DeGEval
Gesellschaft für Evaluation e.V.

...Info

Zeitschrift für Evaluation
19. Jahrgang 2020, Heft 2, S. 372-378
https://doi.org/10.31244/zfe.2020.02.12

Kausalität und Plausibilität – Evaluation zwischen Wissenschaft und Praxis

Frühjahrstagung 2019 des Arbeitskreises ‚Methoden in der Evaluation' der DeGEval – Gesellschaft für Evaluation e.V.

Anna Katharina Kleinoscheg[1]

Evaluierende stehen heute mehr denn je vor der Aufgabe, den wissenschaftlichen und methodischen Ansprüchen an Evaluierung gerecht zu werden. Per Definition müssen diese den Anforderungen wissenschaftlicher Güte entsprechen und sind somit ‚angewandte Wissenschaft'. Die Frage adäquater Methodenanwendung ist in den Evaluierungen somit inhärent und folglich Gegenstand anhaltender Diskussionen. Ein zentrales Thema ist hierbei der Einsatz adäquater Methoden zur Kausalitätsanalyse, die es ermöglichen, qualitativ hochwertige und wissenschaftlichen Standards genügende Designs vor dem Hintergrund eingeschränkter zeitlicher und finanzieller Ressourcen anzuwenden. Neben dem vermeintlichen Gold-Standard experimenteller Ansätze und quasi-experimentellen Designs stehen mittlerweile mit sogenannten theoriebasierten Evaluierungsansätzen auch wissenschaftstheoretische Alternativen zur Verfügung, die nicht auf der Untersuchung eines Kontrafaktums beruhen.

Trotz dieser wissenschaftstheoretischen Fundierung unterscheiden sich Evaluierungen in der Praxis teilweise stark von wissenschaftlichen Studien. Einschränkungen in der Umsetzung anerkannter Methoden werden häufig mit Hinweis auf einen Unterschied zwischen Wissenschaft und Praxis als unumgänglich angesehen und auf ‚Plausibilität' bzw. ‚Plausibilitätsüberlegungen' verwiesen. Plausibilität wird hierbei als ‚alternatives Konzept' zu den in der Wissenschaft angewandten Methoden und Designs bzw. als zulässige Alternative zu gängigen wissenschaftlichen Gütekriterien beschrieben. Durch den Bezug auf Plausibilität soll die Reliabilität und Validität von Evaluierungsergebnissen gewahrt werden. Plausibilität wird jedoch nicht näher definiert und dementsprechend unterschiedlich und in Abhängigkeit der Evaluierenden genutzt. Letztendlich ist ungeklärt, was konkret unter ‚Plausibilität' verstanden wird. Ungeklärt ist auch, welcher konkrete Unterschied zwischen Wissenschaft und (Evaluations-)Praxis existiert, der den Bezug auf Plausibilität erforderlich machen

1 Freie Sozialwissenschaftlerin (Schwerpunkt: Entwicklungszusammenarbeit), Wien

könnte, und inwieweit dies mit einer geringeren methodischen Qualität von Evaluierungen einhergeht.

Diese wissenschaftlichen und methodischen Ansprüche zwischen Praxis und Theorie sind eine zentrale Frage für eine adäquate Methodenanwendung in der Evaluierung, welche im Zentrum der Frühjahrstagung 2019 vom 21. bis 22. Juni stand. Organisiert vom Arbeitskreis ‚Methoden in der Evaluation' in der DeGEval (Prof. Dr. Alexandra Caspari und Dr. Tobias Polak) in enger Zusammenarbeit mit dem Institut für Internationale Entwicklung der Universität Wien (Prof. Dr. Antje Daniel und Dr. Cornelia Staritz) und der Austrian Development Agency (Mag. Sophie Zimm und Mag. Astrid Ganterer) diskutierten Wissenschaftler(innen) verschiedenster universitärer und privater Forschungseinrichtungen mit Praktiker(inne)n der verschiedenen Institutionen und Organisationen für Entwicklungszusammenarbeit vor der übergeordneten Frage, ob und wie sich Evaluierungen in methodischer Hinsicht von wissenschaftlichen Studien unterscheiden.

Im Zentrum der Tagung standen die verschiedenen fachlichen Perspektiven zum Begriff ‚Plausibilität'. Sie stellten sich die Frage, inwieweit Plausibilität wissenschaftlicher Güte entsprechen kann, beziehungsweise welche Mindeststandards für diese gegeben sein müssen.

Durch die Tagung hindurch zeigte sich, dass es durchaus unterschiedliche Definitionsansätze von Kausalität und Plausibilität und ihrem Verhältnis zueinander gibt. Eine gemeinsame Lösung und Definition von Plausibilität wurde von den Tagungsteilnehmenden schlussendlich nicht gefunden.

In einer Fishbowl-Diskussion (moderiert von Dr. Mag. Karin Fischer, ÖFSE) unter dem Titel „Wissenschaft und Praxis: Sichtweisen und Anforderungen von Evaluierungen" trafen aus wissenschaftlicher Perspektive Prof. Dr. Anje Daniel und Mag. Alexandra Heis (IE), aus praktischer Perspektive Mag. Sophie Zimm und Mag. Erwin Künzi (ADA) zusammen und eröffneten die Frühjahrstagung 2019 mit einer ersten Diskussionsrunde.

Kausalität und Plausibilität in experimenteller und quasi-experimenteller Evaluierung

Prof. Dr. Conny Wunsch von der Universität Basel stellte als erste Vortragende wissenschaftliche Evaluierungen anhand von experimenteller und quasi-experimenteller Evaluierung dar. Am Beispiel der Fragestellung: „Verkürzt eine Ausbildung auf dem Arbeitsmarkt die Dauer der Arbeitslosigkeit von Arbeitssuchenden?" erläuterte sie die kausale Wirkung bzw. den Unterschied zwischen Kausalität und Plausibilität.

Das Verständnis von Evaluierung aus wissenschaftlicher Perspektive führte sie vorerst ein, indem sie zuerst auf den Unterschied zwischen Kausalität und Korrelation einging und auch erklärte, dass es zu falschen politischen Schlussfolgerungen kommen kann, wenn der Kausaleffekt nicht isoliert wird.

Den Unterschied zwischen Kausalität und Plausibilität, der im Zentrum der ganzen Tagung stand, beschrieb Dr. Wunsch folgendermaßen: „Eine Evaluierung ist

plausibel, wenn die kausalen Effekte der Politik glaubwürdig isoliert und quantifiziert werden.“ Plausibilität ist nicht nur Kausalität, sondern auch Glaubwürdigkeit, Reliabilität und interne Validität.

Das erste und wichtigste Ziel jeder Evaluierungsstudie ist, Frau Dr. Wunsch zufolge, die Gewährleistung einer hohen internen Validität. Interne Validität meint die Plausibilität und Glaubwürdigkeit der Studie selbst. Externe Validität hingegen meint die Gültigkeit der Evaluierungsergebnisse außerhalb des Kontextes der Studie. Zur Schaffung der internen Validität ist der Ausschluss von Kausaleffekten durch Randomisierung ein Standard der Evaluierung, womit auch der Ausschluss von Drittvariableneffekten (Störfaktoren, die abhängige und unabhängige Variablen beeinflussen können und dadurch Ergebnisse verändern) gemeint ist. Das bedeutet, dass die Veränderung des abhängigen Faktors, in ihrem Beispiel die Dauer der Arbeitslosigkeit von Arbeitssuchenden, eindeutig von dem unabhängigen Faktor, hier Ausbildung, getrennt werden können muss.

Sie führte weiter aus, dass die Wahl der Forschungsmethode für die Erreichung der internen oder externen Validität relevant ist. Eine hohe interne Validität ist durch experimentelle Methoden gegeben. Quasi-experimentelle Methoden nutzen hingegen vorhandene Variationen aus, womit sie eine hohe externe Validität aufweisen. In ihrer Einführung zu den theoretischen Grundlagen der Evaluierungsforschung führte Dr. Wunsch weiter aus, wie der Kausaleffekt isoliert werden kann, und ging genauer auf Randomisierung und Drittvariableneffekte ein.

Als Herausforderung für wissenschaftliche Evaluierung nannte sie die Notwendigkeit des Publizierens sowie die oft mangelnden Ressourcen und den Termindruck. Wissenschaftler(innen) an wissenschaftlichen Instituten führen, ihrer Ansicht nach, eher dann eine Evaluierung durch, wenn diese ein hohes Veröffentlichungspotenzial aufweist, also innovativ ist und somit eine hohe interne Validität verspricht.

Somit liegt Frau Dr. Wunsch zufolge ein grundsätzlicher Unterschied in der Evaluierung zwischen Wissenschaft und Praxis in der internen Validität. Ihrer Ansicht nach haben Praktiker(innen) meistens bessere Kenntnis über Institutionen und die Aspekte, die in der Praxis relevant sind. Mit Praktiker(inne)n meinte sie dabei Wissenschaftler(innen), die reine Auftragsforschung betreiben und gewinnorientierte Evaluierung durchführen. ‚Reine‘ Wissenschaftler(innen) wenden häufig Methoden an, die für politische Entscheidungstragende und Praktiker(innen) sehr komplex und teilweise schwer verständlich sind. Während Wissenschaftler(innen) im Sinne der Wissenschaft forschen und ein gesamtes Bild haben wollen, sind Praktiker(innen) aufgrund des Auftrags auf bestimmte Aspekte besonders fokussiert. Praktiker(innen) sind vor allem aufgrund des Zeitdrucks und den meist relativ engen Fristen manchmal in der internen Validität eingeschränkt.

Sie fasste ihren Vortrag damit zusammen, dass Wissenschaftler(innen) einen anderen Anreiz haben als Evaluator(inn)en, jedoch keiner von beiden aus einem der beiden Gründe besser ist. Das Treffen pragmatischer Entscheidungen für die praktische Anwendung von Evaluierungen sieht sie nicht als verwerflich an, es dürfe jedoch nicht zu Lasten der internen Validität gehen. Nicht kausale Evaluierungen (Non-Causal Evaluations) können ebenfalls Ergebnisse liefen, sollten aber auf keinen Fall für Politikberatung (Policy Advice) verwendet werden.

Plausibilität und kausale Behauptungen in theoretischer Evaluierung

Prof. Dr. Derek Beach, University of Aarhus, eröffnete seine Präsentation mit dem Argument, dass fallbasierte, mechanismusorientierte (Mechanism-Focused) Evaluierung zu grundlegend unterschiedlichen Arten des Erkennens von Kausalzusammenhängen führt, die anhand von sehr unterschiedlichem empirischem Material belegt werden. Daraus ergeben sich seiner Ansicht nach zwei parallele Beweishierarchien für kausale Schlussfolgerungen, die beide Herausforderungen bezüglich der externen Validität aufweisen. Einerseits der varianzbasierte Ansatz (Top-Down), der kontrafaktische Behauptungen über kausale Auswirkungen zwischen Fällen in randomisierten, kontrollierten Studien aufzeigt. Andererseits der fallbezogene Ansatz (Bottom-Up), die mechanistischen Behauptungen über die Verknüpfung von Ursache und Ergebnis innerhalb eines Falls zu prüfen. Diese ermöglichen zwei parallele Beweishierarchien. Sowohl beim varianzbasierten Ansatz als auch beim fallbezogenen Ansatz gibt es Studien, welche sich verstärkt auf externe Validität oder interne Validität berufen.

In seinem Vortrag fokussierte Prof. Dr. Derek Beach auf die fallbezogene, mechanismusfokussierte Evaluierung. Der Fall (Case) ist seiner Ansicht nach ein Beispiel für eine Intervention, die nach der Ursache forscht und damit zu einer Lösung (Outcome) führt. Ursachen arbeiten seiner Ansicht nach immer innerhalb von Fällen, auch wenn es sich um randomisiert-kontrollierte Studien (RCT=Randomised Controlled Trial) handelt. In der praktischen Evaluierung von Fallstudien werden ‚Contribution Analysis' und ‚Theory of Change' verwendet. Das Problem der beiden Methoden für Evaluierungen sieht er in zweierlei Form. Einerseits werden Teile des Prozesses mit Annahmen verknüpft. Andererseits ist es unklar, was Beweise tatsächlich sind. Aus diesem Grund sieht Prof. Dr. Beach ‚Process-Tracing' als das bessere methodische Konzept.

Daher ging er in seinem Vortrag in weiterer Folge auf das Zusammenspiel von Plausibilität und Process-Tracing ein. Dies beginnt er mit der Unterscheidung von interner und externer Plausibilität. Interne Plausibilität ist demnach die Stärke der Schlussfolgerung über den Kausalmechanismus innerhalb des Falls. Theoretisch wird durch die interne Plausibilität eine produktive Kontinuität von Aktivitäten, die verschiedene Teile der Evaluierung miteinander verlinken, klar gemacht. Empirisch zeigt sie eine stark mechanistische Evidenz für alle Aktivitäten. Die Frage ist, wie stark der Grund und das Ergebnis zusammenhängen, da es verschiedene Möglichkeiten gibt, um an das Ergebnis zu kommen. Dies zeigte Prof. Dr. Beach am Beispiel eines Bildungsprogramms für Mütter zur Ernährung ihrer Kinder auf. Die mechanistische Evidenz innerhalb von solchen Fallbeweisen kann in Muster, Sequenz, Pfad und Beschreibung/Bericht unterschieden werden. Die externe Plausibilität hingegen dient vordergründig strategischen Untersuchungen von Zielpopulationen, um festzustellen, ob diese auf die gleiche Weise funktionieren.

Dies bedeutet, dass die Auswertungsmethoden zur Prozessverfolgung sehr unterschiedliche Fragen benötigen, welche anhand von sehr unterschiedlichem Material belegt werden können. Die erforderliche (interne oder externe) Plausibilität hängt von der Evaluierungssituation ab. Prof. Dr. Beach zufolge ist eine hohe interne und

hohe externe Plausibilität in der Evaluierung erstrebenswert. Typisch sind jedoch eine hohe interne und eine moderate externe Plausibilität. Aus diesem Grund sollten in größeren Evaluierungsprojekten randomisiert-kontrollierte Studien und fallbezogene Studien verwendet werden, denn die erstgenannte zeigt die Nettokausaleffekte auf, wohingegen die zweite die Funktionsweisen eines Systems unter bestimmten Bedingungen aufzeigt.

Aus den beiden Vorträgen von Prof. Dr. Conny Wunsch und Prof. Dr. Derek Beach zu Kausalität und Plausibilität zeigte sich sehr gut, dass die Ansätze sehr verschieden sein können, ein Verständnis und die Notwendigkeit von Qualität jedoch unumgänglich ist.

Evaluierungsdesign: ein Spagat zwischen Machbarkeit, Nutzen und wissenschaftlicher Strenge

Dr. Jos Vaessen, IEG Methods Advisor, World Bank Group, sprach gleich zu Beginn seiner Präsentation von Evaluierung als angewandte sozialwissenschaftliche Forschung. Allerdings verwies er auf eine Kluft zwischen Evaluierung und Wissenschaft. Forschung versucht, neues Wissen zu generieren, während Evaluierung Information zur Entscheidungsfindung schafft. Der triangulären Vernetzung von Nützlichkeit, Genauigkeit und Durchführbarkeit stehen Theorien der Evaluierung gegenüber und bedingen sich gegenseitig.

Methodische Qualität (Rigor) bei der Evaluierung kann laut Dr. Vaessen unterschiedlich konzipiert werden. Sie kann sowohl von sachlichen (theoriegeleiteten) oder wertmäßigen Prämissen abhängen. Praktische Evaluierung konzentriert sich seiner Ansicht nach viel expliziter auf letzteres, also wertmäßige Prämissen. Wertmäßige Prämissen versuchen, Qualität und Nützlichkeit von Evaluierungen in Einklang zu bringen. Verschiedene Schulen und Theorien der Evaluierung gehen hier jedoch unterschiedlich vor. Die Rolle der Theorie in der Evaluierung kann sich grob zwischen akademischer Theorie und Programmtheorie (meint die Anwendung zu den Evaluierungen von Programmen in Projekten in der praktischen Entwicklungsevaluierung) bewegen. Durch die verschiedenen Theorieströnge werden unterschiedliche Zielgruppen, Absichten und Verwendungszwecke in der Evaluierung angewandt.

Zur Erklärung dieser wertgemäßen Prämissen erläuterte Dr. Vaessen die „World Bank Group Evaluation Principles“ (April 2019). Die Evaluierungsforschung der World Bank Group beruht auf der Theorie des Leistungsbewertungssystems. Durch die Einbeziehung mehrerer kausaler Optionen und Effekte, wie zum Beispiel Arbeitsmarkteffekte oder nationale Skalierungseffekte, kommt es zu Wissen und Bewusstsein eines Phänomens. Ein zentrales Anliegen der Evaluierungsarbeit der Weltbank ist es, eine Ausgewogenheit zwischen methodischer Qualität und Machbarkeit zu schaffen. Umfang und Tiefe sowie die Notwendigkeit deduktiver und induktiver Untersuchungen sollen schnell herausgefunden werden und der Umgang mit Komplexität und Kausalität geklärt werden. In der Erläuterung der Dimensionen von Evaluierung zeigte Dr. Vaessen die Komplexität dieses Vorhabens auf. Schließlich

setzt sich eine Evaluierung in der Praxis aus Intervention, Institutionen, Kausalität und Veränderungen, der Einbettung, der Natur des Systems und der Evaluierungskriterien als solche zusammen. Um diese methodische Herausforderung der praktischen Evaluierung aufzuzeigen, präsentierte Dr. Vaessen ein systemanalytisches Beispiel durch die Evaluierung eines Projekts der Weltbank im Bereich Stadtentwicklung in Nicaragua.

Dr. Jos Vaessen fasste seinen Vortrag zusammen, indem er auf den Nexus von Forschung und Evaluierung einging. Die Evaluierung folgt methodischen Standards in der akademischen Forschung, soweit dies „machbar und nützlich" ist. Die gründliche Ursachenanalyse kann dabei eine eigenständige Übung oder Teil einer umfassenderen Evaluierung sein. Der Nexus zeigt sich dadurch, dass akademische Forschung eine wichtige Grundlage und somit Teil einer umfassenden Evaluierung ausmacht. Das bedeutet, dass sich akademische Forschung und Evaluierung überschneiden, Evaluierung jedoch auch ihre eigenen (trans-)disziplinären Merkmale aufweist.

Kausalität und Plausibilität: Unterscheidungskriterien zwischen Evaluierung und Wissenschaft?

Der zweite Tagungstag diente der Zusammenfassung und Aufarbeitung der Beiträge der Vortragenden sowie zur Ergebnispräsentation der Arbeitsgruppen vom Vortag. Es ging darum, die große Vielfalt an Beiträgen sowohl in methodischer, als auch in thematischer Sicht in Einklang zu bringen.

Dafür präsentierten als erstes die drei Arbeitsgruppen, thematisch gegliedert nach Berater(inne)n, Auftraggeber(inne)n und Wissenschaftler(inne)n ihre Ergebnisse aus den Diskussionen rund um eine mögliche Definition von Plausibilität. Die Gruppe der Berater(innen) beschäftigte sich zentral mit dem Aushandlungsprozess von Evaluierungen, in dem schwerwiegende Konflikte auftreten können. Die Gruppe der Auftraggeber(innen) sah die Trennung von Evaluierung und Wissenschaft als eine schwierige Herangehensweise. Ihrer Ansicht nach gibt es keinen Grund, Wissenschaft als etwas Separates von Evaluierung zu betrachten. Die Gruppe der Wissenschaftler(innen) beschäftigte sich mit dem Konzept von Plausibilität. Die Herausforderungen innerhalb der Gruppe bezogen sich dabei vor allem auf die Frage der Messmöglichkeiten und Quantifizierbarkeit von Plausibilität. In den Präsentationen der Arbeitsgruppen zeigte sich, dass Evaluierung in der Praxis und in der Forschung unterschiedlich verstanden wird. Am Ende wurde in keiner der drei Gruppen eine Einigung in Hinblick auf eine Definition von Plausibilität gefunden. Dies wurde auch von Frau Dr. Conny Wunsch im weiterführenden Gespräch zwischen den drei Vortragenden mit Dr. Tobias Polak aufgegriffen. Sie wies noch einmal darauf hin, dass über die Tagung hinweg verschiedene Definitionen von Plausibilität verwendet wurden.

Um in der abschließenden Diskussion noch einmal für Denkanstöße zu sorgen, schlug Dr. Tobias Polak eine weitere Definition von Plausibilität vor, die vor allem die Grenzen von Forschung im Kontext der internen Validität aufzeigte. Konkre-

ter: wenn zwar ein Zusammenhang zwischen Ursache und Wirkung aufgezeigt werden kann, aber die Drittvariableneffekte nicht erkennbar sind. Dr. Conny Wunsch zufolge sei dafür vor allem ein Commitment für Transparenz notwendig und die Auseinandersetzung und Definition der Grenzen in der Evaluierungsforschung. Auch Prof. Dr. Derek Beach stimmte diesem Argument zu und fügte der Diskussion hinzu, dass die Agenturen, die Evaluierungen beauftragen, herausfinden müssen, wie plausibel ihre Designs in Hinblick auf interne oder externe Validität sind. Die Vortragenden waren sich in weiterer Folge auch einig, dass es nicht notwendig ist, hier konkrete Schwellenwerte (Thresholds) für Plausibilität und Kausalität zu quantifizieren. Der Schlüssel liege, laut Dr. Jos Vaessen, in der Zuverlässigkeit und Transparenz der Evaluierung. Zentral sei es, eine genaue Fragestellung zu generieren und zu klären, mit welcher Methodenauswahl die Antwort durch die Datensammlung am besten beantwortet werden kann.

Schlussendlich wurden die Vorträge und Argumentationen der Tagung „Evaluierung zwischen Theorie und Praxis“ treffend zusammengefasst, indem darauf hingewiesen wurde, dass Diskrepanzen zwischen Plausibilität und Kausalität in Bezug auf methodische Aspekte möglicherweise nicht immer so groß sind. Für die Evaluierung ist es aber auf alle Fälle relevant, dass die Transparenz über die Methoden klar gegeben sein muss.