Methodische Standards der Evaluation zwischen Schema F und Innovation – Weiterführung einer politikfeldübergreifenden Diskussion

Franziska Heinze[1], Stefanie Reiter[1]

Im Rahmen der 20. Jahrestagung der Gesellschaft für Evaluation (DeGEval) zum Thema „Evaluation (in) der Zukunft“ in Mainz (20.-22.09.2017) organisierte der Arbeitskreis (AK) *Methoden in der Evaluation* in Kooperation mit weiteren Arbeitskreisen in der DeGEval eine Session als Beitrag zur Anregung einer politikfeldübergreifenden Diskussion zu methodischen Standards. Der AK *Methoden in der Evaluation* setzte damit seine auf der DeGEval-Jahrestagung 2016 begonnene Veranstaltungsreihe zur Bestandsaufnahme von Evaluationsansätzen und methodischen Standards in verschiedenen Arbeitsfeldern fort. Die Veranstaltungsreihe wurde im Format einer Fishbowl-Diskussion (Innenkreis-/Außenkreis-Methode) durchgeführt, an der sich jeweils drei weitere Arbeitskreise der DeGEval beteiligten. Diese berichteten zunächst über typische und innovative Evaluationsansätze und Methoden für ihr Feld und diskutierten darauf aufbauend politikfeldübergreifend methodische Standards sowie Innovationspotenziale von Evaluationsmethoden.

Bei der Auftaktveranstaltung 2016 begannen die Arbeitskreise *Kultur und Kulturpolitik*, *Entwicklungspolitik* sowie *Forschungs-, Technologie- und Innovationspolitik* den Austausch von Erfahrungen hinsichtlich der Anwendung von Evaluationsmethoden in den jeweiligen Arbeitsfeldern und diskutierten Gemeinsamkeiten sowie Unterschiede der Ansätze (siehe Altenburg 2017). Die intensive Diskussion brachte als Ergebnis, dass die evaluative Arbeit in den drei beteiligten Themenbereichen maßgeblich von einem Spannungsfeld zwischen Qualitätsanforderungen an Methoden und zur Verfügung stehenden Ressourcen geprägt ist. Die resümierende Feststellung, dass die Wahl der Evaluationsmethoden sich weitgehend an der vorherrschenden Evaluationspraxis im eigenen Handlungsfeld orientiere und selten als Anregung ein Blick auf in anderen Evaluationsfeldern verbreitete Ansätze und Methoden erfolge, bestätigte die Organisatoren aus dem AK *Methoden in der Evaluation*, die 2016 begonnenen Diskussionen mit dem gleichen Format unter Beteiligung weiterer Arbeitskreise fortzusetzen. Entsprechend wurde

1 Deutsches Jugendinstitut e.V., Halle

auf der DeGEval-Jahrestagung 2017 eine gemeinsame Session mit den Arbeitskreisen *Soziale Dienstleistungen*, *Hochschulen* und *Gesundheitswesen* veranstaltet.

Als Diskussionsteilnehmende waren folgende Personen im Innenkreis des Fishbowl-Formats gesetzt:

- PD Dr. Rainer Strobl – Sprecher des AK *Soziale Dienstleistungen*, (proVal – Gesellschaft für sozialwissenschaftliche Analyse – Beratung – Evaluation, Hannover, strobl@proval-services.net),
- Prof. Dr. Philipp Pohlenz – Sprecher des AK *Hochschulen*, (Otto-von-Guericke-Universität Magdeburg, philipp.pohlenz@ovgu.de),
- Marcus Capellaro – von 2011 bis 2017 Sprecher des AK *Gesundheitswesen*, (Capellaro GmbH – Konzeption & Evaluation kommunikativer Maßnahmen, M@Capellaro.de).

Ihre Statements und Diskussionen wurden durch Wortbeiträge weiterer Personen aus dem Außenkreis ergänzt, die in abwechselnder Beteiligung die freien Stühle des Innenkreises besetzten und so zur Diskussion beitrugen. *Prof. Dr. Alexandra Caspari* (Evaluationsforschung, Methoden der empirischen Sozialforschung und Statistik; Frankfurt University of Applied Sciences, Fachbereich 4, Soziale Arbeit und Gesundheit) und *Dr. Jan Tobias Polak (*Evaluator, Evaluierung zivilgesellschaftliche Entwicklungszusammenarbeit, entwicklungspolitische Bildungsarbeit; DEval – Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit gGmbH Bonn) aus dem Sprecherteam des AK *Methoden in der Evaluation* führten wie bereits 2016 die leitfragengestützte Diskussion.

Typische und innovative Evaluationsmethoden – eine Bestandsaufnahme

Rainer Strobl erörterte in seinem Eingangsstatement zunächst die große Heterogenität des Feldes der *Sozialen Dienstleistungen* und eine damit einhergehende Vielfalt von Evaluationsmethoden, welche die Identifizierung von innovativen Ansätzen und typischen Vorgehensweisen ‚nach Schema F' erschwere. Zu den Besonderheiten des Feldes der *Sozialen Dienstleistungen* zählten der Aspekt der Koproduktion[2] sowie eine oftmals sehr geringe finanzielle Ausstattung der Maßnahmen und Evaluationen.[3] Die Ressourcenknappheit trage zu einem hohen Anteil an Selbstevaluationen in diesem Politikfeld bei.

Im Bereich dieser *Selbstevaluation* sind die Möglichkeiten zum Einsatz eines breiten methodischen Spektrums eher begrenzt. Es werden häufig in erster Linie Methoden zur Feststellung von Zufriedenheit mit und Akzeptanz von Maßnahmen eingesetzt, um diese Ergebnisse in die Qualitätssicherung und Qualitätsentwicklung einspeisen zu können. Methodische Designs, welche darüber hinaus Effekte aufzei-

2 Dies bedeutet, dass die Dienstleistungen in enger Kooperation mit den Klientinnen und Klienten produziert werden. Zu den Implikationen von Koproduktionsprozessen siehe auch Halves/Lück-Filsinger/Schmidt 2014.

3 Zu den Besonderheiten des Feldes ist ein Kurzbericht des Arbeitskreises *Soziale Dienstleistungen* der DeGEval veröffentlicht worden (Reiter/Strobl/Buchheit 2017).

gen und Wirkungen nachweisen können, sind hingegen selten, wenngleich die Frage der Wirkungen in diesem Feld häufig debattiert und auch vonseiten der Auftraggeber zusehends Nachweise für Wirksamkeit eingefordert werden. In der Konsequenz wird in einigen Selbstevaluationen zumindest versucht, Wirkungsziele und Indikatoren für Wirkungen festzulegen, anhand derer wahrgenommene Effekte sich bspw. im Rahmen einer (Selbst-)Beobachtung einordnen lassen.

Der sehr heterogene Bereich der *Fremdevaluation* ist nach Einschätzung von *Rainer Strobl* durch den Einsatz qualitativer und quantitativer Methoden gekennzeichnet, die sich an sehr unterschiedlichen Zielgruppen orientieren. Beispielsweise ist im Bereich der frühkindlichen Bildung die Datenerhebung bei Kindern oftmals nur mittels überwiegend qualitativer Verfahren, wie teilnehmender Beobachtungen und Videoanalysen oder unter Einsatz bestimmter Gesprächsformen, möglich. Aus dem Bereich quantitativ ausgerichteter Verfahren werden randomisierte kontrollierte Studien selten eingesetzt, wenngleich es in einigen Feldern wie der Kriminalprävention immer wieder Plädoyers dafür gibt. Möglichkeiten des Einsatzes von (quasi-) experimentellen Designs werden unter Evaluierenden diskutiert und zum Teil realisiert.[4] Schließlich lasse sich als Gemeinsamkeit vieler Ansätze von Fremdevaluation in diesem Arbeitsfeld identifizieren, dass der Einsatz qualitativer und quantitativer Methoden weit verbreitet sei, das Spektrum an triangulativen Verfahren im Rahmen von ‚echten' Mixed-Methods-Designs jedoch – mitunter aus Ressourcengründen – selten gänzlich ausgeschöpft wird.

Philipp Pohlenz kündigte in seinem Eingangsstatement eine kritische Perspektive auf das Feld der *Evaluation im Hochschulbereich* an. Er berichtete eingangs über die Bemühungen in den späten 1990er Jahren, Evaluation in Lehre und Hochschule als Instrument der Legitimation zu verankern, und damit einhergehenden, weit verbreiteten (argumentativen) Abwehrreaktionen seitens Lehrender, welche „durch systematisiertes Beurteilen der Qualität von Lehre und Studium" die Autonomie ihres Lehrhandelns bedroht sahen.[5] Dies führte zu immer wiederkehrenden Diskussionen um den Einsatz gegenstandsangemessener Verfahren, in denen auch die Frage der Repräsentativität von Daten aufgeworfen wurde. Akzeptanzprobleme (auch von qualitativen Daten) haben zur Weiterentwicklung von messbaren Indikatoren und zu nach ‚Schema F' reproduzierbaren Befragungen v.a. von Studierenden „zum Zwecke des Vergleichs einzelner Lehrleistungen" (bspw. im Rahmen der zumeist summativen Evaluation von Lehrveranstaltungen) beigetragen.

Zum jetzigen Zeitpunkt sieht *Philipp Pohlenz* – vor dem Hintergrund von Gewöhnungseffekten und tendenziell geringeren Akzeptanzproblemen von Evaluation – Potenziale zur Etablierung von Innovationen. Entsprechende Innovationsbedarfe von Verfahren und Nutzungsszenarien ergeben sich dabei auch aus Entwicklun-

4 Beispielsweise wurde auf der Frühjahrstagung 2015 des AK *Soziale Dienstleistungen* zum Thema „Methodische Herausforderungen der Wirkungsanalysen bei knappen Ressourcen" intensiv diskutiert, welche ethnischen Bedenken es bei der Bildung von Kontrollgruppen in ausgewählten Settings gibt und welche Alternativen (z.B. Wartegruppen, Cross-Over-Designs) möglich sind (vgl. Reiter et al. 2015).

5 Eine Kurzdarstellung von wesentlichen Entwicklungslinien der Evaluation an Hochschulen ist in einem Bericht des AK *Hochschulen* der DeGEval in der ZfEv 2/2017 erschienen (vgl. Mittauer/Pohlenz/Harris-Huemmert 2017: 275-276).

gen jenseits des Qualitätsdiskurses im engeren Sinne, die von neuen Zielgruppen in der Lehre[6] und sich verändernden gesellschaftlichen Erwartungen an Hochschullehre geprägt seien. Erste innovative Schritte in Richtung neuer Wege der Hochschulevaluation sieht er in den an vielen Orten erkennbaren, gemeinsamen Bemühungen von Evaluierenden und der oftmals an völlig unterschiedlichen Orten verankerten und institutionalisierten Hochschuldidaktik. Diese Wege beziehen sich tendenziell nun „nicht so sehr auf die Frage nach dem Vergleich zwischen Lehrenden X und Lehrender Y", sondern „viel unmittelbarer auf das konkrete Lehrhandeln" (bspw. Überprüfung des eigenen Handelns vor dem Hintergrund des eigenen Professionsverständnisses in der *hochschulinternen Selbstevaluation*).

Mit Blick auf weitere, historisch gewachsene Verfahren der Qualitätsbegutachtung von Lehre und Studium im Rahmen von *hochschulexternen Akkreditierungen* (siehe auch Bologna-Prozess) konstatiert er unterschiedliche Herangehensweisen (u.a. Peer-Review-Verfahren).

Bei den Qualitätsanforderungen an Hochschulen, die durch die verschiedenen Verfahren, sowohl intern im Sinne von Selbstevaluation als auch extern durch Akkreditierungsverfahren betrachtet werden, gibt es ein einigermaßen konsentiertes Qualitätsverständnis, wodurch in der Evaluation zusehends inhaltliche Kriterien von Qualität eine Rolle spielen. So lassen sich laut *Philipp Pohlenz* bspw. in neueren, von außen vorgegebenen Standards für die externe Qualitätssicherung der Hochschullehre, z.B. in den 2015 überarbeiteten *Standards and Guidelines for Quality Assurance* (vgl. European Association for Quality Assurance in Higher Education 2015), bindende inhaltliche Qualitätskriterien finden. Da zuvor Qualitätssicherung und Evaluation viel stärker nur auf Prozesse in der Lehre bezogen waren, könnte dieser *Trend in Richtung Inhaltsqualität* bei aller Vorsicht als Innovation bezeichnet werden.

Marcus Capellaro stützte sein Eingangsstatement neben eigenen Erfahrungen als freier Evaluator auf Ergebnisse einer Onlinebefragung der Mitglieder des AK *Gesundheitswesen*.[7] Er erläuterte zunächst die unterschiedlichen Bereiche des Politikfeldes, die sich von Gesundheitsförderung und Prävention bis hin zu Therapie, Krankheitsbewältigung, Rehabilitation und Pflege erstrecken, sowie die Zusammensetzung des AK, da sich diese Aspekte auf die Befragungsergebnisse auswirkten. Im AK *Gesundheitswesen* sind Politik, öffentliche Verwaltung, Fachhochschulen, Universitäten und nicht universitäre Forschungseinrichtungen relativ gut vertreten, wohingegen Kostenträger (Krankenkassen und Pflegeversicherungen) sowie Leistungserbringende (Ärztinnen und Ärzte, Gesundheits- und Krankenpflegerinnen und -pfleger sowie andere Pflege- und Gesundheitsberufe) nicht repräsentiert sind. Die Onlinebefragung brachte hervor, dass einige Mitglieder die im Politikfeld nach Ein-

6 Zur ansteigenden Heterogenität der Studierendenschaften siehe auch Harris-Huemmert/Mittauer/Pohlenz 2015 sowie Mittauer/Pohlenz/Harris-Huemmert 2017: 275.

7 An der Onlinebefragung 2017 beteiligten sich 29 von 100 auf der Verteilerliste des AK *Gesundheitswesen* eingetragene Personen.

schätzung von *Marcus Capellaro* relativ gut verfügbaren *Sekundärdaten*[8] nutzen, insgesamt jedoch weitaus häufiger Primärdaten erhoben werden. Zu den am häufigsten angewandten Datenerhebungsmethoden zählen hierbei Interviews und standardisiertere Befragungen, des Weiteren werden Beobachtungen im Feld oder im Labor unter den Befragten häufig durchgeführt. Medizinische Untersuchungen (physiologische Messungen etc.) und psychologische Testverfahren werden hingegen von den Befragungsteilnehmenden relativ wenig genutzt.[9]

Die auch in der Gesundheitsförderung und Prävention weit verbreitete *Forderung nach evidenzbasierter Medizin* gehe mit einem Appell zur Anwendung methodisch hochwertiger und randomisierter, kontrollierter Studien einher. Der daraus entstehende Druck auf die Forschenden und Studienanbietenden führe zu Diskussionen hinsichtlich des Nutzens solcher Designs.[10] Eine im Politikfeld kontrovers diskutierte These besagt, dass der Nutzen von diesen randomisierten, kontrollierten Studien aufgrund der zu großen Unterschiede in den Populationen der Intervention und der Kontrollgruppen überschätzt werde. Vor allem im Bereich der präventiven Verhaltensänderung, aber auch bei Therapie müssten vielmehr individuelle Lebensumstände und sozio-kulturelle Erfahrungen („Wir müssen für den Erfolg von Prävention und Therapie den Mensch ganzheitlich betrachten.“) berücksichtigt werden. Innovationen seien dabei neben der *Anpassung von Verfahren und Designs an den Gegenstand* auch in den Bereichen *Digitalisierung* und intersektorielle Gesundheitspolitik *(Health in All Policies*) notwendig.

Politikfeldübergreifende Diskussion zu methodischen Standards

Die anschließende Diskussion rückte verschiedene Aspekte in den Mittelpunkt: Neben der Suche nach *innovativen Evaluationsansätzen* in den thematisierten Politikfeldern bildeten angemessene und realisierbare Verfahren zur Erfassung von *Wirkungen* in Evaluationen sowie Fragen zu (weiteren) *methodischen Standards* die drei zentralen Diskussionsachsen. Hinsichtlich der Frage nach Innovationen von Evaluation zeigten sich in den verschiedenen Politikfeldern recht unterschiedliche Anknüpfungspunkte. Im Bereich Soziale Dienstleistungen wurden hier vor allem methodische Innovationsbedarfe und -potenziale zur Aufklärung von komplexen, nicht linearen Zusammenhängen zwischen Intervention und Outcome benannt, welche von Evaluierenden geleistet werden können. Im Hochschulbereich ging es bei der Identifizierung von Innovationsbedarfen primär um die Neubestimmung bzw. Verän-

8 So gibt es bspw. Abrechnungsdaten von den niedergelassenen Ärztinnen und Ärzten mit Leistungen und Diagnosen, Qualitätsberichte der Krankenhäuser, Daten der epidemiologisch arbeitenden Krebsregister sowie die Studien des Robert Koch-Instituts zur Gesundheit von Kindern und Erwachsenen, die auch zur öffentlichen Nutzung freigegeben sind.

9 Nach Einschätzung von Marcus Capellaro könnte dieser Befund mit der AK-Struktur zusammenhängen, da u.a. Kostenträger und Leistungserbringende nicht vertreten sind.

10 Der AK *Gesundheitswesen* lässt dieser Frage besondere Aufmerksamkeit zukommen und veranstaltete bspw. 2015 eine Frühjahrstagung mit dem Titel „Evidenzbasierung in der Gesundheitsförderung: Anspruch, Wirklichkeit und der Beitrag der Evaluation“, um anhand konkreter Beispiele das Spannungsfeld zwischen Anforderungen und realer Evaluationspraxis aufzuzeigen und zu diskutieren (vgl. Wirtz/Capellaro/Spiel 2017).

derung von Nutzungsszenarien von Lehr- und Studiumevaluationen[11], weg von reinen Kontroll- bzw. Legitimationsinstrumenten, hin zu einem „Kulturwandel", der auf die Integration von Evaluation in das professionelle Selbstverständnis von Lehrenden zielt. Jedoch stelle sich die bislang eher getrennte Zuständigkeit für Evaluation, Hochschuldidaktik und Qualitätsentwicklung im Hochschulbereich als eher hinderlich für Innovationen in diesem Bereich dar. Im Gesundheitswesen speisen sich Innovationen im Bereich Evaluation u.a. aus Fragen der Gegenstandsangemessenheit bzw. aus der Arbeit mit spezifischen Zielgruppen (z.B. Demenzerkrankte). Auch die Einbeziehung von sozialen Medien bzw. die Nutzung von Big Data bergen hier sowie in anderen Bereichen Innovationspotenziale, wenngleich noch wenige Erfahrungen diesbezüglich vorhanden sind. Aus dem erweiterten Diskussionskreis vertrat *Christa Peinhaupt* (EPIG GmbH) darüber hinaus die Auffassung, im Gesundheitswesen werde zunehmend verstärkt in Programmen gedacht, d.h. in mehrteiligen, intentional aufeinander bezogenen Aktivitäten bzw. Interventionen. Daraus ergeben sich ihrer Auffassung nach Innovationen für und in der Gestaltung von Evaluationen, da diese nunmehr mehrere, auf ausgewiesene Ziele ausgerichtete Interventionen parallel in den Blick nehmen müss(t)en. *Marcus Capellaro* verdeutlichte diesen Gedanken am Beispiel des politikfeldübergreifenden „Health in All Policies"-Ansatzes.

Dieser Ansatz diente darüber hinaus als Beispiel, um über Fragen von Wirkungsmodellierung bzw. -messung im Rahmen von Evaluationen zu diskutieren und deren Nützlichkeit sowie Durchführbarkeit zu reflektieren. In allen drei Politikfeldern wurde deutlich, dass Evaluationen mit Anforderungen sowie Fragen der Realisierbarkeit und des Nutzens von *Modellierungen bzw. Messungen von Wirkungen* konfrontiert sind. Diese Anforderungen werden an Evaluierende insbesondere von extern, bspw. von Auftraggebenden im Zuge veränderter Formen der (daten- bzw. evidenzbasierten) Programmsteuerung im Feld der Sozialen Dienstleistungen, herangetragen. In den Bereichen Gesundheitswesen und Soziale Dienstleistungen bearbeiten Evaluierende vor allem Fragen der Wirkungszusammenhänge von Intervention und Outcome, die es zunächst qualitativ (z.B. über die Programmtheoriemodellierung oder die Rekonstruktion von theories of change) aufzuklären gelte (Wirkungsmodellierung). *Rainer Strobl* ergänzte mit Blick auf komplexe (Mehrebenen-)Programme, dass es hier darauf ankäme, insbesondere auch Ziele auf Programmebene möglichst konkret zu formulieren, um darauf aufbauend Aspekte der Zielerreichung durch die Projekte überprüfen zu können.

Gerade schwer messbare Wirkungsziele auf Outcome- und Impactebene würden Evaluierende jedoch vor besondere (methodische) Herausforderungen stellen bzw. Verfahren der Wirkungsmessung ließen sich oft gar nicht umsetzen. *Christa Peinhaupt* regte hier – mit Blick auf ein Spezifikum des Gesundheitswesens – an, die Vielzahl an vorhandenen Sekundärdaten (z.B. Sozialversicherungsdaten, Abrechnungsdaten der Krankenkassen) zu nutzen, um anspruchsvolle Kontrollgruppendesigns umzusetzen. Im Feld der Sozialen Dienstleistungen bestehe im Hinblick auf Fragen nach Wirkungen vor allem auch ein verstärktes Interesse daran, erfolg-

11 Dies schien zunächst widersprüchlich, da gerade Hochschulen bzw. Universitäten als grundsätzlich innovativ bzw. innovationsorientiert gelten.

reiche Strategien der Veränderung bspw. von Einstellungen oder Verhaltensweisen nach einer Intervention bzw. Maßnahme zu identifizieren. Für den Hochschulbereich schlug *Thorsten Braun* (Universität Stuttgart) vor, zum Nachweis von Wirkungen beispielsweise Prüfungen kompetenzorientiert zu gestalten und sie als Instrumente der Outcome-Messung zu nutzen. Gerade die Empirische Bildungsforschung stelle seiner Auffassung nach eine gute Referenz dar, wie kompetenzorientierte Tests, die probabilistische Testtheorie oder die Item-Response-Theorie für die Erfassung von Outputs und Outcomes nutzbar gemacht werden können. Seiner Ansicht nach ließen sich derartige Verfahren auch im Feld der Sozialen Dienstleistungen, bspw. in der politischen Bildung, einsetzen. Als Restriktion für entsprechende Wirkungsmessungen benannte er die oftmals gering(er)en Fallzahlen im Feld der Sozialen Dienstleistungen gegenüber der Evaluation von Lehre und Studium im Hochschulbereich. Auf die Unmöglichkeit der Wirkungsmessung mit Randomized Control Trials (RCT) in komplexen sozialen Programmen wies *Thomas Widmer* (Universität Zürich) hin unter Rekurs auf entsprechende Unternehmungen in den USA in den 1960er und 1970er Jahren. *Alexandra Caspari* schlug vor, anstelle von RCT quasi-experimentelle Designs in kleinen, ausgewählten Settings zu nutzen, um die Plausibilität von Wirkungsannahmen zu stützen. Beispielsweise könnten ausgewählte, einfach messbare Teiloutcomes mit Mixed-Methods-Verfahren gemessen bzw. nachgewiesen werden und darüber wiederum ein Beitrag zur Plausibilität der Gesamt-Outcomes geleistet werden. Aufseiten von *Thomas Widmer, Philipp Pohlenz und Rainer Strobl* wurde dies kritisch gesehen: Vor allem simple Wirkungsketten ließen sich mit quasi-experimentellen Designs bzw. RCT erfassen, komplexe Zusammenhänge in sozialen Programmen jedoch selten adäquat modellieren. Für die Bereiche Soziale Dienstleistungen und Hochschulen wurde darüber hinaus konstatiert, dass Koproduktionsprozesse den sinnvollen Einsatz (quasi-)experimenteller Designs erschwerten. *Thomas Widmer* merkte zudem an, dass die Fokussierung auf einfach messbare Teiloutcomes möglicherweise den Blick für die eigentlich relevanten Zusammenhänge und Wirkungen in sozialen Programmen verstelle.

Politikfeldübergreifend ebenso kontrovers diskutiert wurden *methodische Standards in der Arbeit mit qualitativen Daten* und hier insbesondere die Frage, ob Gespräche, Interviews oder Gruppendiskussionen nicht nur aufgezeichnet, sondern auch (voll-)transkribiert werden sollen. Dieser Punkt wurde – anders als von den gesetzten Diskussionsteilnehmenden – im erweiterten Kreis der Diskutierenden als Mindeststandard hinterfragt. Einerseits wurde eine (Voll-)Transkription vor dem Hintergrund von Aufwand und Kosten als nicht zwingend notwendig argumentiert, andererseits wurden Wissenschaftlichkeitsstandards als Argumente ins Feld geführt. Die Frage der Kosten wurde in Teilen dahingehend relativiert, dass es mittlerweile Spracherkennungssoftware gibt, die zumindest für Standardsprache eine mögliche Alternative zur Transkription bietet. *Susanne Mäder* (Univation) erläuterte, dass qualitativ hochwertige Evaluation und die Arbeit ohne Transkription aus ihrer Perspektive keinen Widerspruch darstellen würden. Sie umriss, wie über eine ausführliche Dokumentation der Daten mit integrierten Interpretationsschritten hinreichende Qualität erzeugt werden könne, insbesondere wenn eine inhaltsbezogene Auswertung im Vordergrund stehe. Vor dem Hintergrund dieser Vorgehensweise reflektierte

sie die Frage nach einer (Voll-)Transkription im Hinblick auf methodische bzw. gegenstandsbezogene Angemessenheit und die Passfähigkeit von Auswertungsmethode und -interesse. Während *Thomas Widmer* weniger die Transkription als vielmehr die ‚Black Box' der Interpretation von qualitativen Daten problematisierte, ergänzte *Christa Peinhaupt* die zusätzliche Herausforderung, dass Auftraggebende einerseits qualitative, differenzierte Aussagen nachfragen, andererseits aber die Ergebnisdarstellung in Tabellen oder Grafiken einfordern und die Länge qualitativer Evaluationsberichte als unpraktikabel bzw. wenig nutzbar bewerten. Diesbezüglich sieht sie es als notwendig an, Überzeugungsarbeit in Richtung Auftraggebende zu leisten.

Wohin bewegen sich Ansätze und Methoden in der Evaluation? – Fazit und Ausblick

Aus den Diskussionen der Jahre 2016 und 2017 ging hervor, dass die sechs betrachteten Politikfelder von einer mitunter großen Heterogenität von Subbereichen gekennzeichnet sind, somit typische Ansätze und auch der Grad der Standardisierung von Methoden nicht nur zwischen den Politikfeldern, sondern auch innerhalb dieser stark variieren können. Dies hängt sowohl mit den jeweiligen Gegenständen und den damit korrespondierenden angemessenen Methoden sowie Ansätzen als auch teilweise mit historischen Entwicklungslinien in den jeweiligen Feldern (bspw. Tradition der Selbstevaluation im Bereich der Sozialen Dienstleistungen, Qualitätsdiskurs durch den Einzug von New Public Management in die Hochschulen im Bereich der Hochschulevaluation) und weiteren Kontextbedingungen zusammen (siehe bspw. Unterschiede im Bereich der Forschungs-, Technologie- und Innovationspolitik in Deutschland und Österreich). Entsprechend muss die Frage nach typischen und innovativen Methoden immer auch im Kontext dieser Subbereiche gestellt und diskutiert werden. Dieser Befund der vorhandenen Heterogenität trug auch dazu bei, dass die in der Session 2017 gestellte abschließende Frage zur pauschalen Einordnung der methodischen Qualität von Evaluationen in den Politikfeldern von den Diskutierenden weitgehend zurückgewiesen wurde. Zugleich formulierten sie für die von ihnen repräsentierten Bereiche Soziale Dienstleistungen, Gesundheitswesen und Hochschule mittlere bis erhöhte Handlungsbedarfe im Hinblick auf die methodische Qualität von Evaluationen. Neben politikfeldspezifischen Besonderheiten (bspw. der 2016 konstatierten geringen Akzeptanz der Evaluation von Projekten im Bereich von Kunst und Kultur), die sich auf die gewählten Vorgehensweisen und Verfahren auswirken[12], zeichneten sich in den Diskussionen auch Kontextbedingungen ab, die für mehrere Politikfelder methodische Herausforderungen aufwerfen (bspw. der Aspekt der Koproduktionsprozesse für den Bereich der Sozialen Dienstleistungen und der Hochschulen). Darüber hinaus wurde aus den Austauschformaten der Jahre 2016 und 2017 ersichtlich, dass Methodenfragen bzw. entsprechende Min-

12 *Vera Hennefeld* aus dem AK *Kultur und Kulturpolitik* wies in der Diskussion 2016 in diesem Zusammenhang darauf hin, dass in der Konsequenz oftmals viel Aufklärungsarbeit über die Ziele von Evaluation – welche explizit nicht eine Bewertung von Kultur an sich intendiere – notwendig sei, um die Akzeptanz in diesem Feld zu erhöhen.

deststandards nicht nur mit Blick auf die Gegenstände von Evaluationen zu diskutieren sind, sondern auch maßgeblich von zur Verfügung gestellten finanziellen und zeitlichen Ressourcen abhängen. Hierbei scheinen – so der Eindruck aus der Diskussionsrunde 2017 – einige Politikfelder, z.B. das Feld der Sozialen Dienstleistungen, tendenziell stärkeren Restriktionen zu unterliegen als andere. Es zeigten sich über die Politikfelder hinweg auch Unterschiede in den Auswirkungen von Ressourcenfragen bzw. im Umgang mit diesen. Während in der Diskussion der Felder Kultur und Kulturpolitik, Forschungs-, Technologie- und Innovationspolitik sowie Entwicklungspolitik 2016 diese Aspekte beispielsweise mit Blick auf die Anforderungen einer transparenten Darstellung von Methoden und auf den Prozess der Erkenntnisgenerierung thematisiert wurden, standen 2017 vor allem Fragen der (Voll-) Transkription von Datenmaterial und der Realisierung aufwendiger Designs (z.B. Kontrollgruppendesigns) im Fokus. In nahezu allen Bereichen spielen Fragen der angemessenen Erfassung und Überprüfung von Wirkungen eine (zusehends) wichtige Rolle, wenngleich die Thematik nicht in allen Bereichen gleichermaßen intensiv diskutiert wird. So scheint der Rückgriff auf Plausibilitätsüberlegungen und die Offenlegung von Grenzen der Erbringung von Wirkungsnachweisen im Bereich der Kultur und Kulturpolitik weit verbreitet zu sein, wohingegen methodische Fragen der Wirkungsmessung in den anderen betrachteten Politikfeldern stärker im Fokus stehen und kontroverser diskutiert werden. Experimentelle Designs können jedoch kaum realisiert werden und auch quasi-experimentelle Designs stellen in keinem der betrachteten Felder einen Standard dar. Vielmehr werden in der Regel, in Abhängigkeit vom Gegenstand und von zur Verfügung stehenden Ressourcen, alternative Wege gesucht, um sich Wirkungsfragen zu widmen. Insgesamt wurde deutlich, dass hierzu nicht nur zwischen den Politikfeldern, sondern auch innerhalb von ihnen, durchaus unterschiedliche Auffassungen dazu bestehen, was Wirkungsmodellierung bzw. -messung leisten soll und wie dies methodisch gut zu bearbeiten ist. Methodische Innovationen speisen sich – wie in den auf der Session 2016 diskutierten Politikfeldern – vorrangig aus den eigenen Politikbereichen und den darin zu bearbeitenden Fragen. Gerade für die Evaluation von Lehre und Studium und für weitere pädagogische Felder können aber auch nahestehende Wissenschaftsbereiche Anregungen bieten. Darüber hinaus wurden in allen drei, 2017 vertretenen Politikfeldern Innovationspotenziale durch zunehmende Digitalisierung, neue Kommunikationsformen wie Social Media und Formen der massenhaften Datengewinnung bzw. -nutzung (Big Data) gesehen, die eher auf gesamtgesellschaftlichen Trends beruhen. Hier sind weitere Entwicklungen zu erwarten, welche auch methodische Fragen aufwerfen und daher in weiteren Sessions thematisiert werden sollten.

Literatur

Altenburg, Thomas (2017): Zwischen Schema F und Innovation. Eine politikfeldübergreifende Diskussion zu methodischen Standards. In: Zeitschrift für Evaluation, 16 (1), S. 210-217.

European Association for Quality Assurance in Higher Education (2015): Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG). Brussels, Belgium. Verfügbar unter: http://www.enqa.eu/index.php/home/esg/ [25.11.2017].

Halves, Edith/Lück-Filsinger, Marianne/Schmidt, Stefan (2014): Evaluation in der Sozialen Arbeit. Entwicklungen und Herausforderungen. In: Böttcher, Wolfgang/Kerlen, Christiane/Maats, Peter/Schwab, Oliver/Sheik, Sonja (DeGEval-Vorstand) (Hg.): Evaluation in Deutschland und Österreich, Stand und Entwicklungsperspektiven in den Arbeitsfeldern der DeGEval – Gesellschaft für Evaluation. Münster: Waxmann, S. 133-138.

Harris-Huemmert, Susan/Mittauer, Lukas/Pohlenz, Philipp (Hg.) (2015): Heterogenität der Studierenden. Herausforderungen für die Qualitätsentwicklung in Studium und Lehre, neuer Fokus für Evaluation? Bielefeld: Universitätsverlag Webler.

Mitterauer, Lukas/Pohlenz, Philipp/Harris-Huemmert, Susan (2017): Aktuelle Trends der Evaluation an Hochschulen. In: Zeitschrift für Evaluation, 16 (2), S. 275-276.

Reiter, Stefanie/Schmidt, Stefan/Strobl, Rainer/Astleithner, Florentina/Froncek, Benjamin/Stepanek, Peter (2015): Methodische Herausforderungen der Wirkungsanalyse bei knappen Ressourcen. Frühjahrstagung 2015 des AK Soziale Dienstleistungen. In: Zeitschrift für Evaluation, 14 (2), S. 319-327.

Reiter, Stefanie/Strobl, Rainer/Buchheit, Frank (2017): Kurzbericht des AK Soziale Dienstleistungen in der DeGEval über Entwicklungen in diesem Feld und die Rolle der DeGEval. In: Zeitschrift für Evaluation, 16 (2), S. 293-296.

Wirtz, Angela/Cappellaro, Marcus/Spiel, Georg (2017): Der Arbeitskreis Gesundheit in der DeGEval: ein Rückblick und eine Standortbestimmung. In: Zeitschrift für Evaluation, 16 (2), S. 270-273.