Zeitschrift für Evaluation
20. Jahrgang 2021, Heft 1, S. 208-211
https://doi.org/10.31244/zfe.2021.01.11

Rezension zu:

Roorda, Mathea Bendino Shulamith: Developing Defensible Criteria for Public Sector Evaluations. Melbourne: The University of Melbourne, 2020. 290 Seiten. Verfügbar unter: https://minerva-access.unimelb.edu.au/handle/11343/239232

Wolfgang Beywl[1]

Evaluation soll glaubwürdige und genaue beschreibende Informationen liefern und systematisch bewerten. Die Methodologie der Evaluation ist bezüglich der empirischen Deskription – aufbauend auf die sich stets weiter entfaltende sozialwissenschaftliche Methodenlehre – weit fortgeschritten. Hingegen findet sich vergleichsweise wenig dazu, wie die für die Bewertungen notwendigen Kriterien systematisch gewonnen und begründet werden. Mathea Roorda will mit ihrer Dissertation einen Beitrag leisten, diese für die wissenschaftliche Güte von Evaluation ausschlaggebende Lücke zu schließen. Die Veröffentlichung beschließt ein mehrjähriges multimethodisches Forschungsprojekt. Am Anfang steht die Fragestellung, wie ein solides Set von Evaluationskriterien für Programmevaluationen entwickelt werden kann.

Der Forschungsbericht ist in drei Hauptteile mit elf Kapiteln gegliedert:

1. Einführung

Kapitel 1 stellt den Entstehungszusammenhang der Studie und deren Anlage vor. Außerdem wird die Ausgangsfragestellung aus eigenen über 20-jährigen Evaluationserfahrungen und Bezügen zur Literatur abgeleitet.

Kapitel 2 enthält eine umfangreiche Aufarbeitung der relevanten Fachliteratur zur Thematik: systematische Bewertung in Evaluationen und Stellenwert von Kriterien in diesem Schlüsselprozess. Zentrale Begriffe wie „Evaluation", „Wertbehauptung" (*evaluative claim*) oder „evaluatives Urteilen" (*evaluative judgement*) werden präzise geklärt. Besondere Aufmerksamkeit gewidmet wird dem Begriff „Kriterien". Hierunter werden Konkretisierungen von Wertdimensionen verstanden, in Bezug auf welche Qualität oder Erfolg beurteilt werden. In der Literatur bestünden zu diesem zentralen theoretischen Konzept viele Unklarheiten. Dies betreffe zum Beispiel die Differenz zwischen Kriteriendimensionen, Kriterien, operationalisierten Kriterien und insbesondere zu den Kriterienpunkten (*standards*). Ähnliches gelte für den Unterschied zwischen den gleichermaßen relevanten expliziten und impliziten Kriterien. Bezüglich der Berechtigung zum Bewerten vertritt die Autorin den Standpunkt, dass es nicht angemessen sei, den Stakeholdern das Bewerten exklusiv zu überlassen.

Als zentral wird der Prozess des „evaluativen Schlussfolgerns" (*evaluative reasoning*), auch „evaluative Logik" genannt, angesehen. Diesen gliedert die Autorin in fünf Schritte:

1. Festlegung der Kriterien – Wertprämisse (*establish criteria – value premise*)
2. Festlegung der Kriterienpunkte (*establish performance standards – value premise*)
3. Messung der tatsächlichen Performanz des Programms (*measure performance – factual premise*)
4. Vergleich Kriterienpunkte – tatsächliche Performanz (*compare performance-standards – measured performance*)
5. Synthese und Integration in eine Beurteilung (*synthesis and integration into a judgement – evaluative claim*).

Die aus verschiedenen maßgeblichen Abhandlungen destillierte Übersicht (Tabelle 2) enthält 14 Quellen (*sources*), aus denen Kriterien geschöpft werden können. Diese umfasst wesent-

1 Univation GmbH, Köln, und Fachhochschule Nordwestschweiz

lich „deskriptive Kriterien“, die üblicherweise durch Auswertung von Dokumenten oder Interviews mit Stakeholdern gewonnen werden. Allerdings seien zusätzlich „präskriptive Kriterien“ erforderlich. Diese könnten „von außen“ gewonnen werden, indem über den Kreis der direkten Stakeholder hinaus weitere Interessen einbezogen werden. Oder die Evaluierenden selbst schlagen solche Kriterien vor. Um damit ausgelöste Wertkonflikte systematisch bearbeiten zu können, elaboriert die Autorin das Konzept der „normativ-ethischen Perspektiven“, verbunden mit drei konkurrierenden Ansätzen der westlichen Philosophie:

1. Konsequentialismus (u.a. Utilitarismus)
2. Deontologie (unter anderem pflichtgemäßes, rechtebezogenes und auf Fairness/Gerechtigkeit ausgerichtetes Handeln)
3. Tugendethik (mit Bezug auf kontextgebundene persönliche Beziehungen, Care-Ethik)

Die Autorin identifiziert drei einschlägige Überblicksstudien zur Fragestellung, wie in Evaluationen Kriterien gewonnen und begründet werden. Deren systematische Auswertung ergibt, dass dieser Schlüsselprozess der Evaluation selten systematisch konzipiert und durchgeführt wird. Das Resümee des ‚theoretischen‘ Kapitels 2 lautet: Während die Evaluationswissenschaft einiges dazu bereithält, wie man mit konfligierenden deskriptiven Kriterien verschiedener Stakeholder umgeht, gibt es kaum etwas dazu, wie divergierende normative Perspektiven in die Begründung von Kriterien einbezogen werden können.

Kapitel 3 stellt die historische Entwicklung und den aktuellen Stand der Programmevaluation in Australien und Neuseeland dar. Nach Einschätzung der Autorin werden Evaluationen in beiden Ländern auf pragmatische Weise und theoretisch eher wenig reflektiert durchgeführt. Die neuseeländische Kultur weise mit der Berücksichtigung der Werte und Rechte der Maori Besonderheiten auf, die sich auch auf das Verständnis von Evaluation auswirkten.

Kapitel 4 erläutert ausführlich das vielgliedrige methodische Mixed-Methods-Vorgehen. Die instruktive Tabelle 4 visualisiert dieses und verbindet es zur Forschungsfragestellung. Das Forschungsdesign hat sich von Anfang an emergent entwickelt: Wenn sich bestimmte Untersuchungen als nicht durchführbar bzw. nicht weiterführend erwiesen, wurde die Forschungsanlage angepasst.

2. Kriterien in der aktuellen Evaluationspraxis

Die Phase 1 der Entwicklungs- und Untersuchungsarbeit wird in den Kapiteln 5 bis 7 dargestellt. Zusammen mit der Literaturauswertung des Kapitels 2 soll geklärt werden, wie mit Kriterien in der Programmevaluationspraxis der beiden Länder umgegangen wird. Wie werden Konzepte wie Programmeffektivität (*program effectiveness*), kulturelle Responsivität (*cultural responsiveness*), Zugänglichkeit (*accessibility*) und Effizienz (*efficiency*) inhaltlich konkretisiert? Welche Quellen nutzen die Evaluierenden? Wie stellen sie die Kriterien in Evaluationsberichten dar und wie begründen sie diese?

Kapitel 5 berichtet über die Ergebnisse einer Onlinebefragung von 137 überwiegend Vollzeitevaluierenden aus Australien und Neuseeland. Wesentliche Ergebnisse sind: Evaluationsfragestellungen sind als Bezugspunkte für Bewertungen meist wichtiger als explizierte Kriterienpunkte. Quelle für Bewertungskriterien sind vorwiegend leicht zugängliche Dokumente, insbesondere bezüglich der Programmziele. Es gibt zwei unterschiedliche Schwerpunktsetzungen bei der Gewinnung von Kriterien: einerseits auf Auftraggebende und andere zentrale Stakeholder; andererseits (auch) auf andere vom Programm Betroffene bzw. Programmitarbeitende sowie Forschungsliteratur.

Kapitel 6 wertet 47 neuere Evaluationsberichte daraufhin aus, welche Rolle Kriterien bei den vorgenommenen Bewertungen spielen. Die Stichprobenziehung erfolgt gemäß maximaler Variation. Lediglich in sechs Berichten würden „Kriterien“ explizit herangezogen, während dies in den anderen implizit erfolge. Gemäß Häufigkeit ihres Vorkommens werden folgende Kriteriendimensionen identifiziert: Effektivität, Outcomes, Impact, Angemessenheit, Effizienz, Alignment (also Konsistenz). Die Bedeutungsinhalte dieser genutzten Begriffe unterschieden sich zwischen den Berichten teilweise erheblich. Oft blieben die Kriterien, welche den Bewertungen zugrunde gelegt werden, implizit. Sie seien in den Berichten oft ungenügend operationalisiert. Dadurch könne es dazu kommen, dass Bewertungsprozesse kaum nachvollziehbar sind. Selten würde der Entwicklungsprozess von Kriterien offengelegt. So sei es für die Lesenden der Berichte kaum nachvollziehbar, wie die gemachten Bewertungen zustande kommen.

Kapitel 7 fasst die Ergebnisse der Phase 1 zusammen. Gemäß der Literaturanalyse sollten Sets von Kriterien für die Evaluation folgende Merkmale aufweisen:

1. Nutzung maßgeblicher (*authoritative*) Quellen, und zwar hinausgehend über üblicherweise leicht zugängliche Dokumente etwa zu den Programmzielen
2. Vollständigkeit, also Repräsentation aller wichtigen Wertperspektiven, über diejenigen der Auftraggebenden oder der Programmverantwortlichen hinaus
3. Vollständige Beschreibung, also explizite Definitionen und Operationalisierungen

Eine Gegenüberstellung dieser Anforderungen aus der Evaluationstheorie mit den empirischen Befunden aus der Umfrage und der Inhaltsanalyse der Evaluationsberichte zeigt: Es besteht „eine Diskrepanz zwischen der Theorie über evaluatives Denken […] und der Praxis, was die Evaluierenden nach eigener Aussage tun, und was in den Programmevaluationsberichten beobachtet werden kann“ (S. 145)

3. Zur systematischen Entwicklung von Kriterien

Die Erkenntnisse aus der ersten Phase führen zu einer neuen Forschungsfragestellung für Phase 2: „Was ist ein evidenz-informierter konzeptioneller Rahmen für verteidigbare Kriterien, die für die Evaluationspraxis im öffentlichen Sektor operationalisierbar sind?“

Kapitel 8 entwickelt aus zwei Traditionen (die eine aus der Evaluationstheorie, die andere aus dem Bereich der Ethik bezüglich der Entwicklung der Biotechnologie) einen konzeptionellen Rahmen für ein umfassendes Set von Kriteriendimensionen. Dieser enthält sowohl deskriptive als auch präskriptive Elemente. Ausgangspunkt sind die im Kapitel 2 aufgearbeiteten Ethik-Theorien. Es ergeben sich fünf mit folgenden Schlagworten betitelte Prinzipien: Konsequenzen, Pflicht, Rechte, Gleichheit, Care-Ethik.

Eine Grafik visualisiert den konzeptionellen Rahmen, in dem für die fünf normativen Prinzipien konkrete Kriterien entwickelt oder vorhandene Kriteriensets auf Vollständigkeit überprüft werden können. Das präskriptive Vorgehen (Ableitung aus den Prinzipien) kann mit dem deskriptiven Vorgehen (Ansprüche der Stakeholder- oder Interessengruppen) kombiniert werden. Mit der Ausfüllung dieses Rahmens wird auf Vollständigkeit hingearbeitet: die Identifizierung aller relevanten Kriterien für einen Evaluationsgegenstand und dessen Evaluation.

Kapitel 9 schildert die mehrstufige, zyklisch angelegte Entwicklung eines Handbuches zu einer „Kriterienmatrix“, die als Werkzeug (*tool*) für die Evaluationspraxis dienen soll. Die Matrix enthält in Spalten die fünf kriterialen Prinzipien und in den Zeilen die Interessengruppen. Deren Perspektiven auf die Prinzipien resultieren in einer Vielzahl von sich teils überschneidenden Wertdimensionen (*value dimensions*) in den Zellen. Am Beispiel eines Programms für Saisonarbeitskräfte wird der Einsatz der Matrix veranschaulicht.

In jede nächste Stufe des Validierungsprozesses des Handbuchs werden die Rückmeldungen der in der vorherigen Stufe beigezogenen Experten eingearbeitet. Zwecks Qualitätssicherung orientiert sich die Validierung des Handbuchs am „*The Checklist Project Validation Process*“ der Universität von Western Michigan, einem weltweit beachteten Verfahren, mit dem Evaluationschecklisten geprüft werden.

Sieben Prüfpunkte werden für diesen umfassenden Validierungsprozess des Handbuchs genutzt. Zwei externe Gutachtende mit Ethik-Expertise checken die beiden Punkte „Genauigkeit“ sowie „Vollständigkeit und Relevanz“. Die drei Checkpunkte „Klarheit der Absicht“, „Organisation des Inhalts“, „Klarheit des Textes“ werden durch neun externe Evaluationsfachleute in zwei Runden überprüft. „Angemessenheit des Inhalts“ (*appropriateness of content*) – also Verbindung zu evaluationsfachlich relevanten Themen – sowie „Literaturreferenzen und Quellen“ werden intern durch die Autorin gesichert. Aus diesen Rückmeldezyklen sind mehrere relevante Klärungen und Anpassungen für die Schlussversion des Handbuchs hervorgegangen.

Kapitel 10 beschreibt die Felderprobung der weiterentwickelten Version des Handbuchs. Diese haben vier Evaluationsfachleute abgeschlossen. Drei von ihnen haben die Dokumentation ihrer Erprobungen zur Veröffentlichung freigegeben. Die Felderprobung zeige auf, dass Evaluierende mithilfe des Handbuchs normativ-präskriptive ethische Perspektiven einnehmen und zur Entwicklung eines vollständigeren Sets von Kriterien nutzen können. Sie würden unterstützt, relevante Kriterien zu identifizieren, die sie sonst möglicherweise ignoriert hätten. Das entwickelte Verfahren könne sowohl prospektiv Kriterien identifizieren als auch als retrospektiv ein vorhandenes Set von Kriterien überprüfen.

Kapitel 11 stellt zunächst die zentralen Ergebnisse der beiden Hauptphasen der Forschung zusammen. Es arbeitet die Bedeutung der Befunde für die Weiterentwicklung von Theorie und Methodologie der Evaluation heraus.

In einem eigenen Unterkapitel werden Begrenzungen angesprochen. Sie resultierten unter anderem aus den teils angefallenen Stichproben in der Onlinebefragung oder bei den ausgewerteten Evaluationsberichten, die vielleicht fassadenhaft geschrieben seien. Es fehlten auch Überle-

gungen dazu, wie Prioritäten zwischen Werten gesetzt werden könnten. Es sei noch zu klären, wie die Belegkraft für Kriterien systematisch bestimmt werden kann, und welche Personen oder Gruppen dies tun sollten.

Die Autorin stellt in ihrer abschließenden Reflexion einen nochmals revidierten konzeptionellen Rahmen zur Entwicklung verteidigbarer Bewertungskriterien vor. Evaluatoren müssten ihre Behauptungen (*claims*) über Wertprämissen verteidigen, indem sie Gründe angeben, die auf maßgeblichen Informationen oder Belegen beruhen. Diese Gründe müssten durch entsprechende Argumente untermauert werden. Hierbei müssten die Belege mit der jeweiligen Wert-Behauptung argumentativ verbunden werden.

Im Anschluss an das das Literaturverzeichnis werden im Anhang zur Verfügung gestellt: der Fragebogen der Online-Befragung und das zugehörige Begleitschreiben; das Codeschema für die Inhaltsanalyse der 47 Evaluationsberichte; die beiden Rückmeldebögen, die in den beiden Validierungsprozessen für die Kriterienmatrix genutzt wurden. Schließlich als Highlight das 18-seitige *Criteria Matrix Handbook*. Dieses bereitet den Forschungsertrag für die Evaluationspraxis auf und liefert eine didaktisch aufbereitete Grundlage zur Durchführung von Workshops zur Entwicklung verteidigbarer Evaluationskriterien.

4. Würdigung

Die Publikation leistet einen maßgeblichen und originären Beitrag zur Weiterentwicklung von Theorie und Methodologie der Evaluation. Zentral sind die Ausführungen zur Anwendung normativ-ethischer Kriterien in Theorie und Praxis der Programmevaluation. Bemerkenswert ist die konsequente Anwendung von Kriterien auf den eigenen Forschungsprozess und das hohe Ausmaß, in dem dieser transparent gemacht wird. Der Text ist für theoretisch in der Evaluation Arbeitende ein Muss, wenn Sie dem – wohl auch in der deutschsprachigen Evaluation – oft vernachlässigten Schlüsselelement der Kriterien den ihnen angemessenen konzeptionellen Stellenwert zukommen lassen wollen. Für die Evaluationspraxis bietet er wertvolle intellektuelle Anregungen bis hin zu konkreten Hilfestellungen für eine kriterienbewusste Praxis.

Weiterführend könnte man fragen: Welche Überschneidungen, welche Unterschiede gelten für die systematische Identifizierung und Begründung von Kriterien für formativ ausgerichtete Evaluationen einerseits, summative andererseits? Muss für beide Evaluationsrollen ähnlicher Aufwand getrieben werden? Sind zu fixierende Kriterienpunkte für beide Evaluationsrollen gleich stark zu fordern, auch wenn ein Evaluationsgegenstand (und dessen Evaluation) in ständiger Entwicklung ist (siehe „evolutive Evaluation“/“*developmental evaluation*”), also nicht mehr in den Status einer formativen, geschweige denn summativen Evaluierbarkeit findet?

Doch selbst, wenn wir unterstellen, dass die Evaluationswissenschaft in den kommenden Dekaden angesichts einer veränderten Weltlage einen tiefgreifenden Wandel vollziehen muss: Als Basis dafür, wie Evaluation zu ihren Kriterien kommt, liefert Mathea Roorda ein unhintergehbares Standardwerk, wie es lange keines mehr in der internationalen Evaluationsliteratur gegeben hat.