LeGes

Stefan Sperlich

Von der Bewertung zur Methode … und zurück

Quantitative Methoden und Methoden-Mix bei der Beurteilung des Evaluationsgegenstandes

Beitragsart: Tagungsberichte

Zitiervorschlag: Stefan Sperlich, Von der Bewertung zur Methode … und zurück, in: LeGes 33 (2022) 3

ISSN 2624-5477, leges.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

EDITIONS WEBLAW

Inhaltsübersicht

1. Einführung

[1] Dieser Tagungsbericht fasst die wesentlichen Punkte des diesjährigen SEVAL Jahreskongresses zusammen, der unter dem im Titel angegebenen Thema stand und am 1. und 2. September 2022 in Fribourg stattfand. Insbesondere der zweite Kongresstag war ganz diesem Thema gewidmet. Dieses sollte den Zusammenhang zwischen der Beurteilung des Evaluationsgegenstandes und der Anwendung quantitativer sowie gemischter Methoden vertiefen. Natürlich hängt die Wahl der (Beurteilungs- oder Evaluations-) Methode sowohl von den Modalitäten der Beurteilung als auch von den verfügbaren Informationen ab. Es stellen sich also unter anderem folgende Fragen: Auf welche Weise gelangt man zu einer Beurteilung oder Bewertung? Welche Rolle spielen quantitative oder gemischte Methoden, um zu evidenzbasierten Beurteilungen zu gelangen? Wie kann man alternative Informationsquellen (oft unter dem Begriff «Big Data» zusammengefasst) einbeziehen? Diesen drei Fragen widmeten sich drei Hauptreferate am Freitagvormittag sowie fünf Workshops am Freitagnachmittag. Nicht weiter erwähnt seien hier andere, vom Kongressthema unabhängige, obgleich interessante Tagesordnungspunkte, ausgenommen der Apéro am Abschluss zwecks weiteren Austauschs, Diskussionen und natürlich auch des Netzwerkens.

2. Der Vortag

[2] Bevor wir im Detail über den zweiten Kongresstag berichten, soll hier auch erwähnt sein, dass bereits am ersten Kongresstag diverse hochinteressante Aktivitäten stattfanden. Dieses waren insbesondere ein intensiver Austausch der verschiedenen Arbeitsgruppen der SEVAL (www.seval.ch/arbeitsgruppen/), sowie zahlreiche Methodenateliers, einige davon ebenfalls von diesen Arbeitsgruppen durchgeführt. Diese umfassten die Themen (jeweils in der Sprache ihrer Ankündigung gelistet): *Keine Evaluation ohne Bewertung und keine Bewertung ohne Kriterien; «Test-retest» in teaching evaluation? Methodological considerations and results of a repeated survey; Einsatz der SEVAL-Standards in der Praxis; Möglichkeiten und Methoden für vertiefte Regulierungsfolgenabschätzungen; Mixed-methods impact evaluations in development cooperation; Evaluer en contexte scolaire – retours d'expérience; Methodische Ansätze zur Evaluation der körperlichen Aktivitäten in der Gesundheitsförderung; Möglichkeit und Nutzung von öffentlich zugänglichen Statistiken und Registerdaten in Evaluationsprojekten; Comprendre les cultures juvéniles en présentiel et en ligne – Enjeux, défis et innovations méthodologiques d'un terrain multisite en Suisse*. Die Ateliers wurden jeweils zwei Mal in insgesamt vier Runden präsentiert. Obgleich sehr interessant und lehrreich, können wir hier angesichts ihrer grossen Anzahl nicht einzeln auf sie eingehen, sondern verweisen auf seval.ch/news-veranstaltungen/veranstaltungen/seval-kongress-2022/.

[3] Alle in den folgenden Sektionen als Zitat gekennzeichneten Abschnitte stammen direkt von den jeweiligen Vortragsfolien, sind aber stets dem Satzbau grammatikalisch angepasst. Darin kursiv markierte Stellen verweisen direkt auf zitierte, im Text angegebene Werke.

3. Wege zur Bewertung

[4] Der erste Vortrag befasste sich mit der Frage nach der Stellung von Werten und Bewertung (values and valuing) in der Evaluation und auf welchem Wege man zu diesen gelangt. Referentin war Daniela Schröter aus den USA, Professorin für öffentliche Verwaltung an der Western Michigan University und ebenso aktive Evaluatorin, mit ihrem Vortrag *Wege zur Bewertung* (auf Englisch gehalten). Darin ging sie der Frage nach, wie und wo Werte und Bewertung in der Evaluationstheorie verankert sind. Zum Auftakt des Vortrags wurde das Publikum befragt, wie und wieweit die Zuhörer Werte und Bewertung beim Durchführen einer Evaluation berücksichtigen. Angesichts der dabei offensichtlich gewordenen Heterogenität wurde anschliessend gefragt, wie Werte und Bewertungen im weiteren Sinne in der Evaluationstheorie verankert sind. Bei den weiteren Betrachtungen wurde zwischen Evaluationstheorie (Modelle und Ansätze) und Evaluationslogik (Bewertungsgerüst) unterschieden.

[5] Der Argumentationsausgang war, dass bei einer Evaluation bereits die Wahl der Evaluationstheorie beeinflusst, auf welche Weise man Werte bestimmt und Bewertungen vornimmt. Dazu wurde angemerkt, dass Ausschreibungen und Richtlinien sich oftmals auf Methodenbeschreibungen und Ziel-basierte Evaluationen beschränken. Evaluationen aber «sollten über Outcome-Messungen zur Rechenschaftspflicht hinausgehen». Hierbei betonte Professor Schröter die Wichtigkeit, dass jede Evaluation einer Evaluationstheorie zugrunde liegen sollte, da letztere uns durch den Methodendschungel leitet, und somit wiederum die Wahl der Werte und Bewertungen beeinflusst, siehe auch Mark (2018). Zudem verwies sie darauf, dass (hinreichende) Kenntnisse der Evaluationstheorien daher sowohl im Evaluation Competency Framework der UN Evaluation Group (UNEG 2016) als auch im Evaluation Capabilities Framework der European Evaluation Society (EES 2014) explizit verlangt werden.

[6] So wie nun aber die Zahl der Evaluationstheorien und ihrer Modelle zunimmt, so gilt das auch für die daraus resultierenden Wege zur Bewertung. Hilfreich könnten dann Klassifizierungen dieser Theorien sein, wozu kurz einige existierende Klassifizierungsvorschläge diskutiert wurden, wie etwa die Einteilung entlang philosophischer Betrachtungen oder eine Einteilung nach Evaluationszwecken. Eine sehr anschauliche Klassifizierung bieten die heute wohlbekannten Klassifizierungsbäume, welche auf diverse Weisen versuchen, mehrere dieser unterschiedlichen Klassifizierungsansätze in einer einzigen Graphik darzustellen. Im weiteren Vortragsverlauf stellte Frau Schröter nun einen von ihr mit Bianca Montrosse-Moorhead und Lyssa Becho jüngst entwickelten Klassifizierungsansatz vor, den sogenannten *Evaluationsgarten*. Bei diesem werden die verschiedenen Evaluationstheorien entlang unterschiedlicher Dimensionen (in den angeführten Beispielen entlang *Activism for social justice*, *Promoting use*, *Engagement in the evaluation process*, *Power dynamics in decision making* etc., aber vor allem eben auch Werte und Bewertungen) als Blätter einer Blüte dargestellt, wobei die Grössen der Blätter den Grad der Wichtigkeit (auf einer Skala von 0 bis 3) jeder Dimensionen in der jeweiligen Evaluationstheorie oder -methode angeben (vgl. Abbildung 1). Da nun jede Evaluationstheorie eine Blüte ergibt, entsteht so insgesamt der Garten.

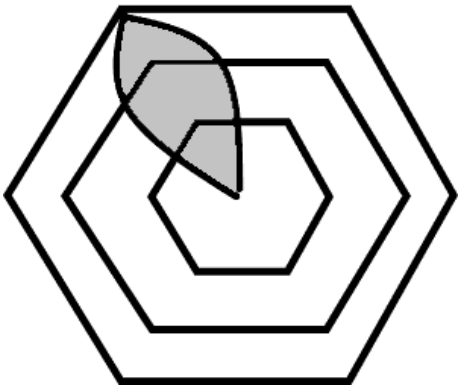

Abbildung 1: Konstruktion einer Blume im Evaluationsgarten

[7] Für die Dimension «Bewertung» bedeutet diese Skala: «1=Bewertung hat keinen Platz in der Evaluation, und es werden keine evaluierenden Beurteilungen vorgenommen; 2=Bewertung ist wichtig, aber der Evaluationsansatz klärt nicht, inwieweit der Evaluator Beurteilungen vornimmt; 3=Bewertung ist wichtig, und es ist definitiv die Aufgabe des Evaluators, solche Beurteilungen auch vorzunehmen.» In gleicher Weise lässt sich ein Blütenblatt für jede andere Dimension anfügen. Anschliessend wurden die so entstandenen Blüten verschiedener Evaluationsansätze beziehungsweise -theorien gezeigt, mit besonderem Augenmerk auf die beiden Blätter «Werte» und «Bewertungen», die in manchen Blüten (Evaluationsansätzen) gänzlich fehlten (Grad 0).

[8] Bei der Diskussion der Bewertung in der Evaluationslogik nahm Prof. Schröter stark auf Scrivens «Key Evaluation Checklist» (KEC) Bezug (Scriven, 2007). Zuerst wurde anhand eines Beispiels das Konzept von «Werten» nach dieser KEC erläutert. Dieses Beispiel diskutierte, wie man vom Zweck, bestimmte Konsumentenbedürfnisse zu befriedigen, zu Wertekriterien gelangen kann. Zur genaueren Betrachtung wurde erst einmal der generelle Evaluationsprozess nach der Evaluationslogik der KEC in zwölf Schritte aus vier Teilen untergliedert: 1) Identifikation der Kriterien, 2) Setzen der Standards, 3) Sammeln und Analysieren von Daten und 4) Formulierung evaluativer, also bewertender Schlüsse. In einer vereinfachenden Gleichung dargestellt hiesse das: *Fakten + Werte = Bewertende Schlüsse*, wobei die Kriterien den Rahmen sowohl für die Werte als auch für die Fakten bilden. Ein einfaches Beispiel hierfür ist, wenn für die Evaluation eines Einführungskurses in die Statistik als «Wert» nicht die Kursnoten definiert wurden, sondern die resultierende technische Fähigkeit, grundlegende deskriptive Statistikmethoden in künftigen Seminararbeiten sinnvoll anzuwenden. Wenn nun die empirisch gemessenen Fakten sagen, dass 95% der Studenten im Einführungskurs die Note B oder höher erhielten, aber eine schlechte Performance bezüglich des definierten Wertes zeigten, dann sind die daraus folgenden, bewertenden Schlüsse offensichtlich.

[9] Ebenfalls mit Verweis auf Scrivens KEC wurde anschliessend eine Liste von achtzehn unterschiedlichen, potentiellen Quellen zur Ableitung der der Evaluation zugrunde liegenden Werte vorgestellt; diese enthielten unter anderem kulturelle Werte, professionelle Standards, Umweltanforderungen, wissenschaftliche oder technologische Verdienste, ethische und gesetzliche Anforderungen, Risikoreduktion, Marktfähigkeit etc., aber auch persönliche, gruppenspezifische oder organisatorische Ziele oder Wünsche.[1]

[1] Siehe Abschnitt B.5 in Scrivens KEC für die vollständige Liste.

[10] Zuletzt diskutierte Frau Schröter einige Strategien und Methoden zum Bewerten und dem Setzen von Benchmarks. In der Summe haben wir gesehen, dass bereits die Wahl der Evaluationstheorie und -methoden unsere Definition der Werte als auch die Art der Bewertungen in der Evaluation in vielfacher Weise beeinflussen – oder eben umgekehrt; dies lässt sich oftmals sicherlich nicht trennen.

4. Der Methodenstreit in den Sozialwissenschaften und die Evaluationsforschung

[11] Udo Kelle, Professor für Methoden empirischer Sozialforschung an der Helmut-Schmidt-Universität Hamburg und derzeit Vorsitzender der DeGEval erläuterte in seinem auf Deutsch gehaltenem Vortrag, warum man speziell in der Evaluation sowohl für qualitative als auch für quantitative Methoden der Sozialforschung offen sein sollte und warum sich häufig ein kombinierter Einsatz qualitativer und quantitativer Verfahren lohnt, ja in manchen Fällen geradezu unabdingbar ist. Die Referenz auf den Methodenstreit diente vornehmlich als Motivation, um das besondere Potential und gegebenenfalls Notwendigkeit des Einsatzes gemischter Methoden in der Evaluation herauszuarbeiten. Dazu ist es nützlich, wesentliche Punkte dieses Methodenstreites genauer zu beleuchten. Am Anfang des Vortrags stand die These: «Der rationale Kern des Methodenstreits (*qualitativ* vs. *quantitativ*) in den Sozialwissenschaften besteht darin, dass in beiden Lagern jeweils unterschiedliche, jedoch gleichermassen legitime Erkenntnisziele (und daraus abgeleitete Qualitätskriterien für gute Forschung) verfolgt werden, die sehr leicht in Konflikt miteinander geraten können.» Das in den Sozialwissenschaften so entstandene methodologische Schisma spiegelt sich zumindest teilweise in der Verwendung zweier oftmals sehr unterschiedlicher Arten von Daten für die jeweiligen Forschungszwecke wider. Auch wenn sie sich nicht klar trennen lassen, haben wir auf der einen Seite oftmals den Hypothesentestenden Ansatz mit experimentellen Daten als golden Standard, wozu man sich hier auf das Quantifizieren und Messen konzentriert. Auf der anderen Seite sieht man eine eher explorative Herangehensweise über offene Fragen, oftmals mit Interviews und Diskussionsrunden. Die Hauptargumentationen auf der jeweiligen Seite kann man in etwa mit den folgenden, wenn auch bereits älteren Zitaten zusammenfassen: laut Schnell, Hill und Esser (1999) sind «für qualitative Verfahren keinerlei methodische Standards und Gütekriterien *jenseits subjektiver Evidenzerlebnisse formulier- und überprüfbar*», wohingegen Filstead (1979) anmahnte, dass «Quantifizierung *mit ihrem Labyrinth der verschiedenen logischen, mathematischen und technischen Anhänge* zu einem *verminderten Verständnis der empirischen sozialen Welt* führt, und sich dort eine *künstliche Auffassung von Realität* durchgesetzt habe».

[12] Mittlerweile werde zwar nicht mehr so heftig gestritten, aber nun gibt es nach Ansicht von Professor Kelle nicht mehr zu viel, sondern zu wenig Methodendebatten. Was der Debatte insgesamt fehle, sei die Bereitschaft, geäusserte Kritik zum Anlass zu nehmen, Schwachstellen des eigenen Ansatzes anzugehen, da doch gute Wissenschaft davon lebe, es besser zu machen, statt sich mit Gegenkritik zu wehren. Als Ergebnis dieser Unterlassung fehlen bislang konstruktive Beiträge in Form neuer Methoden. In der Tat bezeichnen auch die derzeit beliebten, sogenannten *Mixed Methods* eher eine Mischung aus verschiedenen, separaten Methoden statt neuer, tatsächlich gemischter Methoden im Sinne einer Methode, die beide Ansätze sinnvoll nutzt und vermengt. Die nächsten Abschnitte zeigen über ein besseres Verständnis der jeweiligen Grundideen einige Pfade auf, wieso die scheinbar strikt getrennten Ansätze eher als komplementär anzusehen

sind und sich bei geeigneter Anwendung gegenseitig unterstützen könnten – auch wenn dieser letzte Schritt bislang noch auszuarbeiten bleibt.

[13] Ohne Anspruch auf generelle Allgemeinheit skizzierte Herr Kelle zuerst die Unterschiede der Forschungsziele beider Ansätze. Möchte man zum Beispiel allgemeine Phänomene wie die Fertilität in verschiedenen Ländern untersuchen, genügten rein quantitative Erhebungen und Beschreibungen. Wenn hingegen interessiert, welche Bedeutung Elternschaft in diesen Ländern hat, dann sind rein quantitative Verfahren ungenügend. Denn dabei geht es um kulturelle Strukturen und Praktiken, die dem Forscher noch unbekannt sind. Daher kann die Datenerhebung nicht von Beginn an in Form vorgegebener Fragebögen durchgeführt werden; so muss man zuvor etwa die Bedeutung bestimmter Begriffe durch offene Gespräche und Beobachtungen in und mit der Zielgruppe klären. Das heisst, man muss hierzu also nichtstandardisierte, qualitative Daten sammeln. Bereits hier zeigt sich, dass sich beide Ansätze eher ergänzen statt ausschliessen. Dabei soll auf der einen Seite der quantitative Ansatz über Zähl- und Messbarkeit Objektivität (inklusive Beobachterunabhängigkeit) und Wiederholbarkeit (somit auch Verallgemeinerbarkeit) garantieren, der qualitative Ansatz aber dem Umstand Genüge tun, dass der Forschende den betreffenden Wirklichkeitsbereich nicht hinreichend kennt und dieser in seinen Hypothesen somit a priori nicht auftauchen kann. Dies gilt insbesondere für die Perspektiven der Akteure (d.h. der Zielgruppe). Fraglos hat dies eine stark unterstützende Funktion für das Endergebnis, aber die sehr kleinen Fallzahlen bei hoher Heterogenität entbehren in gewissem Grade der empirischen Evidenz und sind interpretationsoffen; vergleiche obige Kriterien Objektivität und Wiederholbarkeit etc. Beide Ansätzen und Sichtweisen wurden im Vortrag mit zahlreichen Zitaten und Beispielen unterlegt.

[14] Dies führt uns zur zweiten These des Vortrags: «der Methodenstreit enthält Potentiale für Methodenentwicklung und -integration, die sich produktiv nutzen lassen, wenn man erkennt, dass die Stärken und Schwächen beider Ansätze komplementär sind. Das heisst, man kann mit den Grenzen und Schwächen beider Ansätze umgehen, indem man die Stärken der jeweils anderen Methodentradition nutzt.» Erfreulicherweise kann man feststellen, dass sich daher Praktiker immer wieder beider Konzepte bedienten, also beide Methodenstränge gemeinsam und parallel in einem Forschungsprojekt eingesetzt und damit Ergebnisse erzielt haben, die die sozialwissenschaftliche Theorieentwicklung nachhaltig beeinflusste; siehe Johnson und Onwuegbuzie (2004): «*Mixed research actually has a long history in research practice because practicing researchers frequently ignore what is written by methodologists when they feel a mixed approach will best help them answer their research questions. It is time that methodologists catch up with practicing researchers*». Ein weiterer Beweis des Potentials ist das seit 2007 erscheinende «Journal of Mixed Methods Research», das heute einen Impact Faktor von über 5.7 aufweist, ein in Sozialwissenschaften beachtlich hoher Wert. Dieser Abschnitt schloss mit der These: «Insbesondere die Evaluation von (politischen, ökonomischen, sozialen) Interventionsprogrammen kann in vielen Fällen eine Kombination qualitativer und quantitativer Methoden erfordern».

[15] Was bedeutet dies für die Evaluationspraxis? Beginnen wir mit den Definitionen von Evaluation und Interventionsprogrammen nach Rossi und Freeman (1993) «Bewertung des Konzeptes, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme», welche definiert sind als «der geplante und gezielte Eingriff politisch einflussreicher Akteure [den *principals*] in bestehende Handlungsabläufe und -routinen, die tatsächlich oder vermeintlich im Einflussbereich derselben liegen, mit dem Ziel, ein erwünschtes Gut zu erlangen oder ein bestehendes Übel zu beseitigen oder zu mindern». Unübersehbar ist der Zusammenhang

mit der uns nicht unbekannten Kausalanalyse (mit input X wird ein output Y erwirkt). Das einfache Modell geht von einer direkten ungestörten Wirkung der Intervention auf den Outcome aus, eventuell mit einigen beobachtbaren sowie unbeobachtbaren Einflüssen verbunden, weswegen man eine Kontrollgruppe benötigt bzw. Experimente durchführt. Ein komplexeres Modell aber berücksichtigt die Heterogenität von Akteuren [hier sind die *agents* gemeint] und Handlungsketten. Dies betrifft nicht nur den Umstand, dass Interventionen bei jedem Akteur (*agent*) aufgrund persönlicher Voraussetzungen unterschiedliche Effekte bewirken können, sondern auch, dass diese die Intervention und ihre Ziele anders einschätzen als der Auftraggeber (hier der *principal*). Es geht also nicht nur darum, dass die Akteure unterschiedlich kreativ oder ungeschickt auf eine Intervention reagieren, sondern oft gar andere Ziele verfolgen als angedacht. Dies kann auch das Auftreten von unerwarteten Nebeneffekten bewirken. Damit ist bereits die objektive, präzise und somit gültige Messung eines standardisierten Outcomes in Frage gestellt. Verbunden damit sind auch die Zweifel, ob der *agent* in einem Fragebogen die Items so versteht und interpretiert wie beabsichtigt, und ob er die Motivation hat, diese Items korrekt zu beantworten.

[16] Diese Betrachtungen wurden mit Beispielen aus der Evaluationspraxis veranschaulicht. Im ersten wurde der Unterschied zwischen einer Validierung im statistischen Sinne zur Feststellung von Konsistenz beziehungsweise Reliabilität, Wiederholbarkeit und statistischer Assoziation auf der einen Seite, und der Validierung eines Messindikatoren im inhaltlichen Sinne auf der anderen Seite herausgestellt. Ersteres kann durch verschiedene statistische Kennzahlen untersucht werden, während Zweiteres durch *think-aloud-Interviews* (Befragte sprechen ihre Gedanken laut aus während sie den Fragebogen ausfüllen) oder *comprehension probing* (Zusatzfragen zum Verständnis werden eingebaut) überprüft werden kann. Hierbei geht es also nicht um die Frage, ob das eine oder andere sinnvoller und wichtiger ist, sondern darum, dass wir erst im Zusammenspiel beider Elemente von einer gründlichen Validierung der Indikatoren reden können. Diskutiert wurde dies am Beispiel des Frageitems «das Seminar ist vermutlich für die spätere Berufspraxis sehr nützlich». Hinzu kommt das Problem, dass auf einem standardisierten, vereinfachten Fragebogen keine Fragen auftauchen, die vom Forschenden nicht antizipiert, also in seinen Hypothesen nicht vorgesehen waren, wie etwa nichtintendierte, unerwartete Nebeneffekte. Diese werden somit nicht erfasst.

[17] Das zweite Beispiel betonte den Umstand, dass bei der rein quantitativen Erfassung von ausschliesslich input und output die kausalen Mechanismen und Pfade, über die die Intervention wirkt, unbekannt und damit eine Blackbox bleiben. Hier sieht man nochmals die Verbindungen zu den beiden anderen Hauptbeiträgen: ohne angemessene Theorie dürfte das Öffnen der Blackbox ein äusserst schwieriges Unterfangen sein; und ohne eine gewisse empirische Unterstützung, nun im induktiven Sinne, entbehrt die Bewertung aber schnell der Evidenz (bleibt also anekdotisch, wenn nicht spekulativ). Das folgende Beispiel stammt aus einer Studie zum Thema Ausbildung und Erwerbssystem in der DDR; siehe Sackmann, Weymann, und Wingens (2000). Untersucht wurde die Effektivität der staatlichen Karrierekontrolle nach dem Studium. Laut der rein quantitativen Datenlage basierend auf über 500 Befragten funktionierte das System ziemlich gut, denn ca. 60% der Befragten nannten die offizielle Vermittlungsstelle als zentrale Informationsquelle für ihre Arbeitsplatzsuche. Bei qualitativen Interviews zeigte sich aber, dass die Akteure in hohem Masse kreativ waren, ihre Karriere dennoch selbst zu bestimmen: «Der beschriebene bürokratische Mechanismus erwies sich in vielen Fällen als eine legitimatorische Fassade für eine individuelle Suche. […] Zunächst boten die Absolventen in Betrieben, wo sie hinwollten, ihre Arbeitskraft an. Da in fast allen Wirtschaftszweigen der DDR stets ein Mangel an qualifizierten

Arbeitskräften bestand, einigte man sich in der Regel schnell über die Einrichtung einer entsprechenden Planstelle, die dieser entsprechende Betrieb an die Vermittlungsbehörde weitermeldete. Die Ausschreibungsliste wurde dort nachträglich ergänzt, und der entsprechende Kandidat fragte dann gezielt diese zusätzliche Stelle an […]»

[18] In seiner Zusammenfassung betonte Herr Kelle nochmals, dass qualitative und quantitative Methoden für sich allein jeweils (die oben aufgezeigten) Stärken und Schwächen aufweisen, bei denen sich beide Ansätze aber oftmals gut ergänzen könnten.

5. Wahrhaftigkeit und Vertrauenswürdigkeit von Daten im Kontext der Evaluation

[19] Abgeschlossen wurde der Reigen durch die französische Präsentation von Diego Kuonen, Gründer und CEO von Statoo Consulting wie auch Professor an der Université de Genève. Er sprach über Wahrhaftigkeit und Vertrauenswürdigkeit von Daten im Kontext der Evaluation. Gleich zu Beginn wurde klar, dass es im Vortrag von Professor Kuonen auch um Entmystifizierung, Herausforderungen und Chancen der sogenannten Datenrevolution ging mit direktem Bezug auf die vielzitierten Schlagwörter wie Big Data, künstliche Intelligenz und/oder Machine Learning.

[20] Er startete mit einem Zitat von Michael Dell, dass Daten wohl die wichtigste natürliche Ressource dieses Jahrhunderts seien. Ihre Bedeutung speziell für die Schweiz steht daher auch im Mittelpunkt der Initiative von digitalswitzerland.com. Dies gilt für Daten (im Sinne von Information und Fakten) im Allgemeinen. Der Terminus «Big Data» beschreibt dabei zumeist nur Datensätze, die aufgrund ihrer Eigenschaften nicht ohne Weiteres mit traditionellen Datenmanagementprozessen oder -werkzeugen gehandhabt und analysiert werden können. Herr Kuonen fasste diese Eigenschaften unter den «drei V» zusammen: volume, velocity und variety. Keine dieser drei Eigenschaften bildet einen Wert oder Mehrwert an sich, sie beschreiben nur die wesentlichen Charakteristika von etwas, was in Fachkreisen als «Big Data» bezeichnet wird. Nun wies Edwards Deming bereits in den 40er Jahren des letzten Jahrhunderts darauf hin, dass Daten nicht zu Museumszwecken gesammelt werden oder werden sollten; einen Wert erhalten sie erst durch gute Analysen, denn Datenanalyse ist die eigentliche Maschine, die Daten selbst nur der Brennstoff. Dank geeigneter Analysen können wir davon reden, von Daten zu lernen oder aus Daten einen Sinn zu machen (populär: *making sense out of data*).

[21] Damit aber Daten (und wie gesagt, nicht nur Big Data) zu einer nützlichen und somit wertvollen Ressource werden können, benötigen diese zuallererst Eigenschaften, die Herr Kuonen im Wort «veracity» zusammenfasste, also Wahrhaftigkeit. Für die sogenannten Big Data muss hierbei der Rahmen für Datenqualität entsprechend erweitert, wenn auch nicht neu definiert werden. In Anspielung auf unser postfaktisches Zeitalter seien Vertrauenswürdigkeit und Zuverlässigkeit in der heutigen Welt wichtiger als jemals zuvor. Anzumerken ist hierbei, dass die Datensätze, auf die man sich im Zusammenhang mit big-data-Debatten oftmals bezieht, entweder keiner gezielten Umfrage noch Studienerhebungen entstammen (oder zumindest mit Blick auf völlig andere Fragestellungen gesammelt wurden als die, welche man selbst eigentlich betrachtet). Nicht selten entstanden sie sogar frei nach dem Prinzip, einfach alles zu sammeln, was beobachtbar und messbar ist.

[22] Im zweiten Abschnitt des Vortrags ging es um die Entmystifizierung von «Data Science» und statistischer Analyse, sowie ihrer Verbindung. Traditionelle statistische Analyse ging oder geht für gewöhnlich von einer a priori Hypothese aus, deren Validität anhand experimenteller oder speziell dafür gesammelter Daten getestet werden soll. Man spricht daher auch von einer top-down Methode oder dem deduktiven Ansatz, bei dem die Idee am Anfang steht. Ein Kritikpunkt ist hieran, dass somit die Daten selbst (im Sinne von Fakten) eben nicht an erster Stelle stehen. Data Science hingegen, hier nur im Sinne einer Neubenennung von «data mining» (Begrifflichkeit aus den 90er Jahren) verstanden, beschäftigt sich typischerweise mit der Analyse sogenannter Sekundärdaten, die also aus anderen Gründen gesammelt wurden (vergleiche auch mit Anmerkung zu Big Data oben). Solche Analysen sollen neue Ideen hervorbringen und gegebenenfalls durch so gemachte Beobachtungen neue Hypothesen generieren. Dies ist somit eher eine bottom-up Methode mit induktivem Ansatz, bei dem nicht die Idee, sondern die Daten an erster Stelle stehen. Daher sagt man auch, Algorithmen künstlicher Intelligenz sind nicht an sich intelligent; die Idee ist vielmehr, dass sie induktiv aus Daten lernen sollen. Ein gern geäusserter Kritikpunkt hieran ist, dass (alle) rein aus Daten abgeleitete Behauptungen – also ohne vorhergehende Idee, Hypothese oder Modell – sehr wahrscheinlich falsch seien; hierfür werden meist Beispiele aus der Kausalanalyse herangezogen, bei denen Anwender aus beobachteten Korrelationen irrtümlich auf bestimmte, direkte Kausalitäten schliessen. Ein anderes beliebtes Beispiel ist das sogenannte p-hacking, bei dem so viele Signifikanztests durchgeführt werden, dass man mit Sicherheit etwas scheinbar Signifikantes findet.

[23] Der daraus entstandene Streit, wenn auch weitaus weniger heftig geführt als der im zweiten Vortrag angesprochene Methodenstreit in den Sozialwissenschaften, erinnert uns dennoch insofern an jenen, als dass für den Aussenstehenden beide Ansätze doch weit mehr komplementär denn konträr sind. Tatsächlich geht die Diskussion über induktive versus deduktive Statistik und deren Zusammenspiel in die 70er Jahre des letzten Jahrhunderts zurück; siehe Box (1976). Man stelle sich dazu einen induktiv-deduktiven Zyklus vor, bei dem es dann meist unerheblich ist, womit er beginnt. Der berühmte Statistiker John W. Tukey meinte dazu, dass weder die explorative noch die konfirmatorische Statistik für sich allein (für datengetriebene Entscheidungen oder datenbasierte Politik) ausreichend seien und keine der beiden durch die andere ersetzbar sei. Dieser Vortragsabschnitt schloss mit einer Rückkopplung auf den ersten Teil: dass eine solide (im obigen Sinne) Datenbasis die allererste Grundvoraussetzung sei[2].

[24] Diese Anmerkung diente dem Übergang zum dritten Vortragsteil, der datenbasierten (Kuonen sprach von *data-informed*) Politikgestaltung. Wieder stand an dessen Anfang ein Zitat, nämlich, dass man sich «ohne Daten auf einem Blindflug befände, auf dem eine evidenzbasierte Politik nicht möglich sei» (Johannes P. Jütting, 2015). Nicht in solcher Totalität, aber vom Prinzip her sieht man Ähnliches auch an den sogenannten Politikpyramiden, die darstellen sollen, wie ein Mehrwert aus Daten für die Politik entstehen kann. Herr Kuonen zeigte im Vortrag eine dieser Illustrationen (als Beispiel, ohne Alleinstellungsanspruch): Das breite Fundament bildet hier immer die Datensammlung mit all ihren Schritten, anschliessend steigt die Pyramide auf über deskriptive, schliessende und schliesslich vorhersagende Statistik, um in der Verbesserung politischer Interventionen zu gipfeln. Auch wenn einen nur die Spitze dieser Pyramide oder des

2 Siehe dazu auch das Strategiepaper *Digitale Schweiz*, zu finden unter goo.gl/T8eJUS und angenommen durch den Schweizer Bundesrat am 5. September 2018.

Eisberges interessieren mag, sollte man sich immer dessen bewusst sein, was darunter liegt (unter Umständen verborgen), nämlich die Themen der beiden vorherigen Vortragsabschnitte.

[25] Von Pyramide und Eisberg wanderten wir zum nächsten Bild. Effektive Politikgestaltung beinhaltet Entwerfen, Monitoring und Evaluation, und jeder Schritt benötigt qualitativ hochwertige Daten, die die richtige Information am (oder zum) richtigen Ort zur richtigen Zeit bereitstellen. Man mag nun kritisieren, dies sei ein zu offensichtlicher aber ebenso zu allgemeiner Slogan; dennoch steht er so explizit im United Nations Secretary-General Report (UN Secretary General, 2014)[3]. Wenn wir die drei Schritte *Entwerfen*, *Monitoring und Evaluation* etwas genauer ausführen und zerlegen, dann kommen wir zum Politikzyklus (siehe Höchtl, Parycek und Schöllhammer, 2016) mit den etwas detaillierteren Schritten: *Setzen einer Agenda*, *Politikdiskussion*, *Politikformation*, *Politikakzeptanz*, *Mittelbereitstellung*, *Implementierung*, *Evaluation*; das Monitoring mag mehrere dieser Schritte begleiten. Traditionell wurden all diese Schritte als sukzessiv verstanden, mit einigen wenigen Rückkopplungen. Der uns vorgestellte revidierte Politikzyklus wurde nun in zweifacher Hinsicht erweitert. Zum einen wurde der ganze Zyklus von einem geschlossenen Kreis umspannt, der eine *kontinuierliche Evaluation* darstellte, und zum anderen zeigte jeder Pfeil in beide Richtungen; also war nun eine permanente Rückkopplung nach jedem Schritt vorgesehen. Genau hierzu können teils automatisierte induktive und deduktive Analysen ihren Beitrag leisten.

[26] Aus all dem lassen sich einige Schlussfolgerungen ziehen. Zuerst ist da die Notwendigkeit der Qualität der Daten, um Wahrhaftigkeit und Vertrauenswürdigkeit garantieren zu können. Als zweites die Wichtigkeit professioneller Datenanalyse, ohne welche Daten Museumsstücke blieben. Als drittes die oben gemachten Anmerkungen zu künstlicher Intelligenz, inklusive dem Verständnis, dass Datenanalysen beim Denken helfen sollen (Engelbart, 1962), es aber nicht ersetzen. Natürlich darf sogenannte künstliche Intelligenz routinierbare Vorgänge automatisieren (siehe oben); die Menschen aber müssen gleichzeitig ihre Kompetenzen stärken, um Nutzen aus der digitalen Transformation ziehen zu können und den Mehrwert von Daten zu erhalten. Dafür gibt es aber oftmals eine kulturelle Kluft zu überwinden.

6. Die Workshops

[27] Fünf Workshops am Freitagnachmittag widmeten sich jeweils einem spezifischen Aspekt des Kongressthemas und vertieften dieses in Form von Referaten, Diskussionen und teilweise Gruppenarbeiten.

Workshop 1 – Ursache und Wirkung: Kausalanalysen verstehen

[28] In Bezugnahme auf den zweiten und dritten Vortrag des Vormittags beschäftigte sich dieser Workshop mit der Frage, die geeignete Methode zu finden, wenn Kausalanalyse gefordert ist. Pirmin Bundi von der Université de Lausanne und Stefan Sperlich von der Université de Genève erläuterten zuerst einige Grundgedanken und -konzepte, was Kausalität meint und wie Ergebnisse folglich interpretiert werden können. Dabei stellten sie einen bislang noch wenig bekannten, obgleich mit den Wirkungsmodellen sehr verwandten Ansatz vor, wie Identifizierbarkeit von Kausaleffekten und daraus resultierenden Schätzmethoden auf anschauliche Art und Weise sicht-

[3] Siehe auch undatarevolution.org/report.

bar und somit verständlicher gemacht werden können. Dies kann auch gerade in Statistik weniger technisch ausgebildeten Evaluierenden helfen, einen Zugang zu diesem komplexen Thema zu verschaffen. Konkret geht es um die Zuhilfenahme graphischer Modelle, die einerseits intuitiv leicht zugänglich sind, für die es aber andererseits auch eine mathematisch rigorose Theorie gibt, die uns sagt, in welchen Situationen welche Effekte identifizierbar sind oder nicht, und zu welchen Methoden es dann zu greifen gilt. Diese Regeln gelten unabhängig von der Art der Information (quantitativ oder qualitativ). Zu den unterschiedlichen Situationen, meist jenseits von Experimentaldaten (aus Experimenten gewonnene Daten), wurden diverse Beispiele besprochen und praktisch durchgeführte Evaluationen vorgestellt.

Workshop 2 – Breiter abgestützte Bewertung dank Methodenmix[4]

[29] Der von Cornelia Hänsli Marrei (Habilis Conseil) und Christian Rüefli (Büro Vatter) vorbereitete und moderierte Workshop griff ebenfalls das Thema des zweiten Hauptreferenten auf. Quantitative und qualitative Methoden haben je spezifische Stärken und Schwächen. Anina Hanimann (Interface Politikstudien) zeigte in einem Referat auf, wie ein geschickter Methodenmix es erlaubt, die jeweiligen Vorteile zu nutzen, um zu einer umfassenderen, breit abgestützten Bewertung eines Evaluationsgegenstandes zu gelangen. Dabei lassen sich drei Funktionen der Kombination von verschiedenen Methoden unterscheiden: Validieren und Vertiefen; Erforschen und Verstärken; und Erkenntnisse erweitern. Anhand konkreter Praxisbeispiele zu jeder Funktion illustrierte sie, wie dies durch einen sequenziellen oder parallelen Einsatz qualitativer und quantitativer Methoden umgesetzt wurde. Im Anschluss daran diskutierten die Teilnehmenden in Gruppen drei Themen: (i) Chancen und Risiken von Mixed Method-Ansätzen, (ii) Voraussetzungen für den Einsatz von Mixed Method-Ansätzen in einer Evaluation sowie (iii) Herausforderungen in der Planung von Evaluationen mit Mixed Method-Ansätzen. Als Nutzen wurde eine Vertiefung der Erkenntnisse gesehen. Risiken bestehen bezüglich Aufwand und Kosten. Als Voraussetzungen orteten die Workshopteilnehmenden unter anderem genügend Ressourcen, ein gutes Verständnis des evaluierten Felds, Offenheit für den jeweils anderen Methodenansatz und ausreichende Methodenkompetenzen im Evaluationsteam, um einen möglichst passenden Methodenmix zu wählen. Konkrete Vorgaben, fehlendes Methodenverständnis von Auftraggebenden oder Vorbehalte gegenüber bestimmten Methoden können die Anwendung von Mixed Method-Ansätzen verhindern. Die Teilnehmenden erwähnten verschiedene Herausforderungen und Fragen in Zusammenhang mit dem Mix verschiedener Methoden, so etwa eine höhere Komplexität der Zusammenarbeit im Evaluationsteam und der Projektplanung, die Präsentation der Ergebnisse verschiedener Methoden oder den Umgang mit widersprüchlichen Ergebnissen. Die Schwierigkeit der Wertung qualitativer Daten lässt sich z.B. über Validierungsgefässe wie Workshops oder Ratingkonferenzen mit Fachpersonen auffangen.

Workshop 3 – Wirkung bewerten: Herausforderungen und Lösungsansätze[5]

[30] Dieser Workshop wurde moderiert von Claudia Peter (Geschäftsfeldleiterin Ecoplan AG sowie Vorstandsmitglied der SEVAL) und durchgeführt von Lilith Wernli, zuständig für Ex-Post Evaluationen in der Sektion Ökonomie des BAFU, und Franziska Müller, Mitglied der Interface Geschäftsleitung.

4 Zusammenfassung der Ausführenden.

5 Zusammenfassung von Philipp Zogg.

[31] Evaluationen haben häufig den Zweck, Wirkungsaussagen zu machen und dazu Wirkungen ab der Outcome-Ebene zu bewerten. Der dritte Workshop drehte sich um verschiedene Herausforderungen, die dabei auftreten können. Konkret wurde diskutiert, was getan werden kann, wenn wichtige Grundlagen fehlen (z.B. Wirkungsmodell), wenn eine Bewertung nach kurzer Interventionsdauer gefordert ist sowie wenn ein besonderes Interesse an der Impactebene besteht. Lilith Wernli (Bundesamt für Umwelt) sowie Franziska Müller (Interface) lieferten dazu anregende Inputs und mögliche Lösungsansätze aus Auftraggeber- sowie Auftragnehmersicht.

[32] Aus den Referaten sowie der anschliessenden Diskussion wurde deutlich, dass in allen Fällen eine klare und offene Kommunikation und damit verbunden auch realistisches Erwartungsmanagement zentral sind. So kann ein gemeinsames Verständnis dazu entwickelt werden, wozu die Evaluation dienen soll und kann sowie in welcher Form sie umsetzbar ist. Notwendige Grundlagen können gemeinsam erarbeitet werden, wobei sich die Teilnehmenden einig waren, dass der frühzeitigen Definition von Bewertungskriterien mehr Gewicht gegeben werden sollte – optimalerweise bereits während der Politikformulierung oder dem Design der Intervention – spätestens im Rahmen eines Evaluationskonzepts.

[33] Wenn Wirkungen noch nicht nachweisbar oder methodisch schwer fassbar sind, könnte verstärkt von Wirkungs*potenzial* gesprochen werden. Dies könnte auf Basis von vergangenen Evaluationen, wissenschaftlicher Literatur oder Expertengesprächen geschehen. In einem präsentierten Beispiel etwa wurde anstelle einer Kosten-Nutzen-Bewertung anhand transparenter Annahmen aufgezeigt, wie gross die Wirkung sein müsste, damit sich eine Massnahme finanziell lohnt. Um die Aussagekraft solcher Bewertungen zu erhöhen, sollten der vorliegende Kontext analysiert und die notwendigen Rahmenbedingungen für eine erfolgreiche Realisierung des Potenzials aufgezeigt werden. Generell waren die Teilnehmenden der Verwendung von Mixed Methods gegenüber positiv eingestellt, um die Resultate validieren und kontextualisieren zu können.

Workshop 4 – Applying Data Science[6]

[34] Die Idee des Workshops war, eine Vorstellung davon zu vermitteln, was *Data Science* jenseits der üblichen Schlagwörter wie unter anderem Machine Learning, Big Data oder künstliche Intelligenz bedeutet. Der Workshop wurde von Stefan Rieder, Präsident der SEVAL, moderiert. Yara Abu Awad, Senior Data Scientist des Bundesamts für Statistik (BFS), stellte zunächst die Grundzüge von Data Science und Big Data vor. Es wurden die Grundkonzepte sowie ausgewählte Anwendungen präsentiert. Data Science lässt sich definieren als *concept to unify statistics, data analysis, informatics, and their related methods in order to understand and analyze actual phenomena with data*. Anschliessend stellte Pauline Maury-Laribière, Data Engineer des BFS, ein Beispiel aus dem Bereich des Bundesamtes für Strassen vor. Mittels automatischer Erkennung von Fehlern bei Messsensoren können fehlende Daten nachträglich rekonstruiert werden. Mittels des Beispiels wurde das Grundkonzept von Data Science veranschaulicht. Im dritten Schritt konnten die Workshopteilnehmenden in Gruppen einen Fall auswählen und versuchen, Data Science darauf anzuwenden. Die Ergebnisse wurden diskutiert und führte zu folgenden Schlussfolgerungen: (i) Die Anwendung von Data Science in der Evaluation weist ein hohes Potential auf. Es lassen sich zahlreiche Felder und Fragestellungen erkennen, in denen Data Science bei Evaluationen eingesetzt werden könnte. (ii) Die Anwendung von Data Science im Kontext der Evaluation bedingt

6 Zusammenfassung von Stefan Rieder.

aber eine klare Fragestellung, die vor der Auswahl der Methoden und Techniken erfolgen muss. Ansonsten wird der Einsatz von Data Science zum Selbstzweck, was im Kontext der Evaluation keinen Mehrwert bietet. (iii) Es braucht Personen, die die Übersetzungsarbeit zwischen den Evaluationsfragen und der Anwendung von Data Science leisten können. Idealerweise verfügen diese Personen über Wissen aus beiden Bereichen. (iv) Der Einsatz von Data Science kann aufwendig und entsprechend teuer sein. Daher braucht es Kriterien, mit deren Hilfe über den Einsatz von data science in Evaluationen entschieden werden kann.

Workshop 5 – Evaluation der COVID-Massnahmen: Methodische Herausforderungen und Zugang zu Daten

[35] Mathias Rickli von der Eidgenössische Finanzkontrolle und Michael Funk von Swiss Economics widmeten sich einem Evaluationsthema aus jüngster Zeit: Die COVID-Krise und die politischen Interventionen. Diese war(en) und ist oder sind Gegenstand zahlreicher Evaluationen, Analysen und Forschungsarbeiten, bei denen oftmals die Evaluation getroffener Massnahmen im Mittelpunkt standen, darunter jene zur Beurteilung des Krisenmanagements als auch jene zur Unterstützung der Wirtschaft. In diesem Workshop wurden Erfahrungen zur Vorgehensweise bei der Durchführung von Evaluationen und methodische Fragen besprochen. Diskutiert wurde, welche methodischen Herausforderungen es gab und welche Ansätze man wählte. Wichtige Punkte spielten auch der Zugang zu Daten sowie die Vorteile und Grenzen quantitativer Analysen, beziehungsweise welchen Mehrwert die Ergebnisse bieten. Als Grundlage für die Diskussion wurde über die Evaluation der Unterstützung zugunsten von Selbstständigerwerbenden berichtet als auch eine Studie im Auftrag des SECO zur Wirkung von nicht-pharmazeutischen Massnahmen auf den Verlauf der Corona-Pandemie. Moderiert wurde der Workshop von Laurent Crémieux, Mitarbeiter des EFK und ebenfalls SEVAL Vorstandsmitglied.

Stefan Sperlich, Professor für Statistik und Ökonometrie an der Université de Genève, Geneva School of Management and Economics, stefan.sperlich@unige.ch.

Literatur

Box, Georg Edward P. (1976): Science and statistics. Journal of the American Statistical Association, 71, 791–799.

EES (2014): Evaluation Capabilities Framework der European Evaluation Society, https://europeanevaluation.org/wp-content/uploads/2020/03/The-EES-Evaluation-Capabilities-Framework-leaflet.pdf.

Engelbart, Douglas C. (1962): Augmenting human intellect, dougengelbart.org/content/view/138.

Filstead, William J. (1979): Qualitative Methods: A Needed Perspective in Evaluation Reserach, in: Thomas D. Cook / Charles S. Reichardt (eds.), Qualitative and Quantitative Methods in Evaluation Research, Bevery Hills (CA), 33–48.

Höchtl, Johann / Parycek, Peter / Schöllhammer, Ralph (2016): Big data in the policy cycle: policy decision making in the digital era, Journal of Organizational Computing and Electronic Commerce, 26, 147–169.

Johnson, R. Burke / Onwuegbuzie, Anthony J. (2004): Mixed Methods Research: A Research Paradigm Whose Time Has Come, Educational Researcher, 33, 14–26.

Kelle, Udo (2022): Der Methodenstreit in den Sozialwissenschaften und die Evaluationsforschung, Folienvortrag auf dem SEVAL Kongress September 2022.

Kuonen, Diego (2022): Veracity and Thrustworthiness of Data in the Context of Evaluation: Demystification, Challenges and Opportunities, Folienvortrag auf dem SEVAL Kongress September 2022.

Mark, Melvin M. (2018): Strengthening Links Between Evaluation Theory and Practice, and More: Comments Inspired by George Grob's 2017 Eleanor Chelimsky Forum Presentation, American Journal of Evaluation, 39, 133–139.

Rossi, Peter Henry / Freeman, Howard E. (1993): Evaluation: A systematic approach, 5. Auflage, Thousand Oaks/ London/New Delhi.

Sackmann, Reinhold / Weymann, Ansgar / Wingens, Matthias (2000): Die Generation der Wende: Berufs- und Lebensverläufe im sozialen Wandel, Wiesbaden.

Scriven, Michael (2007): Key Evaluation Checklist, www.wmich.edu/evalctr/checklists.

Schnell, Rainer / Hill, Paul B. / Esser, Elke (1999): Methoden der empirischen Sozialforschung, 6. Auflage, München.

Schröter, Daniela (2022): Wege zur Bewertung, Folienvortrag auf dem SEVAL Kongress September 2022.

UNEG (2016): Evaluation Competency Framework der UN Evaluation Group, www.unevaluation.org/2016-Evaluation-Competency-Framework.

UN Secretary General (2014): United Nations Secretary-General's Report der Independent Expert Advisory Group on a Data Revolution for Sustainable Development, A Word That Counts: Mobilising The Data Revolution for Sustainable Development, November 6.