Feld. In: Patry, J.-L. (Hrsg.): Feldforschung. Methoden und Probleme sozialwissenschaftlicher Forschung unter natürlichen Bedingungen. Bern: Huber, S. 229-258.

Hammersley, M. (2013): The Myth of Research-Based Policy and Practice. London: Sage.

Hammersley, M. (2015): Against 'Gold Standards' in Research: On the Problem of Assessment Criteria. Frühjahrstagung AK Methoden der DeGEval am 29.05.2015, Saarbrücken. URL: http://www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf; Zugriffsdatum 08.12.2015.

Kelle, U. (2006): Qualitative Evaluationsforschung und das Kausalitätsparadigma. In: Flick, U. (Hrsg.): Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek: Rowohlt-Taschenbuch-Verlag, S. 117-134.

Kriz, J./Lisch, R. (1988): Methoden-Lexikon für Mediziner, Psychologen, Soziologen. München: Psychologie Verlags Union.

Pant, H.A. (2014): Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. In: Zeitschrift für Erziehungswissenschaft 17, Sonderheft 27, S. 79-99.

Wandeler, C./Niggli, A./Villiger, C./Aebischer, M./Leopold, P. (2015): Ein Quasi-Experiment zur Gruppenrallye im Mathematikunterricht: Hält die Methode, was sie verspricht? In: Empirische Pädagogik 29, H. 2, S. 161-188.

Wolff, J. (2015): Evaluation eines komplexen Fortbildungsprogramms für Schulen und ihre Lehrkräfte. Planung – Design – Herausforderungen bei der Umsetzung eines quasi-experimentellen Designs. In: Grimm, A./Schoof-Wetzig, D. (Hrsg.): Was wirklich wirkt!? Effektive Lernprozesse und Strukturen in Lehrerfortbildung und Schulentwicklung. Rehburg-Loccum: Evangelische Akademie Loccum, S. 125-142.

Jutta Wolff, geb. 1964, Wissenschaftliche Referentin für Evaluation im Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Hamburg.
Anschrift: Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Beltgens Garten 25, 20537 Hamburg
E-Mail: Jutta.Wolff@ifbq.hamburg.de